Współczesne metody analizy regresji wspomagane komputerowo

Jacek Syska

Instytut Fizyki, Uniwersytet Śląski, Uniwersytecka 4, 40-007 Katowice, Poland.

Skrypt dla studentów Ekonofizyki, luty 2014, wersja 1. Skrypt współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego. Skrypt jest dystrybuowany bezpłatnie.

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UPGOW – Uniwersytet Partnerem Gospodarki Opartej na Wiedzy
Uniwersytet Śląski w Katowicach, ul. Bankowa 12, 40-007 Katowice, http://www.us.edu.pl

Instytut Fizyki, Uniwersytet Śląski, Uniwersytecka 4, 40-007 Katowice, Poland,

Skrypt dla studentów Ekonofizyki,

luty 2014, wersja 1

Skrypt współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego. Skrypt jest dystrybuowany bezpłatnie.

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UPGOW – Uniwersytet Partnerem Gospodarki Opartej na Wiedzy
Uniwersytet Śląski w Katowicach, ul. Bankowa 12, 40-007 Katowice, http://www.us.edu.pl

Spis treści

Dziękuję Michałowi Czerwikowi, Marcinowi Jaworskiemu, Patrycji Kruczek, Agnieszce Maryniok, Dorocie Mroziakiewicz, Annie Rząsa, Iwonie Kaczmarczyk i Sebastianowi Zającowi za wspólne rozważania, w wyniku których powstało niniejsze opracowanie.

Statystyka stanowi zbiór metod, które służą pozyskiwaniu, prezentacji i analizie danych oraz otrzymaniu użytecznych, uogólnionych informacji na temat zjawiska, którego dotyczą. Dane są pozyskiwane w procesie zwanym badaniem statystycznym poprzez obserwacje statystyczne (bezpośrednio poprzez pomiary lub pośrednio, poprzez obliczenia). Program SAS[1], którego funkcjonowanie w analizie statystycznej zostanie poniżej zaprezentowane, jest jedną z kilku zaawansowanych aplikacji, oferującą szeroką gamę narzędzi analitycznych wykorzystywanych w następujących dziedzinach: w zarządzaniu (analizy finansowe, prognozowanie itp.), przemyśle (kontrola i zarządzanie jakością, badania rynku, analizy sprzedaży), bankowości (analiza kredytowa itp.), ubezpieczeniach (np. badanie rynku), w sektorze publicznym, w nauce (medycyna, ekonomia, fizyka, informatyka, zarządzanie i marketing). Jednocześnie bardzo obszerna pomoc (help) SAS'a ułatwia skuteczne wykorzystanie narzędzi jego analizy statystycznej, które w przejrzystej formie przekształcają dostępne dane w informacje.

[1] SAS, Statistical Analyze System (System Analiz Statystycznych).

Cześć I. Analiza klasyczna.

A. Rozdział 1. Analiza współzależności zmiennych w regresji wielorakiej

Rozdział 1-1. Cel, istota i przykłady badań.

W analizie regresji [1], [2] badania statystyczne mają w ogólności wyjaśniać zależności pomiędzy różnymi cechami badanej populacji. Populację rozumiemy jako zbiór elementów posiadających pewną stałą cechę, która je łączy i wyróżnia spośród innych zbiorów.
Przykładami populacji są następujące zbiory:

1. grupa ludności zamieszkująca pewien określony obszar
   (np.: Europejczycy, Ślązacy, ludność miejska)
2. grupa społeczna (np. studenci, górnicy, lekarze)
3. zbiór podmiotów gospodarczych (np. spółki wchodzące w skład WIG 20, sklepy spożywcze, punkty gastronomiczne)
4. zbiór przedmiotów o podobnej budowie lub właściwościach
   (np.: urządzenia elektroniczne, kryształy, przewodniki)

Przykładowymi cechami (określającymi właściwości elementów populacji) pomiędzy którymi będziemy badać zależności, a które mogą być również zależne od wpływu warunków zewnętrznych, są następujące wielkości, podane kolejno dla powyższych grup:

Dla grupy 1 i 2: wiek, wzrost, dochód, stan zdrowia, wykształcenie, narażenie na emisję spalin, narażenie na hałas.
Dla grupy 3: stopy zwrotu, poziom ryzyka, kondycja finansowa, struktura zatrudnienia, dzienne obroty, ilość klientów, stan prawa podatkowego.
Dla grupy 4: niezawodność, funkcjonalność, twardość, gęstość, kolor, ciężar, przewodność właściwa, cena.

Jedne z powyższych, przykładowych cech (zmiennych losowych) mogą pojawiać się w analizie jako zmienne objaśniane (odpowiedzi), natomiast inne, jako zmienne objaśniające (czynniki), mające wpływ na kształtowanie się rozkładów (warunkowych) cechy objaśnianej. Chociaż nie jest to regułą, to niejednokrotnie zdarza się, że badane cechy oddziałują na siebie wzajemnie. Własność ta ma duży wpływ na interpretację zależności przyczynowo-skutkowej zjawisk i jest ważnym elementem branym pod uwagę przy doborze zmiennych objaśniających. Pojęcie zmiennej losowej zostało przedstawione w Części IV, Rozdział 1.

Badanie to inaczej doświadczenie, zaś zmienne występujące w badaniu są określane jako: zmienna opisywana (Y – zwana zmienną objaśnianą, odpowiedzią lub czasami prognozą) i zmienne objaśniające (X – zwane czynnikami), mające wpływ na zmienną objaśnianą Y.

Badania statystyczne dzielimy na [1]:

- Badania doświadczalne polegające na tym, że osoba badająca współzależność zmiennych może ustalać wartości cech objaśniających. Badania takie mają szerokie zastosowanie przy wyjaśnianiu zjawisk fizycznych gdzie przeprowadzający doświadczenie może kontrolować zmienne (takie jak np. natężenie prądu, temperatura, ciśnienie).
- Badania quasi-eksperymentalne, w których obiekty badań są wyznaczone poprzez warunki losowe.
- Badania obserwacyjne sprowadzające się do opisu przez badacza zależności powstałych w wyniku zachodzących zmian, na które nie może on w żaden sposób wpływać (nie ma możliwości ustalania wartości cech). Badania obserwacyjne mają zastosowanie (teoretycznie) w dociekaniu zależności powstałych w społeczeństwach ludzkich, bądź w procesach rynkowych.

Głównymi celami badań statystycznych w analizie regresji są [1]:

1. Scharakteryzowanie relacji (między innymi jej zasięgu, kierunku i siły).
2. Znalezienie ilościowej zależności redukującej ogólny związek stochastyczny pomiędzy zmienną objaśnianą Y, a zmiennymi objaśniającymi $X_1, X_2, X_3, \ldots, X_k$, do zależności funkcyjnej $f(X_1, X_2, X_3, \ldots, X_k)$ określającej wartość oczekiwaną odpowiedzi Y. Oznacza to określenie modelu matematycznego, który w najbardziej wiarygodny sposób oddaje zachowanie się odpowiedzi. Znajomość takiego modelu daje nam możliwość predykcji wartości odpowiedzi w zależności od zachowania się innych zmiennych.
3. Określenie, które ze zmiennych objaśniających są ważne w analizie współzależności i uszeregowanie tych zmiennych ze względu na siłę wpływu na zmienną objaśnianą.
4. Znalezienie ilościowej i/lub jakościowej relacji pomiędzy odpowiedzią a czynnikami głównymi, gdy są one w populacji pod wpływem zmiennych pobocznych ($C_1, C_2, \ldots, C_m$) oraz uwzględnienie zmiennych pobocznych poprzez wzięcie ich pod kontrolę.
5. Porównywanie różnych modeli dla jednej zmiennej objaśnianej, tzn. porównanie modeli, które składają się z różnych zestawów zmiennych objaśniających.
6. Określenie interakcji zmiennych objaśniających oraz (przy dwukierunkowej zależności) określenie zależności zmiennych objaśniających od zmiennej objaśnianej.
7. Oszacowanie punktowe wartości współczynników regresji (kierunek i siła współzależności oraz istotność statystyczna parametrów wprowadzonych do modelu).

Oto kilka przykładów badań:

1) Określenie wzajemnej relacji pomiędzy produkcją przedsiębiorstwa (Y) a następującymi zmiennymi: X_1 – wydajność pracy , X_2 – środki trwałe przedsiębiorstwa, X_3 – zatrudnienie pracowników.
2) Badania epidemiologiczne polegające na określeniu wpływu: nawyku palenia X_1, klasy społecznej X_2, wieku C_1, wagi C_2 - na ciśnienie krwi Y.
3) określenie współzależności pomiędzy zmienną „satysfakcja pacjenta z opieki medycznej" Y, a zmiennymi: „relacja emocjonalna pacjenta z lekarzem" X_1 oraz „ stopień poinformowania pacjenta przez lekarza" X_2.

W analizie statystycznej badacz powinien ostrożnie analizować otrzymane wyniki, aby uniknąć błędów interpretacyjnych, które mogą wystąpić np. na skutek złej selekcji danych. Uzyskane wyniki powinien weryfikować opierając się o następujące kryteria [1]:

1. Określenie logicznego związku pomiędzy zmiennymi, tzn. sprawdzenie czy uzyskane wyniki nie kolidują z naturą zjawiska.
2. Unikanie czasowej dwuznaczności, czyli sprawdzenie czy przyczyna poprzedza w czasie skutek.
3. Analizę siły związku pomiędzy zmiennymi, a w szczególności zwrócenie uwagi na możliwość uzyskania wysokiej wartości korelacji między zmiennymi, które w rzeczywistości nie oddziałują na siebie.
4. Sprawdzenie czy otrzymany model jest modelem sprawdzającym się w rzeczywistości.
5. Rozpatrzenie spójności wyników.
6. Określenie zgodności wyników z wiedzą teoretyczną oraz doświadczalną, tzn. określenie praktycznej i teoretycznej wiarygodność przyjętych hipotez statystycznych.
7. Określenie specyfikacji związku. Rozpatrzenie możliwości otrzymania badanego skutku, jako przejawu działania różnych przyczyn oraz możliwości wystąpienia kilku skutków jednej przyczyny.

Głównym celem badań statystycznych w analizie regresji jest otrzymanie modelu matematycznego, który w jak najlepszy sposób będzie przedstawiał zależności pomiędzy różnymi cechami (zmiennymi). Jednakże należy zdawać sobie sprawę z tego, że nie jest możliwe uzyskanie idealnego modelu, gdyż większość zależności pomiędzy zmiennymi nie ma charakteru deterministycznego, tylko losowy, co pociąga za sobą uwzględnienie błędów w określeniu relacji.

Obecny rozdział jest związany właśnie z omówieniem metod statystycznych prowadzących do wskazania najlepszego modelu regresji, który przy niezbyt rozbudowanej strukturze daje równanie regresji jak najlepiej opisujące zależności pomiędzy zmiennymi, tzn.:

1) jak najlepiej dopasowujące się do danych empirycznych, a co jest z tym związane,

2) dające jak najsolidniejszą predykcję wartości zmiennej objaśnianej.

Zostaną przedstawione procedury, które na określonym poziomie istotności pozwalają wypowiedzieć się na temat dobroci wspomnianego dopasowania.
Przykłady przedstawione w opracowaniu zostaną przeanalizowane z wykorzystaniem pakietu statystycznego SAS [3].

A. Rozdział 2. Klasyfikacja zmiennych i wybór analizy.

Rozdział 2-1. Klasyfikacja zmiennych.

Do analizy współzależności statystycy posługują się zmiennymi, które należy umieć poprawnie sklasyfikować. Klasyfikacja zmiennych wiąże się z wyborem analizy, dlatego też należy położyć duży nacisk na właściwą ocenę zmiennych.

Podział zmiennych losowych [1]:

1. ze względu na charakter przyjmowanych wartości dystrybuanty:
   - zmienne typu dyskretnego,
   - zmienne typu ciągłego,

   Czasami zmienne typu dyskretnego mogą być traktowane jako ciągłe, a zmienne typu ciągłego, pogrupowane w pewne kategorie, mogą być traktowane jak dyskretne.
2. ze względu na kierunek w opisie zależności:
   - zmienna opisująca (objaśniająca, czynnik),
   - zmienna opisywana (objaśniana, odpowiedź),
3. ze względu na dokładność pomiarową zmiennej:
   - zmienna jakościowa (nominalna, symboliczna, kategoryczna),
   - zmienna porządkowa,
   - zmienna przedziałowa (grupowa, ilościowa).

Wartość przypisana zmiennej *nominalnej* (grupującej wyniki w odpowiednie kategorie) wskazuje różne kategorie, np. zmienna dotycząca płci przyporządkowuje wartość 0 dla płci męskiej, a wartość 1 dla płci żeńskiej. Wyższy poziom miary posiada zmienna *porządkowa*, bo oprócz grupowania wyników w kategorie może je porządkować.
Zmienna *przedziałowa*, oprócz posiadania własności poprzednich zmiennych, nadaje sens mierze odstępu między kategoriami. Musi być ona wyrażona w pewnych standardowych pojęciach i posiada różne skale, według których tworzy się przedziały wartości, jakie dana zmienna przyjmuje.

Tabela 2-1.1. Podział zmiennych jakościowych, porządkowych i przedziałowych na zmienne ciągłe i dyskretne.

Zmienne	ciągłe	dyskretne
Jakościowe	-	×
Porządkowe	×	×
Przedziałowe	×	×

Rozdział 2-2. Kryteria wyboru metody analizy.

Wybór analizy jest jedną z najważniejszych części badania statystycznego, gdyż od niego zależy poprawność analizy. Przy wyborze analizy należy brać pod uwagę następujące kryteria [1]:

1. Cel badania.
2. Matematyczne własności zmiennych.
3. Statystyczne założenia dotyczące zmiennych.
4. Sposób uzyskania danych do analizy.

Przy wyborze metody analizy współzależności zmiennych przydatna może okazać się poniższa Tabela pokazująca możliwe metody analizy zależności pomiędzy zmiennymi.

Tabela 2-2.2. Wybór metody analizy współzależności wielu zmiennych [1].

Metoda analizy	**Zmienna objaśniana**	**Zmienna(e) objaśniająca(e)**	**Ogólne przeznaczenie**
Analiza regresją wieloraką	ciągła	Zmienne ciągłe, ale dopuszcza się także dyskretne.	do opisu zasięgu, kierunku i siły relacji między kilkoma zmiennymi objaśniającymi i ciągłą zmienną objaśnianą.
Analiza wariancji	ciągła	zmienne jakościowe.	do opisu relacji między ciągłą zmienną objaśnianą i zmiennymi objaśniającymi jakościowymi.
Analiza kowariancji	ciągła	Kombinacje zmiennych jakościowych i zmiennych ciągłych (zmienne ciągłe jako zmienne kontrolowane).	do opisu relacji między ciągłą zmienną objaśnianą i zmiennymi objaśniającymi symbolicznymi, mając pod kontrolą ciągłe zmienne objaśniające.
Analiza metodą regresji Poissona	dyskretna	kombinacje różnych typów zmiennych objaśniających.	do badania zależności pomiędzy różnymi zmiennymi, a tempem zmian jakiegoś zjawiska.
Analiza metodą regresji logistycznej	dwuwartościowa	kombinacje różnych typów zmiennych objaśniających.	do badania zależności pomiędzy zmienną objaśnianą przyjmującą tylko dwie możliwe wartości, a innymi zmiennymi różnych typów.

Rozdział 2-3. Wybór postaci równania regresji.

Przypuśćmy, że posiadamy po n -pomiarów dwóch cech w populacji, cechy Y oraz cechy X. Celem jest oszacowanie zależności zmiennej objaśnianej Y od zmiennej objaśniającej X. Dla poszczególnych jednostek w próbce można zapisać wyniki pomiarów zmiennych X i Y w postaci pary liczb $(X_i, Y_i) = (x_i, y_i)$, gdzie i numeruje jednostki w próbce. Tak określone pary liczb możemy nanieść na układ współrzędnych o osiach X i Y, uzyskamy w ten sposób tzw. *diagram punktowy* (*wykres rozproszenia*). Następnie, należy wybrać najodpowiedniejszy model regresji [2], [1], [4] dla zależności opisującej zmianę wartości oczekiwanej odpowiedzi Y wraz ze zmianą wariantu czynnika X, czyli podać postać funkcji matematycznej, która najlepiej „pasuje do zredukowanego" obrazu diagramu punktowego (stąd nazwa „funkcja regresji"). Najczęściej stosowane funkcje regresji mają postać [1], [4]:

- funkcja liniowa $f(x) = a\,x + b$,
- funkcja wielomianowa, najczęściej kwadratowa $f(x) = a\,x^2 + b\,x + c$,
- funkcja logarytmiczna $f(x) = \ln(x)$, (2-3.1)
- funkcja eksponencialna $f(x) = e^{-x}$,
- funkcja logistyczna $f(x) = \dfrac{1}{1+e^{-x}}$.

A. Rozdział 3. Analiza regresji wielorakiej i właściwości macierzy korelacyjnej.

Analiza regresji wielorakiej jest rozszerzeniem prostoliniowej metody analizy regresji z jedną zmienną objaśniającą [2] do analizy regresji, w której występuje większa liczba zmiennych objaśniających. Krótkie omówienie regresji z jednym czynnikiem zostało podane w Rozdziale 4. Pełniejsze omówienie modelu regresji klasycznej można znaleźć w Rozdziale 11.

Analiza regresji wielorakiej jest trudniejsza od analizy regresji liniowej z jednym czynnikiem z następujących powodów [1]:

1. trudno wybrać najlepszy model, gdy występuje kilka możliwych czynników,
2. trudniejsze jest wyobrażenie sobie wybranego modelu, co wynika z niemożliwości narysowania więcej niż trójwymiarowego zbioru danych,
3. interpretacja wyników jest trudniejsza ze względu na trudności w wyjaśnieniu znaczenia najlepiej dopasowanego modelu,
4. obliczenia wymagają użycia szybkich komputerów, aby sprawnie wyliczyć korelacje pomiędzy zmiennymi.

Rozdział 3-1. Model regresji wielorakiej.

W przypadku regresji wielorakiej należy uogólnić model regresji liniowej z jedną zmienną objaśniającą [2] na przypadek większej liczby zmiennych objaśniających:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + E, \qquad (3\text{-}1.2)$$

gdzie: β_0, β_1, …, β_k to współczynniki regresji (parametry strukturalne) równania modelu w populacji, X_1, X_2,…, X_k są zmiennymi objaśniającymi lub funkcjami zmiennych objaśniających, E jest składnikiem losowym.

Graficzna interpretacja regresji wielorakiej.

W przypadku badania współzależności pomiędzy dwoma zmiennymi (jedną zmienną objaśniającą i jedną objaśnianą) w graficznej interpretacji otrzymaliśmy linię na dwuwymiarowym wykresie. W przypadku wielu zmiennych objaśniających liczba wymiarów wynosi k +1, gdzie k jest liczbą tych zmiennych. W przypadku, gdy mamy tylko dwie zmienne objaśniające otrzymujemy wykres trójwymiarowy, na którym model regresji będzie ilustrowany płaszczyzną. W przypadku większej liczby zmiennych objaśniających ilustracja graficzna zależności staje się niemożliwa. Dla liczby zmiennych objaśniających $k \geq 2$ przeprowadza się również badanie korelacji między wszystkimi kombinacjami par zmiennych (np. dla zestawu zmiennych Y, X_1, X_2, X_3 tworzymy następujące kombinacje: (Y, X_1), (Y, X_2), (Y, X_3), (X_1, X_2), (X_1, X_3), (X_2, X_3)).
Zawsze jednak równanie regresji należy rozumieć jako związek podający zależność wartości oczekiwanej (Część IV, Rozdział 1) warunkowej[2] $E(Y \mid X_1, X_2, ..., X_k)$ zmiennej objaśnianej Y od każdej specyficznej kombinacji zmiennych objaśniających. Np. dla każdej pary wartości zmiennych X_1 i X_2 mamy określony rozkład zmiennej Y z określoną wartością oczekiwaną warunkową $\mu_{Y|X_1,X_2} \equiv E(Y \mid X_1, X_2)$ oraz wariancją warunkową $\sigma^2_{Y|X_1,X_2} \equiv E\left((Y - \mu_{Y|X_1,X_2})^2 \mid X_1, X_2\right)$. Dlatego dla odpowiedzi Y i dwu czynników X_1, X_2, równanie regresji jest reprezentowane graficznie przez powierzchnię zależności wartości oczekiwanych warunkowych zmiennej Y od tych czynników.

Funkcja regresji „Pierwszego rodzaju”. Niech ciąg wartości x_1, x_2,…, x_k jest realizacją zmiennych objaśniających X_1, X_2,…, X_k . Warunkowa wartość oczekiwana $E(Y \mid x_1, x_2, ..., x_k)$, traktowana jako funkcja wartości x_1, x_2,…, x_k czynników X_1, X_2,…, X_k, jest nazywana funkcją regresji (Pierwszego rodzaju).

Zadanie. Niech $\mu_X \equiv E(X)$ i $\mu_Y \equiv E(Y)$ oraz $\sigma_X \equiv \sigma(X)$ i $\sigma_Y \equiv \sigma(Y)$ są, kolejno, ogólną wartością oczekiwaną oraz odchyleniem standardowym zmiennej Y i X oraz niech $\rho \equiv \rho_{XY}$ jest współczynnikiem

[2] (lub średniej warunkowej teoretycznej w próbie)

korelacji liniowej Pearsona zmiennych X i Y (określonym w Rozdziale 3-2 oraz 11-2-1) [2]. Pokazać, że gdy rozkład dwuwymiarowy (X, Y) jest normalny (Część IV, Rozdział 2), wtedy $E(Y \mid x)$ jest funkcją liniową x:

$$\mu_{Y|X} \equiv E(Y \mid x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)\,, \tag{3-1.3}$$

natomiast wariancja warunkowa wynosi:

$$\sigma_{Y|X}^2 = \sigma_Y^2\left(1 - \rho^2\right)\,, \tag{3-1.4}$$

co oznacza, że jest ona taka sama (jednorodna) dla wszystkich wariantów x zmiennej X. Parametr $\sigma_Y^2 \equiv E\left((Y - \mu_Y)^2\right)$ jest (ogólną) wariancją zmiennej losowej Y [2].

Wniosek. W analizie regresji zmiennej Y względem X, w przypadku gdy rozkład dwuwymiarowy (X, Y) jest normalny, funkcja regresji $E(Y \mid x)$ jest liniowa, a przed przystąpieniem do analizy należy przeprowadzić test jednorodności wariancji.

Założenia klasycznego modelu regresji wielorakiej (KMRW) **dla metody najmniejszych kwadratów** (MNK) [1]:

1. **Istnienie**: Dla każdej kombinacji wartości zmiennych objaśniających X_1, X_2,…, X_k, zmienna objaśniana Y jest (jednoznaczną) zmienną losową z określonym rozkładem prawdopodobieństwa posiadającym skończoną wartość oczekiwaną i wariancję.
2. **Kontrolowanie wartości czynników**: Tak jak w typowym klasycznym modelu regresji liniowej [2], zmienną losową jest zmienna Y, podczas gdy zmienne X_1,X_2,…,X_k są zmiennymi (nielosowymi) kontrolowanymi.
3. **Liniowość regresji**: Warunkowa wartość oczekiwana $E(Y \mid X_1, X_2, ..., X_k)$ zmiennej Y dla każdej określonej kombinacji zmiennych objaśniających X_1, X_2,…, X_k jest liniową funkcją tych zmiennych:

   $$\mu_{Y|X_1,X_2,...,X_k} \equiv E(Y \mid X_1, X_2, ..., X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{3-1.5}$$

   lub:

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E\,, \tag{3-1.6}$$

   gdzie E jest składnikiem losowym[3] będącym odzwierciedleniem różnic między realizacjami empirycznymi zmiennej Y, a wartościami teoretycznymi średnich $\mu_{Y|X_1,X_2,...,X_k}$ zmiennej Y. Równanie regresji (3-1.5) opisuje tzw. powierzchnię regresji. Stałe β_0,β_1,…,β_k są nieznanymi parametrami populacji, natomiast składnik losowy E jest zmienną losową nieobserwowaną bezpośrednio. Konsekwencją zastosowania MNK dla modelu regresji w populacji jest zerowanie się wartości oczekiwanej składnika losowego $E(E \mid X_1, X_2, ..., X_k) = 0$, co miało swój wyraz w (3-1.5).

[3] W skrypcie oznaczamy składnik losowy literą E (za angielskim: error). Chociaż oznaczenie to pokrywa się z symbolem E wartości oczekiwanej, to biorąc pod uwagę kontekst, nie powinno to prowadzić do nieporozumień.

4. **Niezależność**: Obserwacje zmiennej objaśnianej Y są od siebie niezależne, tzn. poszczególne obserwacje zmiennej Y nie zależą od wartości otrzymanych wcześniej. Wtedy, gdy kilka obserwacji zmiennej Y jest dokonanych na tej samej jednostce zbiorowości [1], założenie to jest na ogół naruszone.
5. **Stałość rozproszenia** (homoscedastyczność): Wariancja (warunkowa) zmiennej Y dla dowolnej ustalonej kombinacji zmiennych X_1, X_2,…, X_k jest taka sama (jednorodna) dla wszystkich rozkładów warunkowych, tzn.:

$$\sigma^2_{Y|X_1,X_2,...,X_k} = Var(Y \mid X_1, X_2,...,X_k) \equiv \sigma^2_E \qquad (3\text{-}1.7)$$

lub:

$$\sigma^2_{E|X_1,X_2,...,X_k} \equiv \sigma^2_E \quad . \qquad (3\text{-}1.8)$$

6. **Normalność:** Dla dowolnej ustalonej liniowej kombinacji zmiennych X_1, X_2,…, X_k, zmienna Y ma rozkład normalny, tzn.

$$Y \sim \mathrm{N}(\mu_{Y|X_1,X_2,...,X_k}, \sigma^2_E) \qquad (3\text{-}1.9)$$

lub równoważnie (dla regresji liniowej):

$$E \sim \mathrm{N}(0, \sigma^2_E) \quad . \qquad (3\text{-}1.10)$$

Dla modelu regresji (wielorakiej) *założenie normalności nie jest konieczne* dla wyznaczenia punktowych oszacowań metody najmniejszych kwadratów (MNK) parametrów modelu regresji, ale na ogół jest wymagane ono do wnioskowania. Wyraźne odejście od rozkładu normalnego daje błędne wyniki. Gdy założenie o normalności jest słabo spełnione, należy poszukać transformacji zmiennej Y (typu: $\log Y$, $\sqrt{Y}$), która w przybliżeniu posiada rozkład normalny. Założenie o normalności rozkładu jest istotne dla estymacji i wnioskowania, co jest spowodowane posługiwaniem się rozkładem t-Studenta i F-Snedecora. Tylko gdy rozkłady warunkowe są normalne, MNK nabiera charakteru probabilistycznego, stając się szczególnym przypadkiem metody największej wiarygodności (MNW) [5].

Macierzowe ujęcie klasycznego liniowego modelu regresji wraz z wyprowadzeniami wynikającymi z zastosowania MNK, zostało przedstawione w Rozdziale 11. W Rozdziałach od 10 do 15 przedstawiono analizę własności reszt klasycznego modelu regresji oraz zaprezentowano działanie testów niezależności reszt i ich normalności. Omówienie testu Goldfelda - Quandt'a jednorodności reszt modelu regresji można znaleźć w [4]. W Rozdziale 16 omówiono test Bartlett'a jednorodności reszt. Rozważania do Rozdział 8, zawierają głównie omówienie podstawowych metod dla selekcji klasycznego modelu regresji.

Rozdział 3-2. Macierz korelacyjna, współczynnik korelacji zupełnej i współczynniki korelacji cząstkowej.

Współczynnik korelacji liniowej Pearsona (zupełnej, całkowitej) pomiędzy zmiennymi X_i, X_j jest zdefiniowany w populacji następująco [2]:

$$\rho \equiv \rho_{X_i X_j} = \frac{\mathrm{cov}(X_i, X_j)}{\sigma(X_i)\,\sigma(X_j)}\ , \tag{3-2.11}$$

gdzie $\sigma(X_i)$ oraz $\sigma(X_j)$ są odchyleniami standardowymi zmiennych X_i oraz X_j w populacji, a $\mathrm{cov}(X_i, X_j) \equiv E\big((X_i - E(X_i))(X_j - E(X_j))\big)$ jest ich kowariancją.

Niech C jest macierzą korelacji [2] dla układu zmiennych Y, X_1, X_2,…, X_k,:

$$C = \begin{bmatrix} 1 & \rho_{YX_1} & \rho_{YX_2} & \cdots & \rho_{YX_k} \\ \rho_{X_1Y} & 1 & \rho_{X_1X_2} & \cdots & \rho_{X_1X_k} \\ \rho_{X_2Y} & \rho_{X_2X_1} & 1 & \cdots & \rho_{X_2X_k} \\ \vdots & \vdots & \vdots & \ddots & \\ \rho_{X_kY} & \rho_{X_kX_1} & \rho_{X_kX_2} & & 1 \end{bmatrix} \tag{3-2.12}$$

gdzie $\rho_{YX_i} = \rho_{X_iY}$ jest współczynnikiem korelacji liniowej Pearsona, pomiędzy zmienną Y, a zmienną X_i, natomiast $\rho_{X_jX_i} = \rho_{X_iX_j}$ jest współczynnikiem korelacji liniowej Pearsona pomiędzy zmiennymi X_i, X_j, (i, j = 1,2,…, k). Z postaci (3-2.11) wynika, że macierz korelacji jest symetryczna.

Uwaga. Nieco więcej informacji dotyczących własności estymatora (empirycznego) współczynnika korelacji liniowej Pearsona $R \equiv \hat{\rho}$, (11-2-1.54), parametru ρ, w tym dotyczących jego rozkładu, można znaleźć w Rozdziale 11-2-1. Zarówno wartość parametru ρ jak i **wartość w próbce** r estymatora R jest liczbą bezwymiarową z przedziału $\langle -1, +1 \rangle$.

Rozdział 3-2-1. Współczynnik korelacji cząstkowej.

Współczynnik Pearsona określa liniową zależność pomiędzy zmiennymi X_i, X_j, ale zależność ta zawiera w sobie również pośredni wpływ pozostałych zmiennych. W celu analizy współzależności pomiędzy, powiedzmy, zmiennymi Y i X_i, przy wyłączonym[4] (zatem kontrolowanym[5]) wpływie zmiennych X_1, X_2,…, X_{i-1}, X_{i+1},…, X_k, oblicza się współczynnik korelacji cząstkowej [2]:

[4] Wyłączony jest wpływ czynników X_1, X_2,…, X_{i-1}, X_{i+1},…, X_k, z korelacji zmiennych Y i X_i .
[5] Po wprowadzeniu zmiennej do analizy, kontrolowane są jej wartości.

$$\rho_{YX_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k} = \frac{-C^d_{YX_i}}{\sqrt{C^d_{YY}\ C^d_{X_iX_i}}} \quad , \tag{3-2-1.13}$$

gdzie $C^d_{X_iX_j}$ jest dopełnieniem algebraicznym elementu $\rho_{X_iX_j}$ wyznacznika det*C*.

Rzędem współczynnika korelacji cząstkowej nazywamy liczbę zmiennych pod kontrolą. Np. współczynniki zerowego rzędu to $\rho_{X_iX_j}$, współczynniki pierwszego rzędu to $\rho_{X_iX_j|X_k}$, współczynniki drugiego rzędu to $\rho_{X_iX_j|X_kX_l}$. Oznaczając, dla jasności zapisu, zmienne kontrolowane następująco X_1=Z_1, X_2=Z_2, itd., a interesujący nas czynnik X_i=X, możemy współczynnik korelacji cząstkowej odpowiedzi Y i wybranego czynnika X, zapisać następująco:

$$\rho_{YX|Z_1Z_2\ldots Z_l} \quad , \tag{3-2-1.14}$$

gdzie l jest liczbą zmiennych kontrolowanych ($l = k - 1$).

Współczynniki korelacji cząstkowej pomiędzy wyróżnionymi zmiennymi (które są po lewej stronie kreski | w indeksie $\rho_{\bullet|}$) określają zależność pomiędzy dwiema zmiennymi przy wyłączeniu działania zmiennych kontrolowanych (które są po prawej stronie kreski | w indeksie $\rho_{|\bullet}$). Zatem współczynniki korelacji cząstkowej mogą przedstawiać faktyczną zależność między badanymi zmiennymi, czego nie można powiedzieć o współczynnikach korelacji zupełnej. Czasami różnice pomiędzy tymi współczynnikami są na tyle duże, że zastosowanie tylko współczynników korelacji zupełnej mogłoby prowadzić do znaczących błędów w analizie współzależności.

Uwaga: Jeśli liczba l zmiennych kontrolowanych jest mniejsza niż k-1, wtedy również można skorzystać z zależności (3-2-1.13), jednakże należy to uczynić dopiero po wcześniejszym skreśleniu z macierzy C (3-2.12) odpowiednich wierszy i kolumn dla zmiennych, które nie są brane pod uwagę jako kontrolowane.

Zachodzi ważne twierdzenie zgodnie, z którym [1], [6]:

$$\rho_{XY|Z_1Z_2\ldots Z_l} = \rho_{X-\mu_{X|Z_1Z_2\ldots Z_l},\, Y-\mu_{Y|Z_1Z_2\ldots Z_l}} \quad . \tag{3-2-1.15}$$

Mówi ono o tym, że cząstkowy współczynnik korelacji dla zmiennych X i Y, przy kontrolowanym wpływie grupy zmiennych Z_1, Z_2, ..., Z_l ($l \le n$ - 1), jest równy współczynnikowi korelacji zupełnej *pomiędzy resztami* pozostałymi z dopasowania zmiennej X do grupy zmiennych Z_1, Z_2, ..., Z_l, a *resztami* pozostałymi z dopasowania zmiennej Y do grupy zmiennych Z_1, Z_2, ..., Z_l. Twierdzenie to dobrze ilustruje nazywanie zmiennych, dla których liczymy korelację jako dostrojonych (dopasowanych równaniem regresji) do zmiennych kontrolowanych.

Ze związku (3-2-1.13) można otrzymać np. następujące wzory na współczynnik korelacji cząstkowej:

a) dla trzech zmiennych (X, Y, Z) [2] (pokazać):

$$\rho_{YX|Z} = \frac{\rho_{YX} - \rho_{YZ}\rho_{XZ}}{\sqrt{(1-\rho_{YZ}^2)(1-\rho_{XZ}^2)}} \tag{3-2-1.16}$$

b) dla czterech zmiennych (X, Y, Z_1, Z_2) [2]:

$$\rho_{YX|Z_1Z_2} = \frac{\rho_{YX|Z_1} - \rho_{YZ_2|Z_1}\rho_{XZ_2|Z_1}}{\sqrt{(1-\rho_{YZ_2|Z_1}^2)(1-\rho_{XZ_2|Z_1}^2)}} \ . \tag{3-2-1.17}$$

Zatem współczynniki korelacji wyższego rzędu można otrzymać ze współczynników korelacji niższego rzędu.

Rozdział 3-2-2. Półcząstkowe współczynniki korelacji cząstkowej.

Współczynniki korelacji cząstkowej nazywane są również „pełnymi współczynnikami korelacji cząstkowej”. Nazwa ta jest związana z tym, że *obie* zmienne, dla których oblicza się korelacje są dopasowane do zmiennych kontrolowanych, w odróżnieniu od tzw. półcząstkowych współczynników korelacji cząstkowej.

Półcząstkowe współczynniki korelacji cząstkowej są to współczynniki, w których tylko jedna zmienna z dwóch zmiennych (dla których oblicza się korelacje) jest dostrojona do zmiennych kontrolowanych.

Półcząstkowy współczynnik korelacji cząstkowej pomiędzy zmiennymi Y i X, gdy tylko zmienna X została dopasowana do zmiennej Z, definiujemy następująco [1]:

$$\rho_{Y(X|Z)} = \rho_{Y,\, X-\mu_{X|Z}} \tag{3-2-2.18}$$

co jest równoważne zależności [1]:

$$\rho_{Y(X|Z)} = \frac{\rho_{YX} - \rho_{YZ}\rho_{XZ}}{\sqrt{1-\rho_{XZ}^2}} \ . \tag{3-2-2.19}$$

Analogicznie można zapisać półcząstkowy współczynnik korelacji cząstkowej pomiędzy zmiennymi X i Y, gdy tylko zmienna Y jest dopasowana do zmiennej Z:

$$\rho_{X(Y|Z)} = \rho_{X,Y-\mu_{Y|Z}} = \frac{\rho_{YX} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{YZ}^2}} \tag{3-2-2.20}$$

Rozdział 3-2-3. Współczynnik korelacji wielorakiej (wielokrotnej, wielowymiarowej).

Współczynnik korelacji wielorakiej określa współzależność pomiędzy zmienną X_i, a *kompletem pozostałych zmiennych.* Istnieje następujący związek pomiędzy współczynnikiem korelacji wielorakiej, a wszystkimi współczynnikami korelacji cząstkowej [2].

$$\rho_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k} =$$
$$= \sqrt{1-(1-\rho^2_{X_iX_1})(1-\rho^2_{X_iX_2|X_1})\cdots(1-\rho^2_{X_iX_{i-1}|X_1X_2\ldots X_{i-2}})(1-\rho^2_{X_iX_{i+1}|X_1X_2\ldots X_{i-1}})\cdots(1-\rho^2_{X_iX_k|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_{k-1}})}$$
(3-2-3.21)

Współczynnik korelacji wielorakiej jest zawsze dodatni. Wyraża on ścisłość związku pomiędzy interesującą nas zmienną, a całokształtem innych uwzględnionych zmiennych. Z powyższej postaci $\rho_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k}$ widać jednokrotne branie pod uwagę wpływu każdej ze zmiennych $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$, na wybraną zmienną X_i. Jeśli jest on bliski 1, to zmienność zmiennych $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$ określa prawie całkowicie zmienność wybranej zmiennej X_i, a wpływ innych zmiennych jest bez większego znaczenia. Jeżeli natomiast jest on daleki od jedności, to oznacza to, że istnieje jeszcze wpływ innych zmiennych, których nie wzięliśmy pod uwagę.

Przez $R_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k}$ będziemy oznaczali estymator współczynnika korelacji wielorakiej $\rho_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k}$. Natomiast przez $r_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k}$ będziemy oznaczali wartość $R_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k}$ przyjętą w próbce.

W dalszej części rozważań zwrócimy uwagę na znaczenie w analizie regresji zarówno współczynnika korelacji wielorakiej $R_{X_i|X_1X_2\ldots X_{i-1}X_{i+1}\ldots X_k}$, gdzie wszystkie zmienne są czynnikami wprowadzonymi do modelu regresji, oraz na znaczenie współczynnika korelacji wielorakiej $R_{Y|X_1X_2\ldots X_i\ldots X_k}$ zmiennej objaśnianej Y z tymi czynnikami.

Rozdział 3-3. Wyznaczanie najlepszych estymatorów równania regresji wielorakiej w MNK.

Sednem MNK jest wyznaczenie oszacowań parametrów modelu regresji, dla których suma kwadratów różnic między wartościami empirycznymi Y_i, a wartościami wyznaczonymi przez model $\hat{Y}_i$ jest minimalna. W klasycznej metodzie najmniejszy kwadratów (KMNK), i stąd w klasycznym modelu regresji, *czynniki nie są zmiennymi losowymi*.

Rozdział 3-3-1. Równanie regresji wielorakiej i metoda najmniejszych kwadratów.

Postać modelu regresji (3-1.2) (zaproponowana) w populacji, a przeniesiona na próbę, jest następująca:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k + \hat{E}\,. \qquad (3\text{-}3\text{-}1.22)$$

Dla *i*-tego pomiaru w próbie, zapisujemy ten model następująco:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} + \hat{E}_i \ , \qquad i = 1, 2,..., n \ , \tag{3-3-1.23}$$

gdzie $\hat{E}_i$ jest tzw. składnikiem resztowym.

Zatem równanie regresji II-go rodzaju, określające postać *teoretycznych średnich warunkowych* $\hat{Y}$ w próbie, ma postać:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \ , \tag{3-3-1.24}$$

co dla *i*-tego pomiaru w próbie, można zapisać następująco:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} \ , \ i = 1, 2,..., n \ . \tag{3-3-1.25}$$

Z KMNK (w której jak wiemy czynniki nie są losowe) wynika również, że punkt $(\overline{X}, \overline{Y})$ również spełnia równanie regresji:

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1 + \hat{\beta}_2 \hat{X}_2 + \ldots + \hat{\beta}_k \overline{X}_k \tag{3-3-1.26}$$

gdzie:

$$\overline{X}_s = \frac{1}{n} \sum_{i=1}^{n} X_{si} \ , \qquad s = 1, 2,..., k \tag{3-3-1.27}$$

są średnimi arytmetycznymi (nielosowych) czynników, kolejno X_1, X_2,…, X_k, oraz

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \ . \tag{3-3-1.28}$$

jest średnią arytmetyczną zmiennych Y_i.

Jeśli oszacowujemy parametry $\beta_0, \beta_1, \ldots, \beta_k$ przy pomocy estymatorów $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, wtedy dla *i*-tej obserwacji w próbie, właściwym oszacowaniem składnika losowego E_i w (3-1.2) jest składnik resztowy:

$$\hat{E}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k Y_{ki}) \ , \qquad i = 1,2,\ldots,n \ , \tag{3-3-1.29}$$

gdzie Y_i jest zmienną losową obserwowaną dla *i*-tego pomiaru w próbie.

Aby wyznaczyć postacie estymatorów $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ należy znaleźć minimum sumy kwadratów:

$$SSE \equiv \sum_{i=1}^{n} \hat{E}_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \ldots - \hat{\beta}_k X_{ki})^2 \ , \tag{3-3-1.30}$$

dla odchyłek $Y_i - \hat{Y}_i$ (tzw. reszt lub „błędów") wartości empirycznych Y_i od teoretycznych średnich warunkowych $\hat{Y}_i$ [1].

Równanie regresji wyznaczone w klasycznej metodzie najmniejszych kwadratów (KMNK) daje teoretyczne średnie warunkowe:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \ , \tag{3-3-1.31}$$

będące liniową kombinacją czynników X_1, X_2,…, X_k, w taki sposób, że $\hat{Y}$ mają możliwie jak największą korelację ze zmienną objaśnianą *Y*. Inaczej mówiąc kombinacja $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$ ma z wszystkich

możliwych liniowych kombinacji $a_0 + a_1X_1 + a_2X_2 + \ldots + a_kX_k$ zmiennych objaśniających, maksymalną wartość $r_{Y,\hat{Y}}$ współczynnika korelacji wielorakiej (wielokrotnego współczynnika korelacji) $R_{Y|X_1,X_2,\ldots,X_k} = R_{Y,\hat{Y}}$, (3-2-3.21):

$$R_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2(\hat{Y}_i - \overline{\hat{Y}})^2}}, \qquad (3\text{-}3\text{-}1.32)$$

zmiennej objaśnianej Y z *kompletem czynników* $X_1, X_2, \ldots, X_k$, gdzie $\hat{Y}_i$ to przewidywana modelem regresji wartość zmiennej Y, a $\overline{\hat{Y}}$ jest średnią zmiennych $\hat{Y}_i$ [1].

Estymatory poszczególnych parametrów strukturalnych w KMNK są *nieobciążone* [2] oraz są liniowe ze względu na wartości zmiennych Y_i, przy czym zgodnie z twierdzeniem Gaussa-Markowa posiadają one, w klasie tych estymatorów, najmniejszą możliwą wariancję, tzn. są estymatorami efektywnymi [7], [2]. Ponieważ każdy estymator $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ jest liniową funkcją wartości zmiennych Y_i, zatem gdy Y_i mają rozkład normalny i są statystycznie niezależne, to na podstawie twierdzenia o addytywności rozkładu normalnego [2], estymatory $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ mają również rozkład normalny.

Zadanie: Pokazać, że $\overline{Y} = \overline{\hat{Y}}$. (3-3-1.33)

Rozdział 3-3-2. Współczynnik determinacji jako miara dopasowania modelu do danych empirycznych.

Nietrudno pokazać, że w MNK zachodzi następujące, fundamentalne równanie rozkładu całkowitej sumy kwadratów:

$$SSY = SSR + SSE. \qquad (3\text{-}3\text{-}2.34)$$

W równaniu tym SSY:

$$SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad (3\text{-}3\text{-}2.35)$$

jest całkowitą (ogólną) sumą kwadratów, określającą ogólną zmienność zmiennej objaśnianej.

SSE (suma kwadratów reszt, błędów) jest zminimalizowaną sumą kwadratów reszt (3-2-1.9) dla badanego modelu:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (3\text{-}3\text{-}2.36)$$

natomiast SSR:

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \qquad (3\text{-}3\text{-}2.37)$$

jest sumą kwadratów modelu regresji, określającą zmienność zmiennej objaśnianej wyjaśnioną funkcją regresji.

Liczba stopni swobody dla sum kwadratów *SS*.

Liczba niezależnych zmiennych, niezbędna do wyznaczenia powyższych sum kwadratów *SS*, czyli ich liczba stopni swobody (*l.st.sw*. lub *df* od 'degrees of freedom'), jest następująca.

a) Ze względu na jedno ograniczenie na zmienne Y_i, płynące z postaci wiążącej je średniej arytmetycznej $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, suma $SSY = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ ma:

$$df_{SSY} = n\text{-}1. \qquad (df_{SSY}) \qquad (3\text{-}3\text{-}2.38)$$

b) Suma $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ ma:

$$df_{SSR} = k. \qquad (df_{SSR}) \qquad (3\text{-}3\text{-}2.39)$$

Stwierdzenie to wynika z tego, że w KMNK dla określenia SSR wystarczy k informacji na temat estymatorów $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ współczynników kierunkowych uzyskanych z próby df_{SSY} [2]. Łatwo sprawdzić powyższe stwierdzenie, gdyż wykorzystując (3-3-1.25) oraz (3-3-1.26) w SSR, widzimy, że SSR zależy jedynie od *k* estymatorów $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$, których wartości trzeba określić z próby.

c) Równanie (3-3-2.34), $SSY = SSR + SSE$, wymusza aby liczba niezależnych zmiennych niezbędna do wyznaczenia występujących w nim sum kwadratów *SS*, była równa po jego prawej i lewej stronie. Dlatego równanie to pociąga za sobą równanie dla liczby stopni swobody (*df*):

$$df_{SSY} = df_{SSR} + df_{SSE}, \qquad (3\text{-}3\text{-}2.40)$$

to znaczy:

$$n - 1 = k + df_{SSE},$$

skąd wynika, że $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ ma:

$$df_{SSE} = (n - k - 1). \qquad (3\text{-}3\text{-}2.41)$$

Współczynnik determinacji. Kwadrat współczynnika korelacji wielorakiej $R \equiv R_{Y,\hat{Y}}$, (3-3-1.32):

$$R^2 \equiv R^2_{Y,\hat{Y}} \qquad (3\text{-}3\text{-}2.42)$$

jest nazywany *współczynnikiem determinacji*. W przypadku jednego czynnika *X*, współczynnik korelacji wielorakiej $R \equiv R_{Y,\hat{Y}}$ sprowadza sie do współczynnika korelacji liniowej Pearsona pary zmiennych X i Y [2] (Rozdział 11-2).

Można pokazać, że w klasycznym modelu regresji, współczynnik determinacji spełnia następujący związek [1]:

$$R^2_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SSY - SSE}{SSY}. \qquad (3\text{-}3\text{-}2.43)$$

Jako kwadrat współczynnika korelacji R, współczynnik determinacji R^2 przyjmuje wartości r^2 z przedziału $\langle 0,1 \rangle$ [2], przy czym wartości bliskie jedynce oznaczają dobre dopasowanie modelu do danych empirycznych. Istotnie, w MNK minimalizowana jest suma kwadratów błędów *SSE* (*która jest w ogólności różna od zera*). *Model idealny* jest określony jako taki, dla którego *SSE* = 0, co podstawiając do wzoru (3-3-2.43) daje:

$$r^2 = \frac{SSY - SSE}{SSY} = \frac{SSY - 0}{SSY} = 1 \qquad \text{(dla modelu idealnego)}. \qquad (3\text{-}3\text{-}2.44)$$

Istnieją pewne podobieństwa ale i różnice w posługiwaniu się współczynnikiem korelacji R i współczynnikiem determinacji R^2. R jest miarą siły związku *liniowego* pomiędzy zmiennymi. ***Tylko*** w następującym zrozumieniu r^2 może być miarą siły związku nieliniowego. Otóż, może się zdarzyć, że r^2 przyjmuje wartości bliskie 1 dla pewnej, *nieliniowej* w zmiennych pierwotnych funkcji regresji dopasowanej do danych empirycznych. Nie stanowi to jednak o sile wspomnianego nieliniowego związku pomiędzy czynnikami a zmienną objaśnianą Y (np. o sile związku kwadratowego), lecz o sile związku *liniowego* zmiennej Y z nowo określonymi zmiennymi, które są tak zdefiniowane, aby wchodziły liniowo w funkcję regresji. Gdy np. zmienna kwadratowa X^2 zostanie zastąpiona nową zmienną $X_1 \equiv X^2$, to wtedy nowa zmienna X_1 wchodzi już liniowo. Po takiej zamianie zmiennych rozważamy już liniową regresję Y względem X_1 i ***przez*** R^2 ***określana jest również siła związku liniowego*** pomiędzy X_1 a Y. Zatem w analizie z wykorzystaniem R^2, nieliniowe funkcje czynników pierwotnych są traktowane jako nowe zmienne wchodzące liniowo.

Uwaga: *W regresji nieliniowej, po wprowadzeniu w miejsce pierwotnych zmiennych, wchodzących nieliniowo w równanie regresji, nowych zmiennych, które wchodzą liniowo w równanie regresji, związek* (3-3-2.43) *okazuje się być również słuszny. Oczywiście po takiej zamianie zmiennych, liczba czynników(z których wszystkie wchodzą liniowo w równanie regresji) na ogół wzrasta.*

Uwaga. Dodatkowe uwagi na temat linearyzacji modeli nieliniowych w czynniku można znaleźć np. w [4].

Uwaga. SAS [3] dysponuje również procedurami analizy modeli nieliniowych w parametrach modelu. Jedną z nich jest PROC NLIN [9].

Rozdział 3-3-3. Test istotności zmiennych w modelu regresji.

Dla modelu regresji z k czynnikami:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E \qquad (3\text{-}3\text{-}3.45)$$

stawiamy hipotezę zerową:

$$H_0\text{:}\ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad (3\text{-}3\text{-}3.46)$$

o nieistotności zmiennych X_i w tym modelu. Hipoteza ta jest pytaniem o nieistotność zależności korelacyjnej pomiędzy zmienną objaśnianą Y, a grupą zmiennych objaśniających $X_1, X_2,\ldots, X_k$.

Istotnie, gdyby bowiem hipoteza zerowa była prawdziwa, to zgodnie z (3-1.5) mielibyśmy, że wszystkie warunkowe wartości oczekiwane $\mu_{Y|X_1,X_2,\ldots,X_k}$ są takie same (i równe ogólnej wartości oczekiwanej μ_Y) dla wszystkich możliwych kombinacji wartości zmiennych objaśniających $X_1, X_2,\ldots, X_k$, co oznacza właśnie brak zależności korelacyjnej zmiennej objaśnianej od zmiennych objaśniających.

Celem testu statystycznego dla hipotezy zerowej (3-3-3.46) jest więc wyeliminowanie z modelu jednocześnie całej grupy zmiennych $X_1, X_2,\ldots, X_k$, o ile występowanie ich nie ma istotnie statystycznego wpływu na zmianę teoretycznej średniej warunkowej zmiennej Y.

Test ten dotyczy również weryfikacji przypuszczenia, że nie istnieją żadne zmienne pośród czynników X_1, $X_2,\ldots, X_k$ wprowadzonych do modelu, które dają istotnie statystycznie lepsze dopasowanie się modelu do danych empirycznych, niż czyni to model $Y = \beta_0 + E$.

Aby przeprowadzić taki test należy obliczyć statystykę F (opierając się o tzw. tablicę ANOVA dla modelu regresji) [1]:

$$F = \frac{MSR}{MSE} = \frac{SSR/df_{SSR}}{SSE/df_{SSE}} = \frac{(SSY - SSE)/k}{SSE/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}, \qquad (3\text{-}3\text{-}3.47)$$

gdzie R^2 jest współczynnikiem determinacji (3-3-2.42) [1]. Zakładając, że zmienne Y_i mają rozkład normalny (jak to czynimy w normalnym, klasycznym modelu regresji), można pokazać, że przy prawdziwości hipotezy zerowej H_0 statystyka F ma rozkład F-Snedecora z liczbą stopni swobody licznika $df_{SSR} = k$, (df_{SSR}), i mianownika $df_{SSE} = n\text{-}k\text{-}1$, (3-3-2.41), tzn. ma rozkład $F_{k,n\text{-}k\text{-}1}$.

Wyznaczoną na podstawie obserwacji (*obs*) w próbce wartość statystyki F porównujemy z wartością krytyczną $F_{k,\ n\text{-}k\text{-}1,1\text{-}\alpha}$, gdzie w indeksie dolnym *α oznacza przyjęty poziom istotności*, $k = df_{SSR}$ oraz $n\text{-}k\text{-}1 = df_{SSE}$. Gdy w próbce[6] $F = F_{obs} \geq F_{k,\ n\text{-}k\text{-}1,\ 1\text{-}\alpha}$ (co oznacza, że F_{obs} wpadła do zbioru krytycznego $W = \langle F_{k,n-k-1,1-\alpha}, +\infty)$, wtedy odrzucamy hipotezę H_0 na rzecz hipotezy alternatywnej H_1 i wnioskujemy o właściwym doborze zmiennych objaśniających. Oznacza to, że grupa zmiennych X_i (i = 1,2,...k) istotnie statystycznie wpływa na zmienność Y, a dokładnie rzecz ujmując, grupa zmiennych X_i (i = 1,2,...k) (tzn.

[6] Jeśli to będzie jasne, to zamiast pisać w próbce F^{obs}, będziemy pisać po prostu F.

przynajmniej jedna z nich) wpływa istotnie statystycznie na zależność od ich wartości teoretycznych średnich warunkowych $\hat{Y}_i$.

W przeciwnym przypadku, tzn. gdy w próbce $F < F_{k,n-k-1,1-\alpha}$, wtedy nie mamy podstaw aby odrzucić hipotezę zerową H_0, o braku korelacji pomiędzy zmienną zależną Y, a całą grupą czynników X_i ($i = 1,2,...k$).

Alternatywnym, na ogół w skrypcie stosowanym sposobem weryfikacji hipotezy H_0 jest obliczenie *empirycznego poziomu istotności* (tzw. *p*-value), określonego jako prawdopodobieństwo [9]:

$$p = P(F \geq F^{obs}) . \qquad (3\text{-}3\text{-}3.48)$$

Wartość p jest podana przez pole pod krzywą rozkładu zmiennej losowej $F_{k,n-k-1}$, na prawo od wartości F^{obs} będącej obserwowaną w próbce wartością statystyki F. W przypadku gdy $p \leq \alpha$, wtedy odrzucamy hipotezę H_0 na rzecz hipotezy alternatywnej H_1, natomiast gdy $p > \alpha$, nie mamy podstaw do odrzucenia hipotezy zerowej H_0.

Uwaga. Przypadek wnioskowania, że co najmniej jedna ze zmiennych objaśniających jest w modelu zbędna (statystycznie nieistotna) i należy ją wyeliminować z równania regresji, ustalając tym samym nowy skład zmiennych objaśniających, pozostawiamy na później (Rozdział 4-1-2-2, Rozdział 6).

A. Rozdział 4: Wielomianowa analiza regresji.

Analiza za pomocą regresji wielomianowej jest stosowana w przypadku, gdy zmienna objaśniana jest co prawda zależna tylko od jednej zmiennej objaśniającej, jednak model regresji liniowej może nie być dokładny w wymaganym stopniu. Wówczas do liniowego modelu regresji można dodać zmienną wyższego rzędu (X^2, X^3, itd.) tak, że model ma postać:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \ldots + \beta_k X^k + E \quad . \qquad (4.1)$$

Następnie, w równaniu modelu zamieniamy zmienne wyższego rzędu na nowe zmienne postaci: $X^2=X_2$, $X^3=X_3$, …, $X^k=X_k$, otrzymując:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + E \quad . \qquad (4.2)$$

W powyższym modelu zmienne X_2, X_3, …, X_k nie są dowolnymi zmiennymi liczbowymi, tylko funkcjami zmiennej podstawowej X: $X_i=X^i$.

Parametry strukturalne $\beta_0, \beta_1, \cdots \beta_k$ modelu (4.2) są już współczynnikami modelu wielorakiej regresji liniowej.

Uwaga: Model regresji wielomianowej (4.1) niesie za sobą trudności obliczeniowe polegające na tym, że w modelu (4.2) występuje silna korelacja pomiędzy zmiennymi X_1, X_2, …, X_k.

Najprostszym modelem regresji wielomianowej jest model kwadratowy:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + E \ . \qquad (4.3)$$

Rozdział 4-1. Metody obliczania parametrów strukturalnych modelu wielomianowego.

Rozdział 4-1-1. Procedura najmniejszych kwadratów dla modelu parabolicznego.

Procedura najmniejszych kwadratów dla modelu wielomianowego ma na celu minimalizację odchyleń wartości empirycznych od wartości na krzywej regresji. Rozważmy model paraboliczny:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 \qquad (4\text{-}1\text{-}1.4)$$

w którym estymatory $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ wyznaczymy metodą najmniejszych kwadratów.

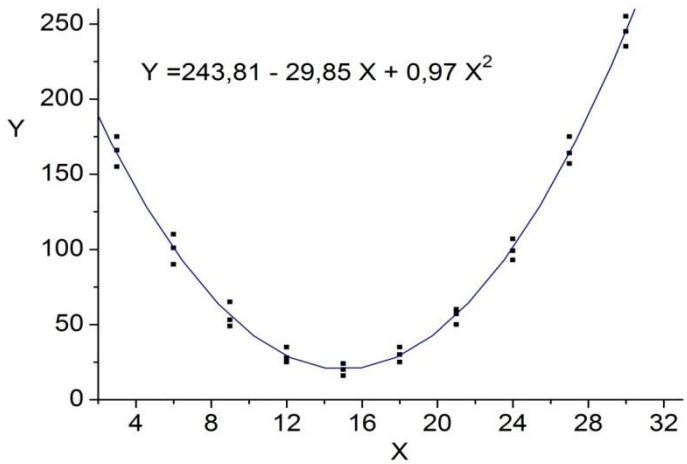
Uwaga 4-1-1.1. Należy pamiętać, że stosując MNK w minimalizacji sumy kwadratów reszt modelu parabolicznego:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2)^2 \quad . \qquad (4\text{-}1\text{-}1.5)$$

równanie (4-1-1.4) traktujemy jako równanie liniowej regresji wielorakiej:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \ , \qquad (4\text{-}1\text{-}1.6)$$

gdzie $X_1=X$, $X_2=X^2$ tak, że to wartości nowych zmiennych X_1 oraz X_2 są wprowadzone jako układ danych do analizy MNK. Pamiętając o tym, będziemy stosowali zapis wielomianowy (4-1-1.4) zamiast (4-1-1.6).



Wykres 4-1.1. Przykład dopasowanie linii regresji modelu parabolicznego do przykładowych danych empirycznych.

Rozdział 4-1-2. Testy dla regresji wielomianowej (na przykładzie modelu parabolicznego).

Rozdział 4-1-2-1. Test istotności modelu regresji wielomianowej.

Hipoteza zerowa H_0: Nie istnieje zależność korelacyjna zmiennej zależnej Y od grupy zmiennych X i X^2. Tzn. nie istnieje istotna statystycznie regresja oparta na zmiennych X i X^2.

Hipotezę tą można sformułować następująco:

$$H_0: \beta_1=\beta_2=0. \qquad (4\text{-}1\text{-}2\text{-}1.7)$$

Podobnie jak w Rozdziale 4, aby zweryfikować tę hipotezę, korzystamy ze statystyki F:

$$F = \frac{MSR}{MSE} \quad , \qquad (4\text{-}1\text{-}2\text{-}1.8)$$

która przy prawdziwości hipotezy zerowej H_0 (4-1-2-1.7) ma rozkład F-Snedecora z liczbą stopni swobody licznika k i mianownika n-k-1 (gdzie dla modelu parabolicznego k = 2). Otrzymaną w próbce wartość F^{obs} porównujemy, dla danego poziomu istotności α, z wartością krytyczną $F_{k,n-k-1,1-\alpha}$ rozkładu F-Snedecora.

Uwaga. Ponieważ badana hipoteza dotyczy ogólnego braku zależności korelacyjnej zmiennej objaśnianej, dlatego statystykę F nazywamy *statystyką ogólną*, lub statystyką dla *testu ogólnego* (w odróżnieniu od statystyki F_p wprowadzonej poniżej dla testów częściowych).

Jako ilościową miarę dokładności dopasowania modelu do danych empirycznych (dokładności modelu), możemy dodatkowo obliczyć współczynnik determinacji R^2 ((3-3-2.43) wraz z Uwagą 5-1-1.1):

$$R^2_{(\text{modelu parabolicznego})} = \frac{\text{SSY - SSE}_{(\text{modelu parabolicznego})}}{\text{SSY}} \qquad (4\text{-}1\text{-}2\text{-}1.9)$$

Rozdział 4-1-2-2.Test celowości dodawania zmiennej objaśniającej wyższego stopnia.

Aby sprawdzić celowość rozbudowy modelu wielomianowego stawiamy następującą hipotezę zerową:

H_0: „Model wyższego rzędu nie dopasowuje się istotnie lepiej do danych empirycznych” lub
H_0: β_{k+1}=0, dla rozszerzenia modelu o k-zmiennych do modelu z k'= k + 1 zmiennymi.

Powyższa hipoteza zerowa oznacza, że „dodanie wyższego stopnia zmiennej objaśniającej do modelu nie zmienia znacząco predykcji zmiennej Y w porównaniu do modelu niższego rzędu”.

W przypadku tym obliczamy w próbce wartość statystyki częściowej (indeks p) F-Snedecora zadaną wzorem [1]:

$$F_p \equiv F(X^{k+1}|X, X^2,..., X^k) = \frac{(SSR_{(dla\,k+1)} - SSR_{(dla\,k)})/1}{MSE_{(dla\,k+1)}} = \frac{SS_{\text{dodanej zmiennej}}/1}{MSE_{(dla\,k+1)}} \ . \qquad (4\text{-}1\text{-}2\text{-}2.10)$$

W przypadku rozszerzenia modelu liniowego (k = 1) do parabolicznego (k'=k +1 = 2), otrzymaną z próbki wartość powyższej statystyki porównujemy z wartością krytyczną $F_{1,\,n\text{-}3,\,1\text{-}\alpha}$ rozkładu F-Snedecora dla stopni swobody licznika 1 oraz stopni swobody mianownika $n - 1 - k' = n - 3$ (która jest liczbą stopni swobody sumy kwadratów reszt SSE modelu parabolicznego).

Liczbę stopni swobody licznika równą 1 otrzymujemy z odjęcia od liczby stopni swobody dla sumy kwadratów $SSR_{\text{modelu parabolicznego}}$ modelu parabolicznego (tzn. k + 1), liczbę stopni swobody dla sumy kwadratów $SSR_{\text{modelu liniowego}}$ modelu liniowego (tzn. k).

Uwaga: Licznik statystyki (5-8) jest miarą poprawy dopasowania się do danych empirycznych modelu z (k + 1) zmiennymi w stosunku do modelu z k zmiennymi objaśniającymi.

Uwaga: Alternatywny sposób polega na wyznaczeniu statystyki t-Studenta:

$$t = \hat{\beta}_{k+1} / S(\hat{\beta}_{k+1}), \qquad (4\text{-}1\text{-}2\text{-}2.11)$$

gdzie $S(\hat{\beta}_{k+1})$ jest estymatorem odchylenia standardowego parametru strukturalnego β_{k+1}. Oznacza to, że testowana hipoteza o nieistotności rozszerzenia np. modelu liniowego do parabolicznego (rozważanego w celu poprawy dopasowania do danych empirycznych), jest *w tym przypadku* równoważna testowi na nieistotność różnicy od zera wartości **ostatniego** estymatora $\hat{\beta}_{k+1}$ parametru strukturalnego, co odpowiada hipotezie zerowej:

$$H_0\text{: } \beta_{k+1} = 0\,. \qquad (4\text{-}1\text{-}2\text{-}2.12)$$

Otrzymaną wartość statystyki t porównujemy z wartościami krytycznymi rozkładu t-Studenta dla stopni swobody sumy kwadratów reszt SSE modelu wyższego.

Okazuje się, że (wyjątkowo) w przypadku testu omawianej hipotezy zerowej H_0: $\beta_{k+1} = 0$, *czyli testu dla ostatniego współczynnika kierunkowego, zachodzi* [1]:

$$t^2 = F_p\ . \qquad (4\text{-}1\text{-}2\text{-}2.13)$$

Rozdział 4-1-2-3. Test braku dopasowania zastosowanego modelu (Lack-of-fit Test).

Test braku dopasowania przeprowadza się w celu nabycia informacji, czy zaproponowany model jest wystarczająco dokładny. Przeprowadzenie odpowiedniego testu polega na porównaniu zaproponowanego modelu z **modelem pełnym** (zawierającym pozostałą część zmiennych objaśniających pominiętych w proponowanym modelu).

Rozważana hipoteza zerowa jest więc postaci:

$$H_0\text{: } \beta_j = 0, \text{ dla wszystkich } j = k+1, k+2, \ldots, m \qquad (4\text{-}1\text{-}2\text{-}3.14)$$

gdzie k jest stopieniem wielomianu użytego w modelu, a m najwyższym możliwym stopieniem wielomianu dla danej zmiennej objaśniającej.

Prawdziwość tej hipotezy oznaczałaby, że model podstawowy daje dobre dopasowanie funkcji regresji do danych eksperymentalnych.

Statystyka (częściowa) F_p dla testu tej hipotezy ma następującą postać[1]:

$$F_p = F(X^{k+1}, X^{k+2}, \cdots, X^m | X, X^2, \cdots, X^k)$$

$$= \frac{[SSR(X, X^2, \cdots, X^k, X^{k+1}, ..., X^m) - SSR(X, X^2, \cdots, X^k)]/df_{LOF}}{SSE(X, X^2, \cdots, X^k, X^{k+1}, ..., X^m)/df_{PE}} \qquad (4\text{-}1\text{-}2\text{-}3.15)$$

$$= \frac{MS_{LOF}}{MS_{PE}},$$

gdzie:

$$MS_{PE} = SSE(X, X^2, ..., X^m) / df_{PE} \qquad (4\text{-}1\text{-}2\text{-}3.16)$$

jest sumą kwadratów ***czystych*** reszt *SSE* podzieloną przez liczbę stopni swobody df_{PE} dla *SSE*.

Reszty czyste są wyznaczone w modelu z maksymalnym stopniem wielomianu *m*. Żaden model nie da więc mniejszej wartości MS_{PE} niż model maksymalny.

W liczniku (4-1-2-3.15) statystyka MS_{LOF} jest dodatkową sumą kwadratów wynikłą z dodania do proponowanego modelu niższego (mniejszego), wszystkich zmiennych wyższego rzędu od X^{k+1} do X^m.

Przy prawdziwości rozważanej hipotezy zerowej (4-1-2-3.14), statystyka (4-1-2-3.15) ma rozkład F-Snedecora ze stopniami swobody licznika: $df_{LOF} = m - k$, oraz mianownika: $df_{PE} = n - 1 - m$.

Uwaga: Licznik statystyki (4-1-2-3.15) jest miarą poprawy dopasowania się do danych empirycznych modelu maksymalnego w stosunku do modelu proponowanego (podstawowego).

Warunki przeprowadzenia testu (4-1-2-3.15) [1].

- Aby można było wykonać test (4-1-2-3.15), należy wyznaczyć średnią MS_{PE}, co (ze względu na mianownik w MS_{PE}, (4-1-2-3.16)) oznacza, że musi zachodzić relacja $df_{PE} = n\text{-}1\text{-}m > 0$. Z zależności tej natychmiast wynika, że minimalna liczba pomiarów (obserwacji), którą należy dokonać wynosi $n > m + 1$. W praktyce przyjmuje się, że $n \geq 10 + m + 1$. Jeszcze inna zasada głosi, że liczba obserwacji zmiennej objaśnianej przypadająca na jedną zmienną objaśniającą, nie może być mniejsza niż 5 (zatem $n \geq 5\,m$).
- Istnieje jeszcze ograniczenie na sumę kwadratów reszt stojącą w liczniku (4-1-2-3.16), $MS_{PE} = SSE_{(X,X^2 \ldots, X^m)} / df_{PE}$.

  Otóż, aby można było wykonać powyższy test (4-1-2-3.15), musi zachodzić nierówność $SSE_{(X,X^2 \ldots, X^m)} > 0$, co w praktyce oznacza, że liczba r_i tzw. *replik* (czyli liczby różnych wartości zmiennej objaśnianej pomniejszona o 1) *dla każdego ustalonego i – tego* zestawu wartości zmiennej objaśniającej, nie może być mniejsza niż 1. Zatem dla konkretnego zestawu wartości zmiennych objaśnianych, liczba obserwacji zmiennej objaśnianej nie może być mniejsza niż 2.

Skoro liczba wszystkich replik wynosi $r = \sum_{i=1}^{l} r_i$ (gdzie l jest liczbą poziomów zmiennej X), to minimalny stopień wielomianu (modelu maksymalnego), który dopasuje się do danych z minimalnym możliwym błędem (czyli błędem czystym MS_{PE}) wynosi $m = n - 1 - r$, gdzie n jest liczbą pomiarów w próbce.

Gdyby liczba replik r była równa 0, to co prawda wielomian stopnia $n - 1$ w modelu maksymalnym dopasował by się do danych empirycznych w próbce w sposób idealny ($SSE_{(X,X^2 \ldots , X^m)} = 0$), jednakże wykonanie testu braku dopasowania byłoby niemożliwe (jak to wynika z postaci (4-1-2-3.15)).

- Można by pomyśleć, że aby $SSE_{(X,X^2 \ldots , X^m)} > 0$, wystarczy mieć jedną replikę dla pewnego, konkretnego zestawu wartości zmiennych objaśnianych. Jednakże gdyby się ograniczyć tylko do tego warunku, to **hipoteza o jednorodności wariancji** zmiennej objaśnianej dla różnych zestawów wartości zmiennych objaśniających (z powodu niewystępowania replik zmiennej objaśnianej dla niektórych poziomów), byłaby w sposób oczywisty odrzucona.

Rozdział 4-2. Stosowanie modeli wielomianowych wyższych rzędów i problemy z tym związane

Modele wyższych rzędów są obliczane w sposób analogiczny do modelu parabolicznego. Ważnym problemem jest celowość zwiększania stopnia krzywoliniowości regresji. Pewnej, czysto statystycznej odpowiedzi dostarcza nam wspomniany powyżej test braku dopasowania.

Z drugiej strony, chociaż wraz ze wzrostem stopnia wielomianu wzrasta dokładność dopasowania się modelu do danych empirycznych (w próbce!), jednakże wzrasta też ilość ekstremów lokalnych krzywej regresji, co oznacza zmniejszenie się funkcjonalności modelu zarówno na skutek komplikacji powstałych w obliczeniach jak i trudności interpretacyjnych. Np. w badaniach w obszarze ekonomii jesteśmy zazwyczaj zainteresowani modelami, w których występuje monotoniczność stosowanych wielomianów. Dodatkowo spada dokładność predykcji przyszłych wartości zmiennej objaśnianej dla modelu z wyższym stopniem wielomianu.

Rozdział 4-3. Wielomiany ortogonalne.

Do tej pory mieliśmy do czynienia z *wielomianami zwykłymi* tzn. każda z niezależnych zmiennych była zadana wielomianem zwyczajnym będącym sumą jednomianów typu X^i. W tym punkcie wprowadzone zostaną wielomiany ortogonalne.

Podstawowym powodem stosowania wielomianów ortogonalnych jest niwelacja współliniowości zmiennych objaśniających. Niestety nabycie tej własności łączy się z komplikacją struktury modelu.

Mając wielomian zwyczajny zmiennych X, X_1, X_2, …, X_k , wprowadzimy ortogonalne zmienne wielomianowe, które składają się z liniowych kombinacji jednomianów zwyczajnych.

Uzyskujemy następujące liniowe kombinacje [1]:

$$
\begin{aligned}
X_1^* &= \alpha_{01} + \alpha_{11} X \\
X_2^* &= \alpha_{02} + \alpha_{12} X + \alpha_{22} X^2 \\
&\vdots \\
X_k^* &= \alpha_{0k} + \alpha_{1k} X + \alpha_{2k} X^2 + \cdots \alpha_{kk} X^k
\end{aligned}
\tag{4-3.17}
$$

gdzie: α_{ij} -stałe, które są tak dobrane aby zmienne X_i^* były ze sobą parami nieskorelowane, tzn. cov(X_i^*, X_j^*) = 0, dla każdej pary indeksów $i \neq j$.

Otrzymane ortogonalne zmienne wielomianowe X_i^* są nowymi zmiennymi stosowanymi do prognozy zmiennej objaśnianej.

Poprzez transformację odwrotną, również wielomiany zwyczajne możemy zapisać w postaci liniowych kombinacji wielomianów ortogonalnych:

$$
\begin{aligned}
X &= \beta_{01} + \beta_{11} X_1^* \\
X^2 &= \beta_{02} + \beta_{12} X_1^* + \beta_{22} X_2^* \\
&\vdots \\
X^k &= \beta_{0k} + \beta_{1k} X_1^* + \cdots + \beta_{kk} X_k^*
\end{aligned}
\tag{4-3.18}
$$

gdzie: β_{ij} - stałe.

Nie tracąc informacji możemy zapisać początkowy model wielomianowy:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + E \tag{4-3.19}$$

w postaci:

$$Y = \beta_0^* + \beta_1^* X_1^* + \beta_2^* X_2^* + \cdots + \beta_k^* X_k^* + E \tag{4-3.20}$$

Uwaga 1: Zysk z zastosowania wielomianów ortogonalnych jest oczywisty: wielomiany zwyczajne są ze sobą mocno skorelowane, a wielomiany ortogonalne są parami nieskorelowane.

Uwaga 2. **O niezmienniczości niektórych statystyk**: Chociaż parametry i czynniki (zmienne objaśniające) dla modelu ze zmiennymi zwyczajnymi i ortogonalnymi mają różne interpretacje, to można wykazać, że kwadrat wielokrotnego współczynnika korelacji (współczynnik determinacji) R^2 dla zmiennej zależnej oraz statystyka F w teście ogólnym są dla obu regresji **takie same**.

Nawet niektóre statystyki częściowe F_p , a mianowicie te, związane z dodawaniem zmiennej najwyższego stopnia na końcu, są dla obu regresji **takie same**, co oznacza, że zachodzi:

$$F_p(X_{i+1}^* \mid X_i^*, X_{i-1}^*, ..., X_1^*) = F_p(X^{i+1} \mid X^i, X^{i-1}, ..., X^1). \tag{4-3.21}$$

Minusem ortogonalizacji jest skomplikowany związek pomiędzy zmiennymi początkowymi (których znaczenie jest na ogół jasne), a zmiennymi ortogonalnymi.

Reasumując wielomiany ortogonalne posiadają dwie podstawowe własności [1]:

1. Zawierają tę sama informacje, co wielomiany zwyczajne. Własność ta daje nam możliwość badania zależności za pomocą modeli ortogonalnych wielomianów, zgodnie z Uwagą 1.
2. Wielomiany ortogonalne nie są ze sobą skorelowane. Własność ta oznacza całkowitą likwidację korelacji (i współliniowości) zmiennych objaśniających.

W przypadku testowania z użyciem statystyki F_p, posługiwanie się wielomianami ortogonalnymi daje większą pewność dokładności obliczeń (ze względu na brak korelacji pomiędzy czynnikami) niż w przypadku wielomianów zwykłych.

Wstępna uwaga o selekcji wprzód i wstecz: Aby otrzymać właściwy model, należy zastosować test częściowy F_p począwszy od wielomianu najwyższego stopnia i schodzić do wielomianów niższych stopni. Testowanie kończymy wówczas, gdy wartość statystyki częściowej F_p dla testowanego wielomianu jest istotna statystycznie, tzn. na tyle duża, że wpada w obszar krytyczny, wskazując na wystarczająco dużą istotność badanego modelu.

Czasami w praktyce idzie się od dołu, dodając kolejne zmienne wyższego stopnia (X^i gdy posługujemy się wielomianami zwykłymi, bądź X_i^*, gdy posługujemy się wielomianami ortogonalnymi) i czekając, aż któreś rozszerzenie modelu okaże się statystycznie nieistotne (tzn. dopasuje się do danych empirycznych w sposób nieistotnie lepszy niż najbliższy model niższego stopnia). Jednakże procedura taka może prowadzić do mylnych wniosków, o czym będzie mowa poniżej (Rozdział 4-4).

Transformacja wielomianów zwyczajnych do wielomianów ortogonalnych.

Transformację wielomianów zwyczajnych do wielomianów ortogonalnych przeprowadza się przy użyciu tabeli przeliczania wielomianów Tabela 5-1.1 [1].

Tabela ta może zostać zastosowana tylko, jeśli [1]:

- Kolejne wartości pierwotnej zmiennej objaśniającej X są jednakowo od siebie oddalone.
- Ta sama liczba obserwacji (więc i replik) pojawia się dla każdego poziomu (wariantu) l zmiennej X.

Gdy warunki te nie są spełnione wówczas Tabela 5-1.1 nie może być użyta. Alternatywą dla tej metody jest posłużenie się programami komputerowymi do przeliczania wielomianów, np. w Systemie SAS używając do tego celu funkcji ORPOL znajdującej się w procedurze SAS PROC IML.

Tabela 5-1.1 Przeliczania wielomianów zwyczajnych na ortogonalne [1] (*l* - liczba poziomów zmiennej).

l	STOPIEŃ WIELOMIANU	X^*										
		1	2	3	4	5	6	7	8	9	10	$\sum_{i=1}^{l} p_i^2$
3	Pierwszy	-1	0	1								2
	Drugi	1	-2	1								6
4	Pierwszy	-3	-1	1	3							20
	Drugi	1	-1	-1	1							4
	Trzeci	-1	3	-3	1							20
5	Pierwszy	-2	-1	0	1	2						10
	Drugi	2	-1	-2	-1	2						14
	Trzeci	-1	2	0	-2	1						10
	Czwarty	1	-4	6	-4	1						70
6	Pierwszy	-5	-3	-1	1	3	5					70
	Drugi	5	-1	-4	-4	-1	5					84
	Trzeci	-5	7	4	-4	-7	5					180
	Czwarty	1	-3	2	2	-3	1					28
	Piąty	-1	5	-10	10	-5	1					252
7	Pierwszy	-3	-2	-1	0	1	2	3				28
	Drugi	5	0	-3	-4	-3	0	5				84
	Trzeci	-1	1	1	0	-1	-1	1				6
	Czwarty	3	-7	1	6	1	-7	3				154
	Piąty	-1	4	-5	0	5	-4	1				84
	Szósty	1	-6	15	-20	15	-6	1				924
8	Pierwszy	-7	-5	-3	-1	1	3	5	7			168
	Drugi	7	1	-3	-5	-5	-3	1	7			168
	Trzeci	-7	5	7	3	-3	-7	-5	7			264
	Czwarty	7	-13	-3	9	9	-3	-13	7			616
	Piąty	-7	23	-17	-15	15	17	-23	7			2184
	Szósty	1	-5	9	-5	-5	9	-5	1			264
	Siódmy	-1	7	-21	35	-35	21	-7	1			3432
9	Pierwszy	-4	-3	-2	-1	0	1	2	3	4		60
	Drugi	28	7	-8	-17	-20	-17	-8	7	28		2772
	Trzeci	-14	7	13	9	0	-9	-13	-7	14		990
	Czwarty	14	-21	-11	9	18	9	-11	-21	14		2002
	Piąty	-4	11	-4	-9	0	9	4	-11	4		468
	Szósty	4	-17	22	1	-20	1	22	-17	4		1980
	Siódmy	-1	6	-14	14	0	-14	14	-6	1		858
	Ósmy	1	-8	28	-56	70	-56	28	-8	1		12870
10	Pierwszy	-9	-7	-5	-3	-1	1	3	5	7	9	330
	Drugi	6	2	-1	-3	-4	-4	-3	-1	2	6	132
	Trzeci	-42	14	35	31	12	-12	-31	-35	-14	42	8580
	Czwarty	18	-22	-17	3	18	18	3	-17	-22	18	2860
	Piąty	-6	14	-1	-11	-6	6	11	1	-14	6	780
	Szósty	3	-11	10	6	-8	-8	6	10	11	3	660
	Siódmy	-9	47	-86	92	56	-56	-42	86	-47	9	29172
	Ósmy	1	-7	20	-28	14	14	-28	20	-7	1	2860
	Dziewiąty	-1	9	-36	84	-126	126	-84	36	-9	1	48620

Założenie: Niech układ danych będzie następujący: zmienna objaśniająca ma l poziomów (wariantów), a każdemu z nich odpowiada *taka sama liczba obserwacji* zmiennej objaśnianej $\frac{n}{l}$, z liczbą replik *dla każdego poziomu* równą $\frac{n}{l}-1$. Wtedy odpowiedni stopień wielomianu dla modelu maksymalnego jest równy $m = n - 1 - r = n - 1 - l\,(\frac{n}{l} - 1) = l - 1$, gdzie $(\frac{n}{l} - 1)$ jest liczbą replik dla jednego poziomu. Zatem liczba zmiennych ortogonalnych, które należy wziąć pod uwagę wynosi $l - 1$.

Układ Tabeli 5-1.1 jest taki, że kolumna po jej prawej stronie zawiera odpowiednią dla każdej zmiennej ortogonalnej wartość $\sum_{i=1}^{l} p_i^2$ (l – liczba wariantów zmiennej objaśniającej), która jest sumą kwadratów odchyleń wartości zmiennej ortogonalnej od jej wartości średniej (**równej zero**, jak to widać z powyższej Tabeli). Podzielenie zmiennych ortogonalnych przez odpowiednią dla każdej z nich wartość odchylenia standardowego $\sqrt{\sum_{i=1}^{l} p_i^2}$, pozwala na przejście do zmiennych *ortonormalnych*, mających **odchylenie standardowe**, dla każdej z nich, **równe** 1.

W związku z powyższą uwagą, otrzymane wyniki w analizie regresji ze zmiennymi zortonormalizowanymi mają dwie cechy:

- Ulepszona jest numeryczna dokładność poprzez uniknięcie problemu skalowania (o skalowaniu nieco dalej w Rozdziale 5-5).
- Szacowane błędy standardowe wszystkich oszacowywanych współczynników regresji są równe, co upraszcza porównywanie i interpretację współczynników regresji.

Uwaga. Innym, aczkolwiek mniej skutecznym sposobem na zmniejszenie korelacji pomiędzy zmiennymi objaśniającymi jest transformacja zmiennych przez scentrowanie opisane w Rozdziale 5-5 (punkt Ad.4) oraz w Rozdziale 6 „Wybór najlepszego modelu regresji”.

Przykład. Dla zilustrowania Tablicy 5-1.1 przeliczmy wartości zmiennych ortogonalnych w najprostszym przypadku wielomianu drugiego stopnia (z liczbą poziomów zmiennej X oraz X^2 równą $l = 3$). Liczba wszystkich obserwacji w próbce wynosi n. Załóżmy, że liczba obserwacji jest taka sama w każdym wariancie zmiennej X (zatem i w każdym wariancie zmiennej X^2) i wynosi n/l dla każdego wariantu.

Zgodnie z założeniem potrzebnym przy konstrukcji Tablicy 5-1.1, przyjmijmy, że zmienna X ma równo rozstawione warianty, np:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} . \tag{4-3.22}$$

Wtedy:

$$X^2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 9 \end{pmatrix} . \tag{4-3.23}$$

Zgodnie z (5-10a) mamy:

$$\begin{aligned} X_1^* &= \alpha_{01} + \alpha_{11} X \\ X_2^* &= \alpha_{02} + \alpha_{12} X + \alpha_{22} X^2 , \end{aligned} \tag{4-3.24}$$

gdzie $\alpha_{11} \neq 0$, $\alpha_{22} \neq 0$.

Warunek nałożony na zmienne ortogonalne ma postać: *cov*(X_1^*, X_2^*) = 0. Oznacza on, że przy braku innego ograniczenia, *zmienne* X_1^* *i* X_2^* *mogą być wyznaczone jedynie z dokładnością do różnych od zera multiplikatywnych stałych oraz z dokładnością do stałych addytywnych, które można przyjąć jako np. średnie zmiennych* X_1^* *i* X_2^*.

Istotnie:

$$cov(\, a X_1^*, b X_2^*) = a\, b\, cov(\, X_1^*, X_2^*) = 0 \tag{4-3.25}$$

oraz:

$$cov(\, X_1^* - c,\ X_2^* - d\,) = cov(\, X_1^*, X_2^*) = 0. \tag{4-3.26}$$

Zatem jeśli zmienne X_1^* i X_2^* są ortogonalne, to ortogonalne są również zmienne $a X_1^*$ i $b X_2^*$, ($a \neq 0,\ b \neq 0$), oraz zmienne $X_1^* - c$ i $X_2^* - d$ (stałe *c* i *d* są dowolne).

Warunek *cov*(X_1^*, X_2^*) = 0, można zapisać jako:

$$\begin{aligned} cov(\, X_1^*, X_2^*) &= cov(\, \alpha_{01} + \alpha_{11} X,\ \alpha_{02} + \alpha_{12} X + \alpha_{22} X^2) = cov(\, \alpha_{11} X,\ \alpha_{12} X + \alpha_{22} X^2) \\ &= \alpha_{11}\ \alpha_{12}\ cov(\, X, X\,) + \alpha_{11}\ \alpha_{22}\ cov(\, X,\ X^2) \\ &= \alpha_{11} (\, \alpha_{12}\ cov(\, X, X\,) + \alpha_{22}\ cov(\, X,\ X^2)\,) = 0. \end{aligned}$$

Ponieważ z założenia $\alpha_{11} \neq 0$, $\alpha_{22} \neq 0$, zatem powyższa równość jest spełniona wtedy i tylko wtedy, gdy:

$$\alpha_{12} = -\alpha_{22} \frac{\operatorname{cov}(X,\ X^2)}{\operatorname{cov}(X, X)} . \tag{4-3.27}$$

W związku z założeniem (dla naszego przykładu), że dla każdego *i*-tego wariantu, liczba obserwacji jest równa *n*/*l*, mamy w pobranej próbce:

$$\overline{X} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{l}\frac{n}{l}x_i = \frac{1}{l}\sum_{i=1}^{l} x_i = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}[1+2+3] = 2 ,$$

$$\overline{X^2} = \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{l}\frac{n}{l}x_i^2 = \frac{1}{l}\sum_{i=1}^{l} x_i^2 = \frac{1}{3}\sum_{i=1}^{3} x_i^2 = \frac{1}{3}\left[1^2+2^2+3^2\right] = \frac{14}{3} ,$$

$$\operatorname{cov}(X, X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x}) = \frac{1}{n}\sum_{i=1}^{l}\frac{n}{l}(x_i-\overline{x})^2 = \frac{1}{l}\sum_{i=1}^{l}(x_i-\overline{x})^2$$

$$= \frac{1}{3}\sum_{i=1}^{3}(x_i-\overline{x})^2 = \frac{1}{3}\left[(1-2)^2+(2-2)^2+(3-2)^2\right] = \frac{2}{3} ,$$

$$\operatorname{cov}(X, X^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(x_i^2-\overline{x^2}) = \frac{1}{n}\sum_{i=1}^{l}\frac{n}{l}(x_i-\overline{x})\,(x_i^2-\overline{x^2}) = \frac{1}{l}\sum_{i=1}^{l}(x_i-\overline{x})\,(x_i^2-\overline{x^2})$$

$$= \frac{1}{3}\sum_{i=1}^{3}(x_i-\overline{x})(x_i^2-\overline{x^2}) = \frac{1}{3}\left[(1-2)(1^2-\frac{14}{3})+(2-2)(2^2-\frac{14}{3})+(3-2)(3^2-\frac{14}{3})\right] = \frac{8}{3}.$$

Korzystając z powyższych dwóch zależności oraz ze związku (P_5-2), otrzymujemy:

$$\alpha_{12} = -4\,\alpha_{22} \neq 0 .$$

Możemy teraz związek (4-3.24) zapisać następująco:

$$X_1^* = \alpha_{01} + \alpha_{11}X$$
$$X_2^* = \alpha_{02} - 4\,\alpha_{22}\,X + \alpha_{22}X^2 ,$$

skąd dla średnich otrzymujemy:

$$\overline{X_1^*} = \alpha_{01} + \alpha_{11}\overline{X}$$
$$\overline{X_2^*} = \alpha_{02} - 4\,\alpha_{22}\,\overline{X} + \alpha_{22}\overline{X^2} .$$

Wykorzystajmy powyższe zależności do wycentrowania zmiennych X_1^* i X_2^*:

$$X_1^* - \overline{X_1^*} = \alpha_{01} + \alpha_{11}X - (\alpha_{01} + \alpha_{11}\overline{X}) = \alpha_{11}(X - \overline{X}) = \alpha_{11}(X-2)$$
$$= \alpha_{11}\begin{pmatrix}1-2\\2-2\\3-2\end{pmatrix} = \alpha_{11}\begin{pmatrix}-1\\0\\1\end{pmatrix} , \tag{4-3.28}$$

$$X_2^* - \overline{X_2^*} = \alpha_{02} - 4\,\alpha_{22}\,X + \alpha_{22}X^2 - (\alpha_{02} - 4\,\alpha_{22}\,\overline{X} + \alpha_{22}\overline{X^2})$$
$$= -4\,\alpha_{22}\,(X - \overline{X}) + \alpha_{22}(X^2 - \overline{X^2}) = \alpha_{22}[-4(X-2) + (X^2 - \frac{14}{3})] \qquad (4\text{-}3.29)$$
$$= \alpha_{22}\left[-4\begin{pmatrix}1-2\\2-2\\3-2\end{pmatrix} + \begin{pmatrix}1-14/3\\4-14/3\\9-14/3\end{pmatrix}\right] = \alpha_{22}\,\frac{1}{3}\begin{pmatrix}1\\-2\\1\end{pmatrix} \quad .$$

Powyżej pokazaliśmy, że zmienne X_1^* i X_2^* mogą być określone jedynie z dokładnością do niezerowej multiplikatywnej stałej, oraz dowolnej stałej addytywnej. Przyjmijmy więc, że $\alpha_{11} = 1$, $\alpha_{22} = 3$.

Natomiast stałe addytywne α_{01} i α_{02} dobierzmy tak, aby $\overline{X_1^*} = 0$ oraz $\overline{X_2^*} = 0$, tzn.

$$\overline{X_1^*} = \alpha_{01} + \alpha_{11}\overline{X} = \alpha_{01} + 2 = 0$$
$$\overline{X_2^*} = \alpha_{02} - 4\,\alpha_{22}\,\overline{X} + \alpha_{22}\overline{X^2} = \alpha_{02} - 4\cdot 3\cdot 2 + 3\cdot\frac{14}{3} = 0\,, \qquad (4\text{-}3.30)$$

skąd otrzymujemy $\alpha_{01} = -2$, $\alpha_{02} = 10$.

Ostatecznie z (4-3.28) oraz (4-3.29) dostajemy wartości zmiennych X_1^* i X_2^*, które pojawiają się w Tablicy 5-1.1, a mianowicie:

$$X_1^* = \begin{pmatrix}x_{11}^*\\x_{12}^*\\x_{13}^*\end{pmatrix} = \begin{pmatrix}-1\\0\\1\end{pmatrix} \quad , \qquad X_2^* = \begin{pmatrix}x_{21}^*\\x_{22}^*\\x_{23}^*\end{pmatrix} = \begin{pmatrix}1\\-2\\1\end{pmatrix} \quad , \quad c.n.d. \qquad (4\text{-}3.31)$$

Rachunki dla zmiennych ortogonalnych w wielomianach wyższego stopnia, chociaż bardziej żmudne, przebiegałyby analogicznie.

Rozdział 4-4. Strategie wyboru modelu wielomianowego.

Podczas wyboru modelu wielomianowego należy dopasowywać model poprzez odejmowanie zmiennych mniej znaczących (procedurą eliminacji wstecz omówioną poniżej, Rozdział 6-3), tzn. należy wybór zacząć od modelu możliwie najpełniejszego i upraszczać go do momentu, w którym okaże się, iż wszystkie pozostawione zmienne mają istotny wpływ na dokładność dopasowania się linii regresji modelu do danych empirycznych.

Powód możliwego problemu przy stosowaniu selekcji wprzód: Jak wspomnieliśmy poprzednio, błędnym podejściem do wyboru zmiennych w przypadku regresji wielomianowej jest konstruowanie modelu poprzez *dodawanie* zmiennych wyższego rzędu do modelu mniej skomplikowanego.

Strategia taka (tzn. strategia selekcji wprzód), może prowadzić do wyznaczenia modelu błędnego (z wyjątkiem jej „ostrożnego stosowania"). Strategia selekcji wprzód może bowiem doprowadzić do wybrania modelu, w którym zostanie pominięta istotna zmienna, czego przyczyną jest to, że średni kwadrat reszt *MSE* składnika błędu modelu (będący estymatorem wariancji σ_E^2 składnika losowego), występuje w mianowniku testów częściowych F [1]:

$$F_p = F(X^i \big| X, X^2, \cdots, X^{i-1}) = \frac{[SSR(X^i \mid X, X^2, \cdots, X^{i-1})]/1}{MSE(X, X^2, \cdots, X^i)}, \qquad (4\text{-}4.32)$$

w związku z czym, *jeśli stopień wielomianu jest za niski, prowadzi to do dużej wartości dla MSE* w pobranej próbce, a w konsekwencji do małej wartości statystyki F_p, wiec wartość ta (wraz z wprowadzonym do modelu parametrem strukturalnym dla dodanej nowej zmiennej) może zostać uznana za nieistotną statystycznie.

Rozdział 4-5. Przeprowadzenie wstępnej diagnostyki modelu.

Mając wyszczególniony model musimy przejść do analizy modelu. Analiza modelu:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + E, \qquad (4\text{-}5.33)$$

ma po pierwsze (1) przetestować model pod względem jego przydatności do badania aktualnie dostępnych danych oraz po drugie (2) wypowiedzieć się na temat jego przydatność do badania innych zbiorów danych tej samej populacji lub populacji o podobnym charakterze. Problem (2) zostanie przedstawiony w Rozdziale 6. Poniżej odnieśmy się do punktu (1).

Diagnostyka modelu składa się z następujących części:

1. Analiza reszt:
   a) analiza ogólna reszt
   b) analiza wartości ekstremalnych (skrajnych) reszt.
2. Analiza współliniowości zmiennych objaśniających.
3. Analiza skalowania.

Ad.1a Analiza ogólna reszt.

Dokładniejszą analizę reszt klasycznego modelu regresji przedstawimy w Rozdziałach od 10 do 14. Tytułem wprowadzenia podkreślmy, że do tej części badania modelu przystępujemy z założeniem o normalności rozkładu reszt, przy czym wartość średnia reszt wynosi 0, a wariancja jest skończona i stała.
Ilościowa analiza reszt polega na przeprowadzeniu testów dotyczących:
(a) normalności rozkładu składnika losowego (np. testy zgodności χ^2 –Pearsona [2], λ-Kołmogorowa-(Smirnowa) [2], [9], Rozdziały 15-1 oraz 15-3). Ważnym elementem analizy reszt jest ich analiza graficzna, w której określamy rozkład reszt na wykresie (w SAS'ie „normal probability-probability plot" dostępny w aplikacji Analyst [10] w opcji: Solutions -> Analysis->Analyst->(i po wczytaniu danych, korzystając z „Open By SAS Name" w zakładce File [11]) -> Statistics->Regresion->Simple->Plots->Residual).

(b) jednorodności wariancji dla składnika losowego dla różnych wariantów zmiennych objaśniających (np. test Bartletta [9], Rozdział 16-1-1),

(c) braku autokorelacji składnika losowego (np. test Durbina-Watsona [12] , Rozdział 11-3-1).

Model uznajemy za prawidłowy, gdy po założeniu hipotez zerowych dotyczących powyższych punktów (a), (b) i (c) i po przeprowadzeniu testów nie mamy podstaw do ich odrzucenia.

Ad.1b Analiza wartości ekstremalnych (skrajnych) reszt.

Wartości skrajne są to wartości oddalone od średniej reszt (równej zero) o trzy lub więcej odchylenia standardowe. W przypadku, gdy w zbiorze danych znajdują się takie wartości należy sprawdzić przyczynę tych odchyłek.
Jeżeli przyczyną zaobserwowania skrajnej wartości są błędy w rejestracji obserwacji lub ustawieniu aparatury pomiarowej, wtedy odrzucamy taką obserwację i do dalszej analizy przyjmujemy dane nie wykazujące takich osobliwych zachowań. Postąpilibyśmy tak samo (tzn. odrzucili obserwację), gdybyśmy mieli uprzednio wiedzę o tym, że skrajne dane pojawiły się w próbce na skutek zajścia zdarzeń mało prawdopodobnych.
Natomiast w innych przypadkach, gdy mamy podejrzenie, że duża odchyłka (nie jest następstwem błędów pomiarowych lub wynikiem zajścia rzadkiego zdarzenia ale) jest następstwem własności populacyjnych zmiennej objaśnianej, które są związane z zależnością tej zmiennej od innych czynników, których istnienie zostało zlekceważone przez badacza przy konstrukcji modelu, wtedy zaleca się staranne badanie zjawiska powstania dużych odchyłek. Ich odrzucenie mogłoby bowiem prowadzić do błędnego określenia modelu w populacji (tzn. pominięcia istotnych czynników).

Ad.3 Analiza współliniowości zmiennych objaśniających.

Głównym wskaźnikiem (siły) korelacji pomiędzy dwoma zmiennymi (w tym przypadku zmiennymi objaśniającymi) jest kwadrat współczynnika korelacji $R^2(X_1,X_2)$, który w próbce przyjmuje wartość $r^2(X_1,X_2)$. Oprócz tego oblicza się *współczynnik inflacji wariancji VIF*, którego wartość w próbce jest równa:

$$vif = \frac{1}{1-r^2(X_1, X_2)} \quad . \tag{4-5.34}$$

Współczynniki r^2 oraz vif wskazują na występowanie silnej korelacji (i współliniowości), o ile ich wartości są zbliżone do następujących:

- $r^2 = 1$ dla kwadratu współczynnika korelacji,
- $vif \to \infty$ dla współczynnika inflacji wariancji,

które oznaczają zależności funkcyjne pomiędzy zmiennymi X_1 i X_2. O dużej współliniowości możemy mówić gdy współczynnik *vif* osiągnie wartość = 10.

W przypadku większej liczby zmiennych określa się kwadrat *wielokrotnego współczynnika korelacji* $R_j^2 \equiv R^2(X_j|X_1,\ldots,X_{j-1}, X_{j+1},\ldots,X_k)$, (porównaj (3-2-3.21)), gdzie k jest liczbą zmiennych objaśniających, oraz korespondujący z nim współczynnik inflacji wariancji (nadęcie inflacji w polskim opisie SAS'a) [1]:

$$VIF_j = \frac{1}{1-R_j^2}, \tag{4-5.35}$$

gdzie $j = 1,2,\ldots, k$. Interpretacja tych współczynników jest taka sama jak odpowiadających im współczynników w analizie pary zmiennych.

Oprócz tych dwóch miar współzależności stosuje się również współczynnik tolerancji [1]:

$$Tolerancja_j = \frac{1}{VIF_j} = 1 - R_j^2, \tag{4-5.36}$$

gdzie $j = 1,2,\ldots, k$.

W przypadku współczynnika tolerancji, silne zależności pomiędzy czynnikami, powodują skupienie się wartości tolerancji wokół zera.

Ad.3-1. Związek współczynnika inflacji wariancji z estymatorem wariancji estymatora parametru strukturalnego modelu.

Jest jeden zasadniczy powód, dla którego występowanie współliniowości pomiędzy zmiennymi objaśniającymi powoduje duże problemy w analizie zależności korelacyjnej dla zmiennej objaśnianej. Otóż,

można pokazać, że estymatory wariancji parametrów strukturalnych $\hat{\beta}_j$ są proporcjonalne do VIF_j i mają postać [1] (porównaj 11-1.38):

$$S^2_{\hat{\beta}_j} = c_j \cdot VIF_j \ , \qquad \text{dla} \qquad j = 1, 2, ..., k \ , \tag{4-5.37}$$

gdzie c_j są odpowiednimi współczynnikami zależnymi od danych.

Im większa jest korelacja zmiennej X_j z którąś z pozostałych zmiennych objaśniających, tym bliżej liniowej współzależności jest z nią zmienna X_j [7]. Związek (4-5.37) oznacza zatem, że im bliżej zmienna X_j jest liniowej zależności od chociażby tylko niektórych z pozostałych zmiennych objaśniających (zatem im bardziej zmienna X_j jest mocniej skorelowana z chociażby tylko niektórymi z pozostałych zmiennych objaśniających), to tym większa jest w próbce wartość VIF_j, a wartość współczynnika R_j^2 bliższa jedności i w konsekwencji tym większe w próbkach są wartości estymatora wariancji $S^2_{\hat{\beta}_j}$. To z kolei oznacza, że rozproszenie możliwych wartości estymatora parametru strukturalnego $\hat{\beta}_j$ jest na tyle duże (a odpowiedni dla β_j przedział ufności na tyle szeroki), że gwałtownie spada jakość predykcji modelu. Taką zmienną objaśniającą należałoby z modelu usunąć.

Na koniec rozważmy problem przesunięcia (*intercept*) β_0. Przesunięcie w analizie współliniowości wymaga osobnego potraktowania. Jeśli model ma postać (4-5.33):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + E \ , \tag{4-5.38}$$

to korzystając np. z MNK możemy wyznaczyć estymatory $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ parametrów strukturalnych. Estymator $\hat{\beta}_0$ ma wtedy postać:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3 - \ldots - \hat{\beta}_k \overline{X}_k \ , \tag{4-5.39}$$

gdzie $\overline{Y} = \sum_{i=1}^{n} Y_i / n$ jest średnią arytmetyczną w próbie dla zmiennej zależnej, a $\overline{X}_j = \sum_{i=1}^{n} X_{ji} / n$ jest średnią arytmetyczną dla j – tej zmiennej (j = 1,2,...,k).

Z zależności (4-5.39) widać więc, że estymator przesunięcia $\hat{\beta}_0$ jest zależny od pozostałych estymatorów $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$. Problem znika w szczególnym przypadku, gdy średnie zmiennych X_j są równe 0, jak to ma

[7] Przy idealnej współliniowości, macierz planowania $\mathbf{X}$ (11-1.9) nie ma pełnej rangi kolumnowej (gdzie kolumny odpowiadają czynnikom), co pociąga za sobą osobliwość macierzy $\mathbf{X}^T\mathbf{X}$, której odwrotność występuje we wzorze $\hat{\sigma}^2_{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1} MSE$, (11-1.38), na macierz wariancji-kowariancji dla estymatorów parametrów strukturalnych modelu. $S^2_{\hat{\beta}}$ są elementami na diagonalnej macierzy $\hat{\sigma}^2_{\hat{\beta}}$.

miejsce np. w przypadku wycentrowania (tzn. przesunięcia o ich średnie) oryginalnych czynników (Rozdział 6). W przypadku tym średnia $\overline{Y}$ jest oszacowaniem przesunięcia.

Gdyby wprowadzić następujący model regresji:

$$I = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + E \,, \tag{4-5.40}$$

w którym I jest jednostkową zmienną stojącą w równaniu (4-5.33) obok parametru przesunięcia β_0, przyjmującą zawsze wartości $I_i = 1$ ($i = 1,2,...,n$), wtedy można wprowadzić współczynnik inflacji wariancji dla przesunięcia, tak samo jak to uczyniono dla pozostałych zmiennych:

$$VIF_0 = \frac{1}{1 - R_0^2} \,, \tag{4-5.41}$$

gdzie R_0^2 można wyliczyć jako kwadrat współczynnika korelacji wielorakiej (4-25) dla zależności zmiennej I od wszystkich zmiennych $X_1, X_2, ..., X_k$.

Tak jak dla pozostałych estymatorów $\hat{\beta}_j$ otrzymujemy wtedy estymator wariancji dla przesunięcia:

$$S^2_{\hat{\beta}_0} = c_0 \cdot VIF_0 \,, \tag{4-5.42}$$

co oznacza, że interpretacja VIF_0 jest taka sama jak pozostałych VIF_j.

Jednym z prostszych sposobów eliminacji współliniowości jest odpowiednie przeskalowanie danych (między innymi przez wycentrowanie i standaryzację). Zagadnienie to omówiono poniżej.

Ad.4 Analiza skalowania.

Analiza ta polega na odpowiednim wyborze tak jednostek, jak i początku układu współrzędnych dla zmiennych mierzalnych. Np. skalowanie liniowe polega na odpowiednim przemnożeniu zmiennej przez stałą lub dodaniu stałej. Przykładem skalowania zmiennej jest przejście od temperatury podanej w jednostkach Fahrenheit'a do jednostek Celsjusza.

Często przed skalowaniem należy ustalić rząd wielkości wartości zmiennych, aby nie utracić informacji zawartej w danych, którymi dysponujemy.

Ad.4-1. Centrowanie i standaryzacja. Przykładem skalowania liniowego jest centrowanie i standaryzacja zmiennych.

Centrowanie polega na przetransformowaniu zmiennych tak, aby nowe zmienne były rozłożone wokół zera, tzn. aby ich wartość średnia wynosiła zero. Jeśli próba jest n-wymiarowa, wtedy transformacja ta wygląda następująco [1]:

$$X_{ji}^{*} = X_{ji} - \overline{X}_j \ , \qquad (\text{lub } X_j^{*} = X_j - \overline{X}_j \ ) \ , \tag{4-5.43}$$

gdzie:

X_j - zmienna pierwotna (gdzie j = 1, 2,…, k) ,

X_{ji} - zmienna, która daje w próbce i – tą wartość (i=1,2,…,n) zmiennej X_j ,

$\overline{X}_j = \sum_{i=1}^{n} X_{ji} / n$ - wartość średnia zmiennej wyjściowej,

X_j^{*} - nowa zmienna (wycentrowana) .

Standaryzacja jest określona następująco [1]:

$$Z_j = \frac{X_j^{*}}{S_{X_j}} = \frac{X_j - \overline{X}_j}{S_{X_j}}, \tag{4-5.44}$$

gdzie S_{X_j} jest odchyleniem standardowym zmiennej X_j.

Po tych transformacjach otrzymujemy zbiór zmiennych Z_j o wartości średniej równej 0 i odchyleniu standardowym równym 1, a ponadto wszystkie te zmienne są niemianowane.

W zmiennych standaryzowanych (łącznie ze zmienną objaśnianą), model regresji ma postać:

$$(Y_i - \overline{Y}) / S_Y = \beta_1^{*}(X_{1i} - \overline{X}_1) / S_1 + \beta_2^{*}(X_{2i} - \overline{X}_2) / S_2 + ... + \beta_k^{*}(X_{ki} - \overline{X}_k) / S_k + E_i^{*}, \tag{4-5.45}$$

gdzie

$$\beta_j^{*} = \beta_j (S_j / S_Y) \tag{4-5.46}$$

są tzw. *standaryzowanymi współczynnikami regresji*.

Centrowanie i standaryzacja jest prostym sposobem redukcji współzależności pomiędzy zmiennymi. Jej stosowanie zaleca się w modelach wielomianowych najniższego stopnia (jest szczególnie skuteczna w modelach wielomianowych stopnia nie większego niż dwa).

Rozdział 4-6. Analiza współliniowości metodą wartości własnych macierzy korelacji.

Do analizy współzależności pomiędzy zmiennymi objaśniającymi wykorzystuje się również wartości własne macierzy kowariancji (lub korelacji) dla zmiennych objaśnianych.

W celu wyjaśnienia tej metody rozważmy następujący model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + E \ . \tag{4-6.47}$$

Rozważmy macierz kowariancji pomiędzy zmiennymi objaśniającymi X_l oraz X_s, l,s =1,2,...,k:

$$C = \begin{bmatrix} \sigma^2(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_k) \\ \mathrm{cov}(X_2, X_1) & \sigma^2(X_2) & \cdots & \mathrm{cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_k, X_1) & \mathrm{cov}(X_k, X_2) & \cdots & \sigma^2(X_k) \end{bmatrix} \tag{4-6.48}$$

przy czym zmienne te mają wartość oczekiwaną równą zero $E(X_r) = 0$, r =1,2,...,k. Zmienne X_l nie muszą być pod kontrolą (tak jak to jest w modelu regresji klasycznej), tzn. mą być zmiennymi losowymi.

Twierdzenie (o składowych głównych (zasadniczych)) (Część IV, Rozdział 3).

Niech k-wymiarowy wektor losowy $\vec{X} = (X_1, X_2, ..., X_k)^T$ ma wartość oczekiwaną:

$$E(\vec{X}) = \vec{0} \tag{4-6.49}$$

oraz macierz kowariancji:

$$C = E(\vec{X}\,\vec{X}^T)\,. \tag{4-6.50}$$

Istnieje wtedy ortogonalna liniowa transformacja:

$$\vec{Q} = \boldsymbol{\alpha}^T \vec{X} \tag{4-6.51}$$

taka że:

$$E(\vec{Q}\,\vec{Q}^T) = \boldsymbol{\Lambda} \tag{4-6.52}$$

gdzie:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \tag{4-6.53}$$

[Zatem każda r-ta kolumna ($r = 1,2,...,k$) macierzy $\boldsymbol{\alpha} = (\vec{\alpha}^{(1)}, \vec{\alpha}^{(2)}, ..., \vec{\alpha}^{(k)})$ ma postać:

$$\vec{\alpha}^{(r)} = (\alpha_1^{(r)}, \alpha_2^{(r)}, ..., \alpha_k^{(r)})^T \tag{4-6.54}$$

i jest ona kolumną współczynników liniowej kombinacji zmiennych $X_1, X_2, ..., X_k$, gdzie $\vec{Q} = (Q_1, Q_{2,}, ..., Q_k)^T$ są nowymi zmiennymi losowymi.**]**

Wartości λ_r na przekątnej macierzy $\boldsymbol{\Lambda}$ są pierwiastkami równania:

$$\det(C - \lambda_r I) = 0 \tag{4-6.55}$$

(są więc one wartościami własnymi macierzy kowariancji C) spełniającymi relację:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq 0\ . \tag{4-6.56}$$

Kolumna r-ta macierzy $\boldsymbol{\alpha}$, czyli $\vec{\alpha}^{(r)}$, spełnia równanie własne:

$$(C - \lambda_r I)\vec{\alpha}^{(r)} = \vec{0}\ . \tag{4-6.57}$$

Wektory $\vec{\alpha}^{(r)}$ tworzą ortonormalny układ wektorów, tzn.:

$$(\vec{\alpha}^{(r)})^T \vec{\alpha}^{(r')} = \delta_{rr'}\ , \qquad r, r' = 1,2,...,k\,. \tag{4-6.58}$$

Składową *r*-tą wektora $\vec{Q}$ jest następująca zmienna:

$$Q_r = (\vec{\alpha}^{(r)})^T \vec{X} . \qquad (4\text{-}6.59)$$

Liniowa kombinacja Q_r ma maksymalną wariancję pośród wszystkich kombinacji liniowych nieskorelowanych ze zmiennymi $Q_1, Q_{2,}, ..., Q_{r-1}$. Wektor $\vec{Q}$ jest tak zwanym wektorem składowych głównych (zasadniczych) wektora losowego $\vec{X}$. (koniec twierdzenia)

Podsumowanie. Powyższa procedura [6] związana z rozwiązaniem równania własnego $(C - \lambda_r I)\vec{\alpha}^{(r)} = \vec{0}$, $\vec{\alpha}^{(r)} \neq \vec{0}$, $r = 1,2,...,k$, diagonalizuje macierz kowariancji *C*, dając diagonalną macierz kowariancji $\mathbf{\Lambda}$, (4-6.53). Oznacza ona przejście od układu zmiennych oryginalnych { X_r } do nowego układu zmiennych { Q_r } nazywanych *głównymi składowymi* dla zmiennych objaśniających [6]. Główne składowe są liniowymi kombinacjami oryginalnych zmiennych objaśniających i mają następujące własności [1]:

1. stanowią układ nowych zmiennych objaśniających, z taką samą informacją jaka jest zawarta w zmiennych oryginalnych,
2. nie są ze sobą skorelowane,
3. ponieważ zachodzi własność (2), więc ich suma ma maksymalną wariancję,
4. wariancje zmiennych { Q_r } są wartościami własnymi macierzy kowariancji *C* [6], tzn.:

$$\lambda_r = \sigma^2(Q_r) \quad , \qquad r = 1,..., k , \qquad (4\text{-}6.60)$$

5. jeśli zbiór *k* oryginalnych zmiennych objaśniających *nie wykazuje idealnej współliniowości*, to do przekazania tej samej informacji, która jest zawarta w zmiennych oryginalnych potrzebnych jest dokładnie *k* głównych składowych.

   Natomiast jeśli np. jedna z oryginalnych zmiennych objaśniających jest liniową kombinacją pozostałych, to tylko *k* – 1 głównych składowych jest potrzebnych do przekazania pierwotnej informacji.
6. liczba równych zero (albo prawie bliskich zeru) wartości własnych λ_i jest liczbą relacji współliniowości (albo prawie idealnej współliniowości) pomiędzy oryginalnymi zmiennymi objaśniającymi.
7. im większa jest konkretna wartość własna, tym istotniejsza (pod względem niesionej informacji przez zmienne objaśniające) jest związana z nią główna składowa.

Znaczenie składowych głównych polega na znalezieniu takich liniowych kombinacji zmiennych wektora losowego $\vec{X} = (X_1, X_2, ..., X_k)^T$, które mają maksymalną wariancję i są z sobą nieskorelowane. W praktycznych zastosowaniach zdarza się, że liczba zmiennych branych pod uwagę jest za duża. Ponieważ istotną sprawą jest rozrzut wartości czynników, dlatego metoda składowych zasadniczych pozwala na

odrzucenie tych liniowych kombinacji zmiennych wektora losowego $\vec{X}$, które mają małą wariancję i na poddanie analizie kombinacji z dużą wariancją.

Pojawienie się jakiejś wartości własnej równej zero (lub bliskiej zeru) oznacza wystąpienie dokładnej (lub prawie dokładnej) współliniowości pomiędzy niektórymi oryginalnymi zmiennymi objaśniającymi, co wynika z faktu, że równanie:

$$\lambda_r = \sigma^2(Q_r) = \sigma^2((\vec{\alpha}^{(r)})^T \vec{X}) = 0 \qquad (4\text{-}6.61)$$

oznacza

$$(\vec{\alpha}^{(r)})^T \vec{X} = c = \text{const}\,, \qquad (4\text{-}6.62)$$

czyli właśnie wystąpienie idealnej współliniowości. Skłania to do próby eliminacji (przynajmniej) jednej ze zmiennych objaśniających z grupy zmiennych oryginalnych $\{X_1, X_2, ..., X_k\}$, jako będącej kombinacją liniową pozostałych. Np. korzystając z jednej z metod selekcji, można próbować wyeliminować zmienne najmniej istotne statystycznie z nadzieją, że usuniemy również zmienne zależne liniowo od pozostałych.

Przykład. Rozważmy wpływ różnych cech (np. liczba pokoi, łazienek czy lokalizacja) określających standard mieszkania na jego cenę (Przykład z Rozdziału 6). Z cech tych metoda składowych głównych tworzy liniowe ich kombinacje, które jakoś różnicują jednostki badanej zbiorowości mieszkań. Te z kombinacji, które mają największy rozrzut wartości przy zmianie mieszkania, są interesujące. Natomiast kombinacje, które zmieniają się nieznacznie od mieszkania do mieszkania, mówią mało o zmienności pomiędzy mieszkaniami i można je usunąć z analizy.

Indeks warunkowy i liczba warunkowa: W analizie za pomocą wartości własnych korzysta się również z wielkości zwanej indeksem warunkowym (*CI*), który jest zdefiniowany następująco [1]:

$$CI_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}} \qquad (4\text{-}6.63)$$

gdzie:

j - numeruje wartości własne (j = 1,..., k),

λ_j jest j-tą wartością własną,

$\lambda_{\max}$ – maksymalna wartość własna w modelu.

Przyjmuje się, że występowanie wartości indeksu warunkowego przekraczającej 30, oznacza występowanie bardzo silnej współliniowości pomiędzy jakąś zmienną oryginalną, a innymi zmiennymi oryginalnymi. Również w tym przypadku należy przejść do próby wyeliminowania zmiennej najmniej istotnej statystycznie, z nadzieją, że wyeliminujemy zmienną zależną liniowo od pozostałych. Największy z indeksów warunkowych CI_j jest nazywany *liczbą warunkową CN*.

Uwaga. Podobną analizę składowych głównych można przeprowadzić w oparciu o macierz korelacyjną:

$$\rho_{kor} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & \rho_{kk} \end{bmatrix}, \tag{4-6.64}$$

gdzie

$$\rho_{ij} = \frac{\mathrm{cov}(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}, \; i = 1,..., k \; ; j = 1,..., k, \tag{4-6.65}$$

są współczynnikami korelacji Pearsona. Po diagonalizacji macierzy (4-6.64) metodą składowych głównych otrzymujemy macierz diagonalną typu $\mathbf{\Lambda}$, (4-6.53), z tym, że suma wartości na diagonalnej równa jest liczbie k zmiennych oryginalnych $\{X_1, X_2,..., X_k\}$ [1]. [Macierz korelacyjna jest nie tylko miarą kierunku zależności pomiędzy parami zmiennych, ale i (w odróżnieniu od macierzy kowariancji C) miarą siły zależności pomiędzy nimi.]

Zadanie: Udowodnić Twierdzenie o składowych głównych [6].

A. Rozdział 5. Przykłady analizy regresji z jednym czynnikiem.

Rozdział 5-1. Liniowa analizy regresji. Przykład. „Dochód z biletów” (dane i wstęp).

Pewne linie lotnicze otworzyły trzy nowe połączenia. W przeciągu dziewięciu miesięcy zyski (w mln PLN) z nich ukształtowały się następująco:

miesiąc	dochód	Miesiąc	dochód	miesiąc	dochód
1	34,9	1	35	1	34,8
2	38,8	2	38,9	2	38,7
3	41,5	3	41,6	3	41,4
4	45,1	4	45,2	4	45
5	48,3	5	48,4	5	48,2
6	51,2	6	51,3	6	51,1
7	56,6	7	56,7	7	56,5
8	59,9	8	60	8	59,8
9	65,4	9	65,5	9	65,3

Znaleźć model regresji liniowej zależności dochodu linii lotniczych od miesięcy działalności nowo otwartych połączeń.

Do analizy stosujemy funkcję SAS’a znajdującą się w następującej lokalizacji[8] aplikacji Analyst [10]: Solutions->Analysis->Analyst, a następnie (po wczytaniu danych, korzystając z „Open By SAS Name” w zakładce File [11]) ->Statistics->Regression->Linear. W tym miejscu określamy zmienne: objaśnianą (dochód) i objaśniającą (zmienna miesiąc lub jej transformacje), oraz w razie potrzeby określamy potrzebne statystyki (Statistics), testy (Tests), wykresy (Plots) itp.

Raport SAS’a ma postać.

```
                    model liniowy

                    Procedura REG
                    Model: MODEL1
Zmienna zależna: dochod bilety
Wczytano obserwacji          27
Użyto obserwacji             27
```

[8] Odpowiedni program dla liniowego modelu, wykorzystujący procedurę SAS’a REG ma postać:

```
proc reg data=Roboczy.Bilety;
   model DOCHOD = MIESIAC / clb;
run;
quit;
```

Jak widać zbiór danych to Bilety znajdujący się we (wcześniej utworzonej) bibilitece SAS’a [11] o nazwie Roboczy. Polecenie ‘clb’ wyznacza granice $(1-\alpha)\cdot 100\%$-wego przedziału ufności dla oszacowywanych parametrów modelu. Elementy kodów w języku SAS 4GL i ich uruchamianie z okna ‘Enhanced Editor’, zostaną omówione w Rozdziale 2 Część B.

Analiza wariancji

Źródło	St. sw.	Suma kwadratów	Średnia kwadratów	Wartość F	Pr. > F
Model	1	2455.32800	2455.32800	3038.97	<.0001
Błąd	25	20.19867	0.80795		
Razem skorygowane	26	2475.52667			

Pierw. bł. śr.-kw.	0.89886	R-kwadrat	**0.9918**
Średnia zależna	49.07778	Skor. R-kw.	0.9915
Wsp. zmienności	1.83150		

Oceny parametrów

Zmienna	Etykieta	St. sw.	Ocena parametru	Błąd standardowy	Wartość t	Pr. > |t|	Przedział ufności 95%	
Intercept	Intercept	1	30.61111	0.37701	81.19	<.0001	29.83464	31.38758
miesiac	miesiac	1	3.69333	0.06700	55.13	<.0001	3.55535	3.83132

Z powyższego raportu SAS'a wynika, że otrzymana w próbce liniowa funkcja regresji II rodzaju ma dla rozważanego problemu zależności średniej wartości dochodów od miesiąca, następującą postać:

$$\hat{Y} = 30{,}6111 + 3{,}6933X \ . \qquad (5\text{-}1.1)$$

Dopasowanie otrzymanego modelu do danych empirycznych charakteryzuje się wysoką wartością współczynnika determinacji, r^2 = 0,9918, co oznacza, że siła tego dopasowania jest duża, bowiem 99,18% średniej zmienności dochodów jest wytłumaczona zmianami miesiąca.

Rozdział 5-2. Wielomianowa analiza regresji. Przykład. „Dochód z biletów" (c.d.).

Rozważmy dalej powyższy przykład zależności dochodów linii lotniczych od miesięcy działalności. Jako pierwsze analizie zostaną poddane modele wielomianowe zwyczajne drugiego, trzeciego i ósmego stopnia, następnie modele wielomianowe centrowane (tych samych stopni), a na końcu wielomian ortogonalny ósmego stopnia. Ósmy stopień wielomianu pojawia się z rozważań nad stopniem modelu maksymalnego.

Nasz układ danych jest następujący: Liczba wszystkich obserwacji w próbce wynosi n = 27, zmienna objaśniająca (numer miesiąca) ma l = 9 poziomów (wariantów) i każdemu z nich odpowiada taka sama liczba obserwacji zmiennej objaśnianej $n/l = 27/9 = 3$, z liczbą *replik* dla każdego wariantu równą $n/l - 1 = 27/9 - 1 = 2$. Stopień wielomianu dla modelu maksymalnego jest wiec równy $m = n - 1 - r$ = 8, gdzie $r = l(n/l - 1)$ == $n - l$ = 18 jest liczbą wszystkich replik.

Funkcja SAS'a pozwalająca dokonać analizy znajduje w: Solutions->Analysis->Analyst, a następnie: Statistics->Regression->Linear. Jest to ta sama funkcja, co w przypadku regresji liniowej, jednakże w miejsce zmiennych objaśniających wstawiamy wszystkie (rozważane) zmienne badanego wielomianu określonego stopnia.

Rozdział 5-2-1. Wielomiany zwyczajne.

Zestaw danych do analizy Przykładu „Dochód z biletów” za pomocą wielomianów zwyczajnych wygląda następująco:

dochód	miesiąc	miesiąc^2	miesiąc^3	miesiąc^4	miesiąc^5	miesiąc^6	miesiąc^7	miesiąc^8
34,9	1	1	1	1	1	1	1	1
38,8	2	4	8	16	32	64	128	256
41,5	3	9	27	81	243	729	2187	6561
45,1	4	16	64	256	1024	4096	16384	65536
48,3	5	25	125	625	3125	15625	78125	390625
51,2	6	36	216	1296	7776	46656	279936	1679616
56,6	7	49	343	2401	16807	117649	823543	5764801
59,9	8	64	512	4096	32768	262144	2097152	16777216
65,4	9	81	729	6561	59049	531441	4782969	43046721
35	1	1	1	1	1	1	1	1
38,9	2	4	8	16	32	64	128	256
41,6	3	9	27	81	243	729	2187	6561
45,2	4	16	64	256	1024	4096	16384	65536
48,4	5	25	125	625	3125	15625	78125	390625
51,3	6	36	216	1296	7776	46656	279936	1679616
56,7	7	49	343	2401	16807	117649	823543	5764801
60	8	64	512	4096	32768	262144	2097152	16777216
65,5	9	81	729	6561	59049	531441	4782969	43046721
34,8	1	1	1	1	1	1	1	1
38,7	2	4	8	16	32	64	128	256
41,4	3	9	27	81	243	729	2187	6561
45	4	16	64	256	1024	4096	16384	65536
48,2	5	25	125	625	3125	15625	78125	390625
51,1	6	36	216	1296	7776	46656	279936	1679616
56,5	7	49	343	2401	16807	117649	823543	5764801
59,8	8	64	512	4096	32768	262144	2097152	16777216
65,3	9	81	729	6561	59049	531441	4782969	43046721

Rozdział 5-2-1-1. Wielomian zwyczajny drugiego stopnia.

```
                    modele wielomianowe normalne:
                          drugiego stopnia

                          The REG Procedure
                            Model: MODEL1
                  Dependent Variable: dochod dochod

                        Analysis of Variance

                                 Sum of           Mean
        Source            DF     Squares         Square    F Value    Pr > F

        Model              2  2468.75869     1234.37935    4377.25    <.0001
        Error             24     6.76797        0.28200
        Corrected Total   26  2475.52667

                  Root MSE          0.53104    R-Square     0.9973
                  Dependent Mean   49.07778    Adj R-Sq     0.9970
                  Coeff Var         1.08203

                          Parameter Estimates

                         Parameter     Standard
 Variable   Label      DF  Estimate        Error  t Value  Pr > |t|    Type II SS

 Intercept  Intercept   1  32.82143      0.39012    84.13    <.0001    1996.07379
 month      month       1   2.48771      0.17913    13.89    <.0001      54.39131
 month2     month2      1   0.12056      0.01747     6.90    <.0001      13.43069

                          Parameter Estimates

                                              Squared        Squared
                         Standardized  Semi-partial        Partial
      Variable   Label      DF      Estimate  Corr Type II   Corr Type II

      Intercept  Intercept   1             0             .              .
      month      month       1       0.67081       0.02197        0.88934
      month2     month2      1       0.33334       0.00543        0.66493

                        Correlation of Estimates

        Variable       Label            Intercept            month           month2

        Intercept      Intercept          1.0000          -0.9128           0.8210
        month          month             -0.9128           1.0000          -0.9753
        month2         month2             0.8210          -0.9753           1.0000
```

Wnioski z raportu.

1. Równanie modelu:

$$\hat{Y} = 32{,}82 + 2{,}49X + 0{,}12X^2 \qquad (5\text{-}2\text{-}1\text{-}1.2)$$

2. Istotność statystyczna modelu:

Niski empiryczny poziom istotności $p < 0.0001$ oznacza, że wartość statystyki F ($F = 4377.25$) „duża na oko”, jest faktycznie wartością istotnie statystycznie różną od zera. O tym czy wartość statystyki testowej w próbce jest istotna statystycznie decyduje wartość p.

Zatem odrzucamy hipotezę zerową:

H_0 : **o niewystępowaniu ogólnej zależności korelacyjnej zmiennej *Y* od zmiennej *X* w modelu parabolicznym**, czyli o nie występowaniu braku dopasowania w modelu, w którym jest jedynie parametr przesunięcia β_0, w porównaniu z modelem parabolicznym.

Powyższa decyzja statystyczna byłaby słuszna dla każdego poziomu istotności $\alpha \geq p$ (np. dla α = 0,05 lub α = 0,0001).

3. Wysoka wartość współczynnika determinacji (R^2 = 0.997) wskazuje na dobre dopasowanie modelu parabolicznego do danych empirycznych. (Przeczytaj Uwagę na końcu Rozdziału 3-2-2.)

4. Istotność parametrów strukturalnych modelu:

Wszystkie parametry strukturalne modelu (a zatem i odpowiadające im zmienne objaśniające) są istotne, gdyż z testów t przeprowadzonych przez system SAS wynika, że empiryczny poziom istotności dla każdego z parametrów jest niski (p < 0.0001).

5. Analiza macierzy korelacji:

Otrzymana macierz korelacji estymatorów posiada poza diagonalną duże wartości współczynników korelacji dla estymatorów parametrów strukturalnych. Np. dla $\hat{\beta}_1$ oraz $\hat{\beta}_2$ wynosi on $\hat{\rho}_{\hat{\beta}_1\hat{\beta}_2} = -0.9753$, (11-1.39). Implikuje to silną korelację między zmiennymi objaśniającymi, co można wywnioskować z porównania (4-5.37) z (11-1.38) (porównaj tekst poniżej (4-5.37) i przypis 7).

Z przeprowadzonych obliczeń wynika, że model wielomianowy zwyczajny drugiego stopnia jest modelem dobrze dopasowanym do danych empirycznych w próbce, ale jego zdolność do predykcji jest niepewna ze względu na dużą korelację pomiędzy czynnikami.

Rozdział 5-2-1-2. Wielomian zwyczajny trzeciego stopnia.

Odpowiedni raport SAS'a ma postać:

```
                    modele wielomianowe normalne:
                          trzeciego stopnia

                          The REG Procedure
                            Model: MODEL1
                Dependent Variable: dochod dochod

                      Analysis of Variance

                                 Sum of           Mean
Source                  DF      Squares         Square    F Value    Pr > F

Model                    3   2471.15145      823.71715    4330.18    <.0001
Error                   23      4.37522        0.19023
Corrected Total         26   2475.52667

            Root MSE              0.43615    R-Square     0.9982
            Dependent Mean       49.07778    Adj R-Sq     0.9980
            Coeff Var             0.88869
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	Intercept	1	31.26032	0.54444	57.42	<.0001	209.04608
month	month	1	3.98259	0.44643	8.92	<.0001	5.04623
month2	month2	1	-0.23424	0.10106	-2.32	0.0297	0.34062
month3	month3	1	0.02365	0.00667	3.55	0.0017	0.79759

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type II	Squared Partial Corr Type II
Intercept	Intercept	1	0	.	.
month	month	1	1.07391	0.00612	0.99184
month2	month2	1	-0.64763	0.00041279	0.66493
month3	month3	1	0.59536	0.00096657	0.35354

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	30.13406	32.38657
month	month	1	0.00530	**188.58826**	3.05907	4.90611
month2	month2	1	0.00098419	**1016.06133**	-0.44330	-0.02517
month3	month3	1	0.00273	**366.70875**	0.00986	0.03745

Correlation of Estimates

Variable	Label	Intercept	month	month2	month3
Intercept	Intercept	1.0000	-0.9404	0.8689	-0.8085
month	month	-0.9404	1.0000	-0.9802	0.9441
month2	month2	0.8689	-0.9802	1.0000	-0.9899
month3	month3	-0.8085	0.9441	-0.9899	1.0000

Wnioski z raportu.

1. Równanie modelu:

$$\hat{Y} = 31{,}26 + 3{,}98X - 0{,}23X^2 + 0{,}02X^3 \qquad (5\text{-}2\text{-}1\text{-}2.3)$$

2. Istotność statystyczna modelu:

Niska wartości empirycznego poziomu istotności, $p < 0.0001$, wskazuje, że wartość statystyki $F = 4330.18$, jest istotnie statystycznie większa od zera. Oznacza to, że model sześcienny jest istotny statystycznie, tzn. hipoteza zerowa:

H_0: o braku ogólnej zależności korelacyjnej zmiennej objaśnianej od łącznego wpływu wszystkich potęg zmiennej objaśniającej, aż do trzeciego stopnia włącznie,

została odrzucona na każdym poziomie istotności $\alpha \geq p$, np. dla $\alpha = 0.01$.

3. Wysoka wartość współczynnika determinacji ($R^2 = 0.998$) wskazuje na dobre dopasowanie modelu do danych empirycznych.

4. Istotność parametrów strukturalnych modelu:

Wszystkie wartości b_j estymatorów $\hat{\beta}_j$ parametrów strukturalnych modelu (a zatem i odpowiadające im zmienne objaśniające) są istotne, gdyż z testów t (ze statystykami $t = \frac{\hat{\beta}_j - 0}{S_{\hat{\beta}_j}}$) przeprowadzonych przez system SAS widać, że prawdopodobieństwo p dla każdego z parametrów strukturalnych jest niskie, tzn. wynosi:

a) dla parametrów β_0, β_1 , $p < 0.0001$,

b) dla parametrów β_2, β_3 kolejno, $p = 0.0297$, $p = 0.0017$ (przy powszechnie przyjętych poziomach istotności α, powyższe wartości p uważane są na ogół za niskie, chociaż istotność wartości $b_2 = -0.23424$ estymatora oszacowującego parametr β_2, dla którego $0.05 > p = 0.0297 > 0.01$, może być poddana dyskusji).

5. Analiza macierzy korelacji:

Podobnie jak dla modelu drugiego stopnia (porównaj rozważania w odpowiednim miejscu), macierz korelacji dla estymatorów parametrów ma poza diagonalną wysokie wartości współczynników korelacji, co wskazuje na istnienie silnej korelacji (i współliniowości) między zmiennymi objaśniającymi. Wskazują na to również wyjątkowo duże wartości współczynników inflacji wariancji, np. $VIF_2 = 188.59$. Jest to sygnał, że należałoby zastosować ortogonalizację pierwotnych czynników (lub przynajmniej ich wycentrowanie) .

Model zwyczajny trzeciego stopnia wykazuje w ogólności podobne cechy, co wcześniejszy model drugiego stopnia. Jednakże rozpatrywany model nieco lepiej dopasowuje się do danych empirycznych niż model poprzedni, co wynika z większej wartości współczynnika determinacji R^2.

Rozdział 5-2-1-3. Wielomian zwyczajny ósmego stopnia.

Odpowiedni raport SAS'a ma postać:

```
                    modele wielomianowe normalne:
                           ósmego stopnia

                          The REG Procedure
                            Model: MODEL1
                Dependent Variable: dochod dochod

                              Analysis of Variance

                                     Sum of           Mean
        Source                 DF       Squares         Square    F Value    Pr > F

        Model                   7    2474.85984      353.55141    10073.8    <.0001
        Error                  19       0.66683        0.03510
        Corrected Total        26    2475.52667

                     Root MSE              0.18734    R-Square     0.9997
                     Dependent Mean       49.07778    Adj R-Sq     0.9996
                     Coeff Var             0.38172
NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique.
       Some statistics will be misleading. A reported DF of 0 or B means that
       the estimate is biased.
```

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

month8 = - 76160 * Intercept + 198530 * month - 202281 * month2 + 107552 * month3 - 33144.3 * month4 + 6132 * month5 - 671.067 * month6 + 40 * month7

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	Intercept	B	-16.72222	5.62031	-2.98	0.0078	0.31069
month	month	B	116.49919	13.02538	8.94	<.0001	2.80753
month2	month2	B	-99.35541	11.30634	-8.79	<.0001	2.71018
month3	month3	B	43.77085	4.90774	8.92	<.0001	2.79169
month4	month4	B	-10.64252	1.17415	-9.06	<.0001	2.88339
month5	month5	B	1.44632	0.15712	9.21	<.0001	2.97389
month6	month6	B	-0.10273	0.01101	-9.33	<.0001	3.05545
month7	month7	B	0.00297	0.00031431	9.44	<.0001	3.12591
month8	month8	0	0	.	.	.	.

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type II	Squared Partial Corr Type II
Intercept	Intercept	B	0	.	.
month	month	B	31.41415	0.00113	0.80807
month2	month2	B	-274.70349	0.00109	0.80254
month3	month3	B	1101.72016	0.00113	0.80719
month4	month4	B	-2389.81866	0.00116	0.81217
month5	month5	B	2888.93980	0.00120	0.81684
month6	month6	B	-1827.19407	0.00123	0.82085
month7	month7	B	470.75129	0.00126	0.82418
month8	month8	0	.	.	.

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	B	.	0	-28.48566	-4.95878
month	month	B	0.00000115	870148	89.23675	143.76162
month2	month2	B	1.450784E-8	68928233	-123.01986	-75.69096
month3	month3	B	9.29087E-10	1076325582	33.49883	54.04287
month4	month4	B	2.03942E-10	4903366270	-13.10004	-8.18500
month5	month5	B	1.4394E-10	6947357709	1.11746	1.77518
month6	month6	B	3.6969E-10	2704968417	-0.12577	-0.07968
month7	month7	B	5.698042E-9	175498873	0.00231	0.00362
month8	month8	0	.	.	.	.

Correlation of Estimates

Variable	Label	Intercept	month	month2	month3
Intercept	Intercept	1.0000	-0.9981	0.9927	-0.9845
month	month	-0.9981	1.0000	-0.9981	0.9929
month2	month2	0.9927	-0.9981	1.0000	-0.9983
month3	month3	-0.9845	0.9929	-0.9983	1.0000
month4	month4	0.9743	-0.9852	0.9936	-0.9985
month5	month5	-0.9630	0.9759	-0.9870	0.9946
month6	month6	0.9513	-0.9658	0.9791	-0.9890
month7	month7	-0.9395	0.9554	-0.9704	0.9824

Correlation of Estimates

Variable	Label	month4	month5	month6	month7
Intercept	Intercept	0.9743	-0.9630	0.9513	-0.9395
month	month	-0.9852	0.9759	-0.9658	0.9554
month2	month2	0.9936	-0.9870	0.9791	-0.9704
month3	month3	-0.9985	0.9946	-0.9890	0.9824
month4	month4	1.0000	-0.9988	0.9955	-0.9910
month5	month5	-0.9988	1.0000	-0.9990	0.9964
month6	month6	0.9955	-0.9990	1.0000	-0.9992
month7	month7	-0.9910	0.9964	-0.9992	1.0000

Wnioski z raportu.

Z powyższego raportu SAS'a wynika bardzo duża korelacja estymatora $\hat{\beta}_8$ z pozostałymi estymatorami parametrów strukturalnych. Zatem zmienna X_8 jest na tyle mocno skorelowana z grupą zmiennych I, $X_{1,\ldots,}$ X_7, że SAS wykazał *numerycznie* istnienie idealnej współliniowości (podał nawet postać liniowego związku pomiędzy czynnikami). Z tego powodu model 8-wymiarowy (nie mając pełnej rangi macierzy planowania, Rozdział 11), nie mógł być przeliczony. Nie został więc oszacowany parametr stojący przy ósmej potędze zmiennej X (month8). Taki model regresji nie może zostać przyjęty do analizy zależności korelacyjnej.

Rozdział 5-2-2. Wielomiany centrowane.

Zestaw danych do analizy za pomocą wielomianów centrowanych:

miesiąc	dochod	m_center	m_center^2	m_center^3	m_center^4	m_center^5	m_center^6	m_center^7	m_center^8
1	34,9	-4	16	-64	256	-1024	4096	-16384	65536
2	38,8	-3	9	-27	81	-243	729	-2187	6561
3	41,5	-2	4	-8	16	-32	64	-128	256
4	45,1	-1	1	-1	1	-1	1	-1	1
5	48,3	0	0	0	0	0	0	0	0
6	51,2	1	1	1	1	1	1	1	1
7	56,6	2	4	8	16	32	64	128	256
8	59,9	3	9	27	81	243	729	2187	6561
9	65,4	4	16	64	256	1024	4096	16384	65536
1	35	-4	16	-64	256	-1024	4096	-16384	65536
2	38,9	-3	9	-27	81	-243	729	-2187	6561
3	41,6	-2	4	-8	16	-32	64	-128	256
4	45,2	-1	1	-1	1	-1	1	-1	1
5	48,4	0	0	0	0	0	0	0	0
6	51,3	1	1	1	1	1	1	1	1
7	56,7	2	4	8	16	32	64	128	256
8	60	3	9	27	81	243	729	2187	6561
9	65,5	4	16	64	256	1024	4096	16384	65536
1	34,8	-4	16	-64	256	-1024	4096	-16384	65536
2	38,7	-3	9	-27	81	-243	729	-2187	6561
3	41,4	-2	4	-8	16	-32	64	-128	256
4	45	-1	1	-1	1	-1	1	-1	1
5	48,2	0	0	0	0	0	0	0	0
6	51,1	1	1	1	1	1	1	1	1
7	56,5	2	4	8	16	32	64	128	256
8	59,8	3	9	27	81	243	729	2187	6561
9	65,3	4	16	64	256	1024	4096	16384	65536

Rozdział 5-2-2-1. Wielomian centrowany drugiego stopnia.

Odpowiedni raport SAS'a ma postać:

```
                                wielomiany centrowane
                                  drugiego stopnia

                                  The REG Procedure
                                    Model: MODEL1
                            Dependent Variable: dochod dochod

                                Analysis of Variance

                                       Sum of           Mean
        Source                  DF        Squares         Square    F Value    Pr > F

        Model                    2     2468.75869     1234.37935    4377.25    <.0001
        Error                   24        6.76797        0.28200
        Corrected Total         26     2475.52667

                     Root MSE              0.53104    R-Square     0.9973
                     Dependent Mean       49.07778    Adj R-Sq     0.9970
                     Coeff Var             1.08203

                                 Parameter Estimates

                            Parameter     Standard
Variable   Label      DF    Estimate        Error  t Value  Pr > |t|    Type I SS

Intercept  Intercept   1    48.27403      0.15495   311.55    <.0001        65033
m_center   m_center    1     3.69333      0.03958    93.31    <.0001   2455.32800
m_center2  m_center2   1     0.12056      0.01747     6.90    <.0001     13.43069

                                 Parameter Estimates

                                          Squared      Squared       Squared
                           Standardized  Semi-partial      Partial  Semi-partial
 Variable   Label      DF      Estimate   Corr Type I  Corr Type I  Corr Type II

 Intercept  Intercept   1            0             .            .             .
 m_center   m_center    1       0.99591       0.99184      0.99184       0.99184
 m_center2  m_center2   1       0.07366       0.00543      0.66493       0.00543

                                 Parameter Estimates

                                             Variance
      Variable   Label      DF    Tolerance      Inflation      95% Confidence Limits

      Intercept  Intercept   1            .              0      47.95423       48.59382
      m_center   m_center    1      1.00000        1.00000       3.61164        3.77502
      m_center2  m_center2   1      1.00000        1.00000       0.08451        0.15662


                              Correlation of Estimates

        Variable       Label             Intercept          m_center        m_center2

        Intercept      Intercept            1.0000            0.0000          -0.7516
        m_center       m_center             0.0000            1.0000           0.0000
        m_center2      m_center2           -0.7516            0.0000           1.0000
```

Wnioski z raportu.

Ponieważ opis tych wniosków jest bardzo podobny do opisu dla wielomianów drugiego i trzeciego stopnia dla czynników niewycentrowanych, dlatego podamy skróconą jego postać:

1. Równanie modelu:

$$\hat{Y} = 48{,}27 + 3{,}69X + 0{,}12X^2 \qquad (5\text{-}2\text{-}2\text{-}1.4)$$

2. Istotność statystyczna modelu:

   Niska wartości empirycznego poziomu istotności, $p < 0.0001$, wskazuje, że wartość statystyki $F = 4377.25$ jest istotnie statystycznie większa od zera. Oznacza to istotność statystyczną modelu i odrzucenie hipotezy zerowej o braku ogólnej zależności korelacyjnej dochodów od dla zmiennych wycentrowanych X i X^2.
3. Wysoka wartość współczynnika determinacji ($R^2 = 0.997$) wskazuje na dobre dopasowanie modelu do danych empirycznych.
4. Istotność parametrów strukturalnych modelu:

   Wszystkie parametry strukturalne modelu (a zatem i odpowiadające im zmienne objaśniające) są istotne, gdyż z testu t przeprowadzonego przez system SAS, empiryczny poziom istotności p, dla każdego z parametrów jest niski ($p < 0.0001$).
5. Analiza macierzy korelacji:

   Macierz korelacji wskazuje na redukcję (w porównaniu z przypadkiem zmiennych zwyczajnych) korelacji pomiędzy estymatorami parametrów strukturalnych. W szczególności, korelacje pomiędzy estymatorami parametrów stojących przy pierwszym i drugim stopniu zmiennej wycentrowanej (tzn. nie uwzględniając estymatora wyrazu wolnego) wykazują absolutny jej brak. Korelacja pomiędzy czynnikami wycentrowanymi jest więc mniejsza niż pomiędzy czynnikami zwyczajnymi.

Z przeprowadzonych obliczeń wynika, że centrowany, wielomianowy model drugiego stopnia jest modelem równie dobrze dopasowanym, co model wielomianowy zwyczajny tego samego stopnia (tzn. wartości statystyki F, oraz wartość współczynnika R^2 są identyczne dla obu modeli – co wynika z tego, że są one niezmiennicze ze względu na liniowe transformacje). Jednakże, z powodu braku korelacji pomiędzy zmiennymi objaśniającymi, jego *zdolność do predykcji* jest dużo większa niż modelu wielomianowego zwyczajnego.

Rozdział 5-2-2-2. Wielomian centrowany trzeciego stopnia.

Odpowiedni raport SAS'a ma postać:

```
                    wielomiany centrowane
                      trzeciego stopnia

                      The REG Procedure
                        Model: MODEL1
            Dependent Variable: dochod dochod

                    Analysis of Variance

                              Sum of           Mean
Source                 DF        Squares         Square    F Value    Pr > F

Model                   3     2471.15145      823.71715    4330.18    <.0001
Error                  23        4.37522        0.19023
Corrected Total        26     2475.52667
```

Root MSE	0.43615	R-Square	0.9982
Dependent Mean	49.07778	Adj R-Sq	0.9980
Coeff Var	0.88869		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	48.27403	0.12726	379.33	<.0001	**65033**
m_center	m_center	1	3.41423	0.08515	40.10	<.0001	**2455.32800**
m_center2	m_center2	1	0.12056	0.01435	8.40	<.0001	**13.43069**
m_center3	m_center3	1	0.02365	0.00667	3.55	0.0017	**2.39276**

							2471.15145

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II
Intercept	Intercept	1	0	.	.	.
m_center	m_center	1	0.92065	0.99184	0.99184	0.12355
m_center2	m_center2	1	0.07366	0.00543	0.66493	0.00543
m_center3	m_center3	1	0.08143	0.00096657	0.35354	0.00096657

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	48.01077	48.53729
m_center	m_center	1	0.14577	6.86027	3.23809	3.59037
m_center2	m_center2	1	1.00000	1.00000	0.09088	0.15024
m_center3	m_center3	1	0.14577	6.86027	0.00986	0.03745

Correlation of Estimates

Variable	Label	Intercept	m_center	m_center2	m_center3
Intercept	Intercept	1.0000	0.0000	-0.7516	0.0000
m_center	m_center	0.0000	1.0000	0.0000	-0.9242
m_center2	m_center2	-0.7516	0.0000	1.0000	0.0000
m_center3	m_center3	0.0000	-0.9242	0.0000	1.0000

Wnioski z raportu.

1. Równanie modelu:

$$\hat{Y} = 48{,}27 + 3{,}41X + 0{,}12X^2 + 0{,}02X^3 \qquad (5\text{-}2\text{-}2\text{-}2.5)$$

2. Istotność statystyczna modelu:

   Niski poziom istotności, $p < 0.0001$, oznacza, że wartość statystyki $F = 4330.18$ jest istotnie statystycznie różna od zera, co wskazuje na istotność statystyczną badanego modelu sześciennego ze zmiennymi wycentrowanymi.

3. Wysoka wartość współczynnika determinacji ($R^2 = 0.998$) wskazuje na dobre dopasowanie modelu do danych empirycznych.

4. Istotność parametrów strukturalnych modelu:

   Wszystkie parametry strukturalne modelu (a zatem i odpowiadające im zmienne objaśniające) są istotne, gdyż z testu t wykonanego przez system SAS widać, że empiryczne poziomy istotności p dla każdego z parametrów są niskie:

   a) dla parametrów $\beta_0, \beta_1, \beta_2$, $p < 0.0001$,

b) dla parametru β_3, p = 0.0017 (co przy powszechnie przyjętych poziomach istotności, uważane jest zazwyczaj za wartość małą).

5. Analiza macierzy korelacji:

Macierz korelacji wskazuje w dalszym ciągu na małą korelację pomiędzy estymatorami parametrów strukturalnych, a w związku z tym korelacja (i współliniowość) pomiędzy odpowiednimi zmiennymi objaśniającymi jest również mała. Jednakże w modelu tym widać już, że skuteczność wycentrowania zmiennych spada wraz z dodawaniem kolejnych stopni zmiennej *X*. Np. wartość współczynnika korelacji pomiędzy estymatorem $\hat{\beta}_2$ i $\hat{\beta}_3$ jest duża i w pobranej próbce wynosi - 0.9242. Oznacza to, że należałoby posłużyć się lepszą techniką usuwania współliniowości, a mianowicie omówioną powyżej metodą ortogonalizacji.

Sprawdźmy jeszcze czy rozszerzenie modelu ze zmiennymi wycentrowanymi *X* oraz X^2 do modelu ze zmiennymi wycentrowanymi *X*, X^2, X^3 jest statystycznie istotne z punktu widzenia lepszego dopasowania się linii regresji do danych empirycznych. Zgodnie z (5-8) musimy wyznaczyć wartość statystyki częściowej F_p obserwowanej (*obs*) w próbce.

Korzystając z powyższych dwóch raportów (obecnego i w Rozdziale 5.4), mamy:

$$F_p^{obs} = F(X^3 | X, X^2) = \frac{(SSR_{(k'=k+1=3)} - SSR_{(k=2)})/1}{MSE_{(k'=k+1=3)}} = \frac{SS_{\text{dodanej zmiennej}}/1}{MSE_{(k'=k+1=3)}} \qquad (5\text{-}2\text{-}2\text{-}2.6)$$

$$= \frac{(2471{,}15145 - 2468{,}75869)/1}{0{,}19023} = 12{,}57825 \approx 12{,}6 \ .$$

Ponieważ powyższa statystyka częściowa ma, przy prawdziwości hipotezy zerowej o nieistotności rozszerzenia, rozkład F-Snedecora z liczbą stopni swobody licznika równą *k*' - *k* = 3 - 2 = 1, a mianownika równą *n*-1-3=23, zatem empiryczny poziom istotności wynosi (rachunek w Excel'u):

$$p = P(F^{obs} \geq 12{,}6) = 0{,}0017 \ . \qquad (5\text{-}2\text{-}2\text{-}2.7)$$

Możemy zatem stwierdzić, że rozszerzenie modelu parabolicznego do sześciennego jest istotne statystycznie z punktu widzenia poprawy dokładności dopasowania się linii regresji do danych empirycznych (na każdym poziomie istotności α $\geq p = 0{,}0017$).

Zauważmy również, że wartość statystyki t – Studenta związana z weryfikacją w modelu sześciennym hipotezy H_0: $\beta_3 = 0$, wynosi w próbce zgodnie z raportem t^{obs} = 3,55. Otrzymujemy więc, że $(t^{obs})^2 = F^{obs}$ = 12,6 oraz (z raportu) $p = P(t^{obs} \geq 3{,}55)$ = 0,0017, czyli tyle samo co dla powyższego testu F_p, tak jak to powinno być dla tego testu statystycznego ze zmienną X^3 dodaną na końcu [1].

Rozdział 5-2-2-3. Wielomian centrowany ósmego stopnia.

Odpowiedni raport SAS'a ma postać:

wielomiany centrowane
ósmego stopnia

The REG Procedure
Model: MODEL1
Dependent Variable: dochod dochod

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2475.34667	309.41833	30941.8	<.0001
Error	18	0.18000	0.01000		
Corrected Total	26	2475.52667			

Root MSE	0.10000	R-Square	0.9999
Dependent Mean	49.07778	Adj R-Sq	0.9999
Coeff Var	0.20376		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	48.30000	0.05774	836.58	<.0001	65033
m_center	m_center	1	2.55488	0.06740	37.90	<.0001	2455.32800
m_center2	m_center2	1	-0.36664	0.10531	-3.48	0.0027	13.43069
m_center3	m_center3	1	0.57035	0.03391	16.82	<.0001	2.39276
m_center4	m_center4	1	0.24720	0.04010	6.16	<.0001	0.19818
m_center5	m_center5	1	-0.07819	0.00453	-17.24	<.0001	0.17333
m_center6	m_center6	1	-0.03170	0.00471	-6.74	<.0001	0.21097
m_center7	m_center7	1	0.00297	0.00016777	17.68	<.0001	3.12591
m_center8	m_center8	1	0.00113	0.00016245	6.98	<.0001	0.48683

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II
Intercept	Intercept	1	0	.	.	.
m_center	m_center	1	0.68893	0.99184	0.99184	0.00580
m_center2	m_center2	1	-0.22400	0.00543	0.66493	0.00004896
m_center3	m_center3	1	1.96352	0.00096657	0.35354	0.00114
m_center4	m_center4	1	2.56749	0.00008005	0.04530	0.00015351
m_center5	m_center5	1	-4.05336	0.00007002	0.04150	0.00120
m_center6	m_center6	1	-5.40620	0.00008522	0.05269	0.00018330
m_center7	m_center7	1	2.41391	0.00126	0.82418	0.00126
m_center8	m_center8	1	3.14406	0.00019666	0.73007	0.00019666

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	48.17870	48.42130
m_center	m_center	1	0.01223	81.77568	2.41327	2.69649
m_center2	m_center2	1	0.00097588	1024.71375	-0.58788	-0.14539
m_center3	m_center3	1	0.00029644	3373.39253	0.49911	0.64159
m_center4	m_center4	1	0.00002329	42942	0.16296	0.33145
m_center5	m_center5	1	0.00007311	13678	-0.08772	-0.06867
m_center6	m_center6	1	0.00000627	159448	-0.04159	-0.02181
m_center7	m_center7	1	0.00021670	4614.58071	0.00261	0.00332
m_center8	m_center8	1	0.00001989	50266	0.00079215	0.00147

Correlation of Estimates

Variable	Label	Intercept	m_center	m_center2	m_center3	m_center4
Intercept	Intercept	1.0000	0.0000	-0.7805	0.0000	0.6824
m_center	m_center	0.0000	1.0000	0.0000	-0.9334	0.0000
m_center2	m_center2	-0.7805	0.0000	1.0000	0.0000	-0.9794
m_center3	m_center3	0.0000	-0.9334	0.0000	1.0000	0.0000
m_center4	m_center4	0.6824	0.0000	-0.9794	0.0000	1.0000
m_center5	m_center5	0.0000	0.8708	0.0000	-0.9867	0.0000
m_center6	m_center6	-0.6390	0.0000	0.9572	0.0000	-0.9955
m_center7	m_center7	0.0000	-0.8324	0.0000	0.9691	0.0000
m_center8	m_center8	0.6170	0.0000	-0.9426	0.0000	0.9896

Correlation of Estimates

Variable	Label	m_center5	m_center6	m_center7	m_center8
Intercept	Intercept	0.0000	-0.6390	0.0000	0.6170
m_center	m_center	0.8708	0.0000	-0.8324	0.0000
m_center2	m_center2	0.0000	0.9572	0.0000	-0.9426
m_center3	m_center3	-0.9867	0.0000	0.9691	0.0000
m_center4	m_center4	0.0000	-0.9955	0.0000	0.9896
m_center5	m_center5	1.0000	0.0000	-0.9961	0.0000
m_center6	m_center6	0.0000	1.0000	0.0000	-0.9987
m_center7	m_center7	-0.9961	0.0000	1.0000	0.0000
m_center8	m_center8	0.0000	-0.9987	0.0000	1.0000

Wnioski z raportu.

1. Równanie modelu:

$$\hat{Y} = 48{,}30 + 2{,}555\,X - 0{,}367\,X^2 + 0{,}570\,X^3 + 0{,}247\,X^4 - 0{,}078\,X^5 - 0{,}032\,X^6 + 0{,}003\,X^7 + 0{,}247\,X^8$$

(5-2-2-3.8)

2. Istotność statystyczna modelu:

   Niska wartość empirycznego poziomu istotności $p < 0.0001$oznacza, że wartość statystyki $F = 30941{,}8$ jest istotna statystycznie, co wskazuje na istotność ogólnej zależności korelacyjnej w modelu.

3. Wysoka wartość współczynnika determinacji ($R^2 = 0.9999$) wskazuje na bardzo dobre dopasowanie modelu do danych empirycznych.

4. Istotność parametrów strukturalnych modelu:

   Wszystkie parametry strukturalne modelu (a zatem i odpowiadające im zmienne objaśniające) są istotne, gdyż z testu t przeprowadzonego przez system SAS widać, że empiryczny poziom istotności p dla każdego z parametrów jest niski:

   a) dla większości parametrów strukturalnych $p < 0.0001$,

   b) dla parametru β_3, $p = 0.0027$ (co przy powszechnie przyjętych poziomach istotności uważane jest ciągle za niewiele).

5. Analiza macierzy korelacji:

   Z analizy macierzy korelacji wynika, iż w modelach wielomianowych centrowanych, wraz ze wzrostem stopnia wielomianu, zmienne są coraz to częściej mocno skorelowane. Jest to wyraz skuteczności centrowania jedynie w modelach z niskim stopniem wielomianu.

Rozdział 5-2-3. Wielomian ortogonalny ósmego stopnia.

Odpowiedni raport SAS'a ma postać:

wielomian ortogonalny
ósmego stopnia

The REG Procedure
Model: MODEL1
Dependent Variable: dochod dochod

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2475.34667	309.41833	30941.8	<.0001
Error	18	0.18000	0.01000		
Corrected Total	26	2475.52667			

Root MSE	0.10000	R-Square	0.9999
Dependent Mean	49.07778	Adj R-Sq	0.9999
Coeff Var	0.20376		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	49.07778	0.01925	2550.16	<.0001	65033
Xortn1	Xortn1	1	28.60844	0.05774	495.51	<.0001	2455.32800
Xortn2	Xortn2	1	2.11587	0.05774	36.65	<.0001	13.43069
Xortn3	Xortn3	1	0.89308	0.05774	15.47	<.0001	2.39276
Xortn4	Xortn4	1	-0.25702	0.05774	-4.45	0.0003	0.19818
Xortn5	Xortn5	1	0.24037	0.05774	4.16	0.0006	0.17333
Xortn6	Xortn6	1	0.26519	0.05774	4.59	0.0002	0.21097
Xortn7	Xortn7	1	1.02077	0.05774	17.68	<.0001	3.12591
Xortn8	Xortn8	1	0.40283	0.05774	6.98	<.0001	0.48683

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II
Intercept	Intercept	1	0	.	.	.
Xortn1	Xortn1	1	0.99591	0.99184	0.99184	0.99184
Xortn2	Xortn2	1	0.07366	0.00543	0.66493	0.00543
Xortn3	Xortn3	1	0.03109	0.00096657	0.35354	0.00096657
Xortn4	Xortn4	1	-0.00895	0.00008005	0.04530	0.00008005
Xortn5	Xortn5	1	0.00837	0.00007002	0.04150	0.00007002
Xortn6	Xortn6	1	0.00923	0.00008522	0.05269	0.00008522
Xortn7	Xortn7	1	0.03553	0.00126	0.82418	0.00126
Xortn8	Xortn8	1	0.01402	0.00019666	0.73007	0.00019666

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	49.03735	49.11821
Xortn1	Xortn1	1	1.00000	1.00000	28.48714	28.72973
Xortn2	Xortn2	1	1.00000	1.00000	1.99457	2.23716
Xortn3	Xortn3	1	1.00000	1.00000	0.77178	1.01437
Xortn4	Xortn4	1	1.00000	1.00000	-0.37832	-0.13572
Xortn5	Xortn5	1	1.00000	1.00000	0.11907	0.36167
Xortn6	Xortn6	1	1.00000	1.00000	0.14389	0.38648
Xortn7	Xortn7	1	1.00000	1.00000	0.89947	1.14207
Xortn8	Xortn8	1	1.00000	1.00000	0.28154	0.52413

Correlation of Estimates

Variable	Label	Intercept	Xortn1	Xortn2	Xortn3	Xortn4
Intercept	Intercept	1.0000	0.0000	-0.0000	-0.0000	0.0000
Xortn1	Xortn1	0.0000	1.0000	-0.0000	0.0000	-0.0000
Xortn2	Xortn2	-0.0000	-0.0000	1.0000	0.0000	-0.0000
Xortn3	Xortn3	-0.0000	0.0000	0.0000	1.0000	0.0000
Xortn4	Xortn4	0.0000	-0.0000	-0.0000	0.0000	1.0000
Xortn5	Xortn5	0.0000	0.0000	0.0000	0.0000	-0.0000
Xortn6	Xortn6	0.0000	0.0000	-0.0000	-0.0000	-0.0000
Xortn7	Xortn7	-0.0000	0.0000	-0.0000	-0.0000	-0.0000
Xortn8	Xortn8	-0.0000	0.0000	0.0000	-0.0000	-0.0000

Correlation of Estimates

Variable	Label	Xortn5	Xortn6	Xortn7	Xortn8
Intercept	Intercept	0.0000	0.0000	-0.0000	-0.0000
Xortn1	Xortn1	0.0000	0.0000	0.0000	0.0000
Xortn2	Xortn2	0.0000	-0.0000	-0.0000	0.0000
Xortn3	Xortn3	0.0000	-0.0000	-0.0000	-0.0000
Xortn4	Xortn4	-0.0000	-0.0000	-0.0000	-0.0000
Xortn5	Xortn5	1.0000	0.0000	-0.0000	-0.0000
Xortn6	Xortn6	0.0000	1.0000	0.0000	0.0000
Xortn7	Xortn7	-0.0000	0.0000	1.0000	0.0000
Xortn8	Xortn8	-0.0000	0.0000	0.0000	1.0000

Uwaga. *W powyższym raporcie widać, że wartości standardowych błędów ocen wartości standardowych współczynników regresji są równe.*

Wnioski z raportu.

1. Równanie modelu:

$$\hat{Y} = 49{,}08 + 28{,}61X + 2{,}12X^2 + 0{,}89X^3 - 0{,}26X^4 + 0{,}24X^5 + 0{,}27X^6 + 1{,}02X^7 + 0{,}40X^8 \qquad (5\text{-}2\text{-}3.9)$$

2. Istotność statystyczna modelu:

   Niski empiryczny poziom istotności $p < 0.0001$ wskazuje, że wartość statystyki $F = 30941{,}8$ jest istotna statystycznie, co wskazuje na istotność ogólnej zależności korelacyjnej w modelu.

3. Wysoka wartość współczynnika determinacji ($R^2 = 0.9999$) wskazuje na bardzo dobre dopasowanie modelu do danych empirycznych.

4. Istotność parametrów strukturalnych modelu:

   Wszystkie parametry strukturalne modelu (a zatem i odpowiadające im zmienne objaśniające) są istotne, gdyż z testu t przeprowadzonego przez system SAS widać, że empiryczny poziom istotności p dla każdego z parametrów jest niski:

   a) dla większości parametrów $p < 0.0001$,

   b) dla parametrów $\beta_4, \beta_5, \beta_6$ otrzymujemy odpowiednio: $p = 0{,}0003$, $p = 0{,}0006$, $p = 0{,}0002$ (które to wartości przy powszechnie przyjętych poziomach istotności α, uważane są za niskie).

5. Analiza macierzy korelacji:

   Macierz korelacji wskazuje na brak współzależności pomiędzy estymatorami parametrów strukturalnych, a zatem również na brak korelacji (i współliniowości) zortogonalizowanych zmiennych objaśniających.

Powyżej, przy okazji testowania istotności rozszerzenia modelu parabolicznego do sześciennego stwierdziliśmy, że (na każdym poziomie istotności α $\geq p = 0{,}0017$) rozszerzenie modelu parabolicznego do sześciennego jest istotne statystycznie z punktu widzenia poprawy dokładności dopasowania się linii regresji do danych empirycznych.

Gdyby jednak przyjąć poziom istotności α $< p = 0{,}0017$ (np. α = 0,001), wtedy rozszerzenie to nie byłoby istotne statystycznie. Uczyńmy tak chociażby dla celów zaprezentowania testu o niewystępowaniu braku dopasowania. Zatem, chociaż przy prawdziwości hipotezy zerowej o nieistotności rozszerzenia modelu parabolicznego do sześciennego, prawdopodobieństwo pojawienia się w próbce tak dużej fluktuacji (tzn. tak dużej wartości licznika statystyki częściowej (5-2-2-2.6)) dla której $p = 0{,}0017$, nie jest duże, to przyjmijmy, że jest to wynik nieistotny statystycznie. Zdecydujemy się więc na wniosek, że nie było podstaw do odrzucenia modelu parabolicznego.

Pozostaje sprawdzenie, czy w modelu parabolicznym nie ma braku dopasowania funkcji regresji do danych empirycznych w porównaniu z modelem maksymalnego, ósmego stopnia. Całe wnioskowanie przeprowadźmy na wybranym powyżej poziomie istotności α = 0,001.

Wartość testowej statystyki częściowej F_p w próbce wyliczamy ze wzoru (4-1-2-3.15) korzystając z raportów dla modeli z wielomianami drugiego i ósmego stopnia. Obserwowana wartość jest równa:

$$\begin{aligned} F_p^{obs} &= F(X^3,\cdots,X^8 \mid X, X^2) \\ &= \frac{[SSR(X, X^2, X^3,..., X^8) - SSR(X, X^2)]/df_{LOF}}{SSE(X, X^2, X^3,..., X^8)/df_{PE}} \\ &= \frac{MS_{LOF}}{MS_{PE}} = \frac{(2475{,}34667 - 2468{,}75869)/(8-2)}{0{,}01000} = 109{,}79967 \ . \end{aligned} \tag{5-2-3.10}$$

Przypomnijmy, że żaden model nie da mniejszej wartości średniego kwadratu dla reszt, czyli MS_{PE} (dla tzw. czystego błędu), niż model maksymalny.

Ponieważ przy prawdziwości hipotezy zerowej o nie występowaniu braku dopasowania w modelu drugiego stopnia, powyższa statystyka F_p ma rozkład F-Snedecora z liczbami stopni swobody licznika $df_{LOF} = 8-2 = 6$ oraz mianownika $df_{PE} = n-1-8 = 18$, zatem (korzystając z programu Excel) wyznaczmy empiryczny poziom istotności $p = \mathrm{P}(F_p \geq F_p^{obs} = 109{,}8) \approx 3{,}49\times10^{-13}$. Jest to wyjątkowo małe prawdopodobieństwo. Zatem, na każdym poziomie istotności $\alpha \geq p \approx 3{,}49\times10^{-13}$ (więc również dla $\alpha = 0{,}001$), hipoteza o nie występowaniu braku dopasowania w modelu parabolicznym zostaje odrzucona, co oznacza, że istnieją zmienne wyższego stopnia niż 2, których dodanie do grona zmiennych objaśniających w sposób istotny statystycznie poprawiłoby dopasowanie funkcji regresji modelu do danych empirycznych.

W samej rzeczy, korzystając z metody eliminacji wstecz (Rozdział 6-3) dla modelu z wielomianem ósmego stopnia dla zmiennych ortonormalnych, można by się przekonać, że wartości wszystkich estymatorów parametrów strukturalnych są w pobranej próbce istotne statystycznie na każdym poziomie istotności $\alpha \geq 0{,}001$. Oznacza to, że model, który należałoby zastosować, jest modelem ósmego stopnia.

Rozdział 5-3. Ogólne wnioski z przeprowadzonej analizy regresji wielomianowej.

Wyniki przeprowadzonej analizy pozwalają na stwierdzenie, iż najlepsze rezultaty daje zastosowanie modeli wielomianowych ortogonalnych ze względu na brak korelacji pomiędzy zmiennymi objaśniającymi. Modele te nie tylko dają takie samo dopasowanie się funkcji regresji II- rodzaju do danych empirycznych jak modele ze zmiennymi zwyczajnymi, ale ze względu na brak współliniowości pomiędzy zmiennymi, otrzymane np. 95%-owe przedziały ufności (95% Confidence Limits w raportach SAS'a) dla parametrów strukturalnych, są węższe niż w modelach tego samego stopnia ze zmiennymi zwyczajnymi. Węższy przedział ufności dla parametru strukturalnego oznacza mniejsze rozproszenie możliwych wartości estymatora parametru strukturalnego w próbkach, a co za tym idzie lepszą zdolność modelu do predykcji wartości zmiennej objaśnianej.
Drugimi w kolejności są modele centrowane, które chociaż tylko częściowo skuteczne w usuwaniu korelacji czynników, mają jednak pewną zaletę w porównaniu ze zmiennymi zortogonalizowanymi. Otóż, ponieważ ich użycie oznacza jedynie przesunięcie początku układu współrzędnych, więc są one znacznie łatwiejsze w praktycznym zastosowaniu modelu.

A. Rozdział 6: Wybór najlepszego modelu regresji.

Wybór modelu regresji może być podporządkowany jednemu z dwóch celów. Pierwszy z nich wyznacza jako priorytet dokładność predykcji zmiennej objaśnianej, drugi cel to otrzymanie modelu z jak najistotniejszymi współczynnikami regresji.

Aby wybrać najlepsze równanie regresji należy wykonać następujące kroki [1]:

1. określić maksymalny model regresji.
2. określić kryterium wyboru modelu.
3. określić strategię wyboru zmiennych do modelu.
4. przeprowadzić analizę modelu.
5. oszacować wiarygodność wybranego modelu.

Rozdział 6-1. Krok 1. Określenie maksymalnego modelu regresji.

Maksymalny model definiujemy jako model, który zawiera największą liczbę zmiennych objaśniających, wykorzystywanych w selekcji modelu. Wszystkie inne modele mogą zostać utworzone poprzez usuwanie zmiennych z modelu maksymalnego. Modele z usuniętymi zmiennymi nazywamy modelami ograniczonymi (zredukowanymi).

Przyjmijmy, że model maksymalny zawiera *m*-zmiennych, oraz że modele ograniczone zawierają $b \leq m$ zmiennych.

Model maksymalny powinien być wybierany w taki sposób, aby zawierał możliwie jak najwięcej informacji, powinien być bardzo rozbudowany, aby uniknąć możliwości popełnienia błędu drugiego rodzaju, tzn. pominięcia istotnej zmiennej, co miałoby miejsce, gdyby odpowiednia hipoteza zerowa $\beta_j = 0$ nie została odrzucona pomimo jej fałszywości.

Model maksymalny powinien zawierać [1]:

1. wszystkie podstawowe możliwe zmienne objaśniające,
2. zmienne podstawowe wyższego rzędu (X^2, X^3,…),
3. różne transformacje zmiennych np. log*X*, 1/*X*,
4. interakcje pomiędzy zmiennymi, zawierające dwukierunkowe i wyższego rzędu współzależności,

Model ten zawiera bardzo dużo informacji. Jednak stosowanie modelu z takim zestawem zmiennych, stwarza nie tylko trudności ze względu na jego rozbudowaną formę, ale powoduje również możliwość wystąpienia współzależności między zmiennymi objaśniającymi. Problemy te są dla nas jednak drugorzędne z tego powodu, iż model ten ma jedynie posłużyć do wyboru modelu *najlepszego* tzn. takiego, który będzie zawierał możliwie jak najwięcej informacji i jednocześnie będzie posiadał możliwie najprostszą strukturę.

Można by postulować wybór modelu maksymalnego, który posiadałby mniejszą liczbę zmiennych tzn. tylko takich zmiennych, które są według badacza istotne. Zatem w praktyce model maksymalny nie jest np. modelem z możliwym największym stopniem występujących zmiennych, a zwrot „model maksymalny" stanowi określenie modelu uznanego za pierwotny wobec potem następującej selekcji. Jednak takie podejście musi być, ze względu na możliwość pominięcie istotnej zmiennej, stosowane bardzo ostrożnie.

Rozdział 6-2. Krok 2. Określenie kryterium wyboru modelu.

W wyborze najlepszego modelu pomagają tzw. kryteria selekcji. Wybrane kryterium może być stosowane do porównywania modeli kandydujących w celu wyboru najlepszego modelu. Kryteriów oceny modeli może być wiele i w dalszej części zostaną opisane cztery najpopularniejsze. Oczywiście wybór modelu, który byłby najlepszym modelem pod kątem wszystkich kryteriów, jest mało prawdopodobne. Dlatego też doboru modelu powinno się dokonywać ze względu na konkretne kryterium.

Obrane kryterium powinno się wiązać z celem analizy, np. w przypadku, gdy zależy nam na dokładności predykcji, kryterium selekcji powinno być nieco liberalne, aby uniknąć „przywiązania" do którejś ze zmiennych.

Liczne kryteria selekcji modeli zostały zaproponowane przez Hocking'a [13]. Cztery najpopularniejsze kryteria selekcji to: $R^2(p)$, $F_p(p)$, $MSE(p)$ i $C(p)$ (p występujące w argumencie statystyk oznacza model proponowany, natomiast indeks „p" dotyczy tak jak poprzednio testów częściowych). W Rozdziale 8-1 omówiono krótko kryterium informacyjne Akaike'a (AIC).

- Kryterium $R^2(p)$.

Według tego kryterium najlepszym modelem jest ten, dla którego wielokrotny współczynnik korelacji R^2 jest największy. Wartość tego współczynnika w modelu z p - zmiennymi wynika z zależności (3-3-2.43):

$$R^2(p) = R^2(Y \mid X_1, X_2, \ldots, X_p) = 1 - \frac{SSE(p)}{SSY} \ . \qquad (6\text{-}2.1)$$

Niestety kryterium to posiada następujące wady:

1. skłonność do przeszacowania przez $R^2(p)$ odpowiedniej wielkości w populacji, co jest związane z tym, że w skrajnym przypadku wartość R^2 w próbce może być równa 1, pomimo, że wartość ta może nie mieć nic wspólnego z dobrocią dopasowania się funkcji regresji I rodzaju w populacji.
2. dodawanie nawet bezwartościowych czynników nigdy nie powoduje zmniejszenia wartości $R^2(p)$. Przeciwnie, zawsze następuje wzrost wartości $R^2(p)$. Z tego wynika, że najwyższa wartość współczynnika $R^2(p)$ występuje w modelu maksymalnym, który może nie być przydatny w badaniu populacji ze względu na jego rozbudowaną strukturę.

Dlatego też najlepszy model może mieć mniejszą wartość współczynnika R^2, ale w zamian będzie: (1) bliższy właściwemu modelowi w populacji oraz (2) praktyczniejszy w użyciu.

- Kryterium F_p.

W kryterium tym porównujemy model maksymalny z badanym na podstawie statystyki F_p testu częściowego, obliczonej ze wzoru [1]:

$$F_p(p) = \frac{[SSR(m) - SSR(p)]/(m-p)}{SSE(m)/(n-m-1)} = \frac{[SSE(p) - SSE(m)]/(m-p)}{MSE(m)} \ . \qquad (6\text{-}2.2)$$

Przy prawdziwości hipotezy zerowej:

H_0: poprawa dopasowania do danych empirycznych modelu „m" w stosunku do modelu „p" jest nieistotna statystycznie,

zmienna $F_p(p)$ ma rozkład F-Snedeckora z (m - p) i (n – m - 1) stopniami swobody. Statystyka F_p testuje czy różnica pomiędzy sumą kwadratów reszt dla modelu proponowanego z p-zmiennymi i modelu maksymalnego z m-zmiennymi jest istotnie statystycznie różna od zera.

Jeżeli wartość F_p nie jest istotna statystycznie (tzn. empiryczny poziom istotności $P(F_p(p) \geq F_p^{obs}(p)) > \alpha$), to możemy przyjąć proponowany model „*p*" jako dobry ze względu na zbliżoną dobroć (dokładność) dopasowania sie do danych empirycznych jak model maksymalny.

W przypadku, gdy $p = m - 1$, statystyka F_p jest statystyką testową dla hipotezy H_0: $\beta_m = 0$.

- Kryterium *MSE*(*p*).

Korzystając z tego kryterium szukamy modeli o najmniejszych średnich kwadratach reszt (najmniejszych wartościach estymatorów średniej wariancji wewnątrzgrupowej).

Średni kwadrat reszt dla modelu klasycznego MNK z *p*-zmiennymi ma postać (5-50)):

$$MSE(p) = \frac{SSE(p)}{n-p-1} \quad . \tag{6-2.3}$$

Według tego kryterium wybieramy model o małym średnim rozproszeniu wartości empirycznych reszt wokół linii (powierzchni) regresji przy ustalonych wartościach zmiennych objaśniających, czyli model o możliwie jak najmniejszym *MSE*(*p*).

- Kryterium Mallows' *C*(*p*) [14].

Zdefiniujmy tzw. współczynnik Mallows'a[9] *C*(*p*) [1]:

$$C(p) = \frac{SSE(p)}{MSE(m)} - [n - 2(p+1)] \ . \tag{6-2.4}$$

W przypadku gdy model „*p*" ma minimalną możliwą wartość *MSE*, tzn. gdy:

$$MSE(p) = MSE(m) \ , \tag{6-2.5}$$

wtedy[10] wartość:

$$C(p) = p + 1 \ . \tag{6-2.6}$$

Wielkość *C*(*p*) pozwala określić liczbę zmiennych objaśniających, jaka powinna się znaleźć w modelu, co wynika z faktu, że:

$$C(p) \approx p+1 \quad \text{o ile} \quad MSE(p) \approx MSE(m) \ . \tag{6-2.7}$$

Jeśli ważne zmienne objaśniające zostały pominięte, wtedy *C*(p) powinno być większe od $p + 1$.

---

[9] Statystyka *C*(*p*) Mallows'a jest nieobciążonym estymatorem średniego kwadratu błędu przewidywania (*MSPE* - mean squared prediction error) w populacji, zapisanego następująco [15]:

$$E\left(\sum_i \left(\hat{Y}_i - E(Y \mid X_{1i}, X_{2i}, ..., X_{pi})\right)^2 / \sigma_E^2 \right) ,$$

gdzie $\hat{Y}_i$ jest teoretyczną średnią warunkową z modelu regresji dla *i*-tej jednostki, a $E(Y \mid X_{1i}, X_{2i}, ..., X_{pi})$ jest wartością oczekiwaną warunkową odpowiedzi, natomiast σ_E^2 jest wariancją składnika losowego, o której zakłada się, że jest jednorodna.

[10] Wykorzystaj (6-2.5) z (6-2.3) w (6-2.4).

Uwaga. Wielkość $p + 1$ występująca w zależności na współczynnik $C(p)$ jest rozumiana dla modeli zawierających wyraz wolny (przesunięcie) inaczej niż dla modeli, w których pomijamy ten wyraz, a mianowicie:

a) $p + 1$ oznacza dla modelu zawierającego przesunięcie liczbę równą ilości zmiennych objaśniających plus jeden (tzn. plus zmienna jednostkowa I dla przesunięcia),

b) $p + 1$ oznacza dla modelu nie zawierającego przesunięcia liczbę zmiennych objaśniających.

Nietrudno pokazać, że istnieje związek pomiędzy powyższymi statystykami, a mianowicie:

$$F_p(p) = \frac{\left(R_m^2 - R_p^2\right)/(m-p)}{\left(1 - R_m^2\right)/(n-m-1)} \tag{6-2.8}$$

oraz:

$$C(p) = (m-p)F_p(p) + (2p - m + 1) \ . \tag{6-2.9}$$

Znaczenie *SSE*: *O dokładności dopasowania modelu do danych empirycznych decyduje wielkość SSE* (otrzymana dla dopasowania funkcji regresji do danych empirycznych metodą najmniejszych kwadratów). Zatem jeśli $SSE(p) = SSE(m)$, wtedy dopasowanie (a więc i predykcja) modelu proponowanego jest taka jak modelu maksymalnego.

Przy warunku $SSE(p) = SSE(m)$ z (6-2.2) widać, że $F_p(p) = 0$, a zatem $C(p)$ osiąga wtedy swoją **minimalną wartość**, równą:

$C(p) = C_{\min}(p) \equiv 2p - m + 1 \le p + 1$, $\quad (m \ge p)$, przy warunku: $SSE(p) = SSE(m)$. (6-2.10)

Uwaga: Ze względu na ważność kryterium dokładności dopasowania modelu do danych empirycznych, a zatem i jakość predykcji, statystycy preferują często modele z najmniejszą wartością $C(p)$ jako głównym kryterium doboru modelu. *Tzn. te proponowane modele, które mają wartość* $C(p)$ *najbliższą wartości* $2p - m + 1$ *są uznawane za najlepsze*.

Z kolei istnienie różnicy pomiędzy rzeczywistą wartością $C(p)$ modelu a wartością minimalną $C_{\min}(p) \equiv 2p - m + 1$, wskazuje na liczbę brakujących zmiennych w modelu (co jest przejawem tego, że $SSE(m)$ jest mniejsze niż $SSE(p)$).

Aby wybrać najlepszy model musimy zdecydować się, które z powyższych kryteriów uznajemy za nadrzędne, gdyż niemożliwe jest aby wszystkie kryteria były spełnione przez jeden model. Możemy otrzymać nawet tyle najlepszych modeli, ile jest kryteriów wyboru. Niemniej, nie można kierować się tylko jednym kryterium. Najlepszym modelem jest ten, który znajduje się wysoko w każdym z kryteriów.

Uwaga. W [4] omówiono również metodę Hellwiga doboru czynników.

Rozdział 6-3. Krok 3. Określenie strategii wyboru zmiennych do modelu.

W trzecim kroku wybieramy strategię wyboru zmiennych, określającą jak wiele i które zmienne będziemy używać w modelu. Są następujące główne strategie wyboru zmiennych [1]:

- Porównywanie wszystkich możliwych modeli regresji.
- Poprzez dodawanie nowych zmiennych do modelu mało rozbudowanego (metoda doboru wprzód – forward selection procedure).
- Poprzez odejmowanie zmiennych z modelu bardzo rozbudowanego (metoda eliminacji wstecz – backward elimination procedure).
- Strategia krocząca – stepwise regression procedure.

Procedura porównania wszystkich możliwych modeli regresji. Procedura ta polega na zestawieniu wszystkich możliwych modeli regresji w tabeli, a następnie wyboru spośród nich modelu, który (według kryteriów podanych w kroku 2 Rozdziału 6-2), powinien w idealnym przypadku mieć: największe $R^2(p)$, najmniejsze $MSE(p)$, najmniejsze $C(p)$, i wszystkie wartości statystyk częściowych $F_p(p)$ statystycznie istotne. Ten sposób daje największe gwarancje wyboru odpowiedniego modelu, ale przy dużej ilości zmiennych objaśniających jest bardzo niepraktyczny, gdyż ilość modeli do porównania może być bardzo duża.
Maksymalna ilość modeli wynosi:

$$\sum_{p=0}^{m} \binom{m}{p} - 1 = 2^m - 1 \qquad (6\text{-}3.11)$$

dla m-zmiennych. Np. dla m = 10 liczba modeli do porównania wynosi 2^{10}-1=1023. Zatem wybór modelu polega na znalezieniu takiego modelu, który posiada najodpowiedniejsze z punktu widzenia badacza wartości $R^2(p)$, $MSE(p)$, $C(p)$, F_p . Systaem SAS daje taką możliwość dla podstawowych kryteriów selekcji modelu. Odpowiednią procedurę aplikacji Analyst można wywołać po uruchomieniu procedury REG dokonując ciągu wyborów Solutions->Analysis->Analyst->Statistics->Regresion->Linear->Model->Mallows'Cp .

Procedura doboru „w przód". W procedurze tej przechodzimy następujące etapy:

1. Jako *pierwszą zmienną wchodzącą* do modelu przyjmujemy tę, która ma największą wartość kwadratu współczynnika korelacji R^2 ze zmienną objaśnianą Y. Zatem wyznaczamy odpowiedni model regresji biorąc pod uwagę jedynie tę zmienną. Następnie obliczamy wartość ogólnej statystyki F (tzn. statystyki, którą testujemy hipotezę o niezależności korelacyjnej zmiennej objaśnianej od zmiennej objaśniającej) dla tego modelu regresji. Jeśli otrzymana wartość F nie jest istotna statystycznie, zatrzymujemy się i wnioskujemy, że żadna ze zmiennych objaśniających nie jest istotna statystycznie. Jeśli jednak test był istotny statystycznie (tzn. wartość statystyki testowej była istotna statystycznie), wtedy włączamy tą zmienną objaśniającą do modelu i przystępujemy do punktu drugiego omawianej procedury.

Dla procedury tej istotna jest statystyka:

$$SSR(\textit{ostatnia dodana zmienna} \mid \textit{zmienne poprzednie}), \qquad (6\text{-}3.12)$$

którą w SAS'ie podaje statystyka **Type I SS**.

2. Opierając się o modele regresji, w których jest zmienna wybrana w punkcie (1), oraz po kolei i z osobna dodana nowa zmienna, określamy wartości częściowych statystyk F_p oraz wartości empirycznych poziomów istotności p (które są z nimi związane), dla każdej z nowych zmiennych objaśniających dodanych na końcu.
3. Skupiamy się na zmiennej z największą wartością częściowej statystyki F_p (najmniejszym empirycznym poziomie istotności p) wyznaczoną w punkcie drugim. Jeśli otrzymana największa wartość F_p jest istotna statystycznie, wtedy dodajemy odpowiadającą jej zmienną objaśniającą jako czynnik do modelu. Jeśli F_p nie jest istotne statystycznie, wtedy w modelu pozostawiamy jedynie zmienną z punkcie pierwszego.
4. O dalszym doborze zmiennych decydują wartości częściowych statystyk F_p zmiennych jeszcze nie ujętych w modelu. Jeśli wynik odpowiedniego testu jest istotny statystycznie, to dodajemy zmienną do modelu.

Procedura eliminacji wstecz (BACKWARD ELIMINATION). W procedurze tej przechodzimy następujące etapy:

1. Określamy równanie regresji zawierające wszystkie zmienne objaśniające.
2. Obliczamy wartości częściowego F_p (lub wartość empirycznego poziomu istotności p) dla każdej zmiennej w modelu. Dla procedury tej istotna jest statystyka:

$$SSR(\textit{jedna z rozważanych zmiennych} \mid \textit{zmienne pozostałe}), \qquad (6\text{-}3.13)$$

   którą w SAS'ie podaje statystyka **Type II SS**.
3. Zwracamy uwagę na najniższe wartości statystyki częściowej F_p lub najwyższe wartości p.
4. Porównujemy najwyższą wartość empirycznego poziomu istotności p z wartością wcześniej wybranego poziomu istotności α (na pozostanie zmiennej w modelu) i decydujemy o usunięciu bądź zostawieniu rozważanej zmiennej.
5. Jeżeli w punkcie (4) decydujemy się na usunięcie zmiennej, to powtarzamy (bez tej jednej zmiennej) punkty 1, 2, 3, 4, aż dojdziemy do układu zmiennych, z których żadnej nie usuwamy. Wówczas otrzymujemy szukany model.

Strategia krocząca. Procedura ta jest modyfikacją procedury selekcji „w przód". W procedurze tej przy każdym kroku sprawdzane są wartości częściowego F_p, aby możliwe było wyeliminowanie zmiennych, które „utraciły" swoją istotność po wprowadzeniu nowych zmiennych silnie z nią skorelowanych. Usunięciu podlegają zmienne o zbyt małym częściowym F_p.

Rozdział 6-3-1. Procedura wyboru najlepszego modelu regresji na przykładzie metody eliminacji wstecz.

Procedurą tą przeprowadzimy dla danych przedstawionych poniżej w Przykładzie „Ceny mieszkań".

Procedura doboru zmiennych poprzez eliminację w programie SAS określana jest w: Solutions->Analysis->Analyst->Statistics->Regression->Linear ->Model. W tym miejscu możliwy jest także wybór innej metody np. metody kroczącej (Stepwise Regression Procedure). Następnie należy określić szczegóły przeprowadzanej procedury tj. w szczególności określić *poziom istotności* α *na pozostanie zmiennej w modelu*, co jest możliwe w zakładce „Criteria".

Przykład: „Ceny mieszkań".

W pewnym mieście przeprowadzono badania dotyczące ceny mieszkań (Y w 10000 PLN). W tym celu zebrano następujące parametry trzydziestu losowo wybranych mieszkań:

- X_1 - powierzchnia mieszkania w 10 m^2,
- X_2 - liczba łazienek,
- X_3 - liczba pokoi,
- X_4 - wiek budynku,
- Z – lokalizacja.

Należy dokonać wyboru tych cech mieszkań, które mają najistotniejszy wpływ na ich cenę (będącą zmienną objaśnianą).

Dane otrzymane dla próbki 30 losowo wybranych mieszkań są następujące [1]:

Mieszkanie	**Y**	**X_1**	**X_2**	**X_3**	**X_4**	**Z**
1	84,0	13,8	3	7	10	0
2	93,0	19	2	7	22	1
3	83,1	10	2	7	15	1
4	85,2	15	3	7	15	1
5	85,2	12	3	7	8	1
6	85,2	15	3	7	12	1
7	85,2	12	3	7	8	1
8	63,3	9,1	3	6	2	1
9	84,3	12,5	3	7	11	1
10	84,3	12,5	3	7	11	1
11	77,4	12	3	7	5	0
12	92,4	17,9	3	7	18	0
13	92,4	17,9	3	7	18	0
14	61,5	9,5	2	5	8	0
15	88,5	16	3	7	11	0
16	88,5	16	3	7	11	0
17	40,6	8	2	5	5	0
18	81,6	11,8	3	7	8	1
19	86,7	16	3	7	9	0
20	89,7	16,8	2	7	12	0
21	86,7	16	3	7	9	0
22	89,7	16,8	2	7	12	0
23	75,9	9,5	3	6	6	1
24	78,9	10	3	6	11	0
25	87,9	16,5	3	7	15	0
26	91,0	15,1	3	7	8	1
27	92,0	17,9	3	8	13	1
28	87,9	16,5	3	7	15	0
29	90,9	15	3	7	8	1
30	91,9	17,8	3	8	13	1

2. Model maksymalny zawiera następujące czynniki:
   - X_1 - powierzchnia mieszkania w 10 m^2,
   - X_2 - liczba łazienek,
   - X_3 - liczba pokoi,
   - X_4 - wiek budynku,
   - Z - lokalizacja, która przyjmuje wartości:
     - 0 - dla mieszkań w centrum miasta,
     - 1 - dla mieszkania na przedmieściach
   - Y – cena mieszkania w 10000 PLN (jest zmienną objaśnianą *Y*)

Pomijamy w poniższej analizie zarówno wyższe potęgi zmiennych objaśniających jak i człony oddziaływania typu $X_3 \cdot Z$, które należałoby potraktować jako nowe zmienne objaśniające.

Zatem model maksymalny ma następującą strukturę:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 Z \ . \qquad (6\text{-}3\text{-}1.14)$$

W programie SAS wybieramy opcję „Backward Elimination”. Ścieżka dostępu do tej opcji w SAS’ie jest następująca: Solution-> Analysis-> Analyst-> a następnie (po wczytaniu danych, korzystając z „Open By SAS Name” w zakładce File [11]) -> Statistics-> Regression-> Linear-> (określenie zmiennej objaśnianej „Dependent” -w przykładzie jest to „cena”- i zmiennych objaśniających „Explanatory” - w przykładzie są to X_1, X_2, X_3, X_4, Z) -> Model (Method)-> Backward elimination.

Wybór tej metody jest uzasadniony małym prawdopodobieństwem pominięcia, przy jej zastosowaniu, ważnej zmiennej w wyborze ostatecznego modelu. Zatem wprowadzamy wszystkie dostępne zmienne do modelu maksymalnego i ustawiamy poziom istotności na pozostanie zmiennej w modelu (w naszym przypadku wybraliśmy wartość 0,05 dla poziomu istotności α na pozostanie zmiennej w modelu). Następnie wykonujemy obliczenia. Przeanalizujmy teraz sposób postępowania wygenerowany przez program. Dla wygody i czytelności raport został podzielony na poszczególne kroki procedury.

1. **Krok zerowy.** Procedura eliminacji (Backward elimination).

```
                    Procedura eliminacji

                      The REG Procedure
                        Model: MODEL1
                   Dependent Variable: Y Y

                 Backward Elimination: Step 0

     All Variables Entered: R-Square = 0.8321 and C(p) = 6.0000

                    Analysis of Variance

                              Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F

Model                    5    2957.41937      591.48387      23.79    <.0001
Error                   24     596.59030       24.85793
Corrected Total         29    3554.00967

                 Parameter     Standard
       Variable   Estimate        Error   Type II SS  F Value  Pr > F

       Intercept  -6.39664     11.87684      7.21052     0.29  0.5951
       X1          0.91200      0.62682     52.62170     2.12  0.1586
       X2          3.39757      2.71086     39.04684     1.57  0.2222
       X3          8.95816      2.74738    264.27966    10.63  0.0033
       X4          0.55188      0.32370     72.25478     2.91  0.1011
       Z           0.45446      2.26230      1.00313     0.04  0.8425

            Bounds on condition number: 4.5338, 67.232
```

Analiza raportu.

Określono model maksymalny, w którym znajdują się wszystkie możliwe zmienne (m = 5). Model maksymalny możemy zapisać równaniem:

$$\hat{Y} = -6{,}397 + 0{,}912\ X_1 + 3{,}397 X_2 + 8{,}958\ X_3 + 0{,}552\ X_4 + 0{,}454\ Z$$

Obliczono współczynnik determinacji R^2 = 0.8321, statystykę F = 23,79 oraz wartość współczynnika Mallows'a $C(p)$ = 6 dla modelu maksymalnego (p = m). Kierując się jedynie tymi rezultatami można wnioskować o tym, że otrzymany model jest najlepszy (co jest zrozumiałe gdyż jest to model maksymalny). Jednakże naszym celem jest otrzymanie modelu możliwie jak najlepszego, zarówno pod względem dokładności jak i możliwie najmniej rozbudowanej struktury.

Dla każdej ze zmiennych został przeprowadzony test istotności, którego wynikiem są następujące wartości empirycznych poziomów istotności p dla testów z odpowiednią (dla wskazanego parametru) zmienną dodaną na końcu:

- dla parametru β_0 , p = 0,5951,
- dla parametru β_1, p = 0,1586,
- dla parametru β_2, p = 0,2222,
- dla parametru β_3, p = 0,0033,
- dla parametru β_4, p = 0,1011,
- dla parametru β_5 (dla zmiennej Z), p = 0,8425 .

Zauważmy, iż największa wartość p występuje przy weryfikacji hipotezy o nieistotności dodania parametru stojącego przy zmiennej Z. *Zatem w kolejnym kroku zostanie ona wyeliminowana*, co widać w następnej części raportu. (Przypomnijmy, że wartość p = 0,8425 oznacza, że odrzucając hipotezę H_0: β_4 = 0, na poziomie istotności $\alpha = p$ = 0,8425 pomylilibyśmy się ponad 84 razy na sto.)

2. **Krok pierwszy.** Procedura eliminacji (Backward elimination).

```
                    Backward Elimination: Step 1

          Variable Z Removed: R-Square = 0.8319 and C(p) = 4.0404

                          Analysis of Variance
                                 Sum of           Mean
Source                  DF        Squares         Square    F Value    Pr > F

Model                    4     2956.41623      739.10406      30.92    <.0001
Error                   25      597.59344       23.90374
Corrected Total         29     3554.00967

                  Parameter     Standard
       Variable    Estimate        Error   Type II SS  F Value  Pr > F

       Intercept   -7.35535     10.66502     11.36972     0.48  0.4968
       X1           0.84653      0.52506     62.13433     2.60  0.1195
       X2           3.38486      2.65760     38.77649     1.62  0.2145
       X3           9.25743      2.26368    399.77637    16.72  0.0004
       X4           0.56114      0.31419     76.24869     3.19  0.0862

                    Backward Elimination: Step 1

             Bounds on condition number: 3.3082, 38.176
```

Zatem zmienna Z została wyeliminowana z równania regresji modelu. Na tym etapie, model możemy zapisać następująco:

$$\hat{Y} = -7{,}355 + 0{,}847X_1 + 3{,}385X_2 + 9{,}257X_3 + 0{,}561X_4 \ . \qquad (6\text{-}3\text{-}1.15)$$

Model ten charakteryzuje się następującymi wartościami współczynników: $R^2(p)$ = 0,8319, F = 30,92, $C(p)$ = 4,0404. W tym kroku analizując wyniki, można by spodziewać się, że usunięciu ulegnie estymator wyrazu wolnego, gdyż empiryczny poziom istotności dla weryfikacji hipotezy o nieistotności parametru β_0 jest najwyższy i wynosi p = 0,4968. Jednak nasze domysły okazałyby się błędne, gdyż program SAS nie usuwa estymatora parametru przesunięcia automatycznie. Zatem w następnym kroku usunięciu ulegnie parametr β_2, dla którego p w teście częściowym F_p wynosi 0,2145.

3. **Krok drugi.** Procedura eliminacji (Backward elimination).

```
                    Backward Elimination: Step 2

        Variable X2 Removed: R-Square = 0.8209 and C(p) = 3.6003

                         Analysis of Variance

                                  Sum of           Mean
Source                   DF      Squares         Square    F Value    Pr > F

Model                     3   2917.63974      972.54658      39.74    <.0001
Error                    26    636.36993       24.47577
Corrected Total          29   3554.00967

                  Parameter     Standard
       Variable    Estimate        Error   Type II SS  F Value  Pr > F

       Intercept   -4.82468     10.60293      5.06783     0.21  0.6529
       X1           0.84448      0.53130     61.83456     2.53  0.1240
       X3          10.49267      2.06973    629.04683    25.70  <.0001
       X4           0.42756      0.29969     49.81885     2.04  0.1656

                 Bounds on condition number: 3.3082, 22.202
```

Zgodnie z przewidywaniami wyeliminowana została zmienna X_2, a otrzymany dotychczas model jest następujący:

$$\hat{Y} = -4{,}825 + 0{,}844\, X_1 + 10{,}493\, X_3 + 0{,}428\, X_4 \ . \qquad (6\text{-}3\text{-}1.16)$$

Współczynniki modelu są następujące: $R^2(p)$ = 0,8209, F = 39,74, C(p) = 3,6003. Analizując wartości prawdopodobieństw dla poszczególnych zmiennych (z wyłączeniem wyrazu wolnego) wnioskujemy, że następną wyeliminowaną zmienną będzie zmienna X_4, gdyż empiryczny poziom istotności p, związany z testem na nieistotność parametru β_4 , wynosi 0,1656 i jest ono większy od przyjętego poziomu istotności na pozostanie zmiennej w modelu α= 0,05. Uznajemy więc, przy przyjętym poziomie istotności α = 0,05, że odrzucenie hipotezy zerowej H_0: β_4 =0 jest ciągle obarczone za dużym prawdopodobieństwo pomyłki, a wartość statystyki F_p = 2,04 uznajemy za nieistotną statystycznie (tzn. uznajemy, że mogła pochodzić z „fluktuacji”) i w konsekwencji, nie pozostawiamy zmiennej X_4 w modelu.

4. **Krok trzeci.** Procedura eliminacji (Backward elimination).

```
                    Backward Elimination: Step 3

          Variable X4 Removed: R-Square = 0.8069 and C(p) = 3.6044

                         Analysis of Variance

                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F

Model                    2    2867.82088     1433.91044      56.42    <.0001
Error                   27     686.18878       25.41440
Corrected Total         29    3554.00967

                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F

       Intercept    -4.21356     10.79550      3.87161     0.15  0.6994
       X1            1.30268      0.43128    231.86141     9.12  0.0055
       X3           10.14196      2.09411    596.10738    23.46  <.0001

                 Bounds on condition number: 2.0994, 8.3975
```

Zgodnie z przewidywaniami zmienna X_4 została wyeliminowana z równania modelu. Na tym etapie doboru zmiennych do modelu zauważmy, że wartości empirycznych poziomów istotności p dla wszystkich parametrów (pomijając β_0) są niższe od przyjętego poziomu istotności α= 0,05 i w związku z tym, na tym kroku eliminacja została zakończona. W ostatniej części raportu znajduje się podsumowanie całej procedury.

5. **Krok czwarty (podsumowanie).** Procedura eliminacji (Backward elimination).

```
      All variables left in the model are significant at the 0.0500 level.

                    Summary of Backward Elimination

      Variable            Number    Partial     Model
Step  Removed     Label  Vars In   R-Square   R-Square    C(p)     F Value   Pr > F

  1   Z           Z          4      0.0003     0.8319     4.0404      0.04   0.8425
  2   X2          X2         3      0.0109     0.8209     3.6003      1.62   0.2145
  3   X4          X4         2      0.0140     0.8069     3.6044      2.04   0.1656

                         Analysis of Variance

                                  Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F

Model                    2    2867.82088     1433.91044      56.42    <.0001
Error                   27     686.18878       25.41440
Corrected Total         29    3554.00967

             Root MSE              5.04127    R-Square     0.8069
             Dependent Mean       83.49667    Adj R-Sq     0.7926
             Coeff Var             6.03769

                          Parameter Estimates

                   Parameter     Standard
Variable   Label      DF    Estimate        Error  t Value  Pr > |t|    Type I SS

Intercept  Intercept   1    -4.21356     10.79550    -0.39    0.6994       209151
X1         X1          1     1.30268      0.43128     3.02    0.0055   2271.71350
X3         X3          1    10.14196      2.09411     4.84    <.0001    596.10738
```

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II
Intercept	Intercept	1	0	.	.	.
X1	X1	1	0.37008	0.63920	0.63920	0.06524
X3	X3	1	0.59340	0.16773	0.46487	0.16773

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	-26.36410	17.93699
X1	X1	1	0.47633	2.09938	0.41776	2.18760
X3	X3	1	0.47633	2.09938	5.84520	14.43871

Podsumowanie zawiera kolejno:

1. informacje o przyjętym poziomie istotności pozostania zmiennej w modelu (w powyższym raporcie 0.05),
2. informacje o czynnikach usuniętych z modelu (w raporcie Z, X_2, X_4), wraz z wartościami statystyk użytych przy eliminacji,
3. dane o wybranym modelu, ilość i rodzaj danych uzależniony jest od ich wyboru na początku procedury. Zostają wydrukowane wybrane wartości, statystyki, ewentualnie wykresy.

Ostateczny model wyselekcjonowany według kryterium F_p ma następującą charakterystykę:

1. funkcja regresji II rodzaju dla modelu:

$$\hat{Y} = -4{,}213 + 1{,}303X_1 + 10{,}142X_3 \qquad (6\text{-}3\text{-}1.17)$$

2. wartość współczynnika determinacji:

   $R^2(p) = 0{,}8069$, gdzie $p = 2$,

3. wartość statystyki ogólnej F, oraz empiryczny poziom istotności (określający -w tym przypadku dużą- istotność zależności korelacyjnej *ceny mieszkania Y* od *powierzchni* X_1 i *liczby pokoi* X_3 w modelu):

   $F = 56{,}42$

   $p < 0{,}0001$

4. wartość współczynnika Mallows'a $C(p)$:

   $C(p) = 3{,}6044$,

   Powyższą wartość C(p) wyznaczył SAS, przyjmując p +1 = 3 jako liczbę wszystkich zmiennych w badanym modelu (*tym razem z przesunięciem*). Sprawdźmy ten wynik, korzystając z wartości zawartych w raporcie:

$$C(p) = \frac{SSE(p)}{MSE(m)} - [n - 2(p+1)] = \frac{686{,}18878}{24{,}85793} - [30 - 2\cdot(2+1)] = 3.60442 .$$

Porównując tą wartość $C(p)$ z wartością minimalną $C_{\min}(p) \equiv 2p - m + 1 = 2 \cdot 2 - 5 + 1 = 0$ widzimy, że wartość $C(p)$ = 3,6044 jest daleka od wartości minimalnej 0, co oznacza, że badany model nie najlepiej dopasowuje się do danych empirycznych (w porównaniu z modelem maksymalnym z m = 5, w którym $C(m)$ = 6, co jest również wartością minimalną kryterium Mallows'a przyjętego modelu maksymalnego). Wskazuje to na brak jakiś istotnych zmiennych objaśniających w badanym modelu i sugeruje rezygnację z tak daleko posuniętej procedury eliminacji wstecz, i powrót do modelu z jedynie usuniętą zmienną Z, dla którego $C(p)$ = 4.0404 jest wartością bardzo bliską wartości minimalnej równej $C_{\min}(p) \equiv 2p - m + 1 = 2 \cdot 4 - 5 + 1 = 4$.

5. w wyselekcjonowanym modelu według kryterium F_p, wartości empirycznych poziomów istotności dla testów o nieistotności parametrów strukturalnych są następujące:
   - dla parametru strukturalnego β_1 przy zmiennej X_1, p = 0,0055,
   - dla parametru strukturalnego β_3 przy zmiennej X_3, p < 0,0001,
   - dla parametru strukturalnego β_0 z wyrazu wolnego, p = 0,6994.

Zatem wybrany model charakteryzuje się:

- stosunkowo wysokim współczynnikiem determinacji,
- istotnością statystyczną wszystkich parametrów stojących przy zmiennych objaśniających. Tzn. na średnią cenę mieszkania mają w sposób istotny statystycznie wpływ: powierzchnia mieszkania X_1 i liczba pokoi X_3,
- w modelu znajduje się statystycznie nieistotny estymator wyrazu wolnego.

W celu otrzymania modelu, w którym nie będzie uwzględniony estymator wyrazu wolnego należy zmodyfikować procedurę w następujący sposób.

Jeżeli, w którymś kroku procedury, osoba badająca zauważy nieistotność estymatora wyrazu wolnego, powinna odnotować zestaw zmiennych aktualnego modelu, a następnie ponowić procedurę. Wznowiona analiza powinna być przeprowadzona dla modelu odnotowanego z wyłączeniem wyrazu wolnego. W naszym przypadku po pierwszym wykonaniu kroku, powinniśmy odnotować zmienne wchodzące do modelu pomijając wyraz wolny (**do not include intercept -** w komendzie SAS'a umieszczonej w opcji Regression> Linear> Model (Method)). Zatem do dalszego badania przechodzą następujące zmienne: X_1, X_2, X_3, X_4 (oczywiście bez wyrazu wolnego). Dalsze postępowanie nie różni się już od wcześniej opisanego. Ostatecznie otrzymujemy następujący raport:

Raport. Procedura eliminacji (Backward elimination) po wyłączeniu wyrazu wolnego.

```
Procedura po wyłączeniu wyrazu wolnego
The REG Procedure
Model: MODEL1
Dependent Variable: Y Y
```

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.9971 and C(p) = 4.0000

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	212096	53024	2263.89	<.0001
Error	26	608.96315	23.42166		
Uncorrected Total	30	212705			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X1	0.96679	0.49024	91.08893	3.89	0.0593
X2	3.04339	2.58461	32.47461	1.39	0.2497
X3	8.11446	1.52635	661.95478	28.26	<.0001
X4	0.53956	0.30946	71.20242	3.04	0.0930

Bounds on condition number: 140.55, 1161.4

Backward Elimination: Step 1

Variable X2 Removed: R-Square = 0.9970 and C(p) = 3.3865

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	212063	70688	2975.46	<.0001
Error	27	641.43776	23.75695		
Uncorrected Total	30	212705			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X1	0.92634	0.49252	84.03905	3.54	0.0708
X3	9.63131	0.82456	3241.28068	136.44	<.0001
X4	0.42205	0.29502	48.62264	2.05	0.1640

Bounds on condition number: 64.089, 359.12

Backward Elimination: Step 2

Variable X4 Removed: R-Square = 0.9968 and C(p) = 3.4625

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	212015	106007	4301.37	<.0001
Error	28	690.06040	24.64501		
Uncorrected Total	30	212705			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
X1	1.36912	0.39022	303.38597	12.31	0.0015
X3	9.39243	0.82243	3214.30507	130.42	<.0001

Bounds on condition number: 38.78, 155.12

```
            All variables left in the model are significant at the 0.0500 level.

NOTE: No intercept in model. R-Square is redefined.


                              Summary of Backward Elimination

         Variable           Number    Partial     Model
   Step  Removed    Label   Vars In   R-Square   R-Square    C(p)     F Value   Pr > F

     1   X2         X2          3      0.0002     0.9970     3.3865      1.39   0.2497
     2   X4         X4          2      0.0002     0.9968     3.4625      2.05   0.1640

                    NOTE: No intercept in model. R-Square is redefined.

                                   Analysis of Variance

                                          Sum of           Mean
       Source                   DF        Squares         Square    F Value    Pr > F

       Model                     2         212015         106007    4301.37    <.0001
       Error                    28      690.06040       24.64501
       Uncorrected Total        30         212705


                    Root MSE              4.96437    R-Square     0.9968
                    Dependent Mean       83.49667    Adj R-Sq     0.9965
                    Coeff Var             5.94560

                                  Parameter Estimates

                            Parameter     Standard
Variable   Label      DF    Estimate        Error  t Value  Pr > |t|    Type I SS

X1         X1          1     1.36912      0.39022     3.51    0.0015       208800
X3         X3          1     9.39243      0.82243    11.42    <.0001   3214.30507

                                  Parameter Estimates
                                          Squared      Squared       Squared
                            Standardized  Semi-partial      Partial  Semi-partial
 Variable   Label      DF      Estimate   Corr Type I  Corr Type I  Corr Type II

 X1         X1          1       0.23519       0.98164      0.98164       0.00143
 X3         X3          1       0.76553       0.01511      0.82326       0.01511


                                  Parameter Estimates

                                                Variance
      Variable   Label      DF    Tolerance      Inflation       95% Confidence Limits

      X1         X1          1      0.02579       38.78026        0.56979        2.16845
      X3         X3          1      0.02579       38.78026        7.70776       11.07710
```

Pomijamy analizę poszczególnych kroków postępowania, gdyż przebiega ona jak we wcześniejszym przypadku z włączonym „na sztywno” przesunięciem. Wypiszemy jednak charakterystykę otrzymanego modelu:

1. równanie regresji dla modelu:

$$\hat{Y} = 1{,}369X_1 + 9{,}392X_3 \qquad (6\text{-}3\text{-}1.18)$$

2. wartość współczynnika determinacji:

$R^2(p) = 0{,}9968,$

3. wartość statystyki F oraz empiryczny poziom istotności dla zależności korelacyjnej ceny mieszkania od powierzchni i liczby pokoi w modelu:

   $F = 4301{,}37$

   $p < 0{,}0001$

4. wartość współczynnika $C(p)$:

   $C(p) = 3{,}4625$,

Powyższą wartość $C(p)$ liczymy przyjmując $p + 1 = 2$ jako liczbę wszystkich zmiennych w badanym modelu, otrzymując po skorzystaniu z wartości zawartych w raporcie:

$$C(p) = \frac{SSE(p)}{MSE(m)} - [n - 2(p+1)] = \frac{690{,}06040}{23{,}42166} - [30 - 2 \cdot 2] = 3{,}46249 .$$

Widać, że wartość $C(p) = 3{,}4625$ jest daleka od wartości minimalnej $C_{\min}(p) \equiv 2p - m + 1 = 2 \cdot 1 - 3 + 1 = 0$ dla modelu, co oznacza, że model kiepsko dopasowuje się do danych empirycznych (w porównaniu z modelem maksymalnym, **w którym** tym razem $m = 3$ i dla którego $C(m) = m + 1 = 4$, jak to jest widoczne na górze raportu dla modelu, w którym **są zmienne** X_1, X_2, X_3, X_4). Wskazuje to na brak jakiś istotnych zmiennych objaśniających w modelu i sugeruje rezygnację z tak daleko posuniętej procedury eliminacji wstecz, i powrót do jednego z modeli wyższych.

5. wartości empirycznych poziomów istotności dla testów o nieistotności parametrów strukturalnych w modelu:
   - dla parametru strukturalnego β_1 przy zmiennej X_1, $p = 0{,}0015$,
   - dla parametru strukturalnego β_3 przy zmiennej X_3, $p < 0{,}0001$,

Na koniec zauważmy, że chociaż wyselekcjonowany przez procedurę eliminacji wstecz model bez wyrazu wolnego (przesunięcia) lepiej dopasowuje się do danych empirycznych, $R^2(p) = 0{,}9968$, niż model z przesunięciem, $R^2(p) = 0{,}8069$, to tak w jednym jak i w drugim przypadku należałoby zdecydować się raczej na modele wyższe, dla których $C(p)$ jest bliższe odpowiednim wartościom minimalnym. Na dodatek okazuje się, że obydwa powyższe modele, jako niewycentrowane, wykazują dużą korelację (i współliniowość) pomiędzy zmiennymi objaśniającymi (tą część raportu pominięto z powodu podobnej analizy przeprowadzonej poniżej dla modelu maksymalnego i wyselekcjonowanego ostatecznie według kryterium $C(p)$).

Rozdział 6-4. Przykład analizy współliniowości dla modelu maksymalnego z niewycentrowanymi zmiennymi.

Przeprowadźmy analizę współliniowości modelu maksymalnego dla przykładu „Ceny mieszkań”. W programie SAS wybieramy ścieżkę dostępu do opcji „Full model”: Solution-> Analysis-> Analyst-> a następnie (po wczytaniu danych, korzystając z „Open By SAS Name” w zakładce File [11]) -> Statistics-> Regression-> Linear-> (zmiennymi objaśniającymi „Explanatory” są w modelu maksymalnym X_1, X_2, X_3, X_4, Z) -> Model (Method)-> Full model.

Odpowiedni raport SAS’a ma postać (część 1):

Analiza współliniowości w modelu maksymalnym

The REG Procedure
Model: MODEL1
Dependent Variable: Y Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2957.41937	591.48387	23.79	<.0001
Error	24	596.59030	24.85793		
Corrected Total	29	3554.00967			

Root MSE	4.98577	R-Square	0.8321
Dependent Mean	83.49667	Adj R-Sq	0.7972
Coeff Var	5.97122		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	-6.39664	11.87684	-0.54	0.5951	209151
X1	X1	1	0.91200	0.62682	1.45	0.1586	2271.71350
X2	X2	1	3.39757	2.71086	1.25	0.2222	193.20196
X3	X3	1	8.95816	2.74738	3.26	0.0033	415.25208
X4	X4	1	0.55188	0.32370	1.70	0.1011	76.24869
Z	Z	1	0.45446	2.26230	0.20	0.8425	1.00313

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	-30.90923	18.11596
X1	X1	1	0.22056	4.53383	-0.38170	2.20569
X2	X2	1	0.70471	1.41903	-2.19738	8.99251
X3	X3	1	0.27068	3.69441	3.28784	14.62848
X4	X4	1	0.44346	2.25501	-0.11621	1.21996
Z	Z	1	0.64760	1.54417	-4.21469	5.12362

W powyższej części raportu zamieszczone są podstawowe dane o modelu i jego parametrach, oraz takie wielkości wykorzystywane w analizie współliniowości jak współczynnik tolerancji i współczynnik inflacji wariancji. Wartość współczynnika tolerancji (4-5.36), $Tolerancja_1 = 1 - R_1^2 = 0{,}22056$(dla $R_1^2 = 0{,}77944$), co świadczy o występowaniu „raczej” dużej korelacji zmiennej X_1 z pozostałymi czynnikami. Natomiast

wydaje się, że współczynnik inflacji wariancji (tzw. nadęcie wariancji), (4-5.35), nie daje tak negatywnej prognozy, gdyż najwyższa jego wartość wynosi $VIF_1 = 1/(1-R_1^2) = 4{,}53383$ (podczas gdy wartość sygnalizująca definitywne występowanie silnej współzależności wynosi co najmniej 10). Zatem, przyjrzymy się macierzy korelacji dla estymatorów parametrów strukturalnych oraz wartościom własnym λ_j, (4-6.53), w układzie składowych głównych modelu (dla analizy związanej z macierzą korelacji dla czynników modelu).

Raport dla analizy współliniowości w modelu maksymalnym (część 2).

Correlation of Estimates

Variable	Label	Intercept	X1	X2
Intercept	Intercept	1.0000	0.4687	-0.1612
X1	X1	0.4687	1.0000	0.0147
X2	X2	-0.1612	0.0147	1.0000
X3	X3	-0.7812	-0.7006	-0.3725
X4	X4	-0.1475	-0.5550	0.3270
Z	Z	0.4018	0.5199	0.0233

Correlation of Estimates

Variable	Label	X3	X4	Z
Intercept	Intercept	-0.7812	-0.1475	0.4018
X1	X1	-0.7006	-0.5550	0.5199
X2	X2	-0.3725	0.3270	0.0233
X3	X3	1.0000	0.0424	-0.5422
X4	X4	0.0424	1.0000	-0.1425
Z	Z	-0.5422	-0.1425	1.0000

Badając macierz korelacji estymatorów parametrów trudno jednoznacznie stwierdzić występowanie dużej korelacji (i współliniowości) pomiędzy czynnikami, gdyż wartości bezwzględne współczynników korelacji dla estymatorów parametrów strukturalnych (czyli elementów poza diagonalną) przyjmują maksymalnie wartość 0,7812, na ogół przyjmując wartości znacznie niższe.

Przejdźmy więc do ostatniej części wydruku, w której dokonano analizy współzależności metodą wartości własnych macierzy korelacji czynników (4-6.48).

Raport dla analiza współliniowości w modelu maksymalnym (część 3).

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	**5.40935**	**1.00000**
2	**0.46687**	**3.40389**
3	**0.09806**	**7.42715**
4	**0.01485**	**19.08764**
5	**0.00921**	**24.23423**
6	**0.00166**	**57.04421**

Collinearity Diagnostics

Number	Intercept	X1	X2	X3	X4	Z
	----------Proportion of Variation----------					
1	0.00019642	0.00033339	0.00046773	0.00007887	0.00178	0.00654
2	0.00016246	0.00097496	0.00019968	0.00004538	0.00662	0.60496
3	0.00582	0.00095773	0.02956	0.00090709	0.37060	0.04976
4	0.06979	0.45765	0.02362	0.00057339	0.45976	0.03896
5	0.21104	0.00462	0.88086	0.01666	0.14525	0.00166
6	0.71300	0.53547	0.06530	0.98174	0.01599	0.29811

W ostatniej części raportu zawarte są wartości własne macierzy korelacji, oraz wartości indeksów warunkowych CI_j , (4-6.63). Przypomnijmy, że wartości wskazujące na występowanie dużej korelacji pomiędzy zmiennymi, to bliskie zeru wartości własne λ_j macierzy kowariancji czynników (4-6.48) lub dla indeksu warunkowego CI_j , wartości przekraczające 30.

Otrzymany wydruk pozwala w końcu na wyciągnięcie wniosku o istnieniu dużej korelacji (i współliniowości) pomiędzy zmiennymi objaśniającymi, gdyż występują w nim, co najmniej dwie niskie wartości własne (tzn.: λ_5 = 0.00921, λ_6 = 0.00166) (dla numerowania wartości własnych zaczynającego się od 1), oraz występuje wartość indeksu warunkowego przekraczająca 30 (CI_6 = 57.04421).
Podsumowując, dochodzimy do wniosku, że nie można kierować się tylko jednym wskaźnikiem korelacji (współliniowości) zmiennych, gdyż pociąga to za sobą możliwość popełnienia błędu w ocenie tej własności modelu. Wystąpienie choćby jednego sygnału o możliwości wystąpienia korelacji powinno skłonić badacza do zastosowania chociażby wycentrowania zmiennych (Rozdział 5) oraz przyjęcia jednej z opisanych wcześniej metod doboru czynników do modelu.

6-5. Przykład eliminacji współliniowości poprzez centrowanie i standaryzację (przeliczyć).

W obecnym rozdziale podamy przykład przeprowadzenia analizy ze zmiennymi standaryzowanymi (więc automatycznie wycentrowanymi) dla danych przykładu „Ceny mieszkań”. Przypomnijmy, że z przeprowadzonej powyższej analizy współliniowości modelu maksymalnego ze zmiennymi zwykłymi wynika, że czynniki mogą być ze sobą silnie skorelowane, a zatem i wyselekcjonowany model, może dawać znaczną niepewność co do dokładności opartych o niego przewidywań (cen mieszkań).
Przeprowadźmy więc procedurę eliminacji wstecz (Backward Elimination) dla danych standaryzowanych. Otrzymane wyniki będziemy analizować zasadniczo pod kątem dwu kryteriów doboru zmiennych do modelu, a mianowicie:

- Kryterium F_p jako kryterium standardowego w procedurze selekcji w systemie SAS,
- Kryterium Mallows’a $C(p)$ jako kryterium preferowanego przez wielu statystyków [1], gdyż pozwala ono do pewnego stopnia na wypowiedź o liczbie brakujących czynników w modelu (z różnicy pomiędzy $C(p)$ modelu a wartością minimalną dla modelu).

Standaryzowane dane dla przykładu: „Ceny mieszkań”.

Mieszkanie	Y	X_1c	X_2c	X_3c	X_4c	Zc
1	84,0	-0,1049	0,4916	0,2573	-0,2251	-0,9832
2	93,0	1,5485	-1,9664	0,2573	2,5689	0,9832
3	83,1	-1,3132	-1,9664	0,2573	0,9391	0,9832
4	85,2	0,2766	0,4916	0,2573	0,9391	0,9832
5	85,2	-0,6773	0,4916	0,2573	-0,6907	0,9832
6	85,2	0,2766	0,4916	0,2573	0,2406	0,9832
7	85,2	-0,6773	0,4916	0,2573	-0,6907	0,9832
8	63,3	-1,5994	0,4916	-1,2866	-2,0877	0,9832
9	84,3	-0,5183	0,4916	0,2573	0,0078	0,9832
10	84,3	-0,5183	0,4916	0,2573	0,0078	0,9832
11	77,4	-0,6773	0,4916	0,2573	-1,3892	-0,9832
12	92,4	1,1987	0,4916	0,2573	1,6376	-0,9832
13	92,4	1,1987	0,4916	0,2573	1,6376	-0,9832
14	61,5	-1,4722	-1,9664	-2,8304	-0,6907	-0,9832
15	88,5	0,5946	0,4916	0,2573	0,0078	-0,9832
16	88,5	0,5946	0,4916	0,2573	0,0078	-0,9832
17	40,6	-1,9491	-1,9664	-2,8304	-1,3892	-0,9832
18	81,6	-0,7409	0,4916	0,2573	-0,6907	0,9832
19	86,7	0,5946	0,4916	0,2573	-0,4579	-0,9832
20	89,7	0,8490	-1,9664	0,2573	0,2406	-0,9832
21	86,7	0,5946	0,4916	0,2573	-0,4579	-0,9832
22	89,7	0,8490	-1,9664	0,2573	0,2406	-0,9832
23	75,9	-1,4722	0,4916	-1,2866	-1,1564	0,9832
24	78,9	-1,3132	0,4916	-1,2866	0,0078	-0,9832
25	87,9	0,7536	0,4916	0,2573	0,9391	-0,9832
26	91,0	0,3084	0,4916	0,2573	-0,6907	0,9832
27	92,0	1,1987	0,4916	1,8012	0,4734	0,9832
28	87,9	0,7536	0,4916	0,2573	0,9391	-0,9832
29	90,9	0,2766	0,4916	0,2573	-0,6907	0,9832
30	91,9	1,1669	0,4916	1,8012	0,4734	0,9832

W poniższych wydrukach litera „c” przy zmiennych oznacza, iż jest to zmienna standaryzowana zgodnie z (4-5.44) (zatem i wycentrowana). Pełny raport Systemu SAS w tym przypadku wygląda następująco.

Raport. Procedura eliminacji (Backward elimination).

```
Procedura eliminacji
 dane standaryzowane

The REG Procedure
  Model: MODEL1
Dependent Variable: Y Y

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8321 and C(p) = 6.0000
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2957.41937	591.48387	23.79	<.0001
Error	24	596.59030	24.85793		
Corrected Total	29	3554.00967			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.49667	0.91027	209151	8413.85	<.0001
X1c	2.86825	1.97136	52.62170	2.12	0.1586
X2c	1.38226	1.10288	39.04684	1.57	0.2222
X3c	5.80237	1.77953	264.27966	10.63	0.0033
X4c	2.37033	1.39030	72.25478	2.91	0.1011
Zc	0.23112	1.15049	1.00313	0.04	0.8425

Bounds on condition number: 4.5338, 67.232

Backward Elimination: Step 1

Variable Zc Removed: R-Square = 0.8319 and C(p) = 4.0404

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2956.41623	739.10406	30.92	<.0001
Error	25	597.59344	23.90374		
Corrected Total	29	3554.00967			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.49667	0.89263	209151	8749.71	<.0001
X1c	2.66235	1.65132	62.13433	2.60	0.1195
X2c	1.37709	1.08121	38.77649	1.62	0.2145
X3c	5.99621	1.46623	399.77637	16.72	0.0004
X4c	2.41012	1.34944	76.24869	3.19	0.0862

Backward Elimination: Step 1

Bounds on condition number: 3.3082, 38.176

Backward Elimination: Step 2

Variable X2c Removed: R-Square = 0.8209 and C(p) = 3.6003

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2917.63974	972.54658	39.74	<.0001
Error	26	636.36993	24.47577		
Corrected Total	29	3554.00967			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.49667	0.90325	209151	8545.22	<.0001
X1c	2.65590	1.67095	61.83456	2.53	0.1240
X3c	6.79630	1.34060	629.04683	25.70	<.0001
X4c	1.83639	1.28717	49.81885	2.04	0.1656

Bounds on condition number: 3.3082, 22.202

Backward Elimination: Step 3

Variable X4c Removed: R-Square = 0.8069 and C(p) = 3.6044

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2867.82088	1433.91044	56.42	<.0001
Error	27	686.18878	25.41440		
Corrected Total	29	3554.00967			

Backward Elimination: Step 3

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.49667	0.92041	209151	8229.62	<.0001
X1c	4.09695	1.35640	231.86141	9.12	0.0055
X3c	6.56914	1.35640	596.10738	23.46	<.0001

Bounds on condition number: 2.0994, 8.3975

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Zc	Zc	4	0.0003	0.8319	4.0404	0.04	0.8425
2	X2c	X2c	3	0.0109	0.8209	3.6003	1.62	0.2145
3	X4c	X4c	2	0.0140	0.8069	3.6044	2.04	0.1656

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2867.82088	1433.91044	56.42	<.0001
Error	27	686.18878	25.41440		
Corrected Total	29	3554.00967			

Root MSE	5.04127	R-Square	0.8069
Dependent Mean	83.49667	Adj R-Sq	0.7926
Coeff Var	6.03769		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	83.49667	0.92041	90.72	<.0001	209151
X1c	X1c	1	4.09695	1.35640	3.02	0.0055	2271.71350
X3c	X3c	1	6.56914	1.35640	4.84	<.0001	596.10738

Z powyższego raportu wynika, że model, w który wchodzą czynniki X_1c, X_3c oraz wyraz wolny jest najlepszym, ale tylko z punktu widzenia kryterium F_p doboru zmiennych z parametrami strukturalnymi istotnymi statystycznie.

Istotnie, dokonajmy analizy współczynnika $C(p)$. Dla poszczególnych modeli otrzymanych w każdym z kroków procedury eliminacji wstecz, porównamy otrzymaną z (6-2.4) wartość współczynnika $C(p)$ z wartością minimalną dla danego modelu. Wartość minimalna jest jak zwykle wyznaczona z $C_{\min}(p) \equiv 2p - m + 1$, (6-2.10).

Wyniki analizy są przedstawione w Tabeli 6-1.1.

Tabela 6-5.1. Porównanie modeli na podstawie współczynnika $C(p)$.

Krok procedury	Zestaw zmiennych	Współczynnik $C(p)$ modelu	Optymalny współczynnik $C(p)$
Zerowy, $p = m = 5$	X_1c X_2c X_3c X_4c Zc	$C(p) = 6{,}0000$	$C(p) = 6{,}0000$
Pierwszy, $p = 4$	X_1c X_2c X_3c X_4c	$C(p) = 4{,}0404$	$C(p) = 4{,}0000$
Drugi, $p = 3$	X_1c X_3c X_4c	$C(p) = 3{,}6003$	$C(p) = 2{,}0000$
Trzeci, $p=4$	X_1c X_3c	$C(p) = 3{,}6044$	$C(p) = 0{,}0000$

Z wyników zawartych w powyższej tabeli wynika, iż najlepszym modelem w rozważanym zagadnieniu jest model składający się ze zmiennych X_1c, X_2c, X_3c, X_4c, ponieważ wartość współczynnika $C(p) = 4{,}0404$ tego modelu różni się nieznacznie (o mniej niż 0,5) od wartości minimalnej $C_{\min}(p) \equiv 2p - m + 1 = 4$ tego współczynnika, co oznacza, że nie ma w wyselekcjonowanym modelu miejsca nawet na jedną dodatkową zmienną objaśniającą. Przeprowadzimy teraz podstawową analizę zaproponowanego modelu.

Analiza modelu charakteryzującego się najlepszym współczynnikiem Mallows'a.

Odpowiedni raport SAS'a ma postać:

```
                    dane standaryzowane
                  model z najlepszym C(p)

                    The REG Procedure
                      Model: MODEL1
                 Dependent Variable: Y Y

                   Analysis of Variance

                            Sum of           Mean
Source                DF       Squares         Square    F Value    Pr > F

Model                  4    2956.41623      739.10406      30.92    <.0001
Error                 25     597.59344       23.90374
Corrected Total       29    3554.00967


          Root MSE              4.88914    R-Square     0.8319
          Dependent Mean       83.49667    Adj R-Sq     0.8050
          Coeff Var             5.85550
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	Intercept	1	83.49667	0.89263	93.54	<.0001	209151
X1c	X1c	1	2.66235	1.65132	1.61	0.1195	2271.71350
X2c	X2c	1	1.37709	1.08121	1.27	0.2145	193.20196
X3c	X3c	1	5.99621	1.46623	4.09	0.0004	415.25208
X4c	X4c	1	2.41012	1.34944	1.79	0.0862	76.24869

Parameter Estimates

Variable	Label	DF	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Partial Corr Type I	Squared Semi-partial Corr Type II
Intercept	Intercept	1	0	.	.	.
X1c	X1c	1	0.24049	0.63920	0.63920	0.01748
X2c	X2c	1	0.12439	0.05436	0.15067	0.01091
X3c	X3c	1	0.54165	0.11684	0.38128	0.11249
X4c	X4c	1	0.21771	0.02145	0.11316	0.02145

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	.	0	81.65826	85.33508
X1c	X1c	1	0.30228	3.30822	-0.73861	6.06330
X2c	X2c	1	0.70509	1.41826	-0.84971	3.60389
X3c	X3c	1	0.38341	2.60817	2.97646	9.01597
X4c	X4c	1	0.45264	2.20924	-0.36911	5.18935

Correlation of Estimates

Variable	Label	Intercept	X1c	X2c	X3c	X4c
Intercept	Intercept	1.0000	0.0000	-0.0000	-0.0000	-0.0000
X1c	X1c	0.0000	1.0000	0.0031	-0.5834	-0.5688
X2c	X2c	-0.0000	0.0031	1.0000	-0.4284	0.3338
X3c	X3c	-0.0000	-0.5834	-0.4284	1.0000	-0.0419
X4c	X4c	-0.0000	-0.5688	0.3338	-0.0419	1.0000

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	2.28623	1.00000
2	1.19648	1.38232
3	1.00000	1.51203
4	0.32462	2.65381
5	0.19267	3.44471

Collinearity Diagnostics

Number	Proportion of Variation: Intercept	X1c	X2c	X3c	X4c
1	0	0.05019	0.00910	0.05548	0.05150
2	0	0.00269	0.41765	0.02260	0.07948
3	1.00000	0	0	0	0
4	0	0.01392	0.56173	0.39335	0.54853
5	0	0.93320	0.01151	0.52857	0.32049

Analizując powyższy raport możemy stwierdzić, że model, wyselekcjonowany według kryterium Mallows'a $C(p)$ ma postać:

$$\hat{Y} = 83{,}497 + 2{,}662\ X_1c + 1{,}377\ X_2c + 5{,}996\ X_3c + 2{,}410\ X_4c\ . \qquad (6\text{-}5.19)$$

Ma on następującą charakterystykę. W modelu tym estymator wyrazu wolnego jest statystycznie istotny (p <0,0001). Ponadto, analizując macierz korelacji estymatorów parametrów strukturalnych modelu widać, że największa wartość bezwzględna współczynnika korelacji wynosi $|\hat{\rho}_{13}| = 0{,}5834$. Oznacza to, że również pomiędzy (standaryzowanymi) czynnikami w modelu nie występuje ani silna korelacja (ani współliniowość). Również wyniki otrzymane metodą analizy wartości własnych nie wskazują na występowania dużej korelacji. Najmniejsza wartość własna jest równa λ_5 = 0,19267, a wartości indeksów warunkowych CI_j nie różnią się bardzo, leżąc pomiędzy 1 a 3.44471.

Podsumowując stwierdzamy, że chociaż powyższy model, wyselekcjonowany przez kryterium Mallows'a $C(p)$, nie spełnia kryterium F_p na pozostanie w modelu tylko tych czynników, dla których wartości odpowiednich częściowych statystyk F_p są istotne statystycznie, to ze względu na przyjętą nadrzędność kryterium $C(p)$ związaną *z dobrocią predykcji* (*cen mieszkań*), wybieramy go, jako model najlepszy. Model ten jest również nieco prostszy niż model maksymalny.

Rozdział 6-6. Przykład procedury porównania wszystkich możliwych modeli regresji.

Na koniec przedstawmy raport SAS'a z porównania selekcji kilku modelu regresji dla badanego przykładu „Cena mieszkań" z wykorzystaniem wszystkich rozważanych kryteriów, gdzie jako model maksymalny został przyjęty model z wszystkimi zmiennymi standaryzowanymi (więc wycentrowanymi).

Odpowiedni raport ma postać:

```
Ceny mieszkań          12:58 Monday, February 17, 2014   1

                          The REG Procedure
                            Model: MODEL1
                       Dependent Variable: Y

                          C(p) Selection Method

                  Number of Observations Read          30
                  Number of Observations Used          30
```

Number in Model	C(p)	R-Square	Adjusted R-Square	AIC	MSE	SSE	Variables in Model
3	3.6003	0.8209	0.8003	99.6375	24.47577	636.36993	X1_cent X3_cent X4_cent
2	3.6044	0.8069	0.7926	99.8987	25.41440	686.18878	X1_cent X3_cent
4	4.0404	0.8319	0.8050	99.7514	23.90374	597.59344	X1_cent X2_cent X3_cent X4_cent
2	4.0878	0.8035	0.7890	100.4194	25.85943	698.20449	X3_cent X4_cent
3	4.5399	0.8144	0.7930	100.7189	25.37414	659.72777	X2_cent X3_cent X4_cent
3	5.1077	0.8104	0.7885	101.3540	25.91700	673.84213	X1_cent X2_cent X3_cent
3	5.4489	0.8080	0.7859	101.7291	26.24314	682.32158	X1_cent X3_cent Z_cent
4	5.5708	0.8211	0.7925	101.6029	25.42549	635.63714	X1_cent X3_cent X4_cent Z_cent
3	5.6347	0.8067	0.7844	101.9316	26.42084	686.94176	X3_cent X4_cent Z_cent
5	6.0000	0.8321	0.7972	101.7010	24.85793	596.59030	X1_cent X2_cent X3_cent X4_cent Z_cent
4	6.1169	0.8173	0.7881	102.2369	25.96848	649.21200	X2_cent X3_cent X4_cent Z_cent
4	6.9067	0.8118	0.7817	103.1307	26.75380	668.84508	X1_cent X2_cent X3_cent Z_cent
1	10.9319	0.7417	0.7325	106.6316	32.78751	918.05019	X3_cent
2	11.2387	0.7535	0.7353	107.2237	32.44298	875.96039	X3_cent Z_cent
2	12.9259	0.7417	0.7226	108.6268	33.99634	917.90130	X2_cent X3_cent
3	13.2384	0.7535	0.7251	109.2235	33.69053	875.95374	X2_cent X3_cent Z_cent
4	14.6316	0.7578	0.7190	110.7024	34.43480	860.86997	X1_cent X2_cent X4_cent Z_cent
3	15.0900	0.7406	0.7106	110.7598	35.46078	921.98026	X1_cent X2_cent Z_cent
2	18.1001	0.7055	0.6837	112.5609	38.76009	1046.52236	X1_cent Z_cent
3	18.1228	0.7194	0.6870	113.1177	38.36038	997.36981	X1_cent X2_cent X4_cent
3	19.7042	0.7083	0.6747	114.2774	39.87230	1036.67977	X1_cent X4_cent Z_cent
2	19.8127	0.6936	0.6709	113.7571	40.33682	1089.09421	X1_cent X2_cent
1	25.5850	0.6392	0.6263	116.6563	45.79629	1282.29616	X1_cent
2	26.9060	0.6439	0.6176	118.2588	46.86729	1265.41676	X1_cent X4_cent
2	39.4231	0.5564	0.5235	124.8542	58.39138	1576.56739	X2_cent X4_cent
3	40.0979	0.5657	0.5156	126.2208	59.37021	1543.62550	X2_cent X4_cent Z_cent
2	62.7459	0.3933	0.3483	134.2489	79.86379	2156.32236	X4_cent Z_cent
1	64.5808	0.3664	0.3438	133.5466	80.41610	2251.65091	X4_cent
1	101.2014	0.1103	0.0785	143.7325	112.92725	3161.96292	X2_cent
2	102.5800	0.1147	0.0491	145.5856	116.53767	3146.51718	X2_cent Z_cent
1	114.9018	0.0145	-.0207	146.8013	125.09024	3502.52667	Z_cent

Raport ten otrzymano z wykorzystaniem aplikacji Analyst SAS'a po uruchomieniu procedury REG dokonując ciągu wyborów Solutions->Analysis->Analyst-> (i po wczytaniu danych, korzystając z „Open By SAS Name" w zakładce File) ->Statistics->Regresion->Linear->Model->Mallows'Cp.

Zadanie. Dokonać selekcji z pośród modeli przebadanych w powyższym raporcie, posługując się np. kryteriami C(p), Adjusted R-Square (R^2_{adj}) oraz AIC (Rozdział 8).

Uwaga. Według informacyjnego kryterium AIC, im jego wartość jest mniejsza tym model jest bardziej preferowany.

Przypomnijmy, że aby skorzystać z kryterium Mallows'a, należy wartość C(p) modelu porównać z jego wartością minimalną $C_{\min}(p)$, pozostawiając jedynie modele, w których różnica między tymi wartościami nie przekracza, powiedzmy, 0,5 (lub ewentualnie 1), tak aby w wyselekcjonowanym modelu nie było już miejsca na dodatkowy czynnik, z punktu widzenia dokładności jego dopasowania do danych empirycznych.

6-7. Krok 5. Określenie solidności wybranego modelu

Model wybrany w czterech poprzednich krokach jest modelem najlepszym do analizy zależności dla danych, które pobraliśmy w konkretnej próbce. Jednakże nie wiemy czy model ten będzie również dobrym dla innej próbki.

Model, który nadaje się do analizy innych próbek nazywamy modelem godnym zaufania (modelem solidnym).

Ponieważ pobranie kolejnej próbki może być z jakiś powodów utrudnione, dlatego powszechnie używane metody testowania solidności modelu polegają na podzieleniu próbki na dwie mniejsze. Metody te polegają bądź na porównaniu modeli wyselekcjonowanych w obu podpróbkach, bądź na sprawdzeniu skuteczności modelu uzyskanego z jednej podpróbki dla danych w drugiej podpróbce.

W pierwszym sposobie analizuje się różnice między modelami otrzymane w obu podpróbkach. Konstrukcję podpróbek omówiono poniżej, przy okazji omawiania drugiego sposobu. Jeżeli różnic jest bardzo dużo, to model nie może być uznany za solidny. Jakakolwiek różnica w wyselekcjonowanych zmiennych jest wskazówką niesolidności modelu. Tak się zazwyczaj składa, że różnice takie pojawiają się, co sprawia, że same metody selekcji zmiennych do modelu nie są uznawane za solidne, co jest to następnym powodem przyjęcia jako ostatecznego kryterium raczej wartości współczynników Mallows'a $C(p)$ niż testów częściowych F_p.

Z drugiej strony model jest uważany za solidny, jeżeli zastosowanie go do obu podpróbek daje podobne wyniki. 2.

Drugi sposób przebiega następująco: Dzielimy pobraną próbkę (o liczebności n) na dwie mniejsze podpróbki.

Sposób podziału:

a) wybieramy z pierwotnej próbki jednostki, które dla tych samych wartości zmiennej X mają różne wartości zmiennej objaśnianej Y (podobnie w przypadku układów wartości dla kilku zmiennych objaśniających, z pierwotnej próbki wybieramy jednostki, które mają takie same układy wszystkich wartości zmiennych objaśniających),

b) dla jednostek z tą samą wartością zmiennej X (podobnie dla kilku zmiennych) dokonujemy losowania tych jednostek do dwóch wspomnianych podpróbek, które na skutek tej procedury zawierają, odpowiednio n_1 i n_2 jednostek z próbki pierwotnej ($n = n_1 + n_2$).

1. Dla pierwszej podpróbki („1”) wyznaczamy równanie regresji:

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_j X_j . \qquad (6\text{-}6.20)$$

2. Obliczamy kwadrat współczynnika korelacji wielokrotnej:

$$R^2(1) \equiv R^2(Y_1 \mid X_1, X_2, ..., X_j) = r^2(Y_1, \hat{Y}_1) \quad . \qquad (6\text{-}6.21)$$

gdzie $r^2(Y_1, \hat{Y}_1)$ jest kwadratem współczynnika korelacji pomiędzy wartościami teoretycznymi średnich warunkowych $\hat{Y}_{1i}$ a wartościami eksperymentalnymi Y_{1i} (i = 1,2,..., n_1) zmiennej Y w podpróbce nr1,

3. Wykorzystujemy równanie regresji (6-6.20) wyznaczone w podpróbce nr1 do wyznaczenia wartości przewidywanych $\hat{Y}^*_{2i}$ ($i = n_1+1,\ldots,n_1+n_2$) dla podpróbki nr2.

4. Obliczamy *współczynnik korelacji krzyżowej* [1]:

$$R^2_*(2) = r^2(Y_2, \hat{Y}^*_2) \ , \qquad (6\text{-}6.22)$$

gdzie liczymy korelację pomiędzy wartościami $\hat{Y}_{2i}$ a wartościami empirycznymi Y_{2i} ($i = n_1+1,\ldots, n_1+n_2$) zmiennej Y w podpróbce nr2 .

5. Następnie obliczamy różnicę [1]:

$$R^2(1) - R^2_*(2) \ , \qquad (6\text{-}6.23)$$

nazywaną *współczynnikiem ściśnięcia korelacji krzyżowej (ściśnięcie)*, który określa różnice pomiędzy oboma współczynnikami korelacji. Różnica ta przyjmuje wartości z przedziału $\langle 0,1 \rangle$. Model można uznać za solidny, gdy różnica ta jest jak najmniejsza, przy czym nie ustalono wartości granicznej. Przyjmuje się często, że model ze ściśnięciem 0,9 lub większym jest niesolidny, a model, dla którego ściśnięcie wynosi 0,1 lub mniej jest solidny, tzn. mamy nadzieję, że równie dobrze stosowałby się do analizy danych dla innych próbek pobranych z populacji.

A. Rozdział 7: Wnioski i dalsze metody analizy.

Wnioski dotyczące analizy współzależności zmiennych metodą regresji:

1. badania statystyczne powinny być prowadzone metodą analizy dostosowaną do zadanego celu badania (Rozdział 2, Tabela 2-2.2),
2. dobór zmiennych powinien być ściśle podporządkowany celowi badania, uwzględniając przy tym statystyczną istotność wartości estymatorów parametrów strukturalnych stojących przy zmiennych (Rozdział 6, Procedury doboru zmiennych),
3. najlepszym modelem statystycznym jest taki, który przy niezbyt rozbudowanej strukturze daje równanie regresji jak najlepiej opisujące zależności pomiędzy zmiennymi, tzn. jak najlepiej dopasowujące się do danych empirycznych, a co jest z tym związane, jak najsolidniejszą predykcję wartości zmiennej objaśnianej,
4. zmienne objaśniające modelu powinny (w miarę możliwości) nie wykazywać między sobą współliniowości (Rozdział 5, Rozdział 6).
5. w celu uniknięcia współliniowości zmiennych objaśniających należy przeprowadzić centrowanie lub standaryzację tych zmiennych, a w przypadku modeli wielomianowych może okazać się niezbędna ich ortogonalizacja (Rozdział 5),
6. wybrana procedura doboru zmiennych powinna prowadzić do selekcji modelu, w którym nie została pominięta żadna istotna zmienna, dlatego metodą preferowaną w tym względzie jest metoda eliminacji wstecz (Rozdział 6),
7. przy doborze zmiennych objaśniających do modelu statystycznego należy przedkładać kryterium Mallows'a $C(p)$ ponad kryterium $F(p)$.

   Dokładniej rzecz ujmując stosowanie $F(p)$ jako jedynego kryterium doboru modelu ma ograniczenia, których nie ma kryterium Mallows'a $C(p)$, pozwalające nie tylko na dobór modelu pod względem jego dokładności dopasowania się do danych empirycznych (podobnie jak to czyni $F(p)$), ale jednocześnie umożliwiające podjęcie decyzji o tym ile zmiennych objaśniających pozostawić w modelu końcowym (Rozdział 6, Przeprowadzona analiza przykładu „Ceny mieszkań"),
8. model statystyczny powinien spełniać wszystkie warunki i wymagania przedstawione powyżej, a ponadto powinien być modelem solidnym, tzn. takim, który będzie można zastosować do danych otrzymanych z innej próbki, pobranej niezależnie w sposób reprezentacyjny z populacji (Rozdział 6-5).

   Na sam koniec podkreślmy fakt, że pogłębiona analiza regresji zwraca baczną uwagę na tzw. diagnostykę regresji związaną z resztami występującymi w badanym modelu. Jest to kolejny, obszerny temat badań (Rozdziały 10 do 15).

A. Rozdział 8: Uzupełnienia.

Rozdział 8-1. Uzupełnienia. Kryterium R^2, R^2_{adj} i kryterium Akaike'a.

W modelach regresji współczynnik determinacji:

$$R^2 = \frac{SSR}{SSY} = 1 - \frac{SSE}{SSY} \tag{8-1.1}$$

mierzy stosunek zmienności zmiennej objaśnianej wyjaśniony regresją do zmienności ogólnej tej zmiennej. Alternatywą do stosowania R^2 jest tzw. *dopasowane* R^2 (adjusted R^2), które uwzględnia liczbę parametrów w modelu.

Dopasowane R^2 (R^2_{adj}) jest zdefiniowane następująco:

$$\mathrm{ADJRSQ} \equiv R^2_{adj} = 1 - \frac{n-i}{n-p}(1-R^2) = 1 - \frac{n-i}{SSY}\frac{SSE}{n-p} = 1 - \frac{n-i}{SSY}MSE\ , \tag{8-2.2}$$

gdzie n jest liczbą obserwacji wykorzystywaną przy dopasowywaniu modelu, p jest liczbą parametrów w modelu (włączając w to przesunięcie), natomiast i jest równe 1 gdy model zawiera przesunięcie oraz 0 gdy przesunięcia nie zawiera.

Widać, że R^2_{adj} zaczyna spadać, gdy w modelu jest za dużo parametrów i następuje przefitowanie modelu. To znaczy, sytuacja taka miałaby miejsce, gdyby zmniejszenie sumy kwadratów dla błędu *SSE* następowało wolniej wraz ze wzrostem liczby p parametrów niż spadek wartości liczby stopni swobody $n - p$ dla reszt modelu, co wiązałoby się (niekorzystnie) ze wzrostem (wraz z p) średniej wariancji wewnątrzgrupowej *MSE* i ze względu na stałość w (8-2.2) wartości n, i oraz *SSY*, spadkiem R^2_{adj}. Moment, w którym R^2_{adj} zaczyna spadać, jest więc sygnałem, że nie należy już modelu (z punktu widzenia tego kryterium) bardziej rozbudowywać.

Poza R^2_{adj} oraz kryterium Mallows'a $C(p)$, rozwinięto szereg kryteriów, w szczególności kryterium informacyjne Akaike'a (AIC), które przez wyznaczanie nakładu związanego z wprowadzeniem każdego dodatkowego parametru, próbuje zapobiegać przefitowaniu modelu. Sytuacja ta jest analogiczna do powyżej omówionego zastąpienia (w zwykłych modelach regresji) współczynnika determinacji R^2 współczynnikiem R^2_{adj}, który w przeciwieństwie do R^2 nie zawsze wzrasta przy dodaniu nowej zmiennej do modelu.

Kryterium AIC dla modelu ze „swobodnym" parametrem Θ jest zdefiniowane następująco:

$$AIC(\Theta) = -2\ln L(\hat{\Theta}) + 2p\ , \tag{8-2.3}$$

gdzie $\hat{\Theta}$ jest estymatorem MNW p-wymiarowego parametru Θ.

Wykorzystując AIC, wartość log-wiarygodności modelu jest redukowana poprzez „restrykcję” związaną ze wzrostem liczby parametrów w modelu, w sposób, który pomaga w porównywaniu modeli i testowaniu hipotez. *To znaczy model z mniejszym AIC jest preferowany.*

Np. niech hipoteza zerowa jest następująca $H_0 : \Theta = \Theta_0$ odpowiadając modelowi o wymiarze $p_0 = 0$, tzn. modelowi nie mającemu swobodnych parametrów. Wtedy AIC ma postać:

$$AIC(\Theta_0) = -2\ln L(\Theta_0) \,. \tag{8-2.4}$$

Oparty o kryterium AIC, pełny model ze swobodnym p – wymiarowym parametrem Θ jest preferowany wobec modelu prostszego odpowiadającego hipotezie $H_0 : \Theta = \Theta_0$, wtedy gdy (porównaj (1-1.42) w Rozdziale 1 Część II):

$$AIC(\hat{\Theta}) - AIC(\Theta_0) < 0 \;\; \Rightarrow \;\; 2\ln\frac{L(\hat{\Theta})}{L(\Theta_0)} > 2\,p \;, \tag{8-2.5}$$

co można zapisać jako:

$$\frac{L(\Theta_0)}{L(\hat{\Theta})} < e^{-p} \;. \tag{8-2.6}$$

W sytuacji gdy w próbce zachodzi (8-2.5), wtedy hipoteza H_0 jest odrzucana na rzecz hipotezy alternatywnej wskazującej na model z p-wymiarowym parametrem swobodnym.

Uwaga: Model o mniejszej wartości AIC można rozumieć jako będący bliżej (w znaczeniu entropii względnej Kullbacka-Leibler’a) pewnego modelu „prawdziwego”.

W Rozdziale 8 Części II podano schemat wyprowadzenia kryterium AIC w modelu ARIMA dla szeregów czasowych.

A. Rozdział 9. Nierówność Bonferroni'ego.

W powyższych rozważaniach nie zwrócono uwagi na konsekwencje faktu, że niejednokrotnie wyznaczano przedziały ufności dla kilku parametrów jednocześnie. Podobnie ma się sprawa z weryfikowanymi hipotezami. Czasami zapomina się o wynikających z tego faktu ograniczeniach dotyczących szczegółowych poziomów ufności i poziomów istotności. Podstawowym niedopatrzeniem jest np. konstrukcja szczegółowych przedziałów ufności na takim samym poziomie ufności, co poziom ufności dla łącznego obszaru ufności kilku estymowanych parametrów.

Np. w analizie regresji konstruowane przedziały ufności dla każdego parametru strukturalnego β_j, wyznaczane są czasami osobno na wcześniej przyjętym ogólnym poziomie ufności (1-α). Ponieważ jednak przedziały te były wyznaczane równocześnie, zatem odpowiednia analiza powinna również uwzględnić konstrukcję wspólnego obszaru ufności.

Podobnie ma się rzecz w przypadku testowania hipotez, tzn. czasami każdą hipotezę testuje się osobno na przyjętym ogólnym poziomie istotności α. Takie podejście ignoruje fakt, że testy te są wykonywane jednocześnie, zwiększając tym samym znacznie ogólny (dla przeprowadzenia wszystkich testów łącznie) poziom istotności. Na dodatek sprawę komplikuje fakt, że na ogół stawiane hipotezy nie wykluczają się nawzajem.

Aby przedstawić problem wyprowadźmy nierówność Bonferroni'ego [1], która jest podstawą wyznaczenia właściwych szczegółowych poziomów ufności i związanych z nimi szczegółowych przedziałów ufności, oraz szczegółowych poziomów istotności i związanych z nimi szczegółowych zbiorów krytycznych.

Szczególnym przykładem zastosowania metody Bonferroni'ego w analizie wariancji jest metoda Scheffe'go dla wyznaczania przedziałów ufności dla tzw. kontrastów (lub testowania hipotez odpowiednich dla tych kontrastów [1]) w ANOVA (Rozdział 16).

Nierówność Bonferroni'ego, wynikająca z aksjomatów Kołmogorowa ma postać [1]:

$$P\left(\bigcap_{j=1}^{g} A_j\right) \geq 1 - \sum_{j=1}^{g} P(\overline{A}_j) \ , \tag{9.1}$$

gdzie A_j oraz $\overline{A}_j$ są zdarzeniami wykluczającymi się wzajemnie, tak jak np. zdarzenia, że dla określonego parametru θ_j i dla ustalonego prawdopodobieństwa α_j mamy, że $(1-\alpha_j)\cdot 100\%$ - owy przedział ufności wyznaczony dla parametru θ_j, i $\alpha_j \cdot 100\%$-owy zbiór krytyczny dla hipotezy dotyczącej wartości parametru θ_j, są rozłączne a ich suma pokrywa wartość tego parametru (która to wartość może być stawiana w hipotezie zerowej).

Dowód dla (9.1). (||…|| oznacza wtrącenie):

$$P\left(\bigcap_{j=1}^{g} A_j\right) = 1 - P\left(\overline{\bigcap_{j=1}^{g} A_j}\right) =$$

$$= \left\| \overline{\bigcap_{j=1}^{g} A_j} = \bigcup_{j=1}^{g} \overline{A_j} \right\| = 1 - P(\bigcup_{j=1}^{g} \overline{A_j}) = \left\| P\left(\bigcup_{j=1}^{g} \overline{A_j}\right) \le \sum_{j=1}^{g} P(\overline{A_j}) \right\| \ge 1 - \sum_{j=1}^{g} P(\overline{A_j}),$$

skąd otrzymujemy nierówność Bonferroni'ego (9.1). Zwróćmy uwagę, że ostatnie przekształcenie prowadzące do nierówności pojawiło się, jako skutek nie wykluczania się zdarzeń $\overline{A_j}$.

Proste przekształcenie (9.1) daje:

$$1 - P\left(\bigcap_{j=1}^{g} A_j\right) \le \sum_{j=1}^{g} P(\overline{A_j}) \ , \tag{9.2}$$

Niech wyrażenie $1 - P\left(\bigcap_{j=1}^{g} A_j\right)$ po lewej stronie będzie (chwilowo) prawdziwym ogólnym poziomem istotności. Wtedy $P(\overline{A_j})$ jest prawdziwym szczegółowym poziomym istotności, określającym prawdopodobieństwo, że odpowiedni dla parametru θ_j szczegółowy zbiór krytyczny, pokrywa wartość tego parametru. Widać, więc, że prawdziwy ogólny poziom istotności jest w przypadku niewykluczających się zdarzeń, nie większy niż suma wszystkich możliwych szczegółowych poziomów istotności.

Gdyby więc np. przyjąć, że każdy szczegółowy poziom istotności jest taki sam i wynosi $P(\overline{A_j}) = \frac{\tilde{\alpha}}{g}$, j = 1, 2, …, g, wtedy (9.2) daje:

$$1 - P\left(\bigcap_{j=1}^{g} A_j\right) \le \sum_{j=1}^{g} P(\overline{A_j}) = \sum_{j=1}^{g} \frac{\tilde{\alpha}}{g} = \tilde{\alpha} \ , \tag{9.3}$$

skąd widać, że gdy mamy g szczegółowych hipotez składających się na hipotezę ogólną, to posługiwanie się wartością prawdopodobieństwa $\tilde{\alpha}$ jako ogólnym poziomem istotności, nie jest na ogół poprawne, bowiem zawyża ono wartość ogólnego poziom istotności. Wyraźmy powstały problem następująco: „odległość wartości prawdziwego szczegółowego poziomu istotności od prawdziwego ogólnego poziomu istotności jest, w przypadku nie wykluczania się zdarzeń $\overline{A_j}$, mniejsza niż odległość $\frac{\tilde{\alpha}}{g}$ od $\tilde{\alpha}$".

W końcu, spójrzmy na sprawę w sposób bliższy wykonywanym w praktyce testom i załóżmy teraz, że α jest *właściwym* (prawdziwym) *ogólnym poziomem istotności*:

$$1 - P\left(\bigcap_{j=1}^{g} A_j\right) = \alpha \tag{9.4}$$

i w (rozsądnym) uproszczeniu załóżmy, że wszystkie hipotezy szczegółowe (indeks *s*) są testowane na tym samym poziomie istotności (który oznaczmy) α_s :

$$P(\overline{A}_j) = \alpha_s, \quad j = 1, 2, ..., g \ . \tag{9.5}$$

Nierówność (9.3) możemy teraz zapisać następująco:

$$\alpha = 1 - P\left(\bigcap_{j=1}^{g} A_j\right) \le \sum_{j=1}^{g} P(\overline{A}_j) = \sum_{j=1}^{g} \alpha_s = g \times \alpha \tag{9.6}$$

lub

$$\frac{\alpha}{g} \le P(\overline{A}_j) = \alpha_s, \quad j = 1, 2, ..., g \ . \tag{9.7}$$

Wniosek. Widzimy więc, że (*w przypadku nie wykluczania się hipotez* $\overline{A}_j$) *wartość prawdziwego szczegółowego poziomu istotności* α_s *jest* ***większa niż*** α / g, gdzie α jest prawdziwym ogólnym poziomem istotności. Wynika stąd, że *indywidualny* (*szczegółowy*) *j*-ty *przedział ufności powinien być wyznaczony na poziomie ufności* ***mniejszym niż*** $(1 - \alpha / g)$, tzn. szczegółowe przedziały ufności ulegają zwężeniu, a szczegółowe zbiory krytyczne ulegają poszerzeniu.

Posługiwanie się błędnym (bo zaniżonym), szczegółowym poziomem istotności równym α / g, ma następujące konsekwencje:

1) Gdyby (przy ogólnym prawdziwym poziomie istotności α) policzyć *poprawnie prawdziwe* szczegółowe poziomy istotności (α_s) dla testów szczegółowych, to zbiór krytyczny przesunąłby się w kierunku centrum rozkładu testowej statystyki szczegółowej i szczegółową hipotezę zerową byłoby łatwiej odrzucić niż przy szczegółowym poziomie istotności przyjętym *błędnie jako* α / g.

2) Gdy wartość szczegółowego poziomu istotności przyjmuje się błędnie jako równą α / g, wtedy trudniej jest odrzucić szczegółową hipotezę zerową (niż wtedy gdyby się posłużyć poprawną wartością α_s). Tzn. moc testu posługującego się wartością α / g jest za mała, a prawdopodobieństwo popełnienia błędu II rodzaju, tzn. prawdopodobieństwo błędnego przyjęcia szczegółowej hipotezy zerowej dotyczącej parametru θ_j, jest wtedy za duże. Przykładowo, w przypadku analizy regresji (w której stawia się hipotezę zerową o zerowej wartości jakiegoś parametru strukturalnego), oznaczałoby to nadmierną skłonność do nie uzasadnionego trwania przy hipotezie zerowej i pomijania w modelu, parametru strukturalnego, który powinien w modelu pozostać.

Przykład konstrukcji obszaru ufności dla pary parametrów (μ, σ^2) rozkładu normalnego $\mathrm{N}(\mu, \sigma^2)$ [2], został podany w Części IV, Rozdział 4.

B. Rozdział 10. Diagnostyka reszt.

Rozdział 10-1. Wstęp

Analiza reszt odgrywa istotną rolę w sprawdzeniu solidności modelu, tzn. weryfikacji na podstawie próby zgodności empirycznych własności testowanego modelu z wymogami teoretycznymi modelu. Informacja, że w próbie naruszone zostały w sposób istotny założenia modelowe, przekreśla w zasadzie stosowanie badanego modelu, ale jednocześnie skłania do postulowania nowych modeli. W tej części skryptu badane są te założenia liniowych modeli regresji, które dają się przetestować w oparciu o analizę reszt.

1. W celu zdiagnozowania reszt należy przeprowadzić:
    a. analizę outsiderów
    b. graficzną analizę reszt
2. Analizy outsiderów dokonuje się oceniając np. wartości reszt zwykłych, reszt studentyzowanych, reszt scyzorykowych, współczynnika dźwignięcia i odległości Cook'a. Obserwacja może być podejrzana o to, że jest outsiderem, gdy:
    a. wartość reszty znacząco odstaje od wartości pozostałych reszt
    b. wartość współczynnika dźwignięcia ma dużą wartość
    c. wartość odległości Cook'a jest duża
3. Przeprowadzenie dokładnej analizy outsiderów pozwala wykryć obserwacje wpływowe, a ich (przemyślane) usunięcie powoduje lepsze dopasowanie modelu do danych empirycznych (zwiększa się np. wartość współczynnika determinacji).
4. Graficzna analiza reszt pozwala na podanie prostej (i niejednokrotnie wystarczająco precyzyjnej) odpowiedzi na pytania:
    a. czy występuje zgodność rozkładu reszt z rozkładem normalnym
    b. jaki jest schemat rozkładu reszt

Do analizy graficznej brane są pod uwagę najczęściej reszty scyzorykowe lub studentyzowane. Analiza reszt jest również istotna ze względu na fakt, że o ich postulowane własności normalności, jednorodności i niezależności, opierają się wszystkie procedury estymacyjne parametrów modelu, począwszy od oceny dokładności oszacowań parametrów strukturalnych modelu, a skończywszy na zagadnieniu konstrukcji pasma przewidywania modelu. Nie spełnienie założeń o braku korelacji reszt oraz ich jednorodności powoduje pogorszenie się własności estymatorów wykorzystywanych w np. testowaniu hipotez o braku zależności korelacyjnej pomiędzy zmiennymi, bowiem stosowane estymatory mogą stać się co najmniej nieefektywne. Odejście od założenia o niezależności reszt narusza związek zwykłej metody najmniejszych kwadratów (MNK) ze standardową metodą największej wiarygodności (MNW), która jest podstawą konstrukcji estymatorów w statystyce klasycznej, a naruszenie założenia o normalności rozkładu reszt sprawia, że analiza regresji MNK nie ma oparcia w MNW i przestaje mieć charakter probabilistyczny.

W analizie zaprezentowano wykorzystanie procedur analitycznych i graficznych pakietu SAS (i czasami Excel).

Założenia regresji wielorakiej (zostały podane w Rozdziale 4-1).

Tak jak w typowym klasycznym modelu regresji liniowej, zmienną losową jest zmienna Y, podczas gdy zmienne $X_1, X_2, \ldots, X_k$ są zmiennymi (nielosowymi) kontrolowanymi. Stałe $\beta_0, \beta_1, \ldots, \beta_k$ są nieznanymi parametrami populacji, natomiast składnik losowy E jest zmienną losową nieobserwowaną bezpośrednio. Jeśli oszacowujemy parametry $\beta_0, \beta_1, \ldots, \beta_k$ przy pomocy estymatorów $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ wtedy właściwym oszacowaniem w próbie, zmiennej E_i dla i-tej jednostki jest (4-8):

$$U_i \equiv \hat{E}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k Y_{ki}), \quad i = 1, 2, \ldots, n \ , \qquad (10\text{-}1.1)$$

gdzie Y_i jest zmienną losową obserwowaną dla i-tej jednostki zbiorowości, U_i jest składnikiem resztkowym, a wartości zmiennej U_i nazywane resztami oznaczamy jako u_i. Zbiór wielkości $\{U_i\}$ odzwierciedla wielkość niezgodności pomiędzy wartościami przewidywanymi i obserwowanymi w próbie, jaka pozostaje po dopasowaniu modelu do danych (w rozważanym przypadku, metodą najmniejszych kwadratów). Każde U_i reprezentuje estymator nieobserwowanego błędu E_i występującego w populacji. Zwykle w analizie regresji zakłada się, że błędy $\{E_i\}$ są niezależne, mają średnią równą zero, mają wspólne wariancje σ^2 i podążają za rozkładem normalnym. Jeśli model jest dobrze dobrany do analizowanych danych, wtedy rozsądnym jest spodziewanie się, że zaobserwowane reszty $\{U_i\}$ wykazują właściwości będące w zgodzie z tymi założeniami.

Poniżej przedstawimy analizę reszt w ramach liniowego modelu normalnego. Analiza reszt w ramach innych modeli jest trudniejsza, jednak część stosowanych w nich procedur została rozwinięta w analogi do procedur dla modelu normalnego [1]. Przeprowadzenie analizy reszt dla oszacowania właściwości dopasowania modelu jest obecnie, ze względu na dostępność programów komputerowych, powszechne. Wiele z tych programów nadaje się do graficznej prezentacji reszt i wykresów diagnostycznych dla wszystkich zwykle używanych modeli [4].

Rozdział 10-2. Typy reszt oraz ich własności w modelu liniowym.

Celem analizy reszt jest sprawdzenie czy spełnione są podstawowe założenia modeli regresji omówione w Rozdziale 4. Poniżej przedstawimy metody identyfikacji outsiderów właściwe dla analizy reszt w modelach regresji.

Rozdział 10-2-1. Współczynnik dźwignięcia.

Współczynnik dźwignięcia h_i jest miarą geometrycznej odległości i-tego punktu $\vec{X}_i = (X_{1i}, X_{2i}, \ldots, X_{k\,i})$ czynnika $\vec{X}$ od punktu środkowego $\left(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k\right)$ w przestrzeni k-wymiarowego czynnika $\vec{X}$. Wielkość h_i nazywana współczynnikiem "dźwignięcia" jest miarą ważności i-tej obserwacji przy określaniu dopasowania

modelu. Ich rola jest pomocna w problemach diagnozowania regresji. Zbiór wartości współczynnika dźwignięcia $\{h_i\}$, $i = 1, 2, ..., n$, wzbogaca diagnostykę modelu regresji.

Dla modelu prostoliniowego z jednym czynnikiem:

$$Y_i = \beta_0 + \beta_1 X_i + E_i \tag{10-2-1.2}$$

wartość współczynnika dźwignięcia dla *i*-tej obserwacji przyjmuje formę [1]:

$$h_i = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{(n-1)\hat{S}_X^2}, \tag{10-2-1.3}$$

gdzie

$$\hat{S}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{10-2-1.4}$$

jest wariancją czynnika *X*. Głównym składnikiem wzoru (10-2-1.3) dla współczynnika dźwignięcia jest kwadrat standaryzowanej odległości wartości X_i od średniej wartości $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ tzn.

$$Z_{X_i}^2 = \left(\frac{X_i - \overline{X}}{\hat{S}_X}\right)^2 \tag{10-2-1.5}$$

Zatem dla prostej liniowej regresji z jednym czynnikiem współczynnik dźwignięcia wskazuje odległość obserwacji w zbiorze wartości czynnika *X*.

Bardziej ogólnie, dla wielorakiej regresji wartość współczynnika dźwignięcia mierzy odległość obserwacji w *k*- wymiarowej przestrzeni czynników $X_1, X_2, ..., X_k$.

Dla specjalnego przypadku, w którym wszystkie czynniki $X_1, X_2, ..., X_k$ mają średnią równą zero i są niezależne, zachodzi związek [1]:

$$h_i = \frac{1}{n} + \sum_{j=1}^{k}\frac{(X_{ji} - \overline{X}_j)^2}{(n-1)\hat{S}_j^2}, \qquad i = 1, 2, ..., n, \tag{10-2-1.6}$$

w którym

$$\hat{S}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \overline{X}_j)^2 \tag{10-2-1.7}$$

gdzie X_{ji} jest *i*-tą wartością *j*-tego czynnika.

Interpretacja wielkości wartości współczynnika dźwignięcia jest prowadzona przy skorzystaniu z następujących jego własności.

Po pierwsze i całkiem ogólnie:

$$0 \le h_i \le 1 \qquad \text{(10-2-1.8a)}$$

Jednakże, jeśli model regresji zawiera parametr przesunięcia β_0, wtedy:

$$\frac{1}{n} \le h_i \le 0 . \qquad \text{(10-2-1.8b)}$$

Na przykład dla przypadku jednego czynnika, nierówność (10-2-1.8) (pokazać dla jednego czynnika) można wyprowadzić korzystając z tzw. nierówności Laguerre'a - Samuelson'a:

$$\overline{X} - \sqrt{n-1}\, S_X \le X_i \le \overline{X} + \sqrt{n-1}\, S_X \qquad \text{(10-2-1.9)}$$

gdzie

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \ .$$

Jak można zauważyć wykorzystując (10-2-1.6), jeśli liczba czynników w modelu wynosi *k*, tzn.:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + E$$

to (pokazać):

$$\sum_{i=1}^{n} h_i = k + 1 \qquad \text{(10-2-1.10)}$$

W konsekwencji średnia wartość współczynnika dźwignięcia wynosi:

$$\overline{h} = \frac{k+1}{n} \qquad \text{(10-2-1.11)}$$

Hoaglin i Welsch [16] zalecili dokładne analizowanie jakiejkolwiek obserwacji, dla której:

$$h_i > \frac{2(k+1)}{n} \ . \qquad \text{(10-2-1.12)}$$

Z dźwignięciem związana jest tzw. odległość Mahalanobisa:

$$m_i = \left(h_i - \frac{1}{n}\right)(n-1)$$

<u>Uwaga:</u> Jeśli czynniki mają rozkład Gaussa (tzn. każdy czynnik ma rozkład normalny) i przy założeniu hipotezy zerowej, że *i*-ta obserwacja jest próbką losową o liczebności 1 pobraną z populacji wszystkich wartości czynnika o rozkładzie Gaussa, wtedy zmienna losowa:

$$F_i = \frac{\left[h_i - (1/n)\right]/k}{(1-h_i)/(n-k-1)} \qquad \text{(10-2-1.13)}$$

dla każdego pojedynczego dźwignięcia h_i ma w próbie rozkład *F*- Snedecora z *k* i *n-k-1* stopniami swobody [1]. Zatem test dla największego współczynnika dźwignięcia, może być dokonywany przez porównanie wartości F_i uzyskanej w obserwacji z wartością krytyczną $F_{kr} = F_{k,(n-k-1),1-\alpha/n}$, gdzie dzielenie α/n w

wartości krytycznej pojawia się na skutek podziału Bonferroni'ego (Rozdział 9). W poniższej tablicy podano krytyczne wartości współczynników dźwignięcia:

$$h_{kr} = \frac{n-k-1+F_{kr}\,k\;n}{n\;(n-k-1+F_{kr}\,k)} \qquad (10\text{-}2\text{-}1.14)$$

odpowiadające wartościom krytycznym F_{kr} dla typowych wartości liczby czynników *k* i wielkości próby *n*, i dla α = 0,01.

(Odejście od założenia ustalonych wartości czynników na rzecz pojawienia się ich losowo (powyżej, z rozkładem Gaussa) może obciążyć w niektórych przypadkach estymatory parametrów rozproszenia (np. wariancji składnika losowego; Rozdział: Dwuczynnikowa ANOVA)).

Tablica 10-2-1.1. Krytyczne wartości dla dźwignięcia, *n* = wielkość próby, *k* = liczba czynników dla α=0,01.

k / *n*	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.785	0.875	0.930	0.965	0.986	0.997	1.000	1.000						
15	0.629	0.724	0.792	0.844	0.887	0.921	0.948	0.969	0.984	0.994				
20	0.524	0.612	0.677	0.731	0.777	0.817	0.852	0.883	0.910	0.933	0.996			
25	0.450	0.529	0.589	0.640	0.685	0.724	0.761	0.794	0.824	0.851	0.953	0.997		
30	0.394	0.466	0.521	0.568	0.610	0.648	0.683	0.716	0.746	0.774	0.889	0.964		
40	0.318	0.377	0.424	0.464	0.501	0.534	0.565	0.595	0.622	0.649	0.763	0.855		
60	0.231	0.275	0.310	0.341	0.369	0.395	0.420	0.443	0.465	0.487	0.584	0.668	0.917	
80	0.183	0.218	0.246	0.271	0.293	0.314	0.334	0.353	0.372	0.389	0.471	0.543	0.778	
100	0.152	0.181	0.205	0.225	0.244	0.262	0.279	0.295	0.310	0.325	0.394	0.456	0.666	0.956
200	0.085	0.100	0.113	0.124	0.135	0.145	0.154	0.163	0.172	0.180	0.219	0.255	0.383	0.598
400	0.046	0.054	0.061	0.067	0.073	0.078	0.083	0.088	0.092	0.097	0.118	0.138	0.208	0.330
800	0.025	0.029	0.033	0.036	0.039	0.041	0.044	0.046	0.049	0.051	0.062	0.073	0.110	0.175

Na przykład, gdy α = 0,01 i trzeba wykonać testy dla 100 obiektów (np. 100 testów z hipotezami $H_0^i : E(h_i) = \frac{1}{n}$ dla dźwignięcia h_i, i = 1, 2, …, n; gdzie n = 100 jest wielkością próby), wtedy możemy dla każdego szczegółowego testu uznać, że nie mamy podstaw do odrzucenia hipotezy H_0^i, jeśli wartość *p* nie będzie mniejsza niż $\alpha_i = \frac{0{,}01}{100} = 0{,}0001$. Unikniemy wtedy nieuzasadnionego traktowania niektórych h_i (dla *i*-tych obserwacji) jako outsiderów. Nawet wtedy, gdy wartości czynnika przyjmiemy jako ustalone (tak jak jest to w modelu klasycznym regresji), powyższa statystyka *F* może być pomocna przy przybliżonym wskazaniu kłopotliwych obserwacji.

Rozdział 10-2-2. Własności reszt

Załóżmy, że mamy *n* kompletów obserwacji (Y_i, X_{i1}, X_{i2},…, X_{ik}), gdzie i = 1, 2, …n dla zmiennych (Y, X_1, X_2,…, X_k). Z przedstawionych w Rozdziale 4 założeń modelu regresji wiemy, że analiza regresji jest związana z metodą najmniejszych kwadratów, którą stosujemy w celu dopasowania modelu regresji:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + E_i \quad \text{dla} \quad i = 1,2,...,n \tag{10-2-2.1}$$

do obserwowanych wartości zmiennej opisywanej Y_i, gdzie E_i jest i-tym składnikiem losowym. Ponieważ dopasowany model ma postać $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$, co oznacza, że prognozowana (przewidywana) odpowiedz w i-tym punkcie danych jest następująca:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki} \tag{10-2-2.2}$$

Zgodnie z (10-1.1) i-ta reszta U_i, czyli różnica pomiędzy obserwowaną wartością Y_i i przewidywaną wartością $\hat{Y}_i$ wynosi $U_i = Y_i - \hat{Y}_i$ dla $i = 1,2,...,n$.

Typy reszt

Główna strategia leżąca u podstaw statystycznej procedury nazywaną ogólnie analizą reszt polega na ustaleniu przydatności modelu, na podstawie obserwacji zachowania się zbioru obserwowanych wartości reszt. Naszym celem jest dyskusja metod, które służą do czynienia takich ustaleń. Podana poniżej metodologia może znaleźć zastosowanie w wielu przypadkach, w których dopasowywany jest pewien model, w wyniku, czego uzyskujemy zbiór reszt. Możliwe jest, więc rozszerzenie analizowanej metody na analizę wariancji, wieloczynnikową regresję liniową i regresje nieliniową w czynnikach [4].

Zgodnie z metodą najmniejszych kwadratów, reszty $U_1, U_2, ..., U_n$ i ich funkcje posiadają następujące własności (porównaj założenia modelu regresji, Rozdział 4):

1. Średnia z $\{U_i\}$ jest równa zero (pokazać):

$$\overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i = 0 \tag{10-2-2.3}$$

Z powyższej równości wynika, że reszty $U_1, U_2, ..., U_n$ nie są niezależne.

2. Estymator wariancji błędów E_i w populacji wyznaczony z próby n reszt, czyli średnia kwadratów reszt *MSE*, nazywana też wariancją resztową, ma postać:

$$MSE = \frac{1}{n-k-1}\sum_{i=1}^{n} U_i^2 = \frac{SSE}{n-k-1} \tag{10-2-2.4}$$

O ile model z $p = k + 1$ (szacowanymi) parametrami funkcji regresji jest właściwy, to *MSE* jest nieobciążonym estymatorem wariancji $\sigma^2(E_i) \equiv \sigma_E^2$ składnika losowego E_i, tzn. $E(MSE) = \sigma_E^2$, a $(n - p)$ liczbą stopni swobody dla $\sum_{i=1}^{n} U_i^2$ [1]. Jak to omówimy poniżej, okazuje się, że wygodnie jest wprowadzić do analizy również inne estymatory wariancji składnika losowego, które lepiej niż składnik resztkowy U_i nadają się do analizy reszt modelu i są bardziej skuteczne w identyfikacji outsiderów.

3. Reszty $\{U_i\}$ nie są niezależnymi zmiennymi losowymi. Wynika to z faktu, że reszty $\{U_i\}$ sumują się do zera. Jednak, jeśli liczba reszt n jest duża w porównaniu z liczbą k niezależnych zmiennych, wtedy efekt zależności reszt możemy w praktyce zignorować [1].

Wielkość:

$$Z_i = \frac{U_i}{\sqrt{MSE}} \qquad (10\text{-}2\text{-}2.5)$$

jest nazywana standaryzowaną resztą; często właśnie Z_i a nie U_i są badane w analizie reszt. Tak jak i dla reszt,$\{U_i\}$, suma standaryzowanych reszt jest równa zero i stąd nie są one niezależne. Standaryzowane reszty mają jednostkową wariancję, co oznacza, że:

$$\frac{1}{n-k-1}\sum_{i=1}^{n} Z_i^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\left(\frac{U_i}{\sqrt{MSE}}\right)^2 = \frac{1}{MSE}\left(\frac{1}{n-k-1}\sum_{i=1}^{n} U_i^2\right) = 1$$

Standaryzowane reszty mają rozkłady przybliżone do rozkładu *t-Studenta* z liczbą stopni swobody *n-k-1*. Zatem $\sqrt{MSE}$ w mianowniku standaryzowanych reszt (10-2-2.5), odzwierciedlające dobroć dopasowania modelu, skaluje reszty tak, aby miały jednostkową wariancję.

Podczas gdy najlepszy estymator wariancji dla odpowiedzi $\hat{Y}_i$ ma postać:

$$S^2_{\hat{Y}_i} = MSE\, h_i \ , \qquad (10\text{-}2\text{-}2.6)$$

standardowe odchylenie reszt U_i jest równe:

$$S(U_i) = \sqrt{MSE\,(1-h_i)} \ , \qquad (10\text{-}2\text{-}2.7)$$

o czym wspomnimy jeszcze później (Rozdział 11). Jeśli dla określonego i współczynnik dźwignięcia h_i jest równy 1, wtedy z (10-2-2.7) widać, że standardowe odchylenie reszt U_i znika, co ze względu na $E(\hat{Y}_i) = E(Y_i)$ *oznacza, że* $\hat{Y}_i = Y_i$ i model został zmuszony (dźwignięty) tak, aby dopasować dokładnie i-tą obserwowaną odpowiedź Y_i.

4. Wielkość:

$$R_i = \frac{U_i}{\sqrt{MSE\,(1-h_i)}} = \frac{Z_i}{\sqrt{1-h_i}} \qquad (10\text{-}2\text{-}2.8)$$

jest nazywana resztą (wewnętrznie) ”studentyzowaną”. Nazywana jest tak, ponieważ jeśli tylko dane spełniają zwykle założenia dla regresji wielokrotnej [1], to ma ona w przybliżeniu rozkład *t* - Studenta z *n-k-1* stopniami swobody. Studentyzowane reszty mają średnią bliską zero oraz wariancję:

$$S_R^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} R_i^2 \ , \qquad (10\text{-}2\text{-}2.9)$$

która jest nieznacznie większa niż 1.

5. Wielkość:

$$R_{(-i)} = R_i \sqrt{\frac{MSE}{MSE_{(-i)}}} = \frac{U_i}{\sqrt{MSE_{(-i)}(1-h_i)}} = R_i \sqrt{\frac{(n-k-1)-1}{(n-k-1)-R_i^2}} \qquad (10\text{-}2\text{-}2.10)$$

jest nazywana resztą ”scyzorykową” (lub zewnętrznie ”studentyzowaną”). Reszty te przyjmują wartości z przedziału $(-\infty, +\infty)$. $MSE_{(-i)}$ jest scyzorykową wariancją resztową. Ponieważ, $U_i = Y_i - \hat{Y}_i$, zatem licznik w $R_{(-i)}$ odzwierciedla odległość i-tej oberwanej odpowiedzi Y_i od przewidywanej wartości $\hat{Y}_i$. Celem stosowania $MSE_{(-i)}$ jest zabezpieczenie się przed ukryciem się wpływu outsiderów, co osiąga się przez spadek wartości przyjmowanej przez $MSE_{(-i)}$ w porównaniem z wartością przyjmowaną przez MSE. Stosowanie $MSE_{(-i)}$ powoduje, więc wzrost $R_{(-i)}$ w porównaniu z R_i.

Reszty scyzorykowe mają wariancję

$$S_{(-i)}^2 = \frac{1}{(n-k-1)-1} \sum_{i=1}^{n} R_{(-i)}^2 \qquad (10\text{-}2\text{-}2.11)$$

większą niż 1.

Jeśli zwykłe założenia regresji liniowej są spełnione, wtedy reszt scyzorykowe mają dokładnie rozkład *t-Studenta* z *(n-k-1)-1 = (n-k-2),* stopniami swobody oraz ze średnią równą zero. Wielkość $S^2_{(-i)}$ jest wariancją reszt liczoną przy usuniętej i-tej obserwacji. Dlatego też, podczas gdy standaryzowana reszta jest związana z odchyleniem standardowym liczonym dla n obserwacji, to i-ta reszta scyzorykowa jest standaryzowana z wykorzystaniem odchylenia standardowego liczonego dla n-1 obserwacji (tzn. po usunięciu i-tej obserwacji) oraz funkcji h_i. Istotnymi elementami wpływającymi na $R_{(-i)}$ są U_i, $S_{(-i)}$ oraz h_i.

Jeżeli standardowe założenia dla regresji [1] są spełnione i w przybliżeniu ta sama liczba obserwacji jest zrobiona dla każdej wartości czynnika (zmiennej objaśniającej), wtedy schematy analizy reszt, w których posługujemy się resztami standaryzowanymi, studenckimi czy scyzorykowymi wyglądają bardzo podobnie.

Jednak, jeśli pojawiają się problemy, tzn. analizowane są ”nietypowe” wartości uzyskane w obserwacji, wtedy analiza w oparciu o reszty standaryzowane, a przede wszystkim reszty scyzorykowe jest bardziej skuteczna. Na przykłada, jeśli i-ta obserwacja leży daleko od pozostałych danych, wtedy $MSE_{(-i)}$ będzie dużo mniejsze niż $\sqrt{MSE}$, co powoduje, że $R_{(-i)}$ jest duże w porównaniu z R_i. W ten sposób $R_{(-i)}$ wyróżnia się bardziej niż R_i, bardziej ujawniając outsiderów. Większe wartości h_i (wartość z wysokim wpływem obserwacji) prowadzą również do większych odpowiednich wartości $R_{(-i)}$ niż R_i.

Ponadto, gdy liczba stopni swobody dla błędów (*n-k-1* dla standaryzowanych i *n-k-2* dla scyzorykowych) rośnie znacznie powyżej 30, wtedy rozkłady reszt mogą być coraz dokładniej przybliżone przez standaryzowany rozkład normalny, (dla którego średnia jest równa zero i wariancja równa 1). Informacja ta jest pomocna do oceny wielkości obserwowanych reszt przez odwołanie się do własności standardowego rozkładu normalnego. Na przykład, jeśli reszty reprezentują w przybliżeniu próbkę losową pobraną z

populacji z rozkładem N(0,1) wtedy oczekujemy, że dla nie więcej niż 5% reszt pojawiają się (co do modułu) wartości przekraczające wartość zero o 1,96.

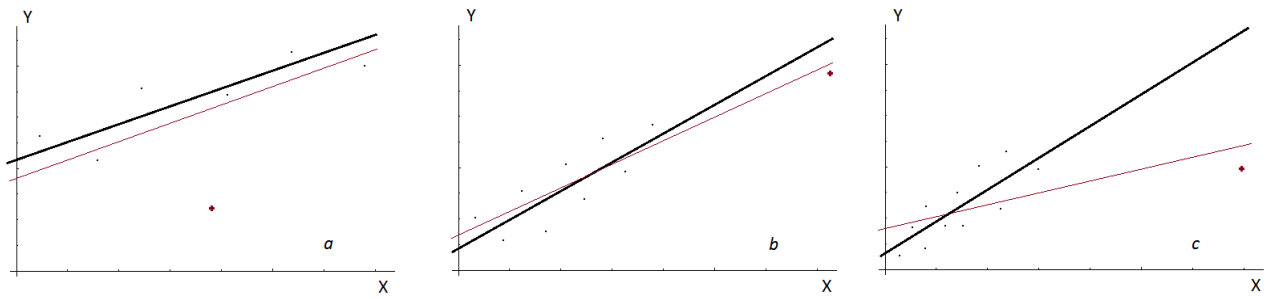
Wykresem stosowanym szeroko do sprawdzania założenia normalności rozkładów błędów jest wykres dla dystrybuant, na którym empiryczna dystrybuanta dla uporządkowanych reszt jest wykreślona na przeciwko dystrybuanty dla rozkładu normalnego ze średnią i wariancją równą średniej i wariancji reszt w próbie (tzw. „normal probability-probability plot"). Wykres ten może być, więc pomocny w wykrywaniu outsiderów. Jeżeli poprawne byłoby założenie o normalności błędów, wtedy wykres powinien przejawiać tendencję liniową z nachyleniem (kąt 45° przy tej samej skali na obu osiach), począwszy od początku układu współrzędnych. Jednak problem w posługiwaniu się takimi wykresami stanowi określenie wielkości dopuszczalnego odstępstwa od zakładanego idealnego przebiegu.

Łatwo jest sprawdzić czy konkretna scyzorykowa reszt różni się istotnie statystycznie od zera. Jeśli zwykłe założenia odnośnie regresji są spełnione [1], wtedy pojedyncza scyzorykowa reszta ma dokładnie rozkład *t*-Studenta z *n - k - 2* stopniami swobody. *Należy jednak pamiętać, że trzeba posłużyć się skorygowaną wartością poziomu istotności, który uwzględnia przeprowadzenie jednocześnie n testów (po jednym dla każdej obserwacji).* Z problemem tym można się zmierzyć odwołując się do zastosowania nierówności Bonferroni'ego [1]. Wynika z niej, że jeśli np. 50 obiektów (np. parametrów) jest szacowanych jednocześnie, a ogólny test ma być przeprowadzony na poziomie istotności 0,05, to test odrzuca hipotezę zerową dla jednego szczególnego obiektu, gdy empiryczny poziom istotności *p* dla tego szczególnego testu wynosi (w przybliżeniu) $\frac{0,5}{50} = 0,001$. Gdy zarówno dodatnie jak i ujemne wartości outsiderów są brane pod uwagę, wtedy w każdym ogonie rozkładu jest wartość 0,025 poziomu istotności α, a wartość *p* w jednym z ogonów rozkładu wynosi wtedy $\frac{0,025}{50} = 0,0005$. Posłużenie się w teście szczegółowym (tzn. dla jednego obiektu) wartością poziomu istotności α = 0,025 a nie α = 0,0005, mogłoby doprowadzić do fałszywego uznania niektórych obserwacji za outsiderów.

Rozdział 10-2-3. Diagnostyka regresji oparta o "odległość Cook'a" D_i.

Bezpośrednia metoda dla ustalenia wpływu obserwacji określa jak zmienia się analiza, kiedy pojedyncze obserwacje są usunięte z danych. Metoda usunięcia pojedynczego przypadku służy często do podkreślenia najbardziej wpływowych obserwacji, które następnie mogą być dokładniej badane. Mówimy, że obserwacja jest "wpływowa", jeżeli ma duży wpływ na dopasowany model. Jest mnóstwo sposobów, na które wpływ ten może się objawić. Chociaż wpływowość i dźwignięcie są pojęciami powiązanymi z sobą, to są one różnymi wielkości. Podczas gdy dźwignięcia, zdefiniowane powyżej, zależą tylko od (macierzy projektu) zmiennych

objaśniających (porównaj (11-1.8)), wpływ obserwacji zależy także od wartości odpowiedzi. Związek ten jest pokazany na wykresie [1].



Rysunek 10-2-3.1 [1] Wykres '*a*' pokazuje wyróżnioną obserwację, której odcięta *x* jest bliska średniej, dając jej niskie dźwignięcie. Punkt ten ma mały wpływ na nachylenie, za to nieco większy na przecięcie z osią rzędnych. W części '*b*' wyróżniona obserwacja ma wysokie dźwignięcie, ale niski wpływ, co jest spowodowane tym, że, podczas gdy wariancja oszacowanego tangensa konta nachylenia wzrasta znacząco, (bo $S^2(X)$ maleje) na skutek nie uwzględnienia obserwacji, to dopasowana linia regresji niewiele się zmienia. W części '*c*' odcięta *x* wyróżnionej obserwacji jest bardzo duża, co wskazuje na duże dźwignięcie. Natomiast rzędna tej obserwacji, która leży blisko starej linii regresji, leży jednak dużo dalej od nowej dopasowanej linii, niż pozostałe obserwacje, co prowadzi do dużego wpływu tej obserwacji. Zatem może się zdarzyć, że obserwacje, dla których x_i jest odległe od średniej mogą mocno wpływać na dopasowanie modelu. Ponieważ h_i jest dla nich duże, dają one mylącą małą zwykłą resztę.

Zatem, jeśli obserwacja jest outsiderem 1) bądź wśród odpowiedzi zmiennej Y, 2) bądź w przestrzeni czynników $X_1, X_2, \ldots, X_k$,, 3) bądź, jeśli mocno wpływa na dopasowanie modelu (jako odzwierciedlenie różnicy pomiędzy MSE i $MSE_{(-i)}$), wtedy obserwacja ta może być kojarzona z mocno odstającą resztą scyzorykową (10-2-2.10). Naturalnie, kombinacje dwóch lub trzech z tych efektów mogłyby również dać duże wartości scyzorykowych reszt.

Szczególnie użyteczna w diagnostyce regresji związanej z wyszukiwaniem *obserwacji wpływowych* jest tzw. "odległością Cook'a". *Odległość Cook'a mierzy zakres, w którym zmieniają się współczynniki regresji, kiedy określona, wskazana obserwacja, jest usuwana.*

W przypadku nieskorelowanych czynników, których średnie są równe zero a wariancje równe, odległość Cook'a D_i dla *i*-tej obserwacji ($i = 1, 2, \ldots, n$) jest proporcjonalna do [1]:

$$\sum_{j=0}^{k}\left[\hat{\beta}_j - \hat{\beta}_{j(-i)}\right]^2 = \left[\hat{\beta}_0 - \hat{\beta}_{0(-i)}\right]^2 + \left[\hat{\beta}_1 - \hat{\beta}_{1(-i)}\right]^2 + \ldots + \left[\hat{\beta}_k - \hat{\beta}_{k(-i)}\right]^2 \qquad (10\text{-}2\text{-}3.12)$$

gdzie $\hat{\beta}_j$ jest oszacowaniem współczynnika regresji przy uwzględnieniu wszystkich danych, a $\hat{\beta}_{j(-i)}$ jest odpowiednim oszacowaniem współczynnika regresji z usuniętą *i*-tą obserwacją. Jeśli czynniki nie mają

średniej równej zero, równych wariancji i nie są nieskorelowane, wtedy odległość Cook'a jest proporcjonalna do ważonej sumy wyrażeń $\left[\hat{\beta}_j - \hat{\beta}_{j(-i)}\right]^2$.

Dla dowolnego zbioru danych, odległość Cook'a D_i dla i-tej obserwacji, może zostać wyrażona poprzez współczynniki dźwignięcia i studentyzowane reszty jako:

$$D_i = \left(\frac{1}{k+1}\right) \cdot R_i^2 \cdot \left(\frac{h_i}{1-h_i}\right) = \frac{U_i^2 h_i}{(k+1) \cdot MSE \cdot (1-h_i)^2} \in <0,+\infty), \quad i = 1,2,...,n. \qquad (10\text{-}2\text{-}3.13)$$

Wyrażenie to przedstawia ścisłą zależność D_i od współczynnika dźwignięcia h_i i studentyzowanych reszt R_i. Widać, że wartość D_i może być duża z dwóch powodów: bądź, dlatego, że 1) obserwacja jest ekstremalnie odległa w przestrzeni czynnika (tzn. h_i jest bliskie wartości 1) bądź, ponieważ 2) obserwacja ma dużą wartość studentyzowanej reszty R_i.

Często w praktyce stosuje się zasadę, według której *obserwacja wpływowa to taka, która ma dużą wartość iloczynu* "$R_i^2 \cdot h_i$", $i = 1,2,...,n$. Według tej zasady i zgodnie z (10-2-3.13) odległość Cook'a nadawałaby się w sposób szczególny do diagnostyki regresji związanej z wyszukiwaniem takich obserwacji. Ponieważ dla czynników posiadających rozkład Gaussa, rozkład statystyki D_i jest z grubsza podobny do rozkładu statystyki F-Snedecora $\mathsf{F}_{k,n-k-1}$ z liczbą stopni swobody k oraz n-k-1 (i im wielkość próby większa tym to przybliżenie jest lepsze), dlatego jeśli model regresji jest dobry, wtedy oczekuje się, że indywidualne wartości D_i powinny być mniejsze niż 1. Stąd Cook i Weisberg [17] zasugerowali, że każda indywidualna obserwacja z wartością większą niż 1 powinna podlegać szczególnej analizie. Przyjmuje się również, że na uwagę zasługuje każda obserwacja większa od mediany statystyki $\mathsf{F}_{k,n-k-1}$. Jeśli chodzi o testy istotności dla D_i, to wspomniane przybliżenie statystyką $\mathsf{F}_{k,n-k-1}$ pracuje dobrze przy ocenie istotności indywidualnej obserwacji. Jednakże nie jest ono wystarczająco dokładne (nawet dla $n > 200$) przy kontroli istotności maksymalnej wartości statystyki D_i [1]. Wartości krytyczne dla maksymalnej wartości $(n-k-1)D_i$ można znaleźć w [1]. Ponadto, w przypadku ustalonych czynników, Obenchain [18] sugeruje w celu wskazania obserwacji wpływowej raczej kontrolę pary R_i oraz h_i niż odległości Cook'a.

Wniosek z powyższych rozważań jest taki, że omówione powyżej statystyki i ewentualne testy istotności użyte w celu wskazania obserwacji wpływowych są tylko pomocą dla badacza, który sam musi zdecydować o tym czy usunąć outsidera z danych pomiarowych (co na ogół poprawia dopasowanie modelu), czy pozostawić go w „przeczuciu" jego znaczenia dla konstrukcji modelu w populacji.

B. Rozdział 11. Macierzowe ujęcie klasycznego modelu regresji i współczynnik dźwignięcia.

Poniższe ujęcie zostało przedstawione jako uzupełnienie powyższych rozważań. Rozważmy model, który może być nieliniowym modelem regresji z błędem E_i, i = 1, 2, ..., n, pojawiającym się addytywnie, gdzie n *jest* wymiarem próby. Własności (nieobserwowanego) losowego składnika losowego E_i są następujące: $E(E_i)=0$, $\sigma^2(E_i)\equiv\sigma_E^2$ dla $i=1,2,...n$, tzn. $\sigma_{\mathbf{E}}^2$ jest stałą wariancją składnika losowego dla każdego poziomu czynników, na którym i-ta ich obserwacja wynosi $(X_{1i}, X_{2i},...,X_{pi})$ oraz $\mathrm{cov}(E_i, E_s)=0$ dla $i \neq s$.

Wprowadźmy oznaczenia (indeks $_{|\mathbf{X}}$ będziemy na ogół pomijać):

$$\vec{\beta}=\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{pmatrix},\quad \vec{X}_i=\begin{pmatrix}X_{1i}\\ X_{2i}\\ \vdots\\ X_{pi}\end{pmatrix},\quad i=1,2,...,n\ ,\quad \mathbf{E}_{|\mathbf{X}}\equiv\mathbf{E}=\begin{pmatrix}E_1\\ E_2\\ \vdots\\ E_n\end{pmatrix}. \tag{11.1}$$

dla wektora $\vec{\beta}\equiv\vec{\beta}_{p\times 1}$ parametrów strukturalnych funkcji regresji, wektora p (będących pod kontrolą) czynników $\vec{X}_i$ dla i-tej obserwacji, i = 1, 2, ..., n, oraz wektora $\mathbf{E}\equiv\mathbf{E}_{n\times 1}$ składnika losowego. W przypadku występowania stałego przesunięcia w funkcji regresji, odpowiada mu współczynnik β_1 (oznaczany wtedy jako β_0 z indeksem „0”), a liczba parametrów $p=k+1$, gdzie k parametrów strukturalnych stoi przy czynnikach (regresorach) modelu regresji.

W przypadku obserwacji n par $\left((X_{i1}, X_{i2},..., X_{ip}), Y_i\right)$, gdzie pierwszym elementem pary jest ciąg p obserwacji zmiennych $(X_1, X_2,..., X_p)$, a drugim odpowiadająca mu obserwacja Y_i zmiennej objaśnianej Y, ogólny model regresji dla zmiennej opisywanej Y można zapisać w postaci układu n równań:

$$Y_i=\eta\left(\vec{X}_i,\vec{\beta}\right)+E_i \qquad \text{dla} \qquad i=1,2,...,n\ . \tag{11.2}$$

Reszty dla tego modelu są zwykle definiowane następująco:

$$U_i=Y_i-\eta(\vec{X}_i,\hat{\vec{\beta}}) \quad \text{dla} \quad i=1,2,...,n \tag{11.3}$$

gdzie $\hat{\vec{\beta}}$ jest estymatorem $\vec{\beta}$ otrzymanym bądź za pomocą metody najmniejszych kwadratów bądź maksymalnej wiarygodności.

Gdy założymy, że liczba parametrów w $\hat{\vec{\beta}}$ jest mała w porównaniu z liczbą obserwacji n, wtedy składowe wektora:

$$\hat{\mathbf{E}} \equiv \mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}, \tag{11.4}$$

mają własności zbliżone do $\mathbf{E} = (E_1, \quad E_2, \quad \cdots, \quad E_n)^T$.

Rozdział 11-1. Wyprowadzenie macierzowego ujęcia klasycznego model regresji.

Dla normalnego modelu liniowego mamy $\eta(\vec{X}_i, \vec{\beta}) = \vec{X}_i^T \vec{\beta}$, gdzie:

$$\vec{X}_i = (X_{1i}, X_{2i}, \ldots, X_{pi})^T = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{pi} \end{pmatrix} \qquad \text{dla} \quad i = 1, 2, \ldots, n. \tag{11-1.5}$$

Model (11.2) przyjmuje więc postać:

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \ldots + X_{pi}\beta_p + E_i = \vec{X}_i^T \vec{\beta} + E_i \tag{11-1.6}$$

dla $i = 1, 2, \ldots, n$, gdzie błędy E_i mają niezależne normalne rozkłady ze średnią równą zero i wariancją σ^2.
Model (11-1.6) można zapisać następująco :

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\, \vec{\beta}_{p\times 1} + \mathbf{E}_{n\times 1}, \tag{11-1.7}$$

gdzie $\operatorname{cov}(\mathbf{E}) = \sigma_E^2\, I$, a $\mathbf{Y} \equiv \mathbf{Y}_{n\times 1}$ jest $n \times 1$ wymiarowym wektorem obserwacji dla zmiennej objaśnianej Y:

$$\mathbf{Y}_{|\mathbf{X}} \equiv \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \tag{11-1.8}$$

natomiast $\mathbf{X} \equiv \mathbf{X}_{n\times p}$ jest znaną macierzą rzędu $n \times p$:

$$\mathbf{X} = \begin{pmatrix} X_{11}, X_{21}, \ldots, X_{p1} \\ X_{12}, X_{22}, \ldots, X_{p2} \\ \vdots \\ X_{1n}, X_{2n}, \ldots, X_{pn} \end{pmatrix} \equiv \begin{pmatrix} \vec{X}_1^T \\ \vec{X}_2^T \\ \vdots \\ \vec{X}_n^T \end{pmatrix}, \tag{11-1.9}$$

tzw. *macierzą planowania* dla zmiennych objaśniających, włączając w nią w razie potrzeby, stałą zmienną przesunięcia równą:

$$X_1 = I = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}. \tag{11-1.10}$$

Równanie (11.1-7) można przepisać następująco:

$$\mathbf{E}_{n\times 1} = \mathbf{Y}_{n\times 1} - \mathbf{X}_{n\times p}\, \vec{\beta}_{p\times 1}, \tag{11-1.11}$$

skąd widać, że suma kwadratów odchyłek wartości pomiarowych od (w ogólności) powierzchni regresji I rodzaju, ma postać:

$$(\mathbf{E}^T)_{1\times n}\mathbf{E}_{n\times 1} = (\mathbf{Y}_{n\times 1} - \mathbf{X}_{n\times p}\,\vec{\beta}_{p\times 1})^T(\mathbf{Y}_{n\times 1} - \mathbf{X}_{n\times p}\,\vec{\beta}_{p\times 1}) \quad . \tag{11-1.12}$$

Analogicznie, dla *n*-wymiarowej próby w miejsce (11.1-6) otrzymujemy *n* równań:

$$Y_i = \hat{E}(Y \mid X_{1i}, X_{2i}, ..., X_{pi}) + U_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_p X_{pi} + U_i = \vec{X}_i^T \hat{\vec{\beta}} + U_i \ , \ i = 1, 2, ..., n, \tag{11-1.13}$$

gdzie $U_i \equiv \hat{E}_i$ jest *i*-tym składnikiem resztowym dla ogólnego modelu regresji wielorakiej II rodzaju w próbie, gdzie *n* jest liczbą par, w których pierwszym elementem pary jest układ czynników $(X_{1i}, X_{2i}, ..., X_{pi})$ a drugim obserwacja Y_i zmiennej *Y*. W próbie, teoretyczne średnie warunkowe $\hat{Y}_i = \hat{E}(Y \mid X_{1i}, X_{2i}, ..., X_{pi})$, $i = 1, 2, ..., n$, są estymatorami warunkowych wartości oczekiwanych $E(Y \mid X_{1i}, X_{2i}, ..., X_{pi})$. Ich wektor ma postać:

$$\hat{\mathbf{Y}}_{|\mathbf{X}} \equiv \hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \mathbf{X}\,\hat{\vec{\beta}} \ , \tag{11-1.13}$$

gdzie $\hat{\vec{\beta}} \equiv \hat{\vec{\beta}}_{p\times 1}$ jest wektorem estymatorów parametrów strukturalnych $\vec{\beta} \equiv \vec{\beta}_{p\times 1}$ modelu regresji:

$$\hat{\vec{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}. \tag{11-1.14}$$

Rozważmy formę kwadratową dla sumy odchyłek wartości pomiarowych, która jest sumą kwadratów reszt modelu regresji:

$$\begin{aligned} W \equiv SSE &= \sum_{i=1}^{n} U_i^{\,2} = \sum_{i=1}^{n} (Y_i - \vec{X}_i^{\,T}\,\hat{\vec{\beta}})^T (Y_i - \vec{X}_i^{\,T}\,\hat{\vec{\beta}}) = \\ &= ((\mathbf{Y}^T)_{1\times n} - (\hat{\vec{\beta}}^T)_{1\times p}(\mathbf{X}^T)_{p\times n})(\mathbf{Y}_{n\times 1} - \mathbf{X}_{n\times p}\,\hat{\vec{\beta}}_{p\times 1}) = (\mathbf{Y}_{n\times 1} - \hat{\mathbf{Y}}_{n\times 1})^T(\mathbf{Y}_{n\times 1} - \hat{\mathbf{Y}}_{n\times 1}) = \mathbf{U}^T\,\mathbf{U} \ , \end{aligned} \tag{11-1.15}$$

gdzie wyrażenie $\mathbf{U}^T\,\mathbf{U}$ jest licznikiem wariancji resztowej (10-2-2.4):

$$MSE \equiv S_{\mathbf{U}}^2 = \frac{SSE}{n-p} = \frac{\mathbf{U}^T\mathbf{U}}{n-p} \ , \tag{11-1.16}$$

będącej nieobciążonym estymatorem wariancji składnika losowego, tzn. $E(MSE) = \sigma_E^2$.

Niech macierz **X** ma pełną rangę kolumnową *p*, tak że macierz $\mathbf{X}^T\mathbf{X}$ jest nieosobliwa. Minimalizując sumę kwadratów odchyłek (11-1.15) po wektorze estymatorów parametrów $\hat{\vec{\beta}}$ otrzymujemy estymator MNK dla $\vec{\beta}$ (pokazać):

$$\hat{\vec{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{11-1.17}$$

Dowód: Wyprowadzenie (11-1.17) jest następujące. Z (11-1.15) otrzymujemy:

$$\frac{\partial(SSE)}{\partial\hat{\vec{\beta}}} = -2\,\mathbf{X}^T\mathbf{Y} + 2\,\mathbf{X}^T\mathbf{X}\,\hat{\vec{\beta}} = \vec{0} \tag{11-1.18}$$

$$\frac{\partial^2(SSE)}{\partial\hat{\vec{\beta}}\,\partial\hat{\vec{\beta}}^T} = 2\,\mathbf{X}^T\mathbf{X} \tag{11-1.19}$$

przy czym ponieważ macierz $\mathbf{X}^T\mathbf{X}$ jest dodatnio określona, dlatego po pierwsze $\mathbf{X}^T\mathbf{X}$ jest nieosobliwa i lewostronne pomnożenie (11-1.18) przez $(\mathbf{X}^T\mathbf{X})^{-1}$ daje (11-1.17), a po drugie, otrzymane rozwiązanie minimalizuje *SSE*.

Zauważmy również, że (11-1.18) daje:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\,\hat{\vec{\beta}}) = (\mathbf{X}^T)_{p\times n}\mathbf{U}_{n\times 1} = \vec{0}, \tag{11-1.20}$$

co oznacza, że *n*-wymiarowy wektor danych dla każdego czynnika jest ortogonalny do wektora reszt (dla stałego wektora przesunięcia I wiemy to już z (10-2-2.3)):

$$X_j \perp \mathbf{U}, \qquad j=1,2,...,p. \tag{11-1.21}$$

Ponieważ $\hat{\mathbf{Y}} = \mathbf{X}\hat{\vec{\beta}}$ zatem $\hat{\mathbf{Y}}^T = \hat{\vec{\beta}}^T\mathbf{X}^T$ skąd ze względu na (11-1.20) otrzymujemy:

$$\hat{\mathbf{Y}}^T\mathbf{U} = \hat{\vec{\beta}}^T\mathbf{X}^T\,\mathbf{U} = 0 \tag{11-1.22}$$

co oznacza, że *n*-wymiarowy wektor teoretycznych średnich warunkowych dla zmiennej objaśnianej (zatem i powierzchnia regresji) jest ortogonalny do wektora reszt:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\vec{\beta}} \perp \mathbf{U}. \tag{11-1.23}$$

Z (11-1.17) widać, że wartości przewidywane modelem są liniową funkcją zmiennej objaśnianej $\mathbf{Y} = \mathbf{X}\hat{\vec{\beta}} + \mathbf{U}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\vec{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \equiv \mathbf{H}\,\mathbf{Y}, \tag{11-1.24}$$

gdzie $n \times n$ macierz dźwignięć **H** (tzw. macierz kapeluszowa od ang. „hat matrix") jest zdefiniowana jako

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T. \tag{11-1.25}$$

Macierz $\mathbf{H}$ określa ortogonalny rzut obserwacji $\mathbf{Y}$ na *n*-wymiarową płaszczyznę obserwacji kombinacji liniowej $\mathbf{X}^T\hat{\vec{\beta}}$ czynników. Stąd jej nazwa „macierz rzutowa". Z postaci (11-1.25) macierzy $\mathbf{H}$ widać, że jest ona symetryczna i (jak przystoi na macierz rzutową) idempotentna, tzn.

$$\mathbf{H}^2 = \mathbf{H}. \tag{11-1.26}$$

Z zależności (11-1.24) widać, że wektor reszt **U = Y - Ŷ** ma następującą postać

$$\mathbf{U} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{M}\,\mathbf{Y}. \tag{11-1.27}$$

gdzie

$$\mathbf{M} \equiv (\mathbf{I} - \mathbf{H}) \tag{11-1.28}$$

jest również symetryczna i idempotentna $\mathbf{M}^2 = \mathbf{M}$, rzutując wektor obserwacji $\mathbf{Y}$ na n-wymiarową płaszczyznę reszt. Z (11-1.24) oraz (11-1.27) wynika rozkład wektora odpowiedzi układu na ortogonalne składowe (porównaj (11-1.23)) :

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{U} = \mathbf{H}\,\mathbf{Y} + \mathbf{M}\,\mathbf{Y} \ . \tag{11-1.29}$$

Macierz kowariancji dla reszt spełnia związek (pokazać):

$$\mathrm{cov}(\mathbf{U}) = (\mathbf{I} - \mathbf{H})\,\sigma_E^2 \ . \tag{11-1.30}$$

Sprawdźmy, że zachodzi

$$E(\mathbf{U}) = \vec{0} \ . \tag{11-1.31}$$

Istotnie, korzystając z postaci funkcji regresji w populacji $E(\mathbf{Y}_{|\mathbf{X}}) = \mathbf{X}\vec{\beta}$, trzymujemy:

$$\begin{aligned} E(\mathbf{U}) &= E(\mathbf{Y}) - E(\hat{\mathbf{Y}}) = (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) = E(\mathbf{Y}) - \mathbf{H}\,E(\mathbf{Y}) = E(\mathbf{Y}) - (\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T)\mathbf{X}\vec{\beta} \\ &= E(\mathbf{Y}) - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}(\mathbf{X}^T\,\mathbf{X})\vec{\beta} = E(\mathbf{Y}) - \mathbf{X}\vec{\beta} = \vec{0} \ . \end{aligned} \tag{11-1.32}$$

Dowód:

Korzystając dla zmiennych nieskorelowanych z $E(Y_iY_s) = E(Y_i)E(Y_s)$, $i \neq s$, z równości $E(\mathbf{Y}\,\mathbf{Y}^T) = \sigma^2(\mathbf{Y}_{|\mathbf{X}})$ $+ E(\mathbf{Y})E(\mathbf{Y}^T)$, gdzie w modelu regresji klasycznej $\sigma^2(\mathbf{Y}_{|\mathbf{X}}) \equiv \sigma^2(Y_i)\,\mathbf{I} = \sigma^2(E_i)\,\mathbf{I} = \sigma_E^2\,\mathbf{I}$, otrzymujemy wykorzystując (11-1.31) oraz symetrię macierzy $(\mathbf{I} - \mathbf{H})$:

$$\begin{aligned} &\mathrm{cov}(\mathbf{U}) \equiv \mathrm{cov}(\mathbf{U},\mathbf{U}) = E((\mathbf{U} - E(\mathbf{U}))(\mathbf{U} - E(\mathbf{U}))^T) = E(\mathbf{U}\,\mathbf{U}^T) = E((\mathbf{I} - \mathbf{H})\,\mathbf{Y}\,\mathbf{U}^T) \\ &= E(\mathbf{Y}\,\mathbf{U}^T) - E(\mathbf{H}\,\mathbf{Y}\,\mathbf{U}^T) = E(\mathbf{Y}\,\mathbf{U}^T) - E(\hat{\mathbf{Y}}\,\mathbf{U}^T) = E(\mathbf{Y}\,\mathbf{U}^T) = E(\mathbf{Y}\,\mathbf{Y}^T)(\mathbf{I} - \mathbf{H})^T \\ &= \left(\sigma^2(\mathbf{Y}_{|\mathbf{X}}) + E(\mathbf{Y})E(\mathbf{Y}^T)\right)\,(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{Y}_{|\mathbf{X}})\,\,(\mathbf{I} - \mathbf{H})^T + E(\mathbf{Y})E(\mathbf{Y}^T)\,\,(\mathbf{I} - \mathbf{H})^T \\ &= \sigma^2(\mathbf{Y}_{|\mathbf{X}})\,\,(\mathbf{I} - \mathbf{H})^T + E(\mathbf{Y})E(((\mathbf{I} - \mathbf{H})\mathbf{Y})^T) = \sigma_E^2\,\mathbf{I}\,(\mathbf{I} - \mathbf{H})^T + E(\mathbf{Y})E(\mathbf{U}^T) \\ &= (\mathbf{I} - \mathbf{H})\sigma_E^2 \ . \end{aligned} \tag{11-1.33}$$

Uwaga: Ze względu na $(\mathbf{I} - \mathbf{H})^2 = (\mathbf{I} - \mathbf{H})$ widać również, że:

$$\mathrm{cov}(\mathbf{U}) = (\mathbf{I} - \mathbf{H})^T\,\Sigma\,(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})\sigma_E^2 \ , \tag{11-1.34}$$

gdzie $\Sigma = \mathbf{I}\sigma_E^2$ jest (diagonalną) macierzą kowariancji dla błędu $\mathbf{E}$ (oraz obserwacji $\mathbf{Y}$).

Ze względu na to, że funkcja regresji w populacji ma postać $E(\mathbf{Y}) = \mathbf{X}\vec{\beta}$, a w próbie ma postać $\hat{\mathbf{Y}} = \mathbf{X}\hat{\vec{\beta}}$ z (11-1.32), $E(\mathbf{U}) = E(\mathbf{Y}) - E(\hat{\mathbf{Y}}) = \vec{0}$, widać również, że estymatory MNK $\hat{\vec{\beta}}$ są nieobciążone, tzn.:

$$E(\hat{\vec{\beta}}) = \vec{\beta} \ . \tag{11-1.35}$$

W końcu korzystając z (11-1.17) oraz (11-1.35) można wyznaczyć postać macierzy wariancji-kowariancji dla estymatorów parametrów strukturalnych $\sigma^2_{\hat{\vec{\beta}}} \equiv \sigma^2(\hat{\vec{\beta}})$ jako równą (pokazać):

$$\sigma^2_{\hat{\vec{\beta}}} = E\left((\hat{\vec{\beta}} - \vec{\beta})\,(\hat{\vec{\beta}} - \vec{\beta})^T\right) = E\left(\hat{\vec{\beta}}\hat{\vec{\beta}}^T\right) - \vec{\beta}\vec{\beta}^T = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2_{\mathbf{E}} \ . \tag{11-1.36}$$

Ponieważ wartość wariancji składnika losowego $\sigma_{\mathbf{E}}^2$ jest na podstawie n-wymiarowej próby oszacowana poprzez wariancję składnika resztowego $S_{\mathbf{U}}^2$:

$$MSE \equiv S_{\mathbf{U}}^2 = \frac{SSE}{n-p}, \tag{11-1.37}$$

zatem oszacowana z próby macierz wariancji-kowariancji dla estymatorów parametrów strukturalnych wynosi:

$$\hat{\sigma}^2_{\hat{\vec{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1} MSE, \tag{11-1.38}$$

a pierwiastki elementów na diagonalnej macierzy (11-1.38) są średnimi błędami $S_{\hat{\beta}_j}$, $j = 1,2,...,p$, oszacowań parametrów strukturalnych modelu.

Macierz korelacyjna dla estymatorów $\hat{\vec{\beta}}$ jest określona następująco:

$$\hat{\rho}_{\hat{\vec{\beta}}} = \mathbf{S}^{-\frac{1}{2}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{S}^{-\frac{1}{2}}, \tag{11-1.39}$$

gdzie macierz:

$$\mathbf{S} = \mathrm{diag}\left((\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{11-1.40}$$

jest diagonalną macierzy $(\mathbf{X}^T\mathbf{X})^{-1}$.

Jak to już omawialiśmy wcześniej, współczynniki dźwignięcia $h_i \equiv h_{ii}$ obserwowane w próbie, wpływają na ogólne dopasowanie modelu. Są one zdefiniowane jako elementy przekątnej macierzy **H**, (11-1.25). Ponieważ zgodnie z (11-1.24) $\hat{Y}_i$ może zostać wyrażone jako:

$$\hat{Y}_i = h_{ii}Y_i + \sum_{s \neq i} h_{is}Y_k, \qquad \text{gdzie } h_i \equiv h_{ii} \tag{11-1.41}$$

zatem jeśli h_{ii} jest dużo większe od innych elementów w i-tym wierszu macierzy $\mathbf{H} = (h_{is})$, wtedy i-ta dopasowana wartość $\hat{Y}_i$ może być w dużym stopniu określona na podstawie Y_i.

Ponadto, ze względu na (11-1.30), mamy:

$$\sigma^2(U_i) = (1-h_i)\sigma_{\mathbf{E}}^2, \tag{11-1.42}$$

zatem obserwacja z wysokim dźwignięciem h_{ii} będzie mieć resztę U_i, której dyspersja jest niewielka. W ten sposób przypadki z wysokim dźwignięciem mogą nie wyróżniać się na wykresach reszt. Z (11-1.42) widać, że oszacowana z próby wariancja reszt ma postać:

$$S^2(U_i) \equiv \hat{\sigma}^2(U_i) = (1-h_i)\sigma_{\mathbf{U}}^2 \equiv (1-h_i)MSE. \tag{11-1.43}$$

Korzystając z $\hat{\mathbf{Y}} = \mathbf{X}\hat{\vec{\beta}}$, następnie z $\sigma^2_{\hat{\vec{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_{\mathbf{E}}^2$, (11-1.36), prawa propagacji błędów oraz z postaci $\mathbf{H}$, (11-1.25), otrzymujemy:

$$\sigma^2(\hat{\mathbf{Y}}) = \sigma^2(\mathbf{X}\hat{\vec{\beta}}) = \mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1}\sigma_{\mathbf{E}}^2)\mathbf{X}^T = \sigma_{\mathbf{E}}^2\mathbf{H}, \tag{11-1.44}$$

skąd widać, że nieobciążony estymator wariancji dla odpowiedzi $\hat{Y}_i$ ma postać:

$$S^2_{\hat{Y}_i} = MSE\, h_i \ . \tag{11-1.45}$$

W końcu, zgodnie z (10-2-1.10) ślad macierzy **H** wynosi:

$$\mathrm{tr}\mathbf{H} = \sum_{i=1}^{n} h_{ii} = p \tag{11-1.46}$$

czyli jest równy liczbie parametrów regresji, a średnie dźwignięcie jest równe $\bar{h} = p/n$ zgodnie z (10-2-1.11). Jak już wspominaliśmy, obserwacja, której dźwignięcie jest dużo większe od tej wartości, wymaga uwagi.

Uwaga. Rozważania Rozdziału 11-1 dotyczą sytuacji, w której zachowane są założenia KMNK (Rozdział 3-1, Część I). Ich niespełnienie, a w szczególności niespełnienie założenia stałości wariancji lub braku autokorelacji reszt, wymaga uogólnienia powyżej przedstawionego sformułowania modelu regresji [19], [4] do uogólnionej metody najmniejszych kwadratów (UMNK) lub jej szczególnego przypadku, ważonej metody najmniejszych kwadratów, dopuszczającej brak jednorodności wariancji. W pakiecie SAS między innymi procedura GLM wykorzystuje metodę najmniejszych kwadratów dla dopasowywania ogólnych liniowych modeli [20].

Uwaga. Często za nadrzędny cel analizy regresji uważa się możliwość przewidywania wartości zmiennej objaśnianej. Zatem, kolejny krok analizy powinien dotyczyć prognozowania na podstawie wyselekcjonowanego modelu regresji [4].

Rozdział 11-2. Podstawowy wynik KMNK dla jednego czynnika.

Poniżej zostały podane podstawowe formuły dla punktowych oszacowań parametrów strukturalnych liniowego modelu regresji z jednym czynnikiem *X*. Szczegółowe rozważania dla estymacji punktowej i przedziałowej tego modelu można znaleźć np. w pozycji [2]. Dla modelu liniowego z jednym czynnikiem, równanie regresji II-rodzaju ma postać:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{11-2.47}$$

Zgodnie z MNK przedstawioną powyżej, minimalizując ze względu na $\hat{\beta}_o$ oraz $\hat{\beta}_1$ sumę kwadratów reszt:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \ , \tag{11-2.48}$$

otrzymujemy *układ równań normalnych,* po rozwiązaniu którego otrzymujemy estymatory parametrów strukturalnych modelu (11-2.47) (pokazać) [4]:

$$\begin{cases} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \\ \hat{\beta}_1 = \dfrac{\mathrm{cov}(X,Y)}{S^2(X)} \end{cases} \tag{11-2.49}$$

gdzie kowariancja zmiennych X oraz Y, zapisana następująco:

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{11-2.50}$$

pokrywa się z estymatorem metody największej wiarygodności (MNW) macierzy kowariancji, której szczególnym przypadkiem dla rozkładu normalnego reszt jest MNK, natomiast:

$$S^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{11-2.51}$$

jest estymatorem wariancji zmiennej X metody MNW. Odpowiednie nieobciążone estymatory kowariancji i wariancji, miałyby w mianownikach dzielenie przez (n-1).

Rozdział 11-2-1. Współczynnik korelacji liniowej Pearsona.

Rozważmy współczynnik korelacji liniowej Pearsona pomiędzy zmiennymi X, Y, (3-2.11), w populacji:

$$\rho \equiv \rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\,\sigma(Y)}\ , \tag{11-2-1.52}$$

gdzie $\sigma(X)$ oraz $\sigma(Y)$ są odchyleniami standardowymi zmiennych X oraz Y w populacji. Wartość współczynnika korelacji ρ jest liczbą bezwymiarową z przedziału (pokazać):

$$\rho \in \langle -1, +1 \rangle\,. \tag{11-2-1.53}$$

Estymatorem parametru ρ jest (empiryczny) współczynnik korelacji liniowej (Pearsona) $R \equiv \hat{\rho}$ dla zmiennej objaśnianej Y i objaśniającej X, zdefiniowany w próbie (dla danych reprezentowanych graficznie przez diagram punktowy), następująco [2]:

$$R \equiv R_{(n)} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\ \sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{SSXY}{\sqrt{SSX}\ \sqrt{SSY}}\ , \tag{11-2-1.54}$$

gdzie:

$$SSXY = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\ ,\ SSX = \sum_{i=1}^{n}(X_i - \overline{X})^2\ ,\ SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2\ . \tag{11-2-1.55}$$

Współczynnik R jest estymatorem zgodnym [2] parametru ρ, tzn.:

$$\forall \varepsilon > 0,\quad \lim_{n\to\infty} P(|R_{(n)} - \rho| < \varepsilon) = 1. \tag{11-2-1.56}$$

Warto zauważyć, że R jest jedynie asymptotycznie nieobciążonym estymatorem [2] parametru ρ, tzn.:

$$\lim_{n\to\infty} E(R_{(n)}) = \rho\,, \tag{11-2-1.57}$$

natomiast dla skończonego n, jest on estymatorem obciążonym [2], tzn. $E(R_{(n)}) \neq \rho$.

Współczynnik R można zapisać następująco:

$$R = \frac{S_X}{S_Y}\hat{\beta}_1\,, \tag{11-2-1.58}$$

gdzie empiryczne odchylenia standardowe:

$$S_X = \sqrt{S^2(X)} \quad \text{oraz} \quad S_Y = \sqrt{S^2(Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\ , \tag{11-2-1.59}$$

przyjmują w próbce wartości s_X oraz s_Y.

Tak jak w przypadku ρ, wartość współczynnika korelacji r jest liczbą bezwymiarową z przedziału:

$$r \in \langle -1, +1 \rangle\,. \tag{11-2-1.60}$$

Wartość r empirycznego współczynnika korelacji R określa rodzaj i siłę związku pomiędzy zmiennymi [2]. Silne związki liniowe pomiędzy zmiennymi są odzwierciedlone w wartości *bezwzględnej* współczynnika korelacji bliskiej 1, a brak zależności prostoliniowej (tzn. bądź owalne rozmycie diagramu punktowego bądź silna korelacja krzywoliniowa) jest odbity w jej wartości bliskiej 0.

Znak r i znak wartości estymatora $\hat{\beta}_1$ w próbce korespondują ze sobą, tzn. zachodzi:

$$\begin{aligned} r < 0 &\Rightarrow \hat{\beta}_1 < 0 \\ r = 0 &\Rightarrow \hat{\beta}_1 = 0\ , \\ r > 0 &\Rightarrow \hat{\beta}_1 > 0 \end{aligned} \tag{11-2-1.61}$$

Jednak jak wynika z (11-2-1.58), wartość r nie określa modułu wartości $\hat{\beta}_1$.

W klasycznym modelu regresji, wartość modułu współczynnika korelacji Pearsona (11-2-1.58) pokrywa się z wartością współczynnika korelacji wielorakiej (Rozdział 3) dla zmiennych X i Y. Wartość współczynnika determinacji r^2 nie daje informacji o wartości bezwzględnej oszacowania $\hat{\beta}_1$ (podobnie jak i r), ale nie informuje również o kierunku zależności pomiędzy zmiennymi (pozytywnej bądź negatywnej), jak to czyni r.

Twierdzenie (o asymptotycznym rozkładzie statystyki R).

Jeśli dwuwymiarowy rozkład łączny zmiennych X, Y w populacji jest rozkładem *dowolnym*, wtedy [2]:

a) pod warunkiem, że istnieją skończone momenty drugiego rzędu zmiennych X i Y, wartość oczekiwana R jest dla $n \to \infty$ równa:

$$E(R) \underset{n\to\infty}{=} \rho \tag{11-2-1.62}$$

b) pod warunkiem, że istnieją skończone momenty czwartego rzędu zmiennych X i Y, wariancja R jest dla $n \to \infty$ równa:

$$\sigma^2(R) \underset{n\to\infty}{=} \frac{(1-\rho^2)^2}{n}, \qquad (11\text{-}2\text{-}1.63)$$

i rozkład asymptotyczny statystyki R jest rozkładem normalnym:

$$N\left(\rho, \frac{(1-\rho^2)^2}{n}\right). \qquad \text{(koniec Twierdzenia)} \qquad (11\text{-}2\text{-}1.64)$$

W przypadku, gdy dwuwymiarowy rozkład łączny zmiennych X, Y w populacji jest rozkładem normalnym, wtedy współczynnik R z próby n-elementowej ma rozkład o gęstości prawdopodobieństwa [2]:

$$f(r) = \frac{n-2}{\pi}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}} \int_0^1 \frac{du}{\sqrt{1-u^2}} \frac{u^{n-2}}{(1-\rho u r)^{n-1}}, \qquad (11\text{-}2\text{-}1.65)$$

gdzie $-1 \le r \le 1$.

Wykorzystując (11-2-1.65) można pokazać, że $E(R) = \int_{-1}^{1} r\, f(r)\, dr \approx \rho$ oraz $\sigma^2(R) = \int_{-1}^{1} (r - E(R))^2\, f(r)\, dr \approx \frac{(1-\rho^2)^2}{n}$. Zatem w przypadku rozkładu normalnego, $E(R)$ oraz $\sigma^2(R)$ są zadane w *przybliżeniu* przez kolejno (11-2-1.62) oraz (11-2-1.63), dla dowolnego, skończonego n.

Korzystając z (11-2-1.64) można, dla wystarczająco dużej próby (n - kilkaset), dokonać estymacji przedziałowej współczynnika korelacji liniowej ρ, bądź przeprowadzić weryfikację hipotezy zerowej odnośnie jego wartości w populacji [2].

Rozdział 11-3. Uzupełnienie. Testy niezależności reszt.

Jednym z założeń klasycznego modelu regresji wielorakiej (Rozdział 3-1) jest założenie o niezależności obserwacji zmiennej objaśnianej Y. Dlatego testom nie występowania autokorelacji reszt w modelu regresji poświęca się sporo uwagi przy sprawdzaniu poprawności wyselekcjonowanego modelu. Jednym z testów nie występowania autokorelacji reszt jest test Durbina-Watsona [21] omówiony poniżej. Test ten wykrywa jedynie autokorelację pierwszego rzędu. Innym testem testującym brak autokorelacji reszt jest np. test Breuscha-Godfrey'a (dostępny w SAS'ie) wykrywający również autokorelacje wyższych rzędów [22].

Rozdział 11-3-1. Test Durbina-Watsona.

W teście Durbina-Watsona [21] weryfikuje się hipotezę o zerowaniu się współczynnika autokorelacji. (Korelacja może być np. autokorelacją $r(Y_t, Y_{t-1})$ pomiędzy pomiarami zmiennej Y w chwilach czasu t i t-1.) W przypadku istnienia autokorelacji składnika losowego należy zmienić postać modelu bądź spróbować dokonać odpowiedniej transformacji zmiennych.

Założeniami wymaganymi przy stosowaniu testu Durbina-Watsona są:

1. Nielosowość czynników.
2. Brak jawnej, opóźnionej zmiennej objaśnianej występującej w charakterze zmiennej objaśniającej (np. z pewnym przesunięciem „lag" typowym w szeregach czasowych).
3. Występowanie wyrazu wolnego w modelu regresji.
4. Normalność rozkładu składnika losowego.
5. Liczba obserwacji $n > 15$ (im większa jest próba, tym węższy jest przedział niekonkluzywny testu).

Rozpatrzmy parę hipotez, gdzie hipoteza zerowa o niezależności reszt, oznacza nie występowanie autokorelacji *pierwszego rzędu* składnika losowego modelu.
Zatem rozważamy hipotezę zerową:

$$H_0:\ \rho = 0 \qquad \text{(brak autokorelacji pierwszego rzędu)} \qquad (11\text{-}3\text{-}1.1)$$

wobec alternatywnej:

$$H_1:\ \rho < 0 \text{ lub } \rho > 0\ , \text{ (występuje autokorelacja pierwszego rzędu)} \qquad (11\text{-}3\text{-}1.2)$$

gdzie ρ jest wartością współczynnika autokorelacji rzędu pierwszego (3-2.11) w populacji. Jego estymatorem w próbie jest współczynnik autokorelacji w próbie $\hat{\rho}$:

$$\hat{\rho} = \frac{\sum_{i=2}^{n}\left(U_i - \overline{U}\right)\cdot\left(U_{i-1} - \overline{U}\right)}{\sqrt{\sum_{i=2}^{n}\left(U_i - \overline{U}\right)^2}\cdot\sqrt{\sum_{i=2}^{n}\left(U_{i-1} - \overline{U}\right)^2}}\ . \qquad (11\text{-}3\text{-}1.3)$$

Gdy $\overline{U} = 0$ (jak to jest w MNK), wtedy estymator parametru ρ ma postać:

$$\hat{\rho} = \frac{\sum_{i=2}^{n} U_i U_{i-1}}{\sqrt{\sum_{i=2}^{n} U_i^2}\cdot\sqrt{\sum_{i=2}^{n} U_{i-1}^2}}\ . \qquad (11\text{-}3\text{-}1.4)$$

Statystyka testowa dla hipotezy zerowej (11-3-1.1) jest dana wzorem:

$$DW \equiv d = \frac{\sum_{i=2}^{n}(U_i - U_{i-1})^2}{\sum_{i=1}^{n} U_i^2} \ . \qquad (11\text{-}3\text{-}1.5)$$

Przy prawdziwości H_0 ma ona rozkład Durbina-Watsona [21]. Jej związek z estymatorem $\hat{\rho}$ jest następujący:

$$d \approx 2(1-\hat{\rho}) \ . \qquad (11\text{-}3\text{-}1.6)$$

Dla ustalonego poziomu istotności α i dla liczby czynników k (czyli liczby szacowanych parametrów k+1) oraz liczebności próby n, odczytujemy z tablic rozkładu Durbina-Watsona [23] dwie wartości krytyczne, (dolną) d_l i (górną) d_u.

Metoda weryfikacji zależy od miejsca w przedziale (0, 4), w który wpada obliczana na podstawie obserwacji wartość statystyki d.

a) W przypadku gdy w próbce $\hat{\rho} > 0$ ($d \in (0,2)$), wtedy hipotezą alternatywną jest hipoteza o dodatniej autokorelacji reszt:

$$H_1 : \rho > 0 \ . \qquad (11\text{-}3\text{-}1.7)$$

Wartości statystyki d porównujemy z wartościami krytycznymi d_l i d_u.. Jeśli $d < d_l$ to H_0 odrzucamy na korzyść H_1 i wnioskujemy, że autokorelacja jest dodatnia. Jeśli $d > d_u$ to nie ma podstaw do odrzucenia H_0 i wnioskujemy, że nie ma autokorelacji (dodatniej) reszt.

b) W przypadku gdy w próbce $\hat{\rho} < 0$ ($d \in (2,4)$),wtedy hipotezą alternatywną jest hipoteza o ujemnej autokorelacji reszt:

$$H_1 : \rho < 0 \ . \qquad (11\text{-}3\text{-}1.8)$$

W przypadku tym obliczamy wartość statystyki:

$$d' = 4 - d \ . \qquad (11\text{-}3\text{-}1.9)$$

Wartości statystyki d' również porównujemy z wartościami krytycznymi d_l i d_u.. Jeśli $d' < d_l$ to H_0 odrzucamy na korzyść H_1 i wnioskujemy, że autokorelacja jest ujemna. Jeśli $d' > d_u$ to nie ma podstaw do odrzucenia H_0 i wnioskujemy, że nie ma autokorelacji (ujemnej) reszt.

Uwaga. Różnica pomiędzy d_l a d_u. wynika z pośredniego wpływu macierzy planowania $\mathbf{X}$, (11-1.9).

Jeśli dla powyższego przypadku (a) zachodzi $d_u \geq d \geq d_l$ lub dla powyższego przypadku (b) zachodzi $d_u \geq d' \geq d_l$, wtedy powyższy test nie pozwala podjąć decyzji statystycznej dotyczącej występowania autokorelacji reszt.

Uwaga. Jeśli mamy dwie cechy mierzalne X i Y, które mają dwuwymiarowy rozkład normalny, wtedy w przypadku braku konkluzji po przeprowadzeniu testu Durbina-Watsona, można odwołać się do statystyki t [2]:

$$t = \frac{R}{\sqrt{1-R^2}} \cdot \sqrt{n-2} \ . \qquad (11\text{-}3\text{-}1.10)$$

Statystykę t można by (przy spełnieniu odpowiednich dla niej założeń) stosować również w pozostałych, konkluzywnych przypadkach. Przy prawdziwości hipotezy zerowej (11-3-1.1), statystyka t ma rozkład t-Studenta z n-2 stopniami swobody. Jeśli t_α jest kwantylem rzędu $(1-\alpha/2)$ rozkładu t-Studenta to, gdy w obserwacji (obs) $|t_{obs}| \geq t_\alpha$, wtedy na poziomie istotności α odrzucamy H_0 na korzyść H_1 i wnioskujemy, że istnieje autokorelacja. Jeśli $|t_{obs}| < t_\alpha$, to nie ma podstaw do odrzucenia H_0.

B. Rozdział 12. Graficzna analiza reszt.

Istnieją trzy proste techniki statystyczne, służące do ich interpretacji danych pomiarowych i wykrywania obserwacji nietypowych:

a. Wykorzystanie nierówności Czebyszewa dla dowolnej zmiennej losowej Z:

$$P\big(|Z - E(Z)| \geq m\sigma(Z)\big) \leq \frac{1}{m^2} \ . \qquad (12.1)$$

W oparciu o nierówność Czebyszewa wiadomo, że dla zmiennej o dowolnym rozkładzie aż 75 % (88,8(9)%) wszystkich obserwacji w populacji mieści się w granicach dwóch (trzech) odchyleń standardowych wokół wartości oczekiwanej tej zmiennej, tzn. w przedziale $\mu \pm 2\sigma$ ($\mu \pm 3\sigma$).

b. Wykorzystanie formuł empirycznych. W oparciu o własności rozkładu normalnego, można przy pewnym wyczuciu stosować empiryczną zasadę, że dla rozkładu z niewielką asymetrią i o kształcie dzwona, około 68% obserwacji w populacji mieści się w przedziale $\mu \pm \sigma$, 95% w przedziale $\mu \pm 2\sigma$, natomiast 99,9% w przedziale $\mu \pm 3\sigma$. Jednak znaczne odejście od wysmukłości rozkładu normalnego narusza skuteczność stosowania tej zasady i aby nazwać obserwację „nietypową” musi się ona pojawić z większym odchyleniem niż 3σ.

c. Wykorzystanie wykresu pudełkowego (z wąsami). Oznaczmy przez M medianę, przez Q_1 i Q_3, pierwszy i trzeci kwanty, przez $IQR = Q_3 - Q_1$. Oznaczmy wewnętrzne „płoty” jako $Q_1 - 1.5\ IQR$ i $Q_3 + 1.5\ IQR$, a $Q_1 - 3\ IQR$ i $Q_3 + 3\ IQR$, jako zewnętrzne „płoty”. Obserwację znajdującą się pomiędzy wewnętrznym i zewnętrznym płotem podejrzewa się o to, że jest nietypowa, natomiast tą na zewnątrz płotu „zewnętrznego” klasyfikuje się jako nietypową, czyli outsidera.

Posługiwanie się miarami tendencji centralnej (lokalizacji) i rozproszenia, oraz powyższymi trzema zasadami, daje najprostsze narzędzie identyfikacji outsiderów dla analizy rozkładu jednej zmiennej. Ich prostota połączona jest ze skutecznością posługiwania się nimi w analizie danych.

Często najbardziej bezpośrednią i odkrywczą drogą badania grupy reszt jest zrobienie serii wykresów reszt. Dwoma podstawowymi najbardziej użytecznymi rodzajami wykresów są:

- wykresy jednowymiarowe
- wykresy z wykorzystaniem wartości prognoz oraz „wielowymiarowe”

Pierwszy używany jest jedynie do rozpatrywania własności i relacji obserwowanych reszt między sobą, podczas gdy drugi ujmuje relacje reszt z innymi zmiennymi (takimi jak odpowiedź, czynniki i prognoza). W graficznych analizach, naruszenie założeń modelu (np. niezależności, normalności albo jednorodności wariancji) jest czasami bardziej widoczne na jednych typach wykresów niż na drugich.

Wykresy jednowymiarowe.

Wykresy jednowymiarowe są najprostszymi możliwymi wykresami. Stanowią jednak mocne wstępne narzędzie w analizie typu rozkładu badanej zmiennej (np. reszt). Trzy rodzaje jednowymiarowych wykresów reszt są szczególnie użyteczne: histogramy (szczególnie wersje ”stem and leaf”), wykresy pudełkowy z wąsami, tzw. (Box and whiskers plot) i wykresy prawdopodobieństwa z rozkładem normalnym (normal probability-probability plot).

Wszystkie poniższe wykresy oraz wartości numeryczne podstawowych charakterystyk opisowych można otrzymać stosując UNIVARIATE Procedure w SAS’ie, wywoływaną z poziomu Solutions->Analysis->Analyst->(File: Open by SAS name: „file name”)-> Descriptive->Distributions.

Wykresy z wykorzystaniem wartości prognoz oraz „wielowymiarowe”.

Kreślenie obserwowanych wartości odpowiedzi v.s. wartości czynników jest dobrym sposobem sprawdzania ważności założeń regresji. Posługując się pojedynczym czynnikiem można nanosić wartości odpowiedzi *Y* albo wartości reszt v.s. wartości czynnika *X*. Kiedy wykorzystywana jest większa ilość czynników, wówczas sytuacja jest bardziej złożona [1].

Przykład. Załóżmy, że temperatura powietrza ma wpływ na odpowiedź przyrządu. Zależnie od obserwowanych schematów zależności kombinacji dwóch czynników, poziomu temperatury i zanieczyszczeń, wykres odczytu przyrządu v.s. poziom zanieczyszczeń może błędnie sugerować np. niejednorodność wariancji.

Zwykle doradza się, aby wykreślić reszty nie tylko v.s. każdego czynnika, ale także v.s. spodziewanych prognoz, jak również wyrysowanie obserwowanych wartości odpowiedzi Y v.s. spodziewanych prognoz. Okazuje się, ze natura prognozowanych wartości $\hat{Y}$ pomaga wyjaśnić tendencje widoczne na wykresach, które bez wykorzystania wykresów z $\hat{Y}$ mogłyby być błędnie interpretowane.

Dla regresji z pojedynczym czynnikiem mamy:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{12.2}$$

a dla wielokrotnej regresji z k-czynnikami:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k \quad . \tag{12.3}$$

Z powodu zastosowania metody najmniejszych kwadratów do estymacji parametrów strukturalnych w (12.2) lub (12.3), prognoza $\hat{Y}$ reprezentuje liniową kombinację zmiennych X, która jest najbardziej skorelowana z Y [1].

Dla regresji z jednym czynnikiem kwadrat współczynnika korelacji Y i $\hat{Y}$, jest równy:

$$r^2\left(Y,\hat{Y}\right) = r^2\left(Y,X\right) \quad . \tag{12.4}$$

Powyższa zależność mówi, że siła liniowego związku pomiędzy Y i $\hat{Y}$ jest taka sama jak pomiędzy Y i X.
Natomiast dla regresji wielorakiej mamy:

$$r^2\left(Y,\hat{Y}\right) = R^2\left(Y \mid X_1, X_2, \ldots, X_k\right) \quad , \tag{12.5}$$

gdzie $R(Y \mid X_1, X_2, \ldots, X_k)$ jest współczynnikiem korelacji wielokrotnej pomiędzy odpowiedzią Y a naraz całą grupą zmiennych $X_1, X_2, \ldots, X_k$. Stąd zachodzi w ogólności związek:

$$r^2\left(Y,\hat{Y}\right) \geq r^2\left(Y, X_j\right) \tag{12.6}$$

Oznacza to, że współczynniki determinacji $r^2\left(Y, X_j\right)$ są z sobą wzajemnie powiązane. Aby to lepiej zobaczyć rozważmy najprostszy przypadek nieskorelowanych czynników $X_1, X_2, \ldots, X_k$. Otrzymujemy wtedy związek pomiędzy indywidualnymi współczynnikami determinacji:

$$r^2\left(Y,\hat{Y}\right) = R^2\left(Y \mid X_1, X_2, \ldots, X_k\right) = r^2\left(Y, X_1\right) + r^2\left(Y, X_2\right) + \ldots + r^2\left(Y, X_k\right) \tag{12.7}$$

Związek ten pomaga zrozumieć, dlaczego relacja pomiędzy Y i pojedynczym X_j musi być rozważana w świetle innych zmiennych X.
Zatem nawet jeśli wszystkie zmienne X są wzajemnie nieskorelowane, rozważania i tak nie mogą się ograniczyć do wykresów z pojedynczym czynnikiem, co jest spowodowane tym, że pojedyncze obserwacje mogą dawać wkład do tzw. outsiderów regresji wielorakiej, z którą mamy do czynienia gdy zmienne X_j są rozważane razem.

Właściwym sposobem rysowania obserwowanych wartości odpowiedzi Y jest użycie wykresu regresji częściowej dla każdego czynnika z osobna. W takim przypadku rysujemy wykres dla odpowiedzi dostrojonej do grupy k-1 czynników v.s. pozostały jeden czynnik, który został dostrojony do grupy tych samych k-1 czynników.

W szczególności przyjmijmy, że ogólny model, który nas interesuje ma postać:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + E_i . \qquad (12.8)$$

Aby utworzyć wykres regresji częściowej, dla k-tego czynnika, dopasowujemy wpierw dwa modele:

- pierwszy: $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_{k-1} X_{(k-1),i} + E_i \qquad (12.9)$$

oraz

- drugi: $$X_{ki} = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_{k-1} X_{(k-1),i} + E_i \quad . \qquad (12.10)$$

Dopiero teraz nanosimy względem siebie reszty z tych dwóch modeli na wykres, czyli rysujemy:

$$(Y_i - \hat{Y}_i) \text{ v.s. } \left(X_{ki} - \hat{X}_{ki}\right) \text{ dla } i = 1, 2, ..., n \ , \qquad (12.11)$$

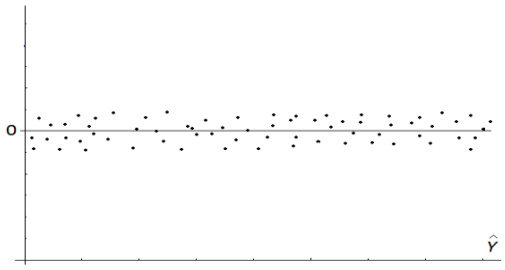
otrzymując k diagramów punktowych.

Najlepiej dopasowaną linię regresji powyższych par reszt (12.11), otrzymuje się metodą najmniejszych kwadratów. Posiada ona przesunięcie równe zero i estymator współczynnika kierunkowego $\hat{\beta}_k$ równy temu z modelu początkowego dla (12.8) [1]. Zwykła korelacja pomiędzy $(Y_i - \hat{Y}_i)$ i $\left(X_{ki} - \hat{X}_{ki}\right)$ jest więc wielokrotną korelacją częściową pomiędzy Y i X_k przy kontrolowanym wpływie zmiennych od X_1 do X_{k-1}.

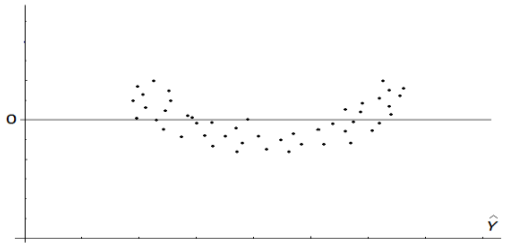
Z różnych typów dwuwymiarowych wykresów, bardziej przydatnymi wykresami dla sprawdzenia założeń wielokrotnej regresji są te, w których rysuje się reszty (szczególnie studentyzowane bądź scyzorykowe) v.s. wartości przewidywane albo wartości czynnika.

Kilka z możliwych schematów, które pojawiają się na wykresach reszt v.s. wartości przewidywane, zostało przedstawionych na poniższych wykresach. Oczywiście, różne typy odstępstw od założeń modelu dają różne, właściwe im, wykresy schematów dla reszt [1].

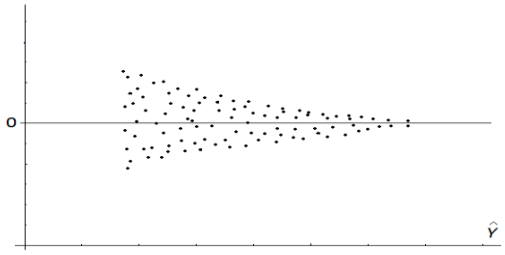
Rysunek 12.1. Reszty v.s. $\hat{Y}$ bądź v.s. czas. Typowe wykresy reszt jako „funkcje” wartości przewidywanej $\hat{Y}$ bądź „funkcje” czasu gromadzenia danych dla pewnych hipotetycznych danych [1].



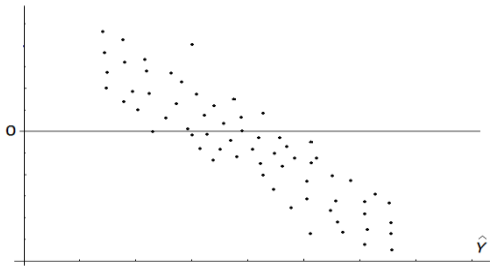
Wykres (a): Dane spełniające wszystkie założenia regresji wielorakiej.



Wykres (b): Odstępstwo danych od liniowości.



Wykres (c): Wariancja reszt maleje ze wzrostem $\hat{Y}$.



Wykres (d): Wykres reszt względem czasu.

Wykres (a) przedstawia schemat, dla którego wszystkie podstawowe założenia modelu wydają się być spełnione: horyzontalnej chmury punktów pomiarowych można się spodziewać, gdy występuje brak systematycznego trendu. Problem ten jest istotny np. w przypadku badania szeregów czasowych [24]. Widać też brak jakiegoś innego, *nielosowego schematu w układzie reszt*, o czym można by również wnioskować przeprowadzając np. nieparametryczny test serii (losowości) Walda-Wolfowitza runs test, weryfikujący hipotezę o losowym pochodzeniu obserwacji zmiennej objaśnianej w próbie [4].

Wykres (b) przedstawia schemat, w którym w danych pierwotnych występuje odstępstwo od liniowości, wskazując tym samym konieczność wprowadzenia regresji nieliniowej.

Wykres (c) reprezentuje schemat, w którym wariancja reszt wzrasta wraz z wzrostem $\hat{Y}$. Oczywiście można wyobrazić sobie schematy, w których zachowanie się wariancji reszt wraz ze wzrostem $\hat{Y}$ jest jeszcze bardziej skomplikowane. Niemniej i w takich przypadkach odpowiednia transformacja reszt często pomaga wyeliminować lub znacznie ograniczyć niejednorodność wariancji reszt (różne typy transformacji podaje [1], [24]). Istotną sprawą jest pobranie tylu ile to tylko możliwe replik dla jak największej liczby wariantów każdego czynnika X. Jeżeli w próbie jest za mała ilość replik, prowadzi to do trudności w rozróżnieniu problemu niejednorodności wariancji reszt od problemu doboru niewłaściwego modelu regresji.

Wykres (d) przedstawia zależność reszt (np. scyzorykowych) do czasu. Liniowy trend zależności reszt od czasu jest wyraźnie obecny. Jeśli istnieją zmienne nie uwzględnione w rozważanym modelu regresji (np. gdy pominięta została zmienna "czas", a dane są zebrane w następstwie czasowym), wtedy sytuacja taka może mieć znaczące konsekwencje przedstawione graficznie na wykresie (d). Jest ona świadectwem istnienia silnej korelacji pominiętej ukrytymi do tej pory zmiennymi, a resztami. Zatem wykres taki niesie niezmiernie istotną informacyjne, potrzebną, aby zbudować modele włączające zmienne wcześniej ukryte.

Istnieje jeszcze inna metoda w badaniu rozkładu reszt, a mianowicie: Ponieważ o resztach studentyzowanych i scyzorykowych zakłada się, że reprezentują próbkę pochodzącą z rozkładu, który jest w przybliżeniu standardowym rozkładem normalnym, zatem oczekujemy, że około 68% standaryzowanych reszt leży w przedziale (-1.00, 1.00), około 95% zawiera się w przedziale (-1.96, 1.96) i tak dalej. Jeśli jednak liczba stopni swobody n-k-1 (n – liczba punktów pomiarowych, k- liczba czynników) dla estymatora wariancji składnika losowego jest mała, wtedy 68-procentowe i 95–procentowe granice przedziałów muszą być wyznaczone z rozkładu t-Studenta, przy czym dla reszt studentyzowanych mamy n-k-1 stopniami swobody a dla scyzorykowych n-k-2 stopni swobody. Występowanie więc zgodnej z podanymi udziałami procentowymi liczby reszt wewnątrz i poza tymi przedziałami, jest wskazówką, że być może mamy raczej do czynienia z outsiderem (outsiderami), a nie z odstępstwem modelu od tego co się dzieje w populacji.

W Rozdziale 13 przedstawione zostaną na przykładach powyższe metody graficzne, ze wskazaniem odpowiednich schematów dla reszt.

B. Rozdział 13. Przykłady diagnostyki reszt.

Rozdział 13-1. Przykład 1. Skurczowe ciśnienie krwi.

Dokonano obserwacji wieku i pomiaru ciśnienia krwi w losowo pobranej próbce 30 osób [1]. Dane zamieszczono w poniższej tabelce. Poniżej przeprowadzimy analizę regresji (w jej podstawowym zakresie) dla zależności ciśnienia krwi od wieku z uwzględnieniem wszystkich pomiarów, następnie analizę reszt, i po wskazaniu ewentualnych outsiderów ponownie analizę regresji, tym razem bez outsiderów.

Tabela 13.1. Dane dla przykładu „skurczowe ciśnienie krwi"

Jednostka	Skurczowe ciśnienie krwi	Wiek
1	144	39
2	220	47
3	138	45
4	145	47
5	162	65
6	142	46
7	170	67
8	124	42
9	158	67
10	154	56
11	162	64
12	150	56
13	140	59
14	110	34
15	128	42
16	130	48
17	135	45
18	114	17
19	116	20
20	124	19
21	136	36
22	142	50
23	120	39
24	120	21
25	160	44
26	158	53
27	144	63
28	130	29
29	125	25
30	175	69

Liczba wszystkich obserwacji w próbie wynosi n = 30.

Tabela zawiera:

- zmienną objaśniającą *X*: wiek
- zmienną objaśnianą (odpowiedź) *Y*: skurczowe ciśnienie krwi

Rozważmy liniowy model regresji:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{ci} \tag{13.1}$$

gdzie $\hat{Y}_i$ oznacza teoretyczne średnie warunkowe dla wartości X_{ci} zmiennej $X_c = X - \overline{X}$, w klasycznym modelu regresji, gdzie wartości X_{ci} są określone przed pomiarem. Przejście od zmiennej X do zmiennej wycentrowanej X_c miało na celu eliminację efektu współliniowości zmiennej ciśnienia oraz jednostkowej zmiennej *I* stojącej przy estymatorze parametrze przesunięcia $\hat{\beta}_0$ [1]. Wycentrowanie w najniższym, liniowym stopniu wielomianu, jest równoważne ortogonalizacji układu zmiennych *I* oraz *X*, gdzie *I* jest zmienną jednostkowa stojącą przy $\hat{\beta}_0$ [1].

Ponieważ rozważany model jest modelem klasycznym regresji, zatem *X* nie jest zmienną losową.

Odpowiednie rachunki przeprowadzono w SAS'ie, posługując się procedurą PROC REG wywoływaną z poziomu Solutions->Analysis->Analyst->(File: Open by SAS name: „file name")->Statistics->Regression->Linear.

Raport SAS'a ma postać:

```
                                                        16:34 Wednesday, May 18, 2005   1

                                  The REG Procedure
                                    Model: MODEL1
                         Dependent Variable: Cisnienie Cisnienie

                                   Analysis of Variance

                                          Sum of           Mean
          Source                   DF        Squares         Square    F Value    Pr > F

          Model                     1     6394.02269     6394.12269      21.33    <.0001
          Error                    28     8393.44398      299.76586
          Corrected Total          29          14787


                      Root MSE             17.31375    R-Square     0.4324
                      Dependent Mean      142.53333    Adj R-Sq     0.4121
                      Coeff Var            12.14716


                                   Parameter Estimates

                             Parameter     Standard                        Standardized
Variable    Label      DF    Estimate      Error     t Value  Pr > |t|     Estimation

Intercept   Intercept   1   142.47084      3.16107    45.07    <.0001      0
Wiek_center Wiek_center 1     0.97087      0.21022     4.62    <.0001      0.65757
```

Correlation of Estimates

Variable	Label	Intercept	Wiek_centr
Intercept	Intercept	1.0000	0.0000
Wiek_center	Wiek_center	0.0000	1.0000

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	--Proportion of Variation- Intercept	Wiek_center
1	1.00000	1.00000	1.00000	0.00000
2	1.00000	1.00000	0.00000	1.00000

Wnioski z raportu:

- Średnie skurczowe ciśnienie krwi w grupie badanych osób wynosi 142,533.
- Liczba stopni swobody dla średniej sumy kwadratów SSR modelu, wynosi $df(SSR) = p - 1 = k = 1$, gdzie k jest liczbą parametrów kierunkowych modelu (u nas tylko β_1).
- Liczba stopni swobody dla sumy kwadratów reszt SSE wynosi $df(SSE) = n - p = 30 - 2 = 28$.
- Suma kwadratów dla modelu (SSR) jest równa 6394,023, suma kwadratów dla reszt (SSE) wynosi 8393,444. Średnia suma kwadratów modelu $MSR = SSR/\ df(SSR) = 6394{,}023$, natomiast średnia suma kwadratów błędów (wariancja resztowa) $MSE = SSE/df(SSE) = 299{,}766$.
- Postawiona hipoteza zerowa H_0: $\beta_1 = 0$ mówiąca o braku zależności korelacyjnej ciśnienia od wieku została odrzucona, tzn. model liniowy uznajemy za istotny statystycznie. Powodem jest istotna statystycznie, na każdym poziomie istotności $\alpha \geq p$, gdzie $p < 0{,}0001$ wartość $F_{obs} = MSR\ /\ MSE = 6394{,}023\ /\ 299{,}766 = 21{,}33$ statystyki testowej F-Snedecora. (Prawdopodobieństwo $p = P(F \geq F_{obs})$ jest empirycznym poziomem istotności.)
- Wartość współczynnika determinacji R^2 jest równa 0,4324 co oznacza, że około 43% zmienności średniej wartości ciśnienia jest w otrzymanym modelu wyjaśniona zmianami wieku. Ponieważ otrzymana wartość R^2 może być uznana co najwyżej za średnią, a R^2 jest miarą dopasowania modelu do danych empirycznych, zatem dopasowanie to nie jest za wysokie. Spróbujemy wskazać na możliwą przyczynę takiego stanu rzeczy (pomijając być może istotność rozszerzenia modelu liniowego do modelu z wyższym stopniem zależności ciśnienia od wieku, bądź do modelu z innymi zmiennymi objaśniającymi obok wieku).
- Otrzymane w pobranej próbce równanie regresji ma następującą postać:

$$\hat{Y} = 142{,}471 + 0{,}971\, X_c \qquad (13.2)$$

Otrzymany wynik oznacza, że zwiększenie wartości zmiennej „wiek” o jednostkę (1 rok) spowoduje wzrost ciśnienia średnio o 0,971 jednostki. Wartości oszacowań parametrów strukturalnych modelu w próbce są istotne statystycznie (p<0,0001) zarówno dla β_0 i β_1.

- Macierz korelacji (poza diagonalną są zera) wskazuje na brak współliniowości pomiędzy przesunięciem (Intercept) a zmienną objaśnianą Wiek_center, czego przyczyną było wprowadzenie do analizy

wycentrowanej zmiennej wieku X_c. Podobnie, przyglądając się wartościom własnym macierzy kowariancji w układzie tzw. składowych głównych (Rozdział 5-6), widać, że ze względu na ich równość możemy stwierdzić, że pomiędzy zmiennymi I i X_c nie ma współliniowości (tzn. nie są one skorelowane). Oznacza to, że otrzymane w próbce wartości $\hat{\beta}_0 = 142{,}471$ i $\hat{\beta}_1 = 0{,}971$ są, z punktu widzenia braku współliniowości, wartościami stabilnymi, tzn. pochodzą z estymatorów o małej wariancji i stąd otrzymany model (13.2) dobrze nadaje się do przewidywania ciśnienia.

Rozdział 13-1-1. Diagnostyka reszt dla modelu. Przykład „Skurczowe ciśnienie krwi".

Całość powyższej analizy zakłóca jednak nie za wysoka wartość R^2 = 0,4324, która może sygnalizować wystąpienie obserwacji nietypowej, czyli outsidera. Z tego powodu przyjrzyjmy się *diagnostyce badanego powyżej modelu*. Odpowiedni raport SAS'a ma postać.

	Wiek_centr	_RESID	_STUDENT	_COOKD	_H	_RSTUDENT
1	-6.068965517	7.4213381555	0.4372220635	0.0038664227	0.0388788427	0.4308167096
2	1.9310344828	75.654375344	4.4454938531	0.346165924	0.0338470006	8.0482592512
3	-0.068965517	-4.403883953	-0.258706298	0.0011540412	0.0333359541	-0.25434871
4	1.9310344828	0.654375344	0.038451465	0.0000258982	0.0338470006	0.0377595861
5	19.931034483	0.178709018	0.010829213	5.906726E-6	0.0915166287	0.0106340983
6	0.9310344828	-1.374754305	-0.0807645	0.0001128504	0.0334440601	-0.079318404
7	21.931034483	6.2369683151	0.3805266582	0.0083874681	0.1038210551	0.374639731
8	-3.068965517	-15.4912729	-0.910716093	0.0149433377	0.0347806431	-0.907852531
9	21.931034483	-5.763031685	-0.351611084	0.0071612003	0.1038210551	-0.346040022
10	10.931034483	0.916542181	0.054333667	0.000078901	0.0507410174	0.0533574147
11	18.931034483	1.1495793694	0.0694430607	0.0002263136	0.0858066672	0.0681976048
12	10.931034483	-3.083457819	-0.1827909	0.0008930044	0.0507410174	-0.179604294
13	13.931034483	-15.99606887	-0.953776356	0.0298986671	0.0616793773	-0.952185232
14	-11.06896552	-21.72431009	-1.288428969	0.0451649123	0.0516058652	-1.304472366
15	-3.068965517	-11.4912729	-0.675560183	0.0082226098	0.0347806431	-0.668860317
16	2.9310344828	-15.31649501	-0.90033108	0.014501894	0.0345447755	-0.897189705
17	-0.068965517	-7.403883953	-0.43494139	0.0032618862	0.0333359541	-0.428554121
18	-28.06896552	-1.219514113	-0.076399286	0.0005150639	0.1500117934	-0.07503043
19	-25.06896552	-2.132125167	-0.131758564	0.0012565406	0.1264545169	-0.12942447
20	-26.06896552	6.8387451844	0.4244528261	0.0139399457	0.1340121079	0.4181518276
21	-9.068965517	2.3339492098	0.1379882334	0.0004551906	0.0456305527	0.1355478512
22	4.9310344828	-5.25823571	-0.309454162	0.0018306197	0.036824829	-0.304398933
23	-6.068965517	-16.57866184	-0.976718294	0.0192949581	0.0388788427	-0.975886951
24	-24.06896552	0.8970044815	0.0552030211	0.0002061867	0.1191917604	0.0542112407
25	-1.068965517	18.566986398	1.0908235461	0.0206360281	0.0335226826	1.0946798819
26	7.9310344828	7.8291532353	0.4621089494	0.0047341378	0.0424561679	0.455522347
27	17.931034483	-15.87955028	-0.956413367	0.0399823826	0.0803915402	-0.954906626
28	-16.06896552	3.1300416699	0.1876358154	0.0013597428	0.0717037503	0.1843706633
29	-20.06896552	2.0135230757	0.1221189926	0.0007653684	0.0930890793	0.1199504177
30	23.931034483	9.2952276122	0.5714306995	0.0216971276	0.1173048196	0.5644346479

Powyższy wydruk przedstawia wartości: reszt zwykłych U_i (_RESID), reszt studentyzowanych R_i (_STUDENT), reszt scyzorykowych $R_{(-i)}$ (_RSTUDENT), odległość Cooka D_i (_COOKD) i współczynnika dźwignięcia h_i (_H) dla przykładu "skurczowe ciśnienie krwi".
Z powyższego raportu widać, że obserwacje 2 i 18 są podejrzana o to, że są outsiderami.

Rozważmy obserwację i=2. Wartości wszystkich typów reszt są duże, a szczególnie reszta scyzorykowa $R_{(-2)}$ = 8,0483 mocno odstaje od pozostałych. Posiada ona wyjątkowo dużą wartości (10-2-3.13) odległości Cook'a D_2 = 0,3462, o co najmniej rząd większej niż dla pozostałych. Ze względu na niską wartość dźwignięcia h_2 = 0,0338 wnioskujemy, że pomiar leży w niedużej odległości od średniej wieku. Właśnie ze względu na tą małą wartość dźwignięcia, obserwacja 2 nie jest szczególnie wpływowa (Rysunek 10-2-3.1 (a)). Zatem w równaniu regresji po usunięciu tego pomiaru, może ulec istotnej zmianie przesunięcie, natomiast wartość współczynnika kierunkowego nie powinna się wyraźnie zmienić.

Rozważmy obserwację i=18. Ma ona największą wartość dźwignięcia, h_{18} = 0,150012. Jest to wartość dźwignięcia, która jest większa niż podana przez Hoaglin i Welsch jako wartość, która podlegać powinna sprawdzeniu (wzór (10-2-1.12)): $h_{18} = 0{,}150012 > \frac{2(k+1)}{n} = 0{,}13(3)$. Zatem jest zalecenie, aby obserwacja została przebadania. Zakładając pomocniczo, że rozkład wieku jest normalny, każde pojedyncze dźwignięcie z F_i (wzór (10-2-1.13)) posiadającym rozkład F-Snedecora z k=1 i n-k-1=28 stopniami swobody.

Sprawdźmy czy wartość statystyki $F_i = \frac{[h_i - (1/n)]/k}{(1-h_i)/(n-k-1)}$ służąca do testowania hipotezy

$$H_0^i : E(h_i) = \frac{1}{n}, \tag{13.3}$$

wpadła w obszar krytyczny.

Ponieważ wartość krytyczna $F_{kr} = F_{k,(n-k-1),1-\alpha/n} = F_{1,28,1-0.05/30}$ = 12,1006, zatem ponieważ wartość statystyki w obserwacji wyznaczona zgodnie z (10-2-1.13) ma wartość $F_{18,\text{obs}}$ = 3,8436 , która nie należy do obszary krytycznego, zatem nie ma w zasadzie powodu, aby obserwację 18 uważać za outsidera. Potwierdza to również mała wartość odległości Cook'a (10-2-3.13), która wynosi D_{18}=0,00052, sugerując, że usunięcie obserwacji 18 ma niewielki wpływu na wielkość zmian współczynników regresji. Innymi słowy uznajemy, że nie jest ona wpływowa.
Istnieje jeszcze obserwacja i=20 z dość dużą wartością dźwignięcia h_{20}=0,134012. Jednak ze względu na niedużą wartość odległości Cook'a D_{20}=0,01394, również nie uznajemy jej za outsidera.

Zatem z danych usuwamy jedynie obserwację 2. Po jej usunięciu z tabeli danych (osoba z wiekiem 47 lat, której skurczowe ciśnienie krwi wynosiło podczas badania 220), otrzymujemy następujący raport SAS'a.

```
                                                        16:34 Wednesday, May 26, 2005   1

                                  The REG Procedure
                                    Model: MODEL1
                         Dependent Variable: Cisnienie Cisnienie

                                  Analysis of Variance

                                         Sum of           Mean
          Source                  DF        Squares         Square    F Value    Pr > F

          Model                    1     6110.10173     6110.10173      66.81    <.0001
          Error                   27     2469.34654       91.45728
          Corrected Total         28     8579.44828


                      Root MSE              9.56333    R-Square     0.7122
                      Dependent Mean      139.86207    Adj R-Sq     0.7015
                      Coeff Var             6.83769


                                  Parameter Estimates

                           Parameter     Standard                        Variance
Variable    Label      DF   Estimate      Error     t Value  Pr > |t|    Inflation

Intercept   Intercept   1   139.86207      1.77587    78.76    <.0001           0
Wiek_center Wiek_center 1     0.94932      0.11614     8.17    <.0001     1.00000


                                Correlation of Estimates

                  Variable       Label           Intercept       Wiek_center

                  Intercept      Intercept        1.0000            0.0000
                  Wiek_center    Wiek_center      0.0000            1.0000


                                Collinearity Diagnostics

                                         Condition    --Proportion of Variation-
                 Number     Eigenvalue       Index      Intercept     Wiek_center

                     1        1.00000      1.00000        1.00000        0.00000
                     2        1.00000      1.00000        0.00000        1.00000
```

Równanie regresji dla modelu ma następującą postać:

$$\hat{Y} = 139{,}862 + 0{,}949\, X_c \tag{13.4}$$

Powyższa część raportu SAS'a potwierdza częściowo nasze przypuszczenia. Wartość estymatora $\hat{\beta}_0$ uległa następującej zmianie: obecnie wynosi $\hat{\beta}_0 = 139{,}862$, w porównaniu z poprzednią 142,471. Wartość estymatora $\hat{\beta}_1$ uległa następującej zmianie: obecnie wynosi $\hat{\beta}_1 = 0{,}949$ w porównaniu z poprzednią 0,971. Okazuje się, że zmiany wartości tych estymatorów nie były bardzo duże, około 2%, a obserwację, chociaż jest być może outsiderem nie należy z tego punktu widzenia uznać za wpływową.

Natomiast poprawiło się znacznie dopasowanie modelu do danych empirycznych. Po usunięciu outsidera, współczynnik determinacji R^2 zwiększył się o 0,2798 i wynosi obecnie R^2 = 0,7122, co oznacza, że 71%

zmienności średniej wartości ciśnienia jest wyjaśniona przez otrzymany model (w porównaniu z 43% poprzednio). Ponieważ pozostała część analizy modelu (13.4) z usuniętą, uznaną za outsidera obserwacją, wygląda podobnie jak poprzednio, zatem pominiemy jej omówienie.
Poniżej przedstawimy jeszcze tylko analizę testu na normalność rozkładu reszt w próbce oraz analizę korelacji pomiędzy zmiennymi dla przypadku z domniemanym outsiderem i bez niego.

Analiza normalności rozkładu reszt

Hipoteza H_0: Teoretyczny rozkład reszt zgodny z rozkładem normalnym.
Metoda analizy: Test Kołmogorowa-Smirnowa (Rozdział 15).
Analizie poddano reszty zwykłe U_i.
Przypadek z „outsiderem" (n=30)

```
cisnienie center z outsiderem
                                    analiza reszt              21:20 Monday, December 9, 2013

                                 The UNIVARIATE Procedure
                             Fitted Distributions for _RESID

                            Parameters for Normal Distribution

                              Parameter   Symbol   Estimate

                              Mean        Mu        0
                              Std Dev     Sigma     17.01262

                        Goodness-of-Fit Tests for Normal Distribution

                  Test                  ---Statistic----   -----p Value-----

                  Kolmogorov-Smirnov    D     0.22573825   Pr > D     <0.010
```

Wartość statystyki D_n dla testu Kołmogorowa-Smirnowa jest równa D_{30} = 0.2257. Ponieważ empiryczny poziom istotności p < 0,010, zatem na każdym poziomie istotności $\alpha \geq p$, (np. dla $\alpha = 0{,}01$) wielkość D_n jest istotna statystycznie, co oznacza, że odrzucamy hipotezę o normalności rozkładu reszt.

Przypadek bez „outsidera" (n=29)

```
                              cisnienie center bez outsidera                                5
                               analiza reszt bez outsidera       12:11 Thursday, June 9, 2005

                                 The UNIVARIATE Procedure
                             Fitted Distributions for _RESID

                            Parameters for Normal Distribution
                              Parameter   Symbol   Estimate

                              Mean        Mu              0
                              Std Dev     Sigma    9.391004

                        Goodness-of-Fit Tests for Normal Distribution

                  Test                  ---Statistic----   -----p Value-----
                  Kolmogorov-Smirnov    D     0.11781868   Pr > D     >0.150
```

Wartość statystyki D_n dla testu Kołmogorowa-Smirnowa jest równa D_{29} = 0.0927. Ponieważ empiryczny poziom istotności p > 0,150, zatem na każdym poziomie istotności $\alpha < p$, (np. dla $\alpha = 0{,}05$) wielkość D_n nie jest istotna statystycznie, co oznacza, że nie mamy podstaw odrzucić hipotezy o normalności rozkładu reszt.

Podsumowanie testu o normalności rozkładu reszt.

Jest to kolejny (obok R^2) wyraźny sygnał, że być może należałoby pominąć obserwację 2 z analizy zależności ciśnienia od wieku. Sygnał ten jest tym bardziej istotny, że pozostawienie „outsidera" w analizie chwieje podstawami teoretycznymi modelu regresji klasycznej (w tym przypadku normalnością rozkładu reszt). Nie usunięcie „outsidera" oznaczałoby więc odejście od założeń modelu i wskazywało być może na ewentualną konieczność dokonania transformacji (np. logarytmicznej) wartości ciśnienia z nadzieją, że to pomoże zachować zgodność z założeniami modelu.

Analiza niezależności reszt dla przykładu (zgodnie z Rozdziałem 11-3-1).

Hipoteza H_0: brak korelacji pomiędzy składnikami losowymi

Metoda analizy: Test Durbina-Watson'a

Analizie poddano reszty zwykłe U_i.

Przypadek z „outsiderem" (n=30)

Raport SAS'a ma postać:

```
                 The REG Procedure
                   Model: MODEL1
Dependent Variable: Cisnienie Cisnienie

  Durbin-Watson D                 1.692
  Number of Observations             30
  1st Order Autocorrelation       0.146
```

Wartość statystyki Durbina-Watsona $d = 1{,}692 \in (0, 2)$ (Rozdział 11-3-1). Oznacza to, że hipoteza alternatywna ma postać $H_1 : \rho > 0$ (Rozdział 11-3-1). Z tablic dla k = 1, n = 30 i α = 0,05 odczytujemy wartości krytyczne dla testu. Wynoszą one d_l = 1,352 i d_u = 1,489. Wartość statystyki $d > d_u$, co świadczy o tym, że nie ma podstaw do odrzucenia hipotezy H_0, która mówi, że nie występuje autokorelacja reszt.

Przypadek bez „outsidera" (n=29)

```
                 The REG Procedure
                   Model: MODEL1
Dependent Variable: Cisnienie Cisnienie

  Durbin-Watson D                 1.331
  Number of Observations             29
  1st Order Autocorrelation       0.284
```

Wartość statystyki Durbina-Watsona $d = 1{,}331 \in (0,2)$. Oznacza to, że hipoteza alternatywna ma postać $H_1 : \rho > 0$ (Rozdział 11-3-1). Z tablic dla k = 1, n = 29 i α = 0,05 odczytujemy wartości krytyczne dla testu. Wynoszą one d_l = 1,341 i d_u = 1,483. Wartość statystyki d=1,331 < d_l=1,341 wpadła, zatem w obszar krytyczny testu i na poziomie istotności α = 0,05 należałoby, zatem odrzucamy hipotezę H_0 na korzyść H_1 i wnioskować, że autokorelacja jest dodatnia. Jednak wartość d=1,331 leży blisko granicy obszaru krytycznego i na poziomie istotności α = 0,01 nie wpadłaby zapewne do odpowiedniego obszaru krytycznego.

Dokonajmy więc dodatkowego sprawdzenia powyższego problemu, badając w pobranej próbce, dla przypadku bez outsidera tzn. dla 29 obserwacji, korelację reszt U_{i-1} z resztami U_i z wykorzystaniem procedury PROC REG w SAS'ie (punktów jest 28, dodatkowy brak jeszcze jednego punktu wynika z braku reszty u_0 w danych). Fragment raportu ma postać:

```
reszty przesuniete z cisnienie center bez outsider                    1
                                                 11:46 Friday, June 10, 2005

                          The REG Procedure
                            Model: MODEL1
                Dependent Variable: _RESID_1 _RESID_1

               Number of Observations Read          28
               Number of Observations Used          28


                         Analysis of Variance

                                  Sum of           Mean
Source                   DF      Squares         Square    F Value    Pr > F

Model                     1    204.48087      204.48087       2.53    0.1241
Error                    26   2105.10884       80.96572
Corrected Total          27   2309.58972


          Root MSE              8.99810    R-Square     0.0885
          Dependent Mean       -0.44356    Adj R-Sq     0.0535
          Coeff Var         -2028.61160

                         Parameter Estimates

                         Parameter       Standard                          Variance
Variable     Label        DF   Estimate          Error    t Value    Pr > |t|    Inflation

Intercept    Intercept     1   -0.33966        1.70174      -0.20      0.8433           0
_RESID       _RESID        1    0.29387        0.18492       1.59      0.1241     1.00000
```

Z raportu tego widać, że linia regresji pomiędzy resztami U_{i-1} a resztami U_i ma współczynnik kierunkowy o wartości = 0.29387 i wartość ta z powodu dużego prawdopodobieństwa p = 0,1241 nie jest istotna statystycznie, zarówno na poziomie istotności 0,01 jak i 0,05. Również wartość statystyki F dla testu o brak zależności korelacyjnej pomiędzy zmiennymi (tu resztami) nie jest istotna statystycznie (F = 2,53 przy p = 0,1241).

Uwaga. Odpowiednia wartości p dla współczynnika kierunkowego wynosi w tym przypadku tyle samo ile wartość p dla testu F. (Test dla współczynnika kierunkowego jako testem na dodanie wieku jako ostatniej

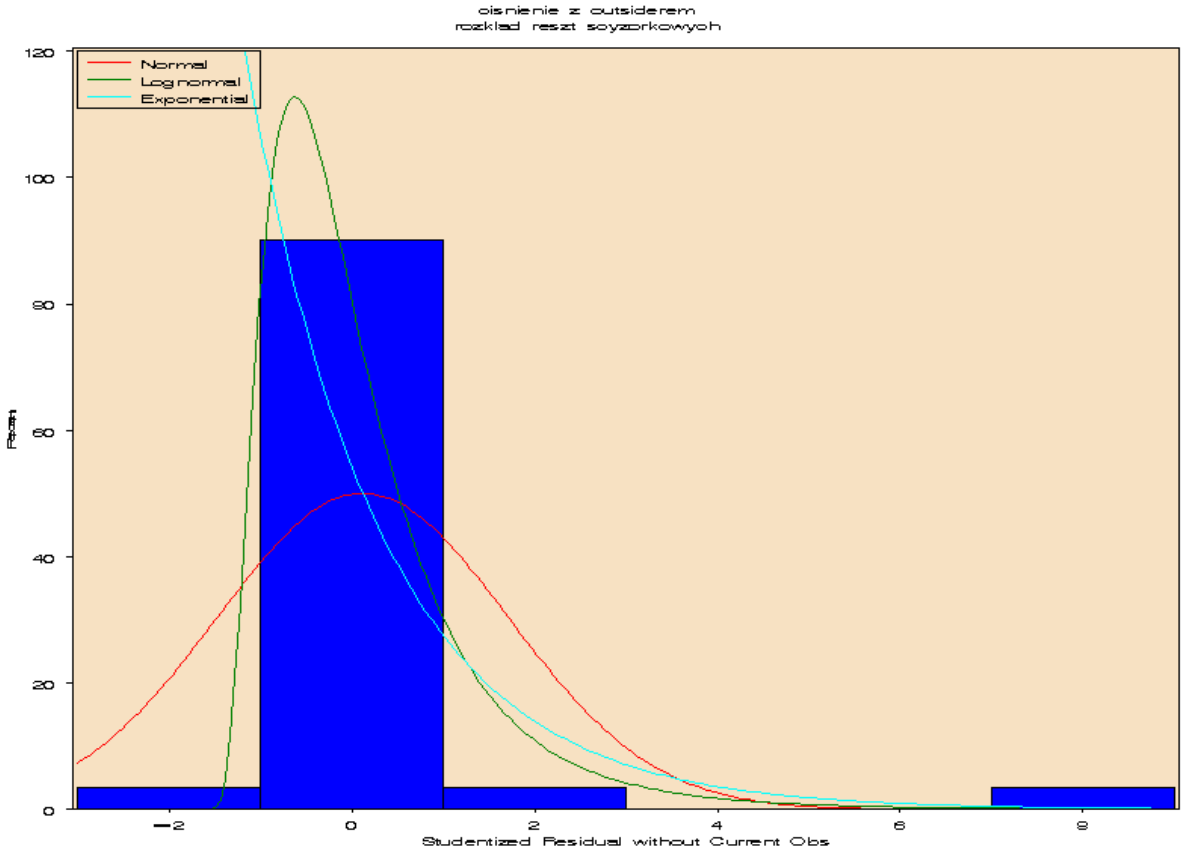
zmiennej, co pokrywa się tym razem z testem F, skąd t^2=F i wartości *p* są nieprzypadkowo takie same (Rozdział 5-1-2) [1]).

Uwaga (dotycząca testu jednorodności wariancji). Analizę jednorodności wariancji składnika losowego pomijamy bowiem jest ona w pobranej próbce naruszona w sposób jawny. Wynika to z tego, że istnieją warianty wieku (zmiennej *X*) z jedną repliką (tzn. z wariancją wewnątrzgrupową $\hat{s}_i^2(Y) > 0$) i bez replik (tzn. z wariancją wewnątrzgrupową $\hat{s}_i^2(Y) = 0$). Nie oznacza to, że przy odpowiedniej liczbie replik model naruszałby również to założenie, ale w celu sprawdzenia go, należałoby pobrać próbę z większą od zera liczbą replik dla każdego wariantu wieku (Rozdział 5-1-2).

Podsumowanie. Powyższa analiza reszt wskazuje, że model zależności ciśnienia osób od ich wieku, otrzymany z pominięciem obserwacji 2 z populacji, którą uznaliśmy za outsidera, jest do zaakceptowania. Model jest w zgodzie z podstawowymi założeniami modelu regresji klasycznej, przy czym z powodu braku wystarczającej liczby replik nie przebadano hipotezy o jednorodności wariancji dla różnych wariantów wieku. Pozostaje pytanie: czy nie usunęliśmy obserwacji, która była świadectwem istnienia innej niż rozważana liniowa zależności ciśnienia od jednego czynnika wieku?

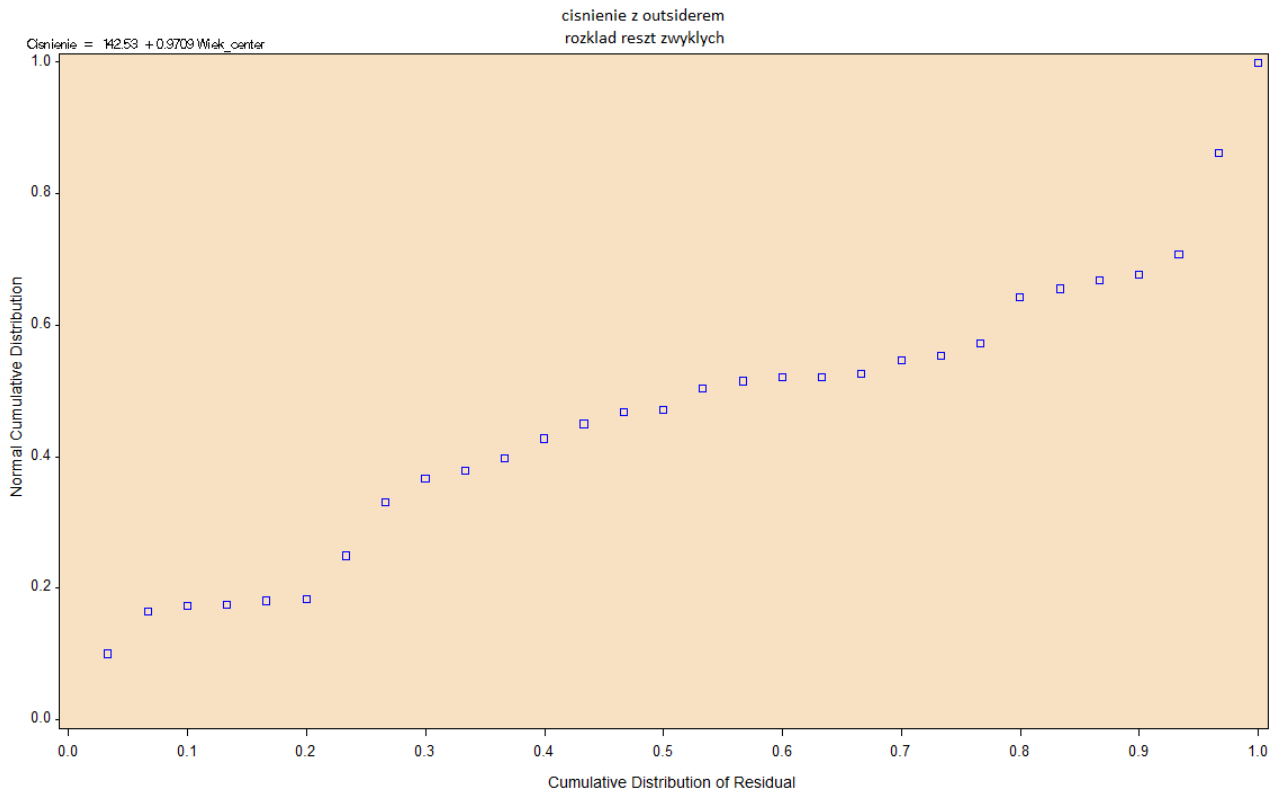
Rozdział 13-1-2. Graficzna analiza reszt dla Przykładu 1 „Skurczowe ciśnienie krwi”.

Dla przykładu skurczowe ciśnienie krwi (z outsiderem).



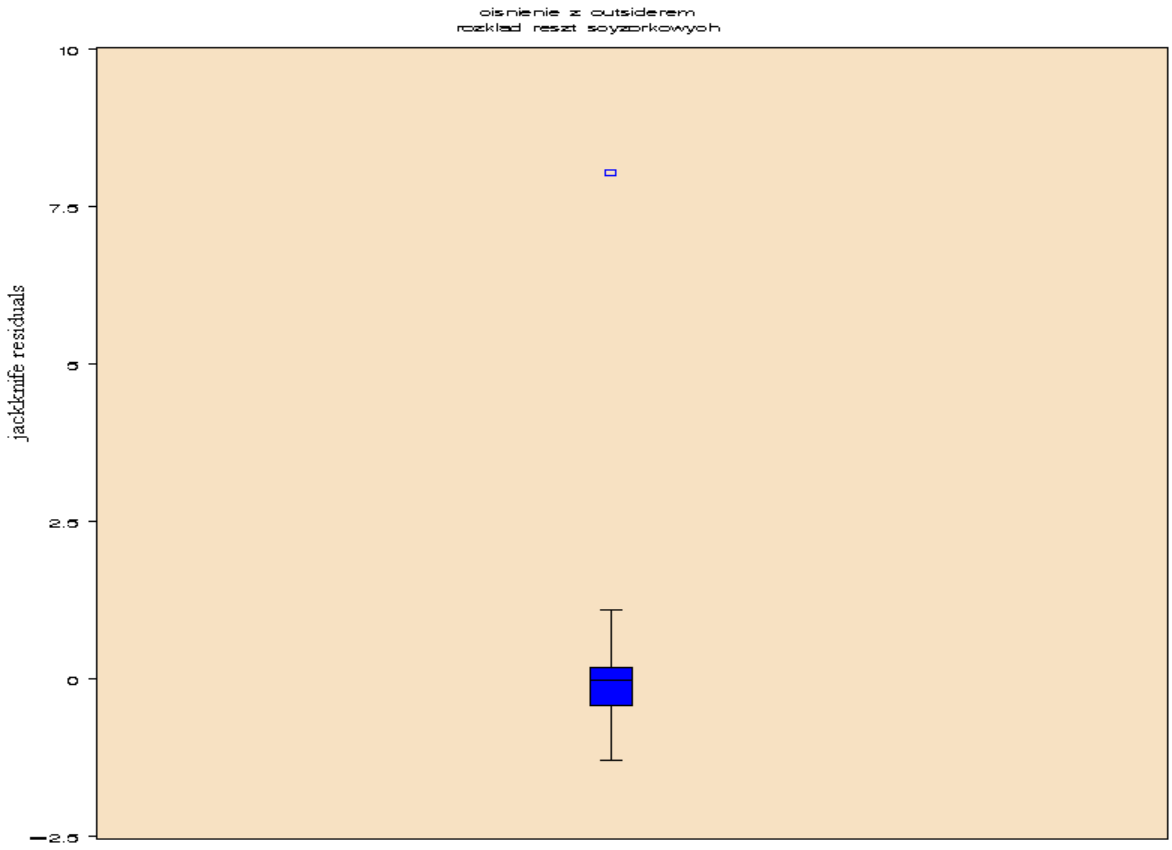
Rysunek 13-1-2.1. Histogram dla reszt scyzorykowych dla przykładu „skurczowe ciśnienie krwi”.

Histogram wskazuje na to, że występuje duża rozbieżność empirycznego rozkładu reszt scyzorykowych z rozkładem normalnym. Rozkład empiryczny jest znacznie bardziej wysmukły niż rozkład normalny. Również w Rozdziale 13-1-1 odrzuciliśmy testem Kołmogorowa-Smirnowa hipotezę zerową o zgodności rozkładu empirycznego reszt (poprzednio zwykłych) z rozkładem normalnym (dla analizy z outsiderem otrzymaliśmy $p < 0{,}01$). Dla porównania na rysunku przedstawiono teoretyczne rozkłady lognormalny (niezła zgodność, co potwierdziłaby dokładniejsza analiza testem Kołmogorowa-Smirnowa) i eksponencjalny.



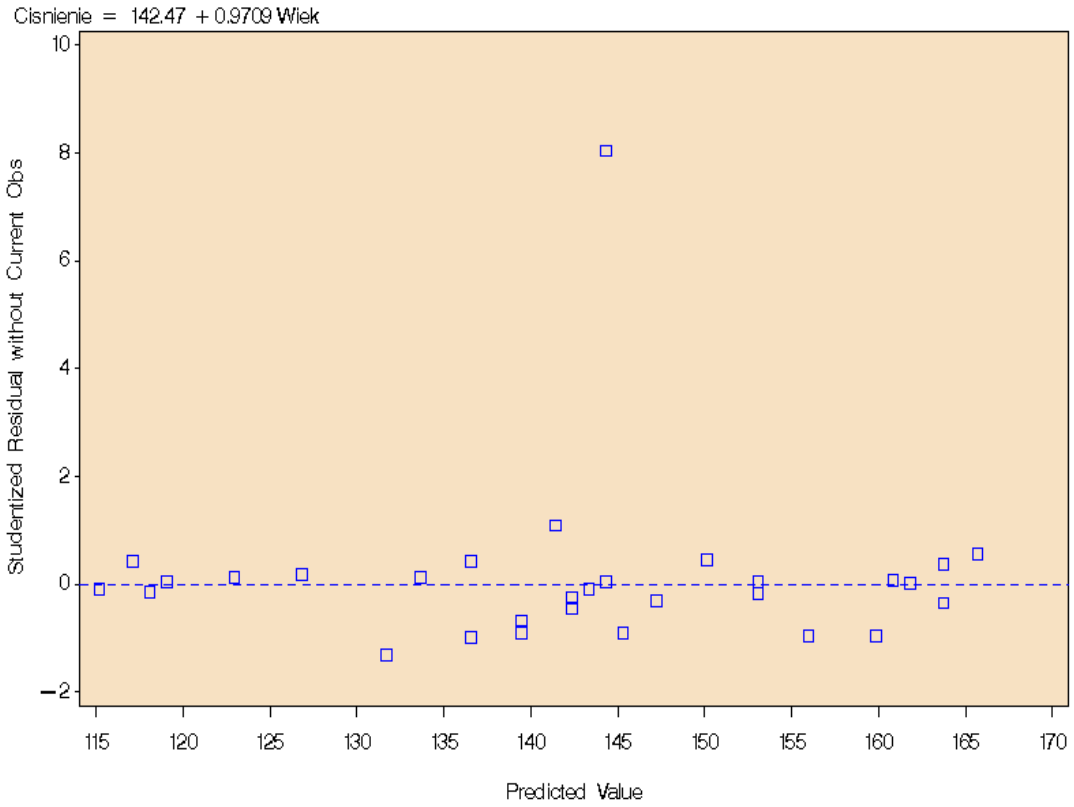
Rysunek 13-1-2.2 Normalny wykres prawdopodobieństwa (Normal probability-probability plot) dla przykładu „skurczowe ciśnienie krwi”

Z powyższego wykresu wynika, że rozkład reszt zwykłych jest niezgodny z teoretycznym rozkładem normalnym. Podobny wniosek otrzymaliśmy analizując powyższy histogram i test Kołmogorowa-Smirnowa w Rozdziale 13-1-1. Przebiegi dystrybuanty empirycznej po minięciu 50-tego percentyla wyprzedza dystrybuantę teoretycznego rozkładu normalnego (na przekątnej).



Rysunek 13-1-2.3 Box plot dla reszt scyzorykowych dla przykładu „skurczowe ciśnienie krwi”.

Pierwszy i trzeci kwantyl rozkładu reszt scyzorykowych są równe: Q1 = - 0.4285541 oraz Q_3 = 0.1844. Mediana M = - 0,0322, natomiast wartość maksymalna $r_{(-i=-2)max}$= 8.0483. Odpowiadała ona obserwacji 2 dla przykładu „skurczowe ciśnienie krwi” omawianego w Rozdziale 13-1-1. Odległość $Q_3 – Q_1$ = 0,61292, zatem różnica $r_{(-i)max}$ - Q_3 > 3*($Q_3 – Q_1$) skąd wynika, że istnieją podstawy, aby obserwację tą rozważyć jako outsidera. Wniosek ten jest zgodny z analizą przykładu z Rozdziału 13-1-1.



Rysunek 13-1-2.4 Wykres reszt scyzorykowych v.s. przewidywana wartość ciśnienia krwi dla Przykładu 1 „skurczowe ciśnienie krwi" (z outsiderem).

Z powyższego wykresu wynika, że tylko 1 wartość (tj. < 5%) z 30 reszt scyzorykowych przewyższa 1,96 co do wartości bezwzględnej. Oznacza to, że obserwacja ta może być outsiderem. Po jej usunięciu (i przeliczeniu modelu) otrzymalibyśmy schemat zgodny z Rysunkiem a (Rozdział 12).

Rozdział 13-2. Przykład 2 „FEV_1 (natężona jednosekundowa objętość)".

Przebadano grupę losowo wybranych 19 osób będących astmatykami, pod względem zależności ich FEV_1 od grupy czynników: wieku, wzrostu, wagi i płci. Dane zaczerpnięto z [1] (za cytowanym tam [25]).

Liczba wszystkich obserwacji w próbie wynosi $n = 19$.

Poniższa tabela zawiera zbiór danych potrzebny do przeprowadzenia analizy:

- zmienną zależną (objaśnianą): FEV_1 - [*l/sek*]
- zmienne niezależne: $X_1 \equiv$ height [*cm*], $X_2 \equiv$ weight [*kg*], $X_3 \equiv$ age [*lata*], $X_4 \equiv$ female – płeć (female dla mężczyzn przyjmuje wartość 0 a dla kobiet 1)
- dodatkowo zmienną zależną standaryzowaną: FEV_{1st}

	HEIGHT	WEIGHT	AGE	FEMALE	FEV_1	FEV_{1st}
1	175	78	24	0	4,7	1,05
2	172	67,6	36	0	4,3	0,59
3	171	98	28	1	3,5	-0,32
4	166	65,5	25	0	4	0,25
5	166	65	26	1	3,2	-0,67
6	176	65,5	22	0	4,7	1,05
7	185	85,5	27	0	4,3	0,59
8	171	76,3	27	0	4,7	1,05
9	185	79	36	0	5,2	1,62
10	182	88,2	24	0	4,2	0,47
11	180	70,5	26	0	3,5	-0,32
12	163	75	29	0	3,2	-0,67
13	180	68	33	1	2,6	-1,35
14	180	65	31	0	2	-2,04
15	180	70,4	30	0	4	0,25
16	168	63	22	0	3,9	0,13
17	168	91,2	27	0	3	-0,90
18	178	67	46	0	4,5	0,82
19	173	62	36	0	2,4	-1,58

Tabela 13-2.1. Tabela danych do przykładu „FEV_1 (natężona jednosekundowa objętość)"

W celu ograniczenia wpływu współliniowości pomiędzy czynnikami analizę przeprowadzono od razu dla zmiennych objaśniających wycentrowanych (indeks *c*).

Rozważmy liniowy model regresji:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \; X_{1i}^c + \hat{\beta}_2 \; X_{2i}^c + \hat{\beta}_3 \; X_{3i}^c + \hat{\beta}_4 \; X_{4i}^c \qquad (13\text{-}2.1)$$

gdzie $\hat{Y}_i$ oznacza teoretyczne średnie warunkowe dla wartości X_{ci} zmiennej $X_c = X - \overline{X}$.

Raport SAS'a:

```
                                  FEV1 na 1 sek            15:25 Friday, June 10, 2005   1
                             zmienne wycentrowane

                                   The REG Procedure
                                     Model: MODEL1
                               Dependent Variable: FEV1 FEV1
                           Number of Observations Read          19
                           Number of Observations Used          19
```

Test braku zależności korelacyjnej:

```
                                   Analysis of Variance
                                          Sum of           Mean
        Source                   DF       Squares         Square    F Value    Pr > F

        Model                     4       2.72245        0.68061       0.86    0.5116
        Error                    14      11.08281        0.79163
        Corrected Total          18      13.80526
```

Miary dopasowania modelu:

```
                     Root MSE              0.88974    R-Square     0.1972
                     Dependent Mean        3.78421    Adj R-Sq    -0.0322
                     Coeff Var            23.51179
```

Estymacja parametrów strukturalnych modelu:

```
                                   Parameter Estimates
                                    Parameter     Standard                      Standardized
   Variable       Label          DF    Estimate       Error   t Value  Pr > |t|    Estimate

   Intercept      Intercept       1     3.78421      0.20412    18.54    <.0001           0
   height_center  height_center   1     0.02111      0.03359     0.63    0.5400     0.16128
   weight_center  weight_center   1     0.01642      0.02114     0.78    0.4502     0.19632
   age_center     age_center      1    -0.00907      0.03723    -0.24    0.8110    -0.06231
   female_center  female_center   1    -0.81977      0.57571    -1.42    0.1764    -0.35068

2                                   The REG Procedure
                                     Model: MODEL1
                               Dependent Variable: FEV1 FEV1

                                 Correlation of Estimates
                                                  height_    weight_    age_    female_
     Variable         Label          Intercept    center     center    center   center

     Intercept        Intercept       1.0000     -0.0000     0.0000    0.0000   -0.0000
     height_center    height_center  -0.0000      1.0000    -0.2161   -0.2938    0.1890
     weight_center    weight_center   0.0000     -0.2161     1.0000    0.2495   -0.1741
     age_center       age_center      0.0000     -0.2938     0.2495    1.0000   -0.0674
     female_center    female_center  -0.0000      0.1890    -0.1741   -0.0674    1.0000
```

Analiza współliniowości z wykorzystaniem wartości własnych:

```
                                  Collinearity Diagnostics
                           Condition    -----Proportion of Variation------
  Number     Eigenvalue      Index        height_     weight_       age_       female_
                                          center      center       center      center

     1        1.33567       1.00000     0.19294       0.08004      0.26484     0.12847
     2        1.12677       1.08876     0.28905       0.48846   0.00050402     0.01002
     3        1.00000       1.15571           0             0            0           0
     4        0.98490       1.16454     0.00118       0.0044       0.27035     0.66261
     5        0.55266       1.55461     0.51683       0.420        0.46431     0.19891
```

Analiza raportu SAS'a:

Wnioski:

- Średnia FEV_1 (natężona jednosekundowa objętość) w grupie badanych ludzi wynosi 3,78421.
- Liczba stopni swobody dla średniej sumy kwadratów SSR modelu, wynosi *df*(SSR) = *p* – 1 = k = 4.
- Liczba stopni swobody dla sumy kwadratów reszt SSE wynosi *df*(SSE) = *n* – *p* = 19 – 5 = 14.
- Suma kwadratów dla modelu (SSR) jest równa 2,72245, suma kwadratów dla reszt (SSE) wynosi 11,08281. Średnia suma kwadratów modelu MSR = SSR/*df*(SSR) = 0,68061, natomiast średnia suma kwadratów błędów MSE = SSE/df(SSE) = 0,79163.
- Postawiona hipoteza zerowa H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ mówiąca o braku zależności korelacyjnej FEV_1 od wszystkich czynników nie została odrzucona, tzn. model liniowy uznajemy za nieistotny statystycznie. Powodem jest nieistotna statystycznie na każdym poziomie istotności $\alpha < p$ = 0,5116 wartość statystyki testowej F-Snedecora F_{obs} = MSR/MSE = 0,68061/0,79163 = 0,86. Nieistotność statystyczna estymatorów parametrów kierunkowych (punkt dalej) mogłaby wskazywać na to, aby z modelu usunąć zależność FEV_1 od wszystkich czynników, pozostawiając tylko wyraz wolny β_0. Niemniej, analizę przeprowadzimy dla oszacowanych wartości parametrów.
- Wartość współczynnika determinacji R^2 jest równa 0,1972 co oznacza, że około 20% zmienności średniej wartości FEV_1 jest wyjaśniona przez model. Ponieważ otrzymana wartość R^2 może być uznana za małą (a R^2 jest miarą dopasowania modelu do danych empirycznych), zatem dopasowanie to jest niskie. Spróbujemy wskazać na możliwą przyczynę takiego stanu rzeczy (pomijając być może istotność rozszerzenia modelu liniowego do modelu z wyższym stopniem zależności FEV_1 od zmiennych objaśniających).
- Otrzymane w pobranej próbce równanie regresji ma następującą postać:

$$\hat{Y} = 3{,}784 + 0{,}021X_{h\_c} + 0{,}016X_{w\_c} - 0{,}009X_{a\_c} - 0{,}82X_{f\_c} \qquad (13\text{-}2.2)$$

Wartości oszacowań parametrów strukturalnych modelu w próbce poza wyrazem wolnym $\hat{\beta}_0 = 3{,}784$ (dla Intercept empiryczny poziom istotności p < 0,0001) są nieistotne statystycznie (wszystkie p > 0,1), co jest w zgodzie z brakiem podstaw do odrzucenia hipotezy zerowej o braku zależności korelacyjnej.

- Macierz korelacji wskazuje na brak współliniowości (zera poza diagonalną) pomiędzy przesunięciem (Intercept) a zmiennymi objaśniającymi: Height_center, Weight_center, Age_center, Female_center. Przyczyną jest wprowadzenie do analizy wycentrowanych zmiennych height X_{h_c}, weight X_{w_c}, age X_{a_c}, female X_{f_c}. Otrzymane w próbce wartości $\hat{\beta}_0 = 3{,}784$, $\hat{\beta}_1 = 0{,}021$, $\hat{\beta}_2 = 0{,}016$, $\hat{\beta}_3 = 0{,}009$, $\hat{\beta}_4 = 0{,}82$ są z powodu braku współliniowości wartościami stabilnymi, tzn. pochodzą z estymatorów o małej wariancji.

Rozdział 13-2-1. Diagnostyka reszt dla modelu. Przykład 2 „FEV_1 (natężona jednosekundowa objętość)".

Całość powyższej analizy zakłóca niska wartość R^2 = 0,1972, która może być spowodowana wystąpieniem obserwacji nietypowej, czyli outsidera.

Z tego powodu przyjrzyjmy się diagnostyce badanego powyżej modelu. Poniższy wydruk przedstawia wartości: zmiennych centrowanych (height, weigh, age, female), reszt zwykłych U_i (_RESID), reszt studentyzowanych R_i (_STUDENT), reszt scyzorykowych $R_{(-i)}$ (_RSTUDENT), odległości Cooka D_i (_COOKD) i współczynnika dźwignięcia h_i (_H) dla przykładu FEV_1 (natężona jednosekundowa objętość).

osoba	height_ center	weight_ center	age_ center	female_ center	FEV1	_PRED	_RESID	_STUDENT	_COOKD	_H	_RSTUDENT
1	0.31579	4.27895	-5.21053	-0.15789	4.70000	4.03783	0.66217	0.78940	0.01559	0.11117	0.77820
2	-2.68421	-6.12105	6.78947	-0.15789	4.30000	3.69491	0.60509	0.74548	0.02241	0.16777	0.73306
3	-3.68421	24.27895	-1.21053	0.84211	3.50000	3.42569	0.07431	0.12975	0.00476	0.58566	0.12510
4	-8.68421	-8.22105	-4.21053	-0.15789	4.00000	3.63360	0.36640	0.46002	0.01049	0.19861	0.44667
5	-8.68421	-8.72105	-3.21053	0.84211	3.20000	2.79654	0.40346	0.61610	0.06423	0.45830	0.60191
6	1.31579	-8.22105	-7.21053	-0.15789	4.70000	3.87187	0.82813	1.05665	0.06449	0.22409	1.06142
7	10.31579	11.77895	-2.21053	-0.15789	4.30000	4.34480	-0.04480	-0.05839	0.00024	0.25641	-0.05627
8	-3.68421	2.57895	-2.21053	-0.15789	4.70000	3.89828	0.80172	0.94682	0.01867	0.09430	0.94307
9	10.31579	5.27895	6.78947	-0.15789	5.20000	4.15644	1.04356	1.34205	0.11139	0.23620	1.38541
10	7.31579	14.47895	-5.21053	-0.15789	4.20000	4.35303	-0.15303	-0.20002	0.00282	0.26067	-0.19302
11	5.31579	-3.22105	-3.21053	-0.15789	3.50000	4.00209	-0.50209	-0.60941	0.01235	0.14254	-0.59519
12	-11.68421	1.27895	-0.21053	-0.15789	3.20000	3.68995	-0.48995	-0.64901	0.03277	0.28007	-0.63502
13	5.31579	-5.72105	3.78947	0.84211	2.60000	3.07776	-0.47776	-0.73847	0.09721	0.47126	-0.72589
14	5.31579	-8.72105	1.78947	-0.15789	2.00000	3.86643	-1.86643	-2.26097	0.16531	0.13918	-2.73442
15	5.31579	-3.32105	0.78947	-0.15789	4.00000	3.96416	0.03584	0.04252	0.00004	0.10221	0.04097
16	-6.68421	-10.72105	-7.21053	-0.15789	3.90000	3.66198	0.23802	0.30788	0.00615	0.24502	0.29769
17	-6.68421	17.47895	-2.21053	-0.15789	3.00000	4.07958	-1.07958	-1.49738	0.23448	0.34336	-1.57449
18	3.31579	-6.72105	16.78947	-0.15789	4.50000	3.72097	0.77903	1.23565	0.30280	0.49789	1.26147
19	-1.68421	-11.72105	6.78947	-0.15789	2.40000	3.62408	-1.22408	-1.52422	0.10567	0.18529	-1.60826

Tabela 13-2-1.1 Wydruk diagnostyki dla przykładu „FEV_1 (natężona jednosekundowa objętość)" z wycentrowanymi czynnikami.

Z powyższego raportu widać, że obserwacje 3, 14 i 18 są podejrzana o to, że są outsiderami. Obserwacja 3 odstaje od pozostałych z wartością dźwignięcia. Natomiast obserwacja 14 odstaje wielkością reszty scyzorykowej. Obserwację 18 omówimy na końcu.

<u>Rozważmy wpierw obserwację i=14</u>. Wartości wszystkich typów reszt są stosunkowo małe, i jedynie reszta scyzorykowa $R_{(-14)}$ = -2,73442 mocno odstaje od pozostałych. Obserwacja 14 posiada dość dużą wartości (trzecią w kolejności) odległości Cook'a, D_2 = 0,16531, odstającą od wielu pozostałych. Ze względu na niską wartość dźwignięcia h_{14} = 0,13918 wnioskujemy, że pomiar leży w niedużej odległości w przestrzeni czynników od kompletu średnich czynników $\{\overline{X}_{h\_c}, \overline{X}_{w\_c}, \overline{X}_{a\_c}, \overline{X}_{f\_c}\} = \{0, 0, 0, 0\}$. Właśnie ze względu

na tą małą wartość dźwignięcia, obserwacja 14 nie jest szczególnie wpływowa (Rysunek 20-2-3.1 (a)). Zatem, po usunięciu tego pomiaru w równaniu regresji może ulec istotnej zmianie przesunięcie, natomiast wartość współczynnika kierunkowego nie powinna się wyraźnie zmienić.

Rozważmy obserwację i=3. Ma ona największą wartość dźwignięcia, h_3 = 0,58566. Jest to wartość dźwignięcia, która jest mniejsza niż podana przez Hoaglin i Welsch jako wartość, która podlegać powinna sprawdzeniu (10-2-1.12):

$$h_{18} = 0{,}58566 < \frac{2(k+1)}{n} = 0{,}631579 \cdot \qquad (13\text{-}2.3)$$

Zatem nie trzeba badać tej obserwacji. Nie ma w zasadzie powodu, aby obserwację 3 uważać za outsidera. Potwierdza to również mała wartość (10-2-3.13) odległości Cook'a, która wynosi D_3= 0,00476, sugerując, że usunięcie obserwacji 3 ma niewielki wpływu na wielkość zmian współczynników regresji. Innymi słowy uznajemy, że nie jest ona wpływowa.

Obserwacja 18. Istnieje jeszcze obserwacja i=18 z dość dużą wartością dźwignięcia h_{18}= 0,49789. Dla tej obserwacji odległość Cook'a ma największą wartość równą D_{18}= 0,30280. Na tej podstawie możemy podejrzewać, że obserwacja ta jest outsiderem (pomimo, że ciągle daleko jej do wartości 1; porównaj Rozdział 10-2-3).

Zatem z danych usuwamy obserwację 14 i 18. Po ich usunięciu z tabeli danych poprawiło się dopasowanie modelu do danych empirycznych. Otrzymujemy wtedy następujący raport SAS'a.

```
                                                    15:04 Sunday, June 12, 2005   1

                          The REG Procedure
                            Model: MODEL1
                     Dependent Variable: FEV1 FEV1

                        Analysis of Variance

                                 Sum of           Mean
Source                  DF      Squares         Square    F Value    Pr > F

Model                    4      3.67214        0.91803       1.73    0.2081
Error                   12      6.37022        0.53085
Corrected Total         16     10.04235


            Root MSE              0.72860    R-Square     0.3657
            Dependent Mean        3.84706    Adj R-Sq     0.1542
            Coeff Var            18.93903
```

Po usunięciu outsidera, współczynnik determinacji R^2 zwiększył się o 0,1685 i wynosi obecnie R^2 = 0,3657, co oznacza, że 37% zmienności średniej wartości FEV_1 jest wyjaśniona przez otrzymany model (w porównaniu z 20% poprzednio). Ponieważ pozostała część analizy modelu z usuniętymi, uznanymi za outsiderów obserwacjami wygląda podobnie jak poprzednio, zatem pominiemy jej omówienie.

Analiza normalności rozkładu reszt.

Poniżej omówimy jedynie przypadek z nie usuniętymi outsiderami.

Hipoteza H_0: Teoretyczny rozkład reszt zgodny z rozkładem normalnym.

Metoda analizy: Test Kołmogorowa-Smirnowa (Rozdział 15).

Analizie poddano reszty zwykłe U_i.

Przypadek z „outsiderami" (n = 19)

```
FEV1 na 1 sek      15:25 Friday, June 10, 2005   4
            rozklad reszt

      The UNIVARIATE Procedure
  Fitted Distributions for _RESID

  Parameters for Normal Distribution

    Parameter   Symbol   Estimate

    Mean        Mu              0
    Std Dev     Sigma    0.784673


 Goodness-of-Fit Tests for Normal Distribution

Test                  ---Statistic----   -----p Value-----
Kolmogorov-Smirnov    D     0.10881504   Pr > D     >0.150
```

Wartość statystyki D_n (4.14) dla testu Kołmogorowa-Smirnowa jest równa D_{19} = 0,1088. Ponieważ empiryczny poziom istotności p > 0,150, zatem na każdym poziomie istotności $\alpha < p$, (np. dla $\alpha = 0{,}05$ lub $0{,}01$) wielkość D_n nie jest istotna statystycznie, co oznacza, że nie mamy podstaw do odrzucenia hipotezy o normalności rozkładu reszt.

Analiza niezależności reszt (Rozdział 11-3-1)

Hipoteza H_0: brak korelacji pomiędzy składnikami losowymi

Metoda analizy: Test Durbina-Watson'a

Analizie poddano reszty zwykłe U_i.

<u>Przypadek z „outsiderem" (n = 19)</u>

Raport SAS'a ma postać:

```
                    FEV1 na 1 sek      15:25 Friday, June 10, 2005
Analiza niezależności reszt:

                       The REG Procedure
                         Model: MODEL1
                 Dependent Variable: FEV1 FEV1

                Durbin-Watson D                1.663
                Number of Observations            19
                1st Order Autocorrelation      0.081
```

Wartość statystyki *Durbina-Watsona* $d = 1{,}663 \in (0, 2)$. Oznacza to, że hipoteza alternatywna ma postać $H_1 : \rho > 0$ (Rozdział 11-3-1). Z tablic dla k = 4, n = 19 i α = 0,05 odczytujemy wartości krytyczne dla testu. Wynoszą one d_l = 0,859 i d_u = 1,848. Wartość statystyki $d_u > d > d_l$, co świadczy o tym, że test Durbina-Watsona nie pozwala nam na podjęcie konkretnej decyzji statystycznej odnośnie hipotezy $H_0 : \rho = 0$.

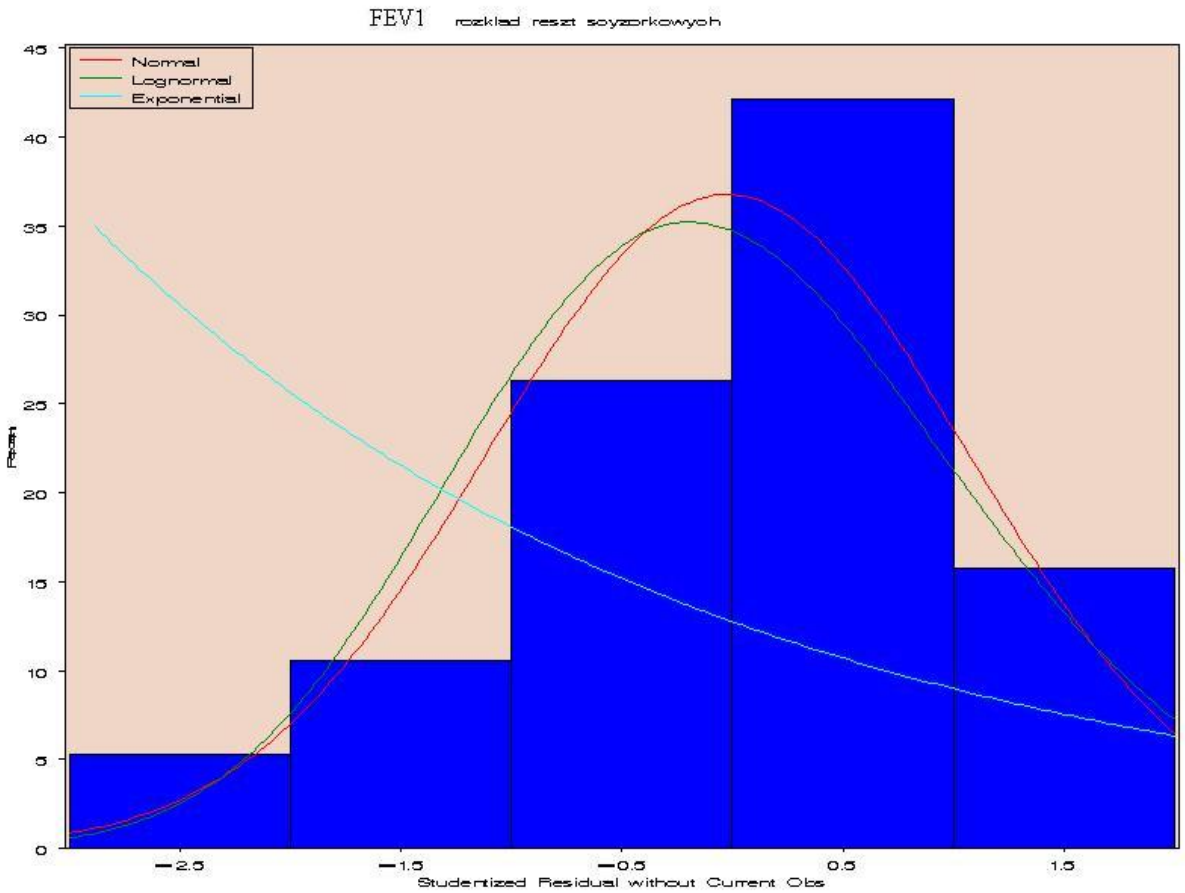
Analizę dokończymy więc w Excel'u korzystając z normalności rozkładu reszt w H_0. Skorzystamy ze statystyki t-Studenta $t = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \cdot \sqrt{n-2}$ (Rozdział 11-3-1, wzór (11-3-1.10)). W próbce otrzymaliśmy wartość $\hat{\rho} = 1 - d/2 = 0{,}169$, skąd statystyka t wyniosła w próbce t_{obs} = 0,705. Wartość krytyczna statystyki testowej t przy α=0,05 i liczbie stopni swobody n-2= 17 wynosi t(0,975; 17) = 2,11, co oznacza, że nie mamy podstaw do odrzucenia hipotezy zerowej o braku korelacji reszt.

Uwaga (dotycząca testu jednorodności wariancji): Analizę jednorodności wariancji składnika losowego pomijamy. Aby przeprowadzić test należałoby pobrać próbę z większą od zera liczbą replik dla każdego kompletu wartości czynników.

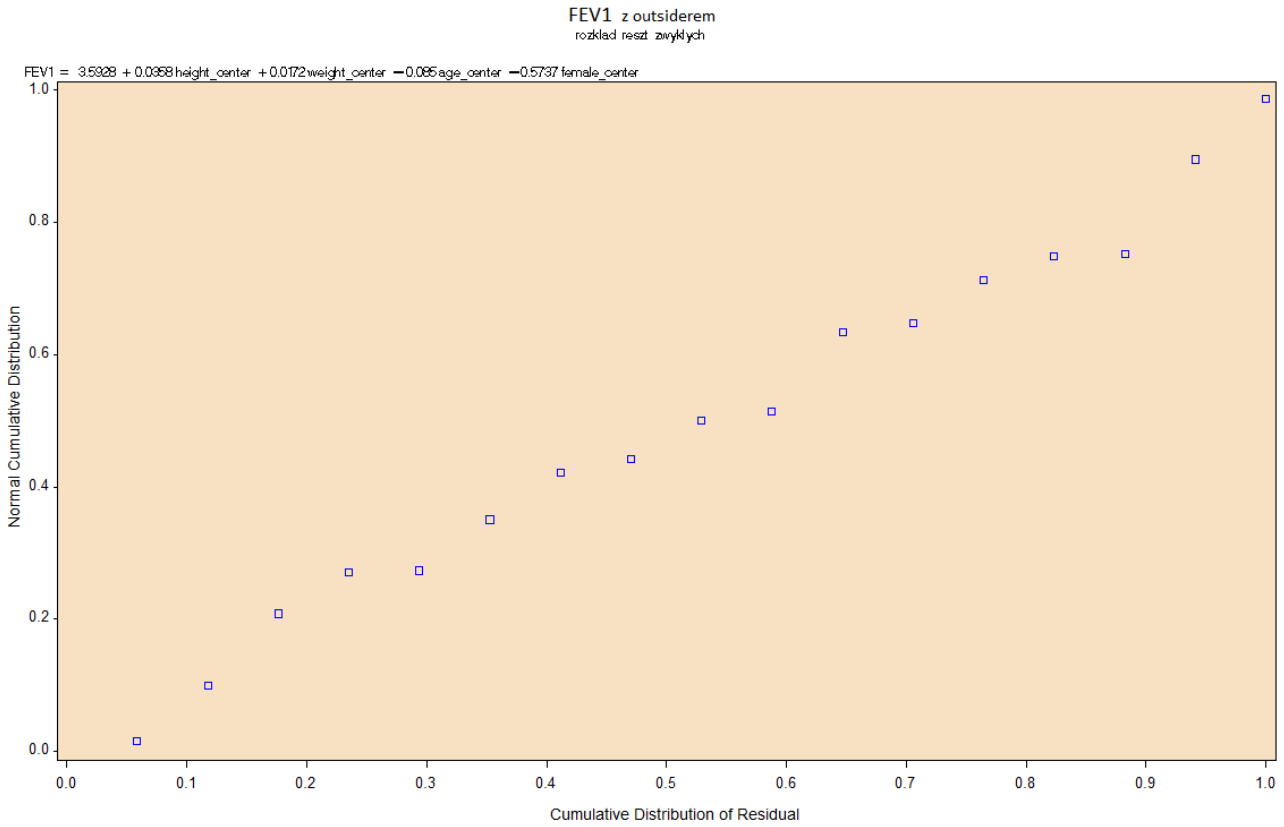
Podsumowanie. Powyższa analiza reszt wskazuje, że nie występuje istotna statystycznie zależność korelacyjna FEV_1 osób od zmiennych objaśniających. Analiza reszt wykazała, ze dwie obserwacje możemy uznać za outsiderów, obserwację 14 i 18. Badany model (13-2.1), chociaż jak się okazało nieistotny statystycznie (z ousiderami czy bez nich), jest w zgodzie z podstawowymi założeniami modelu regresji klasycznej. Usuwając outsiderów przyczyniliśmy się do lepszego dopasowania modelu do danych empirycznych.

Rozdział 13-2-2. Graficzna analiza reszt dla Przykładu 2 „FEV_1 (natężona jednosekundowa objętość)” (z Rozdziału 13.2-1).

Poniższy histogram wskazuje na to, że chociaż występuje zauważalna lewostronna skośność tego rozkładu, to można by uznać zgodność rozkładu reszt scyzorykowych z rozkładem normalnym. Dokładniejsza analiza w Rozdziale 13.2-1 oparta o test Kołmogorowa-Smirnowa potwierdziła taką decyzję. Dla porównania histogramu przedstawiono teoretyczne rozkłady: normalny, lognormalny i eksponencjalny.

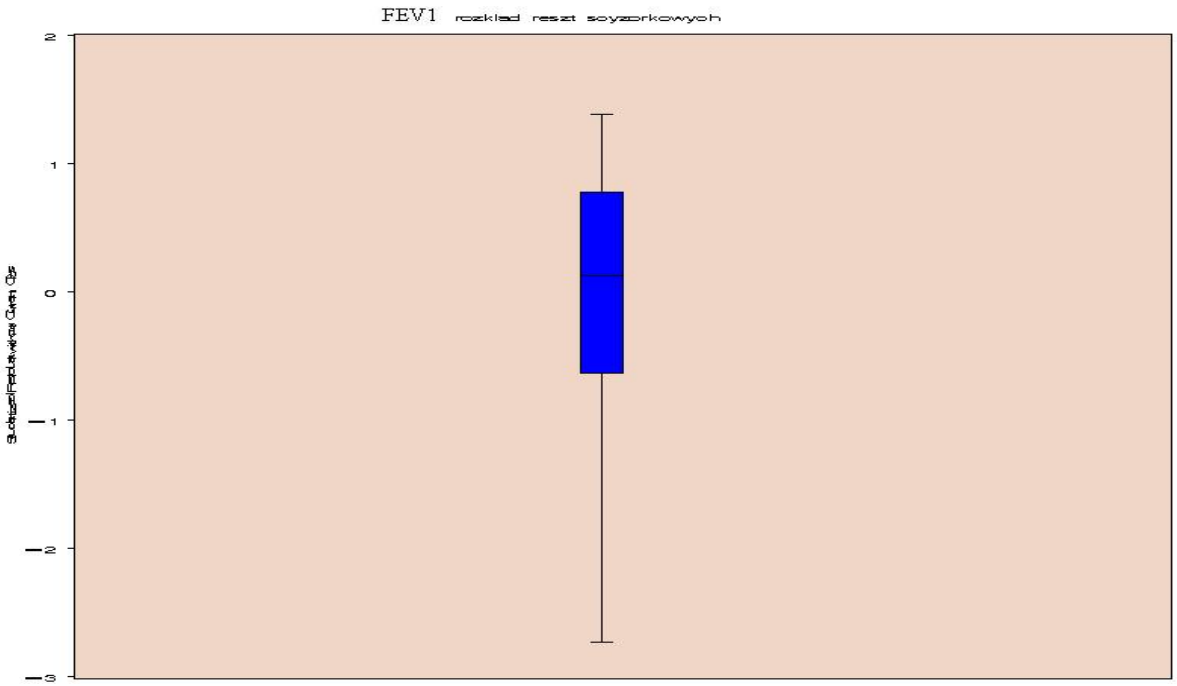


Rysunek 13-2-2.1. Histogram dla reszt scyzorykowych z Przykładu 2 „FEV_1 (natężona jednosekundowa objętość)”



Rysunek 13-2-2.2. Normalny wykres prawdopodobieństwa (Normal probability-probability plot) dla reszt zwykłych dla Przykładu 2 „FEV1 (natężona jednosekundowa objętość)".

Wykres dystrybuanty teoretycznego rozkładu normalnego leży na przekątnej. Z powyższego wykresu wynika, że w (szczególności) w centralnej części, empiryczny rozkład reszt zwykłych jest dość dobrze zgodny z teoretycznym rozkładem normalnym. W lewym ogonie rozkładu widać lekkie opóźnienie dystrybuanty rozkładu empirycznego w stosunku do rozkładu normalnego, co widać na wcześniejszym histogramie i co sugeruje lekką lewostronną skośność rozkładu empirycznego.

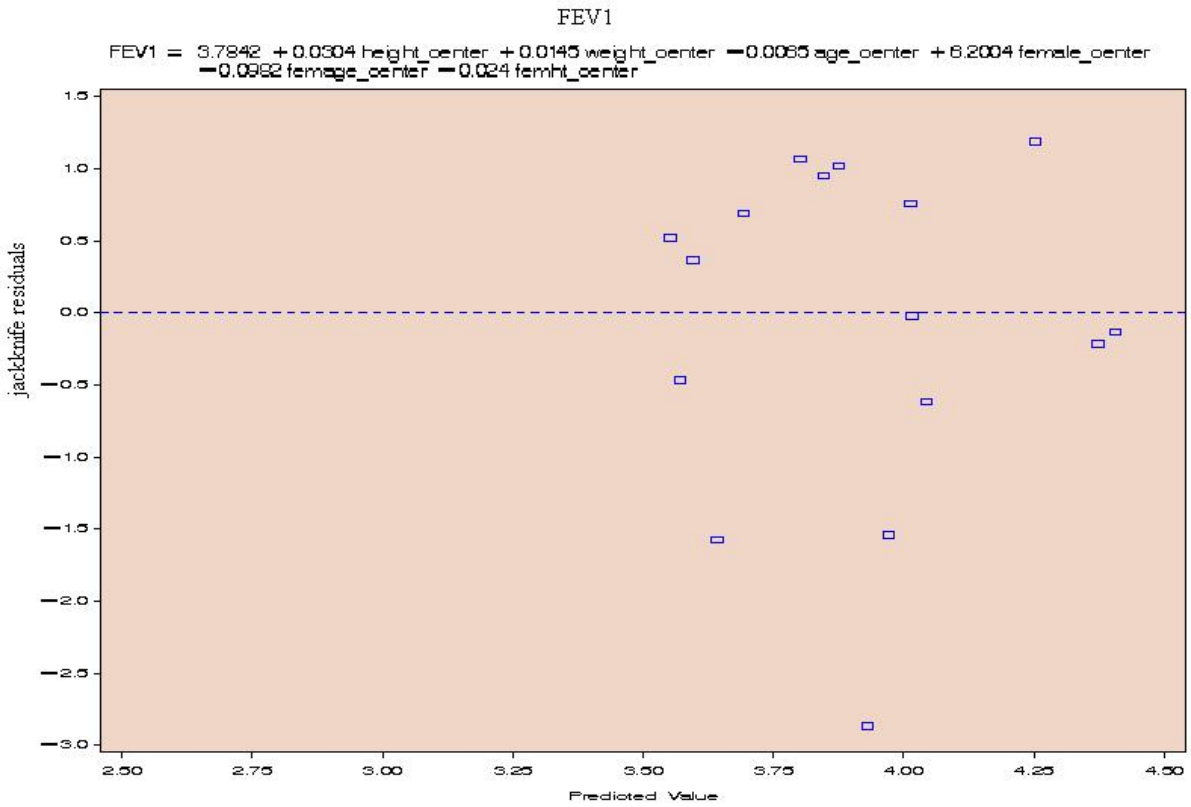


Rysunek 13-2-2.3 Box plot dla reszt scyzorykowych z Przykładu 2 „FEV_1 (natężona jednosekundowa objętość)”

Wykres pudełkowy z wąsami (Box plot) odzwierciedla graficznie następującą sytuację:

- medianę , M = 0.125 – kwartyl II (kreska wewnątrz pudełka) – oznacza to że 50% reszt scyzorykowych ma wartość mniejszą bądź równą wartości mediany a 50% scyzorykowych ma wartość większą bądź równą wartości mediany
- kwartyl I , Q_1= - 0.635 (dół pudełka) – oznacza że 25% reszt scyzorykowych znajduje się poniżej wartości kwartyla I
- kwartyl III , Q3 = 0.778 (góra pudełka) - oznacza że 75% reszt scyzorykowych znajduje się powyżej wartości kwartyla III
- średnią arytmetyczną $\bar{r}_{(-i)}$ = - 0.024 (słabo widoczna, oznaczona na wykresie symbolem +)
- wartość min i max (wartość min reszt scyzorykowych wynosi -2.734 a wartość max równa jest 1.385)

Na podstawie powyższego wykresu trudno jakąś obserwację uznać za zdecydowanego outsidera.



Rysunek 13-2-2.4 Wykres reszt scyzorykowych v.s. przewidywana wartość FEV_1 dla Przykładu 2 „FEV1 (natężona jednosekundowa objętość)".

Z powyższego wykresu wynika, że tylko 1 wartość (tj. < 5%) z 19 reszt scyzorykowych przewyższa 1,96 co do wartości bezwzględnej. Schemat zależności dla reszt jest trudniejszy do sprecyzowania, być może jest to schemat z Rysunku c z Rozdziału 12. Z całą jednak pewnością widać, że należałoby uzyskać większą liczbę replik, i to dopiero pozwoliłoby stwierdzić czy należy dokonać jakiejś transformacji zmiennej *Y*, i czy należałoby dodać do modelu jakieś dodatkowe czynniki (w tym i ich oddziaływania).

Na tym kończymy podstawowe rozważania (poparte przykładami) dotyczące sposobów badania na podstawie pomiarów w próbce spełnienia podstawowych założeń dla liniowych modeli regresji klasycznej.

B. Rozdział 14. Zakończenie.

Jakikolwiek ślad nielosowości obserwowanych reszt jest dowodem pewnego odstępstwa od założonego modelu. Podstawowe rodzaje odstępstw od założonego modelu mogą być wykryte poprzez:

1) Obecność licznych outsiderów. Obecność pojedynczych outsiderów jest na ogół łatwo wykrywana we wszystkich graficznych procedurach, jednakże liczni outsiderzy są czasami przyczyną trudności polegających na niemożności wykrycia outsiderów, powodując odstępstwa od założonego modelu.

2) Wyrysowanie wykresu zależności reszt od zmiennych objaśniających. Niewłaściwa postać zależności od zmiennych objaśniających, tzn. założenie np. modelu liniowego zamiast nieliniowego, może być zauważone po wyrysowaniu wykresu zależności reszt od zmiennych objaśniających.

3) Wyrysowanie reszt w zależności od wartości przewidywanych zmiennej objaśnianej, lub wyrysowanie skumulowanego rozkładu dla reszt wobec dystrybuanty teoretycznego rozkładu tych reszt (np. dystrybuanty rozkładu normalnego).

4) Wyrysowanie wykresów diagramów punktowych dla przesuniętych (o pewien „lag”) reszt. Można w ten sposób zauważyć korelacje pomiędzy obserwacjami. Procedura ta jest powszechna w szeregach czasowych.

5) Wykreślenie zależności reszt badanego modelu od niewprowadzonej do modelu zmiennej. Pominięcie jakiejś zmiennej objaśniającej może być zauważone na wykresie zależności reszt badanego modelu od niewprowadzonej do modelu zmiennej.

B. Rozdział 15. Uzupełnienie. Testy nieparametryczne.

Testy zgodności stanowią wraz z testami jednorodności [2] oraz z testami losowości i niezależności [2], główne działy wnioskowania nieparametrycznego.

Rozdział 15-1. Test zgodności Kołmogorowa – Smirnowa. Wprowadzenie.

Bardziej ilościowe kryteria dla oszacowania ważności założeń o typie rozkładu (np. że rozkład jest normalny), oparte są o standardowe testy statystyczne zgodności rozkładu empirycznego z rozkładem hipotetycznym np. takie jak test chi-kwadrat Pearsona i test Kołmogorowa(-Smirnowa). W testach zgodności porównuje się funkcję rozkładu prawdopodobieństwa badanej cechy w próbie z jej hipotetycznym rozkładem w populacji. Stawiane hipotezy zerowe mogą, w zależności od typu rozkładu i konkretnego zagadnienia, dotyczyć postaci funkcji rozkładu prawdopodobieństwa, funkcji rozkładu gęstości prawdopodobieństwa bądź postaci dystrybuanty.

Test Kołmogorowa (–Smirnowa) jest przeznaczony do testowania hipotez o zgodności rozkładu empirycznego i teoretycznego w przypadku zmiennych losowych typu ciągłego. Hipotezy można sformułować następująco.

Hipoteza zerowa ma postać:

$$\mathrm{H}_0 : F(y) = F_0(y) \tag{15-1.1}$$

i jest stawiana względem hipotezy alternatywnej:

$$\mathrm{H}_1 : F(y) \neq F_0(y) \quad . \tag{15-1.2}$$

W skrócie, zastosowanie testu do weryfikacji hipotezy zerowej (15-1.1) przebiega następująco. Statystyka testowa dla testu Kołmogorowa – Smirnowa jest określona wzorem (Rozdział 15.3):

$$D_n = \max_{1 \le i \le n} | F_n(y_i) - F_0(y_i) | \quad , \tag{15-1.3}$$

gdzie $F_n(y)$ jest dystrybuantą empiryczną (częstością skumulowaną) rozkładu empirycznego (Rozdział 15.2), która jest równa $F_n(y_i) = i/n$ dla i-tej najmniejszej obserwacji y_i w n-elementowej próbce, natomiast $F_0(y_i)$ jest dystrybuantą teoretyczną w populacji wyznaczoną dla tej samej wartości y_i. W teście tym dystrybuanta teoretyczna $F_0(y)$ musi być funkcją ciągłą i musi mieć w pełni określone parametry teoretycznego rozkładu (tzn. nieestymowane z próby). Zatem test Kołmogorowa – Smirnowa nadaje się do testowania zgodności rozkładu empirycznego z np. rozkładem normalnym (standaryzowanym) lub rozkładem t-Studenta z określoną liczbą stopni swobody.

Hipoteza dotycząca postaci dystrybuanty jest odrzucona, jeśli statystyka D_n testu jest większa niż wartość krytyczna uzyskana z tablic [26]. Test D_n jest bardziej czuły na błędy w ogonach rozkładu niż test χ^2-Pearsona [2]. Przykłady zastosowania testu Kołmogorowa – Smirnowa można znaleźć w [2].

Rozdział 15-2. Rozkład empiryczny [27].

Rozkład empiryczny to rozkład prawdopodobieństwa określony z próby dla oszacowania rozkładu prawdziwego w populacji. Załóżmy, że wyniki w próbce są realizacją próby prostej $Y_1, Y_2, ..., Y_n$ n-niezależnych zmiennych losowych z tym samym rozkładem określonym dystrybuantą F oraz niech $Y_{(1)} \le Y_{(2)} ... \le Y_{(n)}$ są odpowiednimi statystykami porządkowymi (powstałymi po uporządkowaniu próbki $y_1, y_2, ..., y_n$ w porządku rosnącym).

Rozkład empiryczny odpowiadający $Y_1, Y_2, ..., Y_n$ jest zdefiniowany jako rozkład dyskretny, który przypisuje każdej wartości y_k prawdopodobieństwo $1/n$. Empiryczna dystrybuanta F_n jest funkcją schodkową z krokami będącymi wielokrotnością $1/n$ w punktach $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$:

$$\begin{cases} 0, & y \le Y_{(1)}, \\ k/n, & Y_{(k)} < y \le Y_{(1)}, \quad 1 \le k \le n-1 \\ 1, & y > Y_{(n)} \end{cases} . \tag{15-2.4}$$

Dla ustalonych wartości $y_1, y_2, ..., y_n$ funkcja $F_n(y)$ ma wszystkie własności zwykłej dystrybuanty. Dla każdego ustalonego, rzeczywistego y, funkcja $F_n(y)$ jest zmienną losową jako funkcja $Y_1, Y_2, ..., Y_n$. Zatem, rozkład empiryczny odpowiadający próbie $Y_1, Y_2, ..., Y_n$ jest zadany przez rodzinę zmiennych losowych $F_n(y)$ zależną od rzeczywistego parametru x, z których każda określona jest rozkładem Bernoulliego [27]:

$$P\left(F_n(y) = \frac{k}{n} \right) = \binom{n}{k} \left(F(y)\right)^k \left(1 - F(y)\right)^{n-k} , \tag{15-2.5}$$

$$E(F_n(y)) = F(y) , \qquad \sigma^2(F_n(y)) = \frac{1}{n} F(y) \left(1 - F(y)\right) . \tag{15-2.6}$$

Dla każdego ustalonego, rzeczywistego x, każda ze zmiennych losowych $F_n(y)$ jest zbieżna stochastycznie do wartości $F(y)$, tzn.:

$$\forall \varepsilon > 0, \qquad \lim_{n \to \infty} P\left(|F_n(y) - F(y)| < \varepsilon \right) = 1. \tag{15-2.7}$$

Zatem $F_n(y)$ jest nieobciążonym i zgodnym estymatorem dystrybuanty $F(y)$.

Ponieważ empiryczna dystrybuanta $F_n(y)$ zbiega się do $F(y)$ w sposób jednostajny oraz z prawdopodobieństwem 1, tzn. zgodnie z twierdzeniem Gliwenki–Cantelli'ego [28], zachodzi:

$$P\left(\lim_{n \to \infty} D_n = 0 \right) = 1 , \tag{15-2.8}$$

gdzie (15-1.3):

$$D_n = \sup_y |F_n(y) - F(y)| . \tag{15-2.9}$$

Wielkość D_n jest miarą odległości $F_n(y)$ od $F(y)$. W roku 1933 A.N. Kołmogorow znalazł dystrybuantę graniczną, która dla ciągłej funkcji $F(y)$ spełnia związek:

$$P(\sqrt{n}\, D_n < \lambda) \underset{n\to\infty}{\to} \sum_{m=-\infty}^{m=+\infty} (-1)^m e^{-2m^2\lambda^2}, \quad \text{gdzie} \quad \lambda > 0. \tag{15-2.10}$$

Jeśli $F(y)$ nie jest znane, wtedy aby zweryfikować hipotezę H_0 (15-1.1), że $F(y)$ jest zadane ciągłą funkcją $F_0(y)$, stosuje się test (nazywany testem Kołmogorowa lub Kołmogorowa-Smirnowa), wykorzystujący statystykę typu D_n. Działanie tego testu zostało w skrócie omówiony powyżej, a poniżej znajduje się nieco bardziej szczegółowe omówienie jego teoretycznych podstaw.

Rozdział 15-3. Test zgodności Kołmogorowa [29]. **Rozwinięcie.**

Test Kołmogorowa przyczynił się do rozwoju statystyki matematycznej, będąc początkiem wielu badań nad nowymi metodami analizy statystycznej leżącej u podstaw nieparametrycznej analizy statystycznej [30]. Jest to test statystyczny wykorzystywany dla testowania prostej nieparametrycznej hipotezy H_0, (15-1.1), zgodnie z którą zmienne losowe niezależne $Y_1, Y_2, ..., Y_n$ posiadające taki sam rozkład (tworzące więc próbę prostą) mają dystrybuantę F, wobec dwustronnej hipotezy alternatywnej H_1 (15-1.2), którą można zapisać następująco:

$$H_1 : |E(F_n(y)) - F(y)| > 0, \tag{15-3.11}$$

gdzie $E(F_n(y))$ jest wartością oczekiwaną dystrybuanty empirycznej F_n (Rozdział 15-2). Zbiór krytyczny testu Kołmogorowa wyraża się nierównością:

$$D_n = \sup_{|y|<+\infty} |F_n(y) - F(y)| \geq \lambda_n \tag{15-3.12}$$

i opiera się on o następujące twierdzenie udowodnione przez A.N. Kołmogorowa w roku 1933 [31].

Twierdzenie. Jeśli hipoteza H_0 jest prawdziwa, wtedy rozkład statystyki D_n nie zależy od F. Ponadto, dla $n \to \infty$ statystyka $\sqrt{n}\, D_n$ ma asymptotycznie rozkład:

$$P(\sqrt{n}\, D_n < \lambda) \to Q(\lambda), \qquad \lambda > 0, \tag{15-3.13}$$

gdzie

$$Q(\lambda) = \sum_{m=-\infty}^{m=+\infty} (-1)^m e^{-2m^2\lambda^2}. \tag{15-3.14}$$

W roku 1948, N.V. Smirnow stablicował dystrybuantę Kołmogorowa $Q(\lambda)$.

W zgodzie z testem Kołmogorowa na poziomie istotności α, gdzie $0 < \alpha < 0{,}5$, hipotezę zerową H_0 odrzucamy, gdy $D_n \geq \lambda_n(\alpha)$, gdzie pierwiastek $\lambda_n(\alpha)$ równania:

$$P(D_n \geq \lambda_n) = \alpha \tag{15-3.15}$$

jest wartością krytyczną testu Kołmogorowa dla zadanego poziomu istotności α.

W celu określenia $\lambda_n(\alpha)$ wykorzystuje się przybliżenie prawa granicznego dla statystyki Kołmogorowa D_n oraz jej granicznej dystrybuanty. Można pokazać [32], że dla $n \to \infty$ oraz $0 < \lambda_0 < \lambda = O(n^{1/3})$, zachodzi:

$$P\left(\frac{1}{18n}(6n\,D_n+1)^2 \geq \lambda\right) = \left(1-Q(\sqrt{\frac{\lambda}{2}}\right)\left(1+O(\frac{1}{n})\right). \qquad (15\text{-}3.16)$$

Zastosowanie przybliżenia (15-3.16) w równaniu (15-3.15) daje następujące przybliżenie wartości krytycznej:

$$\lambda_n(\alpha) \approx \sqrt{\frac{z}{2n}} - \frac{1}{6n}\ , \qquad (15\text{-}3.17)$$

gdzie z jest pierwiastkiem równania:

$$1-Q\left(\sqrt{\frac{z}{2}}\right) = \alpha\,. \qquad (15\text{-}3.18)$$

W praktyce, w celu wyznaczenia wartości statystyki D_n wykorzystuje sie fakt, że:

$$D_n = \max(D_n^+, D_n^-)\,, \qquad (15\text{-}3.19)$$

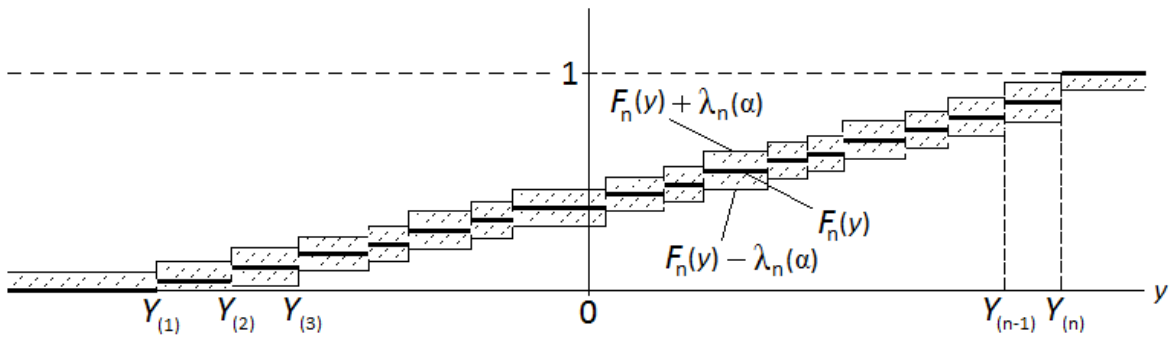
gdzie:

$$D_n^+ = \max_{1\leq m\leq n}\left(\frac{m}{n} - F(Y_{(m)})\right)\ , \qquad (15\text{-}3.20)$$

$$D_n^+ = \max_{1\leq m\leq n}\left(F(Y_{(m)}) - \frac{m-1}{n}\right), \qquad (15\text{-}3.21)$$

a $Y_{(1)} \leq Y_{(2)} \ldots \leq Y_{(n)}$ jest zbiorem statystyk porządkowych utworzonym z próby $Y_1, Y_2, \ldots, Y_n$.

Test Kołmogorowa ma interpretację geometryczną jak na poniższym rysunku:



Rysunek 15-3.1. Interpretacja testu Kołmogorowa. Wykres zbioru funkcji $F_n(y)$, $F_n(y) \pm \lambda_n(\alpha)$ na płaszczyźnie w układzie kartezjańskim. Obszar zakreskowany jest obszarem ufności dla dystrybuanty $F(y)$ wyznaczonym na poziomie ufności $1-\alpha$, co wynika z tego, że o ile hipoteza H_0, (15-1.1), jest prawdziwa, to według twierdzenia Kołmogorowa mamy:

$$P\left(F_i(y) - \lambda_i(\alpha) < F(y) < F_i(y) + \lambda_i(\alpha)\right) \approx 1-\alpha \quad , \quad \text{dla} \quad i = 1,2,\ldots,n\,. \qquad (15\text{-}3.22)$$

Jeśli wykres $F(y)$ nie opuszcza obszaru zakreskowanego, wtedy według testu Kołmogorowa nie ma podstaw do odrzucenia H_0 na poziomie istotności α. W przeciwnym przypadku H_0 jest odrzucona na rzecz hipotezy alternatywnej.

Uwaga. Tak test oparty o statystykę $D_n = \sup_Y |F_n(y) - F(y)|$ jak i o statystykę $\tilde{D}_n = \sup_Y (F_n(y) - F(y))$ bywa nazywany testem Kołmogorowa–Smirnowa [33].

Również pokrewny test dla problemu dwóch prób, oparty o statystyki $D_{m,n} = \sup_Y |F_m(y) - G_n(y)|$ oraz $\tilde{D}_{m,n} = \sup_Y (F_n(y) - G_n(y))$, gdzie $G_n(y)$ jest empiryczną dystrybuantą dla m- wymiarowej próby dla populacji z dystrybuantą G, jest nazywany testem Kołmogorowa–Smirnowa [33]. Ten ostatni jest przykładem nieparametrycznego testu *jednorodności* stosowanego do weryfikacji hipotezy zerowej o zgodności ze sobą dwóch rozkładów empirycznych, gdzie $\mathsf{H}_0 : F(y) = G(y)$, a hipoteza alternatywna $\mathsf{H}_1 : F(y) \neq G(y)$.

Uwaga. Podstawowa baza programowa SAS'a dostarcza kilku testów normalności w ramach procedury UNIVARIATE. W zależności od wymiaru próby, PROC UNIVARIATE wykonuje np. testy Kołmogorowa-Smirnowa, Shapiro-Wilk'a, Anderson'a-Darling'a oraz Cramér-von Mises'a.

W celu dokonania weryfikacji hipotezy mówiącej o tym, że dwie (lub więcej) grupy obserwacji są generowane z identycznych rozkładów, można wykorzystać w SAS'ie procedurę NPAR1WAY, która umożliwia wyznaczenie statystyk dla funkcji rozkładu empirycznego (EDF). Procedura ta wylicza statystykę testową Kołmogorowa-Smirnowa oraz Cramér-von Mises'a. W przypadku, gdy dane są sklasyfikowane w dwóch próbach, dostępny jest również test Kuiper'a. Dokładne wartości empirycznego poziomu istotności *p* są dostępne dla testu Kołmogorowa-Smirnowa dla dwóch prób. Aby uzyskać dostęp do wspomnianych testów, należy zastosować opcję EDF w poleceniach procedury PROC NPAR1WAY [34].

C. Rozdział 16. Analiza wariancji.

Analiza wariancji (ANOVA - analysis of variance) jest metodą statystyczną wykorzystywaną do porównywania wartości średnich zmiennej objaśnianej (odpowiedzi) w kilku populacjach. Jest to technika badania wyników (obserwacji), które zależą od jednego lub kilku czynników działających równocześnie. Czynnikiem w ANOVA nazywamy podstawową zmienną objaśniającą, która przyjmuje różne poziomy odpowiadające poszczególnym kategoriom (wariantom) czynnika. ANOVA pozwala sprawdzić, czy analizowane czynniki wywierają wpływ na obserwowaną zmienną objaśnianą.

Zmienna objaśniana musi być zmienną mierzalną a czynniki mogą mieć charakter zarówno jakościowy jak i ilościowy. Czynnikiem jakościowym może być na przykład stosowany lek, metoda leczenia, płeć, liczebność członków rodziny. Każdy czynnik ma kilka poziomów lub wariantów, którymi, dla wspomnianych czynników, mogą być: różne dawki leku, określone metody leczenia, płeć męska i żeńska, liczba członków rodziny.

Jeśli dany czynnik wpływa na zmienną objaśnianą, to średnie wartości tej zmiennej powinny różnić się istotnie w zależności od tego, jaki jest wariant czynnika. Istotą ANOVA jest pomiar wpływu jakościowej zmiennej objaśniającej (zmiennych objaśniających) na skalę zmienności zmiennej objaśnianej Y. Analiza wariancji jest zatem metodą równoczesnego badania istotności różnic między wieloma średnimi z prób pochodzących z wielu populacji, które łącznie tworzą (jednorodną bądź niejednorodną) populację generalną. Pomiar wpływu zmiennej objaśniającej obejmuje wyznaczenie miar statystycznych opisujących rozkłady warunkowe zmiennej objaśnianej oraz weryfikację hipotez statystycznych.

Wpływ wyróżnionych czynników na zmienną objaśnianą może być rozpatrywany oddzielnie dla pojedynczego czynnika i wtedy mamy do czynienia z modelem jednoczynnikowym, dla którego przeprowadza się tzw. jednokierunkową analizę wariancji (One-way ANOVA). Można też badać dwa lub więcej czynników razem, oceniając, oprócz ich indywidualnego wpływu, ich łączny wpływ na zmienną objaśnianą. Mamy wówczas do czynienia z dwuczynnikowym lub wieloczynnikowym modelem i tzw. dwukierunkową (Two-way ANOVA) lub wielokierunkową analizą wariancji.

Rozdział 16-1. Jednoczynnikowa analiza wariancji (ANOVA- tablica analizy wariancji).

Głównym problemem jednoczynnikowej analizy wariancji jest „zbadanie”, czy wartości oczekiwane zmiennej objaśnianej (odpowiedzi) w populacjach są sobie równe. Aby przeprowadzenie tej analizy było możliwe, trzeba aby były spełnione cztery podstawowe założenia:

- Próbka musi być wyselekcjonowana w sposób losowy z każdej z *k* populacji lub grupy.
- Wartość zmiennej objaśnianej musi być określona dla każdej jednostki w pobranej próbce.
- Zmienna objaśniana ma rozkład normalny w każdej populacji.
- Wariancja zmiennej objaśnianej jest taka sama w każdej populacji.

Ogólna konfiguracja danych w jednoczynnikowej ANOVA jest przedstawiona w poniższej Tabeli.

Tabela 16-1.1. Ogólny układ danych występujących w jednoczynnikowej ANOVA [1].

Numer populacji	Wielkość (pod)próby	Obserwacje	Suma	Średnia z próby
1	n_1	$Y_{11}, Y_{12}, ..., Y_{1n_1}$	$Y_{1\bullet}$	$\overline{Y}_{1\bullet} = Y_{1\bullet} / n_1$
2	n_2	$Y_{21}, Y_{22}, ..., Y_{2n_2}$	$Y_{2\bullet}$	$\overline{Y}_{2\bullet} = Y_{2\bullet} / n_2$
3 ⋮	n_3 ⋮	$Y_{31}, Y_{32}, ..., Y_{3n_3}$ ⋮	$Y_{3\bullet}$ ⋮	$\overline{Y}_{3\bullet} = Y_{3\bullet} / n_3$ ⋮
K	n_k	$Y_{k1}, Y_{k2}, ..., Y_{kn_k}$	$Y_{k\bullet}$	$\overline{Y}_{k\bullet} = Y_{k\bullet} / n_k$
Razem	Wielkość próby $n_\bullet = \sum_{i=1}^{k} n_i$		$Y_{\bullet\bullet}$	$\overline{Y} = Y_{\bullet\bullet} / n_\bullet$

Podstawowe równanie jednoczynnikowej ANOVA opisujące rozkład *całkowitej (T) zmienności zmiennej objaśnianej Y* na *zmienność wyjaśnioną zmianą wariantu czynnika G* oraz *zmienność spowodowaną rozproszeniem wewnątrz ustalonych wariantów (grup) tego czynnika (tzw. błędem E) ma postać:*

$$TSS = SSG + SSE, \tag{16-1.1}$$

gdzie:

a) *SSG* (*G* oznacza „grupę”) jest sumą kwadratów odchyleń średnich grupowych od średniej ogólnej i reprezentuje zmienność międzygrupową. Ma ona postać:

$$SSG = \sum_{i=1}^{k} n_i (\overline{Y}_{i\bullet} - \overline{Y})^2, \tag{16-1.2}$$

gdzie obserwacje w próbie oznaczamy jako Y_{ij}, przy czym $i = 1,\ldots,k$ jest numerem populacji wskazanym przez *i*-ty poziom czynnika *G*, z której pochodzi próba *i*-ta, a $j = 1,\ldots,n_i$ jest numerem obserwacji w *i* – tej próbie (grupie).

Występująca w (16-1.2) średnia dla *i* – tej próby (grupy) jest równa:

$$\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \tag{16-1.3}$$

natomiast średnia arytmetyczna z wszystkich obserwacji w (całkowitej) próbie jest równa:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \quad , \tag{16-1.4}$$

gdzie:

$$n_\bullet = \sum_{i=1}^{k} n_i \tag{16-1.5}$$

jest całkowitą liczebnością próby.

b) *SSE* (*E* od „error" - błąd) jest sumą kwadratów odchyleń wartości cechy od średniej grupowej i reprezentuje zmienność wewnątrzgrupową wynikającą z błędu losowego:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2 \ . \tag{16-1.6}$$

c) *TSS* (*T* od „total") jest całkowitą sumą kwadratów odchyleń od średniej ogólnej we wszystkich grupach (dla wszystkich obserwacji):

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2 + \sum_{i=1}^{k} n_i(\overline{Y}_{i\bullet} - \overline{Y})^2 = SSG + SSE \ . \tag{16-1.7}$$

Powyższe źródła zmienności zmiennej objaśnianej *Y* i związane z nimi sumy kwadratów (*SS*) odchyłek, przedstawia poniższa tabela.

Tabela 16-1.2. Schemat jednoczynnikowej analizy wariancji [1].

Źródła Zmienności *Y*	Liczba stopni swobody (*df*)	Sumy kwadratów Odchyleń (*SS*)	Średnie kwadraty Odchyleń
Czynnik *G* z poziomami dającymi zróżnicowanie międzygrupowe *Y*	$\nu_G = k - 1$	*SSG*	*MSG*
Błąd losowy *E* powodujący zróżnicowanie wewnątrzgrupowe *Y*	$\nu_E = n_\bullet - k$	*SSE*	*MSE*
Ogółem	$\nu = n_\bullet - 1 = \nu_G + \nu_E$	*TSS*	

Centralną hipotezą jednoczynnikowej ANOVA jest hipoteza o równości wartości oczekiwanych w populacjach wyznaczonych przez poziom czynnika *G*:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{16-1.8}$$

wobec hipotezy alternatywnej:

$$H_1 : istnieje\ para\ i \neq j,\ taka\ że\ \ \mu_i \neq \mu_j \ . \tag{16-1.9}$$

Weryfikację hipotezy zerowej H_0 przeprowadza się stosując statystykę testową będącą ilorazem średnich kwadratów odchyleń międzygrupowych *MSG* i wewnątrzgrupowych *MSE*:

$$F = \frac{MSG}{MSE} = \frac{SSG/(k-1)}{SSE/(n_\bullet - k)}, \tag{16-1.10}$$

gdzie średnie sumy kwadratów odchyłek mają postać:

$$MSG = \frac{SSG}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{Y}_{i\bullet} - \overline{Y})^2 = \frac{\sum_{i=1}^{k}(Y_{i\bullet}^2/n_i) - Y_{\bullet\bullet}^2/n_\bullet}{k-1} \tag{16-1.11}$$

$$MSE = \frac{SSE}{n_\bullet - k} = \frac{1}{n_\bullet - k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}(Y_{i\bullet}^2/n_i)}{n_\bullet - k} \tag{16-1.12}$$

przy czym $Y_{i\bullet}$ jest sumą wartości w próbie pobranej z populacji *i*-tej, a $Y_{\bullet\bullet}$ jest łączną sumą wartości zaobserwowanych dla wszystkich *k* populacji.

Statystyka testowa (test) *F* przy prawdziwości H_0 ma rozkład F-Snedecora z $\nu_G = k-1$ stopniami swobody licznika i $\nu_E = n_\bullet - k$ stopniami swobody mianownika. Ponadto, *jeśli hipoteza zerowa jest prawdziwa*, to zarówno *MSG* jak i *MSE*, są nieobciążonymi estymatorami wariancji składnika losowego σ_E^2 w populacji i stąd przy prawdziwości H_0 statystyka ta powinna przyjmować małe wartości, bliskie jedności. Istotna statystycznie wartość statystyki *F* skutkuje przyjęciem hipotezy alternatywnej.

Ponieważ statystyka *F* (16-1.10) jest stosunkiem dwóch estymatorów wariancji, przy czym *licznik* opisuje zmienność międzygrupową, a *mianownik* zmienność wewnątrzgrupową, zatem widać, że postać *F* (16-1.10) w jednoczynnikowej analizie wariancji jest uogólnieniem testu *t*:

$$t = \frac{(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})/\sqrt{1/n_1 + 1/n_2}}{S_p}, \tag{16-1.13}$$

który dotyczy weryfikacji hipotezy o równości wartości oczekiwanych $\mu_1 = \mu_2$ w dwóch populacjach, na przypadek wielu populacji. Dlatego też statystyka S_p w mianowniku jest pierwiastkiem średniej wariancji wewnątrzgrupowej:

$$MSE \equiv S_p^2 = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2 = \frac{(n_1-1)\hat{S}_1^2 + (n_2-1)\hat{S}_2^2}{n_1 + n_2 - 2}, \tag{16-1.14}$$

gdzie $\hat{S}_1^2$, $\hat{S}_2^2$ to wariancje z próby kolejno dla grupy 1 i 2.

Rozdział 16-1-1. Test jednorodności wariancji.

Jest to test, który koniecznie trzeba przeprowadzić zanim przystąpi się do testowania hipotezy o równości średnich w populacjach. *W przypadku odrzucenia* hipotezy o jednorodności wariancji w populacjach nie ma sensu przechodzić do testu o równości średnich, bowiem (i) statystyka *MSE* występująca w mianowniku statystyki *F* nie jest wtedy nieobciążonym estymatorem wariancji składnika losowego σ_E^2 (więcej na ten temat jest dalej) oraz (ii) populacja generalna i tak nie jest jednorodna ze względu na zmianę poziomu czynnika *G*. Do sprawdzenia założenia o jednorodności wariancji w grupach służą takie testy jak np. test Bartlett'a, test Brown'a-Forsythe'a oraz test Levene'go. Krótkie omówienie testu Bartlett'a jest podane poniżej [9].

***Rozdział 16-1-1-1**. Test Bartlett'a.*

Przedmiotem weryfikacji jest hipoteza zerowa o równości wariancji:

$$H_0^{\sigma}: \quad \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \tag{16-1.15}$$

wobec hipotezy alternatywnej:

$$H_1: istnieje\ para\ i \neq j,\ taka\ że \quad \sigma_i^2 \neq \sigma_j^2 . \tag{16-1.16}$$

W teście Bartlett'a wykorzystuję się statystykę:

$$\lambda = \frac{M}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{(n_i - 1)} - \frac{1}{n_\bullet - k}\right)}, \tag{16-1.17}$$

gdzie:

$$M = (n_\bullet - k)\ln MSE - \sum_{i=1}^{k}(n_i - 1)\ln \hat{S}_i^2 , \tag{16-1.18}$$

oraz:

$$MSE = \frac{1}{n_\bullet - k}\sum_{i=1}^{k}(n_i - 1)\hat{S}_i^2 , \tag{16-1.19}$$

przy czym $n_\bullet = \sum_{i=1}^{k} n_i$ jest całkowitą liczebnością próby, a $\hat{S}_i^2$ jest wariancją zmiennej objaśnianej wewnątrz *i*-tej próby:

$$\hat{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet}) . \tag{16-1.20}$$

Jeśli hipoteza zerowa $H_0^{\sigma}: \ \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ jest prawdziwa i spełnione jest założenie o normalności rozkładów to statystyka λ, (16-1.17), ma asymptotycznie rozkład χ^2 z (*k*-1) stopniami swobody, a zbiór krytyczny określony jest relacją: $P(\lambda \geq \chi^2_{\alpha,\nu}) = \alpha$.

Tabela 16-1.3. Zebranie weryfikowanych hipotez oraz testów im odpowiadających [1].

Hipoteza		Testy	Liczba stopni swobody
Hipoteza o równości wartości oczekiwanych w populacjach	$H_0 : \mu_1 = \mu_2 = ... = \mu_k$ $H_1 : istnieje\ para\ i \neq j, \mu_i \neq \mu_j$	Test w ANOVA: $F = \frac{MSG}{MSE}$	$\nu_G = k - 1$ $\nu_E = n_\bullet - k$
Hipoteza o jednorodności wariancji	$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ $H_1 : istnieje\ \sigma_i^2 \neq \sigma_j^2, dla\ i \neq j$	Testy: Barlett'a, Brown'a-Forsythe'a, Levene'go	

Rozdział 16-1-2. Testy szczegółowe. Pojęcie kontrastu. Metoda Scheffe'ego.

Jeśli okaże się, że hipotezę zerową $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ wyrażającą równość wszystkich wartości oczekiwanych w populacjach musimy odrzucić, wtedy w poszukiwaniu przyczyny jej odrzucenia, sprawdzamy hipotezy szczegółowe dotyczące poszczególnych średnich w populacjach, np. o następującej postaci:

$$H_{0(1-2)} : \mu_1 = \mu_2,\ H_{0(2-3)} : \mu_2 = \mu_3,\ H_{0(3-4)} : \mu_3 = \mu_4,$$
$$H_{0(12-34)} : \frac{(\mu_1 + \mu_2)}{2} = \frac{(\mu_3 + \mu_4)}{2},\ H_{0(1-234)} : \mu_1 = \frac{\mu_2 + \mu_3 + \mu_4}{3}. \qquad (16\text{-}1.21)$$

Widzimy więc, że w celu wykrycia przyczyny odrzucenia centralnej hipotezy (16-1.8) w ANOVA, musimy wykonać jednocześnie więcej testów częściowych.

a) Ogólny poziom istotności i metoda LSD (least-significant-difference) dla wielokrotnego testu par.

Metoda LSD (*najmniejszej istotnej różnicy*) dotyczy testu wielokrotnego. Przy testowaniu wszystkich powyższych hipotez odnoszących się do par średnich, przy ogólnej liczbie *k* średnich μ_i, należałoby wykonać $g = {}_kC_2 = \binom{k}{2} = k(k-1)/2$ porównań typu $\mu_i = \mu_j$. Oznacza to, że aby wszystkie testy te były przeprowadzone na wspólnym (ogólnym) poziomie istotności α, to pojedynczy test dla pary $H_0 : \mu_i = \mu_j$ powinien być, zgodnie z nierównością Bonferroni'ego (9.2), wykonany na poziomie istotności nie mniejszym niż α / g (ale ciągle tak, aby ogólny poziom istotności był równy α). Wadą metody LSD jest to, że testy szczegółowe są wykonywane na *zaniżonym* poziomie istotności równym α/g (równanie (9.7) w Rozdziale 9), co oznacza, że jest on dla pojedynczego testu tak mały, że żadna z pojedynczych hipotez nie będzie na ogół

odrzucona (zatem moc pojedynczego testu dla jednej pary jest mała). W związku z tym zostało opracowanych kilka lepszych procedur, między innymi omówiona poniżej metoda Scheffé'go [1].

b) Metoda Scheffe'ego testowania równości par oraz układów wartości oczekiwanych.

Szczególnym przykładem zastosowania metody Bonferroni'ego w analizie wariancji jest np. metoda Scheffe'ego testowania hipotez odpowiednich dla kontrastów [1] (lub wyznaczania przedziałów ufności dla kontrastów). Metoda Scheffe'ego jest zalecana szczególnie w przypadku, gdy zaistnieje jeden z następujących przypadków:

1. Liczebności prób pobieranych z różnych populacji nie są jednakowe.

2. Zachodzą również porównania inne niż proste porównania pomiędzy parami wartości oczekiwanych. Wspomniane ogólniejsze typy porównań nazywamy kontrastami.

Z pomocą metody Scheffe'ego określamy istnienie jakiejkolwiek *istotnej różnicy* pomiędzy układami wartości oczekiwanych na ogólnym poziomie istotności α.

Pojęcie kontrastu.

Kontrast jest funkcją liniową wartości oczekiwanych z k - populacji, którą można zapisać w postaci:

$$L = \sum_{i=1}^{k} c_i \mu_i \,, \tag{16-1.22}$$

przy czym $$\sum_{i=1}^{k} c_i = 0 \,. \tag{16-1.23}$$

Hipotezy zerowe dla kontrastów mają postać:

$$H_0 : L = \sum_{i=1}^{k} c_i \mu_i = 0 \,, \tag{16-1.24}$$

i są stawiane wobec hipotez alternatywnych:

$$H_1 : \sum_{i=1}^{k} c_i \mu_i \neq 0. \tag{16-1.25}$$

Odpowiedni nieobciążony estymator kontrastu L ma postać:

$$\hat{L} = \sum_{i=1}^{k} c_i \overline{Y}_{i\bullet} \quad . \tag{16-1.26}$$

Na przykład gdy w ogólnym teście porównujemy $k = 4$ wartości oczekiwane to przykładem kontrastu jest następująca wielkość:

$$L_{13-24} = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2} \quad , \tag{16-1.27}$$

Równoważne są następujące hipotezy zerowe:

$$H_0 : L_{13-24} = 0 \quad \Leftrightarrow \quad H_0 : (\mu_1 + \mu_3)/2 = (\mu_2 + \mu_4)/2 . \tag{16-1.28}$$

Powyższy kontrast możemy zapisać również w postaci:

$$L_1 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2} = \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4 , \tag{16-1.29}$$

skąd zgodnie z (16-1.22) widać, że:

$$c_1 = -c_2 = c_3 = -c_4 = \frac{1}{2} , \qquad c_1 + c_2 + c_3 + c_4 = 0$$

Metoda Scheffe'ego pozwala wyznaczyć przedziały ufności do szacowania wszystkich możliwych kontrastów. W metodzie tej prawdopodobieństwo, że przedziały ufności zawierają jednocześnie prawdziwe wartości *wszystkich* rozważanych kontrastów wynosi 1 - α (i jest to *ogólny poziom* ***ufności***). Jednocześnie *ogólny poziom* ***istotności*** α jest prawdopodobieństwem błędnego odrzucenia *przynajmniej jednej* z szczegółowych hipotez zerowych dla kontrastów.

Rozpatrzmy hipotezę typową dla jednoczynnikowej ANOVA:

$$H_0 : \mu_1 = ... = \mu_j = ... = \mu_k . \tag{16-1.30}$$

Hipoteza ta jest równoważna postawieniu jednocześnie całej grupy hipotez zerowych dotyczących wszystkich możliwych g kontrastów:

$$H_0 : L_1 = ... = L_j = ... = L_g = 0 . \tag{16-1.31}$$

Jeśli A_j jest zdarzeniem, że wyznaczony przedział ufności dla j-tego kontrastu L_j, pokrywa prawdziwą wartość tego parametru, wtedy lewa strona nierówności Bonferroni'ego (9.1):

$$P\left(\bigcap_{j=1}^{g} A_j\right) \geq 1 - \sum_{j=1}^{g} P(\overline{A}_j) , \tag{9.1'}$$

jest prawdopodobieństwem, że wszystkie g - wyznaczonych przedziałów ufności (j = 1,2,..,g) pokrywa jednocześnie prawdziwe wartości odpowiadających im L_j. Prawa strona nierówności jest wtedy równa $1 - \sum_{j=1}^{g} P(\overline{A}_j)$, gdzie każde ze zdarzeń $\overline{A}_j$ oznacza, że konkretny, szczegółowy przedział ufności wyznaczony dla parametru L_j nie pokrył prawdziwej wartości tego parametru. Tak więc, jeśli szukany ogólny (wspólny) wielowymiarowy obszar ufności dla wszystkich parametrów L_j, ma być wyznaczony na ogólnym poziomie ufności (1 - α), wtedy pojedyncza, j-ta hipoteza zerowa $\mathrm{H}_0 : L_j = 0$, jest (zakładając równy podział poziomu istotności pomiędzy testowane hipotezy) testowana na poziomie istotności α_s większym lub równym α/g, zgodnie z postacią (9.7) nierówności Bonferroni'ego z Rozdziału 9:

$$\frac{\alpha}{g} \le P(\overline{A}_j) = \alpha_s, \quad j = 1, 2, ..., g, \tag{9.7'}$$

a indywidualny *j*-ty przedział ufności jest wyznaczony na poziomie ufności mniejszym lub równym niż (1 – α/g) [1]. Oznacza to, że szczegółowe przedziały ufności ulegają zwężeniu (a prawdziwe, szczegółowe zbiory krytyczne ulegają poszerzeniu w porównaniu z tymi, które byłyby wyznaczone dla błędnie przyjętego szczegółowego poziomu istotności α/g).

Przedział ufności wyznaczony metodą Scheffe'ego dla każdego kontrastu *L* jest następujący (porównaj (16-1.13)- (16-1.14)):

$$\sum_{i=1}^{k} c_i \overline{Y}_i \pm S\sqrt{MSE\,(\sum_{i=1}^{k} \frac{c_i^2}{n_i})}\,, \tag{16-1.32}$$

gdzie $\hat{L} = \sum_{i=1}^{k} c_i \overline{Y_i}$ jest estymatorem nieobciążonym estymowanej wartości parametru *L* oraz:

$$S^2 = (k-1)F_{k-1,n_\bullet-k,1-\alpha}\,, \tag{16-1.33}$$

gdzie $n_\bullet = \sum_{i=1}^{k} n_i$.

Uwaga. W powyższej zależności nie należy mylić statystyki *S* z odchyleniem standardowym.

Pojedynczą hipotezę zerową $H_0 : L = \sum_{i=1}^{k} c_i \mu_i = 0$, (16-1.24), odrzucamy jeśli przedział ufności (16-1.32) nie zawiera wartości 0. Takie sformułowanie testu statystycznego wynika z dopełniania się zbioru krytycznego w stosunku do obszaru ufności, co oznacza, że jeśli wartość statystyki testowej nie wpada w zbiór krytyczny to musi ona wpaść do obszaru ufności (i na odwrót).

Przykład. Gdy dokonujemy porównania jedynie poszczególnych par wartości oczekiwanych, wtedy hipotezą zerową jest:

$$H_0 : L = \mu_i - \mu_j = 0, \ (c_i = 1, c_j = -1)\,. \tag{16-1.34}$$

Przedział ufności (16-1.32) przyjmuje dla ogólnego poziomu ufności równego (1-α), postać:

$$(\overline{Y}_i - \overline{Y}_j) \pm S\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}\,, \tag{16-1.35}$$

a zbiór krytyczny dla szczegółowej hipotezy (16-1.34) leży na zewnątrz przedziału (16-1.35).

Rozdział 16-2. Model regresji dla jednoczynnikowej ANOVA.

Większość procedur dla ANOVA może być wyrażonych w języku analizy regresji. Np. testy F w ANOVA mogą być tak sformułowane aby dotyczyły parametrów stojących przy tzw. zmiennych wskazujących (definicja poniżej) w odpowiednim modelu regresji.

Definicja zmiennej ukrytej (wskazującej, kierunkowej).

Zmienną wskazującą może być każda zmienna w równaniu regresji, która może przyjmować skończoną liczbę wartości. Nazwa „zmienna wskazująca" bierze się stąd, że wartości tej zmiennej nie pojawiają się na skutek pomiaru, ale odpowiadają różnym kategoriom (np. różnym badanym populacjom), którymi jesteśmy zainteresowani w przeprowadzanym badaniu. Np. pewna zmienna wskazująca przyjmuje wartość 1 jeśli przedmiot badania jest płci żeńskiej (populacja kobiet) lub przyjmuje wartość 1, gdy przedmiot badania jest płci męskiej (populacja mężczyzn). Inny przykład zmiennej wskazującej zostanie podany poniżej.

Rozważmy następujący model regresji:

$$Y = \mu + \sum_{i=1}^{k-1} \alpha_i X_i + E, \qquad (16\text{-}2.36)$$

w którym zmienne X_i są tzw. *zmiennymi ukrytymi* (*wskazującymi, kierunkowymi*), określonymi następująco:

$$X_i = \begin{cases} 1 & dla \quad populacji \;\; i-tej \\ -1 & dla \quad populacji \;\; k-tej \\ 0 & w \quad pozostalych \quad przypadkach \end{cases} \qquad i = 1, 2, \ldots, k-1. \qquad (16\text{-}2.37)$$

Parametry powyższego modelu regresji są tak zdefiniowane, aby warunkowe wartości oczekiwane $\mu_{Y|X_1,X_2,\ldots,X_i,\ldots,X_{k-1}} \equiv E(Y|X_1,X_2,\ldots,X_i,\ldots,X_{k-1})$ zmiennej Y wyznaczone w tym modelu miały następującą postać:

$$\begin{cases} \mu_{Y|0,0,\ldots,X_i=1,\ldots,0} = \mu + \alpha_i = \mu_i, \quad dla \quad i = 1, 2, \ldots, k \\ \mu_{Y|-1,-1,\ldots,-1,\ldots,-1} = \mu - (\alpha_1 + \alpha_2 + \ldots + \alpha_{k-1}) = \mu_k, \quad dla \quad i = k \end{cases}. \qquad (16\text{-}2.38)$$

Zatem, odpowiednie warunkowe wartości oczekiwane w modelu regresji są równe wartościom oczekiwanym w populacjach kolejno od 1-szej do *k*-tej.

Wiadomo, że średnia warunkowa empiryczna $\overline{Y}_{i\bullet}$ jest oszacowaniem μ_i natomiast średnia warunkowa teoretyczna w modelu regresji $\hat{Y}_{i\bullet}$ jest oszacowaniem odpowiedniej warunkowej wartości oczekiwanej $\mu_{Y|0,0,\ldots,X_i=1,\ldots,0}$ bądź $\mu_{Y|-1,-1,\ldots,-1,\ldots,-1}$.

Ponieważ również dla próby, w modelu regresji *ze zmiennymi ukrytymi* skonstruowanego dla ANOVA:

$$Y = \hat{\mu} + \sum_{i=1}^{k-1} \hat{\alpha}_i X_i + \hat{E}, \qquad (16\text{-}2.39)$$

średnie warunkowe teoretyczne $\hat{Y}_{i\bullet}$ modelu regresji są (z założenia) równe średnim warunkowym empirycznym $\overline{Y}_{i\bullet}$, zatem suma kwadratów odchyłek związana ze zmiennością zmiennej objaśnianej, która jest wyjaśniona modelem regresji (Reg) ze zmiennymi ukrytymi, jest równa:

$$MS\text{Reg}\,(\text{dla modelu ze zmiennymi ukrytymi}) = MSG(\text{w ANOVA}) \;, \qquad (16\text{-}2.40a)$$

natomiast suma kwadratów odchyłek dla zmienności zmiennej objaśnianej niewyjaśnionej regresją i związana z losowym błędem, jest równa:

$$MSE\,(\text{dla modelu ze zmiennymi ukrytymi}) = MSE(\text{w ANOVA}) \;. \qquad (16\text{-}2.40b)$$

Parametry modelu (16-1.37) można wyrazić w języku wartości oczekiwanych $\mu_1, \mu_2, \ldots, \mu_k$ dla rozważanej liczby *k* populacji w następujący sposób (pokazać):

$$\mu = \frac{\mu_1 + \mu_2 + \ldots + \mu_k}{k} \overset{def}{=} \overline{\mu}^* \qquad (16\text{-}2.41a)$$

$$\begin{aligned}
\alpha_1 &= \mu_1 - \overline{\mu}^* \\
\alpha_2 &= \mu_2 - \overline{\mu}^* \\
&\vdots \\
\alpha_i &= \mu_i - \overline{\mu}^* \\
&\vdots \\
\alpha_{k-1} &= \mu_{k-1} - \overline{\mu}^*
\end{aligned} \qquad (16\text{-}2.41b)$$

skąd:

$$\alpha_k \equiv -(\alpha_1 + \alpha_2 + \ldots + \alpha_{k-1}) = \mu_k - \overline{\mu}^* \qquad (16\text{-}2.41c)$$

Z (16-2.41) nietrudno zauważyć, że hipotezę zerową w ANOVA $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ można (gdy wyrazić μ_i przez α_i) zapisać następująco:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0, \qquad (16\text{-}2.42)$$

czyli jako hipotezę zerową dla modelu regresji (16-1.37)-(16-1.38) z *k* parametrami, tzn. *k*-1 *współczynnikami kierunkowymi* α_i oraz jednym parametrem przesunięcia μ (oczywiście przy H_0 zachodzi również α_k=0).

Ze względu na (16-1.40)-(16-1.41), widać, że statystyka testowa *F* dla hipotezy $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0$ modelu regresji (16-1.37) jest taka sama jak statystyka $F = MSG / MSE$, (16-1.10), w ANOVA dla $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, tzn.:

$$F\ (\text{modelu regresji (16-1.37)}) = \frac{MS\text{Reg}}{MSE} = F = \frac{MSG}{MSE} \,. \qquad (16\text{-}2.43)$$

To z kolei, ze względu na *k*-1 *współczynników kierunkowych* α_i, pozwala zrozumieć dlaczego występująca w liczniku *F* statystyka *SSG* ma *k*-1 stopni swobody.

Zwróćmy uwagę, że $\overline{\mu}^*$ jest średnią nieważoną i jak to wynika z powyższego określenia średniej $\overline{\mu}^*$ (wynikającego z przyjęcia zmiennych ukrytych jak w (16-2.38)), jest ona równa ogólnej wartości oczekiwanej w populacji $\mu_\bullet \equiv E(Y)$ tylko wtedy, gdy populacje są równoliczne. Zatem, gdy populacje *nie są równoliczne*, a kodowanie (16-2.38) miałoby być utrzymane, wtedy estymatorem $\overline{\mu}^*$ jest średnia nieważona:

$$\overline{Y}^* = \frac{\overline{Y}_{1\bullet} + \overline{Y}_{2\bullet} + \ldots + \overline{Y}_{k\bullet}}{k} \quad . \tag{16-2.44}$$

Podobna argumentacja, która dla modelu regresji (16-2.36) doprowadziła do związku $\mu = \overline{\mu}^*$, (16-2.41a), dla parametru przesunięcia μ, doprowadziłaby dla modelu regresji w próbie (16-2.39) do następującej postaci estymatora przesunięcia:

$$\hat{\mu} = \frac{\overline{Y}_{1\bullet} + \overline{Y}_{2\bullet} + \ldots + \overline{Y}_{k\bullet}}{k} = \overline{Y}^* . \tag{16-2.45}$$

Zatem, $\hat{\mu}$ jest równe średniej ogólnej $\overline{Y}_{i\bullet}$ w próbie tylko dla równolicznych wariantów czynnika *G*.

Można sprawdzić, że zmieniając kodowanie zmiennych ukrytych z (16-2.37) na następujące:

$$X_i = \begin{cases} n_k & dla \quad populacji \quad i-tej \quad , \qquad i = 1, 2, \ldots, k-1 \\ -n_i & dla \quad populacji \quad k-tej \\ 0 & w \quad pozostalych \quad przypadkach \end{cases} \quad . \tag{16-2.46}$$

spowodowałoby, że średnie $\overline{X}_i$ byłyby równe zero:

$$\overline{X}_i = \frac{1}{n_\bullet} \sum_{s=1}^{k} n_s X_{is} = \frac{1}{n_\bullet} (n_i \times n_k - n_k \times n_i) = 0 ,$$

gdzie X_{is} jest wartością zmiennej X_i w próbie dla *s*-tego poziomu czynnika. W takiej sytuacji otrzymalibyśmy, że $\mu = \mu_\bullet$. Istotnie:

$$\begin{aligned} \mu_\bullet \equiv E(Y) &= \sum_{s=1}^{k} E(Y_s) \frac{n_s}{n_\bullet} = \frac{1}{n_\bullet} \sum_{s=1}^{k} \left(\mu + \sum_{i=1}^{k-1} \alpha_i X_{is} \right) n_s = \frac{\mu}{n_\bullet} \sum_{s=1}^{k} n_s + \sum_{i=1}^{k-1} \alpha_i \frac{1}{n_\bullet} \sum_{s=1}^{k} n_s X_{is} = \\ &= \mu + \sum_{i=1}^{k-1} \alpha_i \overline{X}_i = \mu \end{aligned} \tag{16-2.47}$$

oraz podobnie (pokazać):

$$\hat{\mu} = \overline{Y}_{\bullet\bullet} . \tag{16-2.48}$$

Rozdział 16-3. Przykład „hipermarket ABC" dla jednoczynnikowej ANOVA.

Wielki hipermarket ABC przeprowadził przegląd średnich tygodniowych wydatków klientów pośród 542 losowo wybranych mieszkańców wielkiej metropolii. Klienci byli przyporządkowani do następujących grup: lojalni wobec supermarketu (L), nowi dla tego marketu (N), odstępujący od niego (D), lojalni wobec konkurencyjnych supermarketów (NL) oraz tych, którzy nie należeli do żadnej z powyższych grup (U). Zweryfikujmy hipotezę o niezależności tygodniowych zarobków od grupy klienckiej.

Dane źródłowe. Poniżej podano dane dla badanego przykładu. Zostały one wygenerowane z wykorzystaniem SAS w następujący sposób.

1) Wybieramy z paska MENU formuły:

Data → Random Variates → Norma (zaczynając od najmniejszej liczebności danej populacji wpisujemy ile wartości ma wygenerować SAS, określamy nazwę danej populacji oraz parametry: średnia (mean), odchylenie standardowe (SD))

2) Aby wygenerować następne wartości musimy dołożyć odpowiednią ilość potrzebnych komórek przez polecenie:

Edit → Mode → Edit, po czym klikając prawym klawiszem myszki na ostatnią komórkę tabeli, klikamy Add lub Duplicate, aż otrzymamy pożądaną ilość komórek.

3) Jeśli mamy już właściwą ilość komórek, postępujemy ponownie jak w 1) i 2).
4) Po wykonaniu powyższych czynności otrzymamy tabelę danych, którą scalamy (ujednolicamy) dzięki następującym poleceniom:

Data → Stack Columns → zaznaczamy dane kolumny i wprowadzamy pod Stack → Ok. Zapisujemy scalone dane w wybranej bibliotece poprzez: File → Save as By Sas Name. Następnie musimy scalone dane otworzyć w oknie analizy: File → Open By Sas Name → Save → wybieramy bibliotekę, w której zostały zapisane scalone dane → Ok.

Tabela danych źródłowych została zamieszczona na końcu Rozdziału 16-3.

Wstępna analiza zbiorowości klienckich hipermarketu ABC z wykorzystaniem charakterystyk opisowych oraz wykresów.

Aby otrzymać raport z analizy przeprowadzonej w SAS dotyczący charakterystyk opisowych wykonujemy następujące kroki. Po uruchomieniu SAS, należy najpierw utworzyć projekt następująco:

a) Z paska MENU wybieramy Solutions → Analysis → Analyst.
b) Po wczytaniu zbioru danych (File → Open By SAS Name) wykonujemy wstępnie analizę w celu otrzymania charakterystyk opisowych, odwołując się do poleceń:

   Statistics → Descriptive →Summary Statistics.

Otrzymany raport SAS'a dla wstępnej analizy charakterystyk opisowych ma postać:

```
Przykład_hipermarket ABC 14:40 Sunday, May 16, 2004

 The ANOVA Procedure

 Level of -----------WYDATKI (_Stack_)-----------
 _Source_      N       Mean           Std Dev

 D             27      83.4388889     9.5473107
 L             84      76.8621429     13.7685118
 N             25      62.1548000     10.6164790
 NL            173     91.0826012     14.7576854
 U             233     82.0094421     13.7085395
```

Wyniki powyższego raportu dla przeciętnych tygodniowych wydatków klientów są zebrane w poniższej tabeli. Dane te wykorzystamy w późniejszej analizie.

Tabela 16-3.1. Dane wykorzystane do obliczeń.

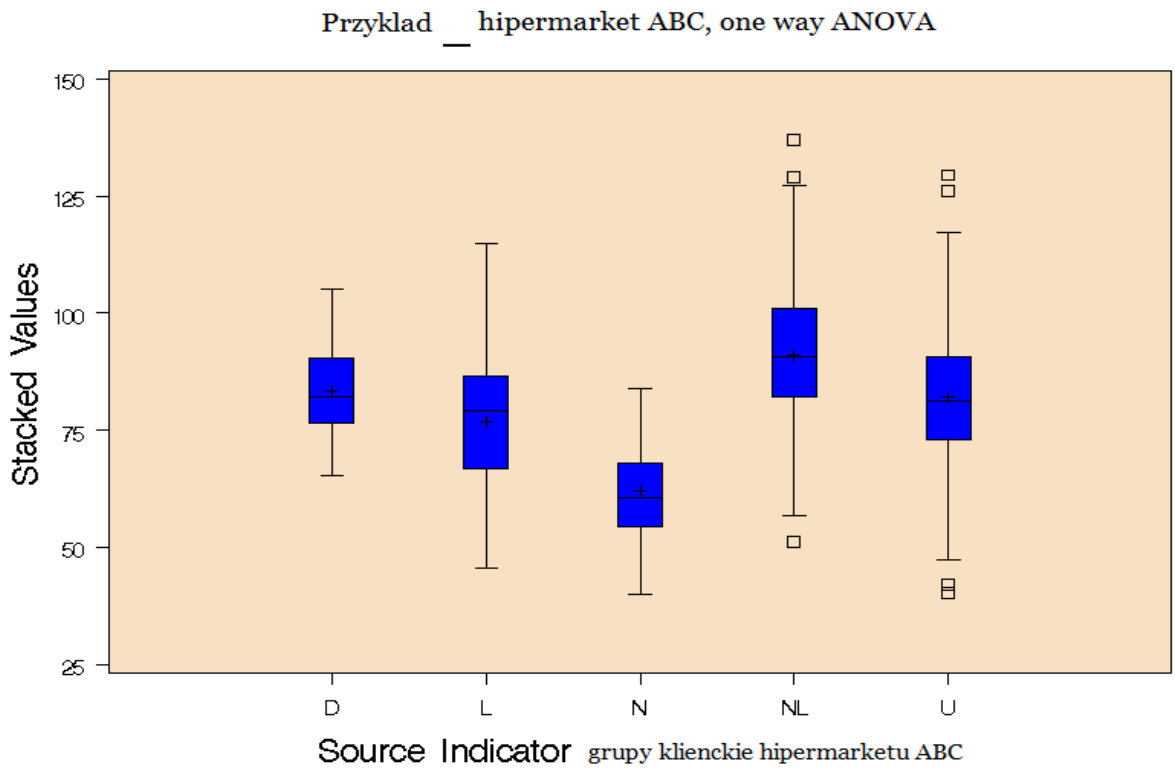
i	**Typ klienta (grupa)**	**Liczebność *n***	**Suma obserwacji** $Y_{i\bullet}$	**Średnia z próby ($)** $\overline{Y}_{i\bullet} = Y_{i\bullet} / n_i$	**odchylenie standardowe (SD) w grupie**
1	L	n_1 = 84	$Y_{1\bullet}$ = 6456,42	$\overline{Y}_{1\bullet} = Y_{1\bullet} / n_1$ = 76,86	13,77
2	N	n_2 = 25	$Y_{2\bullet}$ = 1553,87	$\overline{Y}_{2\bullet} = Y_{2\bullet} / n_2$ = 62,16	10,62
3	D	n_3 = 27	$Y_{3\bullet}$ = 2252,85	$\overline{Y}_{3\bullet} = Y_{3\bullet} / n_3$ = 83,44	9,55
4	NL	n_4 = 173	$Y_{4\bullet}$ = 15757,29	$\overline{Y}_{4\bullet} = Y_{4\bullet} / n_4$ = 91,08	14,76
5	U	n_5 = 233	$Y_{5\bullet}$ = 19108,20	$\overline{Y}_{5\bullet} = Y_{5\bullet} / n_5$ = 82,01	13,71
k=5	**Suma**	$n_\bullet = \sum_{i=1}^{k} n_i$ = 542	$Y_{\bullet\bullet}$ = 45128,63	$\overline{Y} = Y_{\bullet\bullet} / n_\bullet$ = 83,269	

Uwaga. Wartości obserwacji (Y_{ij}) umieszczone są na końcu Rozdziału 16-3 w Tabeli 16-3.3.

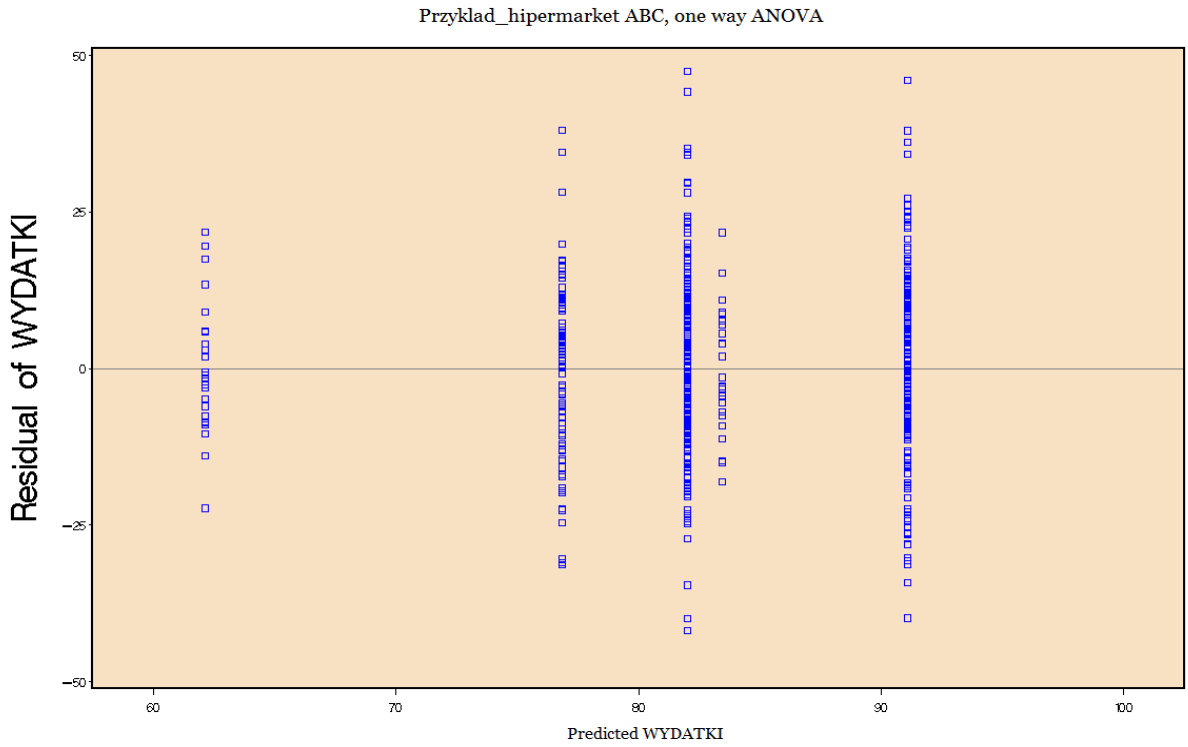
We wstępnym omówieniu sytuacji w pobranej próbce (składającej się z pięciu (pod)próbek grup klienckich L, N, D, NL, U) i odbitej w przedstawionym powyższym raporcie, odwołamy się do wykresów wygenerowanych w pakiecie Analyst (wśród poleceń właściwych dla one-way ANOVA; polecenia te podamy dalej).

Wykresy i ich omówienie.

1) Wykres (1) „pudełkowy z wąsami" (Box-&-whisker plot), na którym pokazano zależność średnich tygodniowych (śr.tyg.) wydatków w hipermarkecie ABC od typu klienta (Source Indicator). Na każdym z wykresów pudełkowych przedstawiono wartości: **kwartyl I –** dół pudełka - (25% śr.tyg. wydatków znajduje się poniżej pierwszego kwartyla – tzn. 25% osób ma śr.tyg. wydatki poniżej tej wartości zmiennej); **kwartyl III –** góra pudełka – (75% śr.tyg. wydatków znajduje się poniżej tej wartości zmiennej); **mediana** – pozioma kreska wewnątrz pudełka - (50% śr.tyg. wydatków ma wartość ≤ medianie lub ≥ miedianie); dodatkowo + oznacza średnią arytmetyczną dla danej zbiorowości typu klienta. Z wykresu można także odczytać wartości maksymalne i minimalne wśród obserwacji (zaznaczone przez końce wąsów lub skrajne punkty (kwadraty), czyli takie, które znacznie odstają od typowego obszaru zmienności śr.tyg. wydatków w badanej zbiorowości).



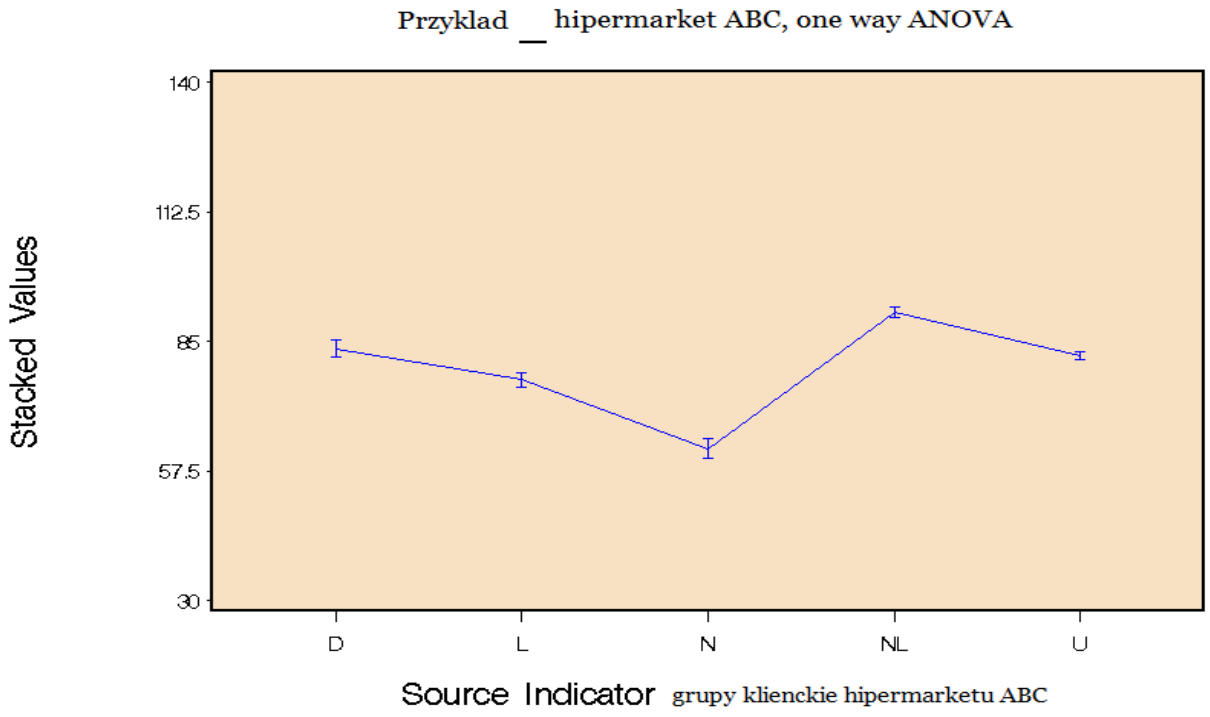
2) Wykres (2) przedstawiający rozrzut wartości reszt wokół wartości średniej, w każdej z pięciu (pod)próbek v.s. wartość średnich wydatków (na osi odciętych) (Residual plot of predicted Y).



Z powyższych wykresów (1) i (2) można wnioskować, że nie ma zasadniczych różnic w rozproszeniu wydatków w ramach każdej z (pod)próbek. Można by mieć jednak pewne zastrzeżenia co do tego

spostrzeżenia, które jak się okazuje mają również swoje odzwierciedlenie w niejednoznaczności poniżej przeprowadzonego testu Bartlett'a. Jednak analiza oparta o testy Levene'go i Brown'a – Forsythe'a potwierdzi optyczną analizę powyższego wykresu rozkładu reszt, prowadząc ostatecznie do przyjęcia wniosku o braku podstaw do odrzucenia hipotezy o jednorodności rozkładu reszt.

3) Wykres (3) (Means plot) zależności średnich tygodniowych wydatków od typu klienta (Source Indicator). Na wykresie zaznaczono empiryczną linię regresji (krzywa łamana łącząca punkty $(i, \overline{Y}_i)$, $i = 1,2,\ldots,5$) oraz odchylenia standardowe w 5 – ciu pobranych (pod)próbkach dla 5 – ciu typów klientów L, N, D, NL, U.



Przyglądając się wykresowi (3), ale i dwóm poprzednim, można wyciągnąć wniosek, że różnice w średnich tygodniowych wydatkach dla pięciu (pod)próbek pobranych z pięciu badanych populacji klientów są istotne statystycznie. Potwierdzi to poniżej przeprowadzona analiza ANOVA. Taka analiza „na oko" pozwala stwierdzić, że głównym źródłem nierówności wartości oczekiwanej wydatków jest populacja klientów nowych (N) dla hipermarketu ABC oraz (w mniejszym stopniu) populacja klientów lojalnych wobec konkurencyjnych supermarketów (NL).

Metoda statystycznej analizy numerycznej przykładu:

a. Zakładamy, że ANOVA jest tworzona dla porównań wartości oczekiwanych tygodniowych wydatków dla różnych typów klientów.

   W rozważanym przykładzie stawiamy hipotezę zerową *o braku istotnego wpływu typu klienta na średnie tygodniowe wydatki na zakupy w hipermarkecie* ABC:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 . \qquad (16\text{-}3.49)$$

   (Indeksy ***i* = 1,2,…,5** odpowiadają kolejno grupom klienckim **L, N, D, NL, U**).

b. Ponieważ ANOVA zakłada konieczność jednorodności (homoskedastyczności) wariancji w populacjach dla różnych typów klientów, więc hipotezę tą weryfikujemy odpowiednimi testami statystycznymi.

c. Precyzujemy model ANOVA. Typ klienta jest czynnikiem ustalonym. (O różnicy pomiędzy czynnikiem ustalonym i losowym powiemy dalej w Rozdziale 16-4).

d. Konstruujemy odpowiedni test statystyczny i tworzymy tablicę ANOVA dla modelu wyznaczonego w punkcie c).

e. Testujemy czy typ klienta ma istotny wpływ na średnie tygodniowe wydatki.

f. Poprzez użycie metody Scheffe'ego określamy jakiekolwiek znaczące różnice pomiędzy parami średnich na poziomie istotności α = 0,05.

Model regresji dla rozważanego przykładu.

W przykładzie $k = 5$ gdyż mamy 5-typów klientów. Równanie modelu regresji (16-2.37) ma postać:

$$Y = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + E , \qquad (16\text{-}3.50)$$

gdzie $\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ to współczynniki modelu regresji ze zmiennymi kierunkowymi, które mogą być wyrażone zgodnie z (16-2.41) poprzez wartości oczekiwane w grupach:

$$\begin{aligned} \mu &= \overline{\mu}^* = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5}{5}, \\ \alpha_1 &= \mu_1 - \overline{\mu}^*, \\ \alpha_2 &= \mu_2 - \overline{\mu}^*, \\ \alpha_3 &= \mu_3 - \overline{\mu}^*, \\ \alpha_4 &= \mu_4 - \overline{\mu}^* \end{aligned} \qquad (16\text{-}3.51)$$

przy czym: $\alpha_5 \equiv -(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) = \mu_5 - \overline{\mu}^*$.

Estymatory powyższych parametrów w popranej próbie mają realizacje:

$$\hat{\mu} = \frac{\overline{Y}_{1*} + \overline{Y}_{2*} + \overline{Y}_{3*} + \overline{Y}_{4*} + \overline{Y}_5}{5} = 79{,}11.$$

$$\hat{\alpha}_1 = \overline{Y}_{1*} - \hat{\mu} = 76{,}86 - 79{,}11 = -2{,}25$$

$$\hat{\alpha}_2 = \overline{Y}_{2*} - \hat{\mu} = 62{,}16 - 79{,}11 = -16{,}95$$

$$\hat{\alpha}_3 = \overline{Y}_{3*} - \hat{\mu} = 83{,}44 - 79{,}11 = 4{,}33$$

$$\hat{\alpha}_4 = \overline{Y}_{4*} - \hat{\mu} = 91{,}08 - 79{,}11 = 11{,}97$$

Liczba stopni swobody modelu regresji ze zmiennymi kierunkowymi wynosi $\nu_G = k - 1 = 5 - 1 = 4$ i jest ona związana z wyznaczonymi z próby czterema oszacowaniami parametrów $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ tego modelu.

Omówienie kolejnych kroków analizy przykładu w programie SAS

Aby otrzymać raport z analizy przeprowadzonej w SAS, trzeba zastosować szereg następujących kroków:

1) Po uruchomieniu SAS, należy najpierw utworzyć projekt następująco:
   Z paska MENU wybieramy Solutions → Analysis → Analyst. Po wczytaniu zbioru danych (File → Open By SAS Name) (co wykonaliśmy już poprzednio, przy okazji otrzymania wcześniejszego raportu dla charakterystyk opisowych).
2) Przechodzimy do jednoczynnikowej analizy wariancji ANOVA:
   Statistics → Anova → One – Way Anova.
   a) "typ klienta" określamy jako zmienną objaśniającą (Independent - niezależną) (u nas: _Source_), zaś obserwowane wartości jako zmienną objaśnianą (Dependent - zależną) (u nas: WYDATKI).
   b) wybieramy testy, które SAS wykona w celu weryfikacji odpowiednich hipotezy o równości wariancji:
   Test → zaznaczamy wybrane testy (Barlett's, Brown – Forsythe, Levene's) → Ok.
   c) wybór metody Scheffe'a: Means → Comparisons → Comparison Metod → Scheffe's multiple – comparison procedure → Significance level (wybieramy poziom istotności, np. 0,05) → zaznaczamy zmienną objaśniającą (Main effects) (u nas: _Source_) → Add → Ok.
   d) wybieramy odpowiednie wykresy zaznaczając w Plots: Box-&-whisker plot, Means plot, Residual plot of predicted Y.
   e) określamy nazwę naszego projektu:
   Titles → Global → wpisujemy nazwę (np. Przyklad_hipermarket ABC, one-way ANOVA)
   Titles → One – Way Anova → wpisujemy nazwę (jak w Global) → Ok.
   f) po wykonaniu czynności (a)-(e) zatwierdzamy → Ok i otrzymujemy raport w SAS z danej analizy.

Ponieważ ANOVA zakłada konieczność jednorodności wariancji, zatem wśród powyższych kroków (krok 2b) i w raporcie SAS'a znalazły się testy dla hipotezy:

$$\mathsf{H}_0^{\sigma}: \quad \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 . \qquad (16\text{-}3.52)$$

Tylko brak podstaw do odrzucenia tej hipotezy nadaje sens (Rozdział 16-1-1) wykonaniu testu hipotezy H_0 o równości wartości oczekiwanych, przeprowadzonego po dokonaniu weryfikacji hipotezy H_0^{σ}, dla której odpowiednie raporty SAS'a (z dodanymi komentarzami) mają postać:

Przykład_hipermarket ABC 14:40 Sunday, May 16, 2004

The ANOVA Procedure

Levene's Test for Homogeneity of WYDATKI(_Stack_) Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Squares	Sum of Mean Square	F Value	Pr > F
Source	4	571450	142862	1.56	0.1827
Error	537	49071298	91380.4		

Wg testu Levene'go, dla α < p (= 0,1827) nie ma podstaw do odrzucenia hipotezy H_0^{σ}: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$.

Brown and Forsythe's Test for Homogeneity of WYDATKI(_Stack_) Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Squares	Sum of Mean Square	F Value	Pr > F
Source	4	513.9	128.5	1.71	0.1470
Error	537	40415.4	75.2614		

Wg testu Brown'a – Forsythe'a, dla α < p (= 0,1470), nie ma podstaw do odrzucenia hipotezy H_0^{σ}: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$

Bartlett's Test for Homogeneity of WYDATKI(_Stack_) Variance

Source	DF	Chi-Square	Pr > ChiSq
Source	4	$\lambda =$ 9.9486	0.0413

Widać, że wg testu Barlett'a empiryczny poziom istotności p = 0,0413, zatem dla $\alpha \geq p$ (np. dla α = 0,05) należałoby odrzucić hipotezę zerową H_0^{σ}: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$, natomiast dla $\alpha < p$ (np. dla α = 0,01) nie ma podstaw do odrzucenia hipotezy zerowej H_0^{σ}. Wartość p = 0,0413 (która leży pomiędzy 0,01 a 0,05) nie pozwalałaby więc w oparciu o test Barlett'a, na podjęcie jednoznacznej decyzji. Wyniki testów Levene'go i Brown'a – Forsythe'a skłaniają nas do podjęcia następującej decyzji statystycznej: obserwując otrzymane wartości empirycznych poziomów istotności *p* dla zastosowanych testów hipotezy o jednorodności wariancji wnioskujemy, że różnica wariancji nie jest istotna statystycznie (z wyjątkiem pewnej statystycznej istotności w teście Barlett'a). Tak więc, ze względu na wyniki testów Levene'go i Brown'a – Forsythe'a decydujemy się na nieodrzucenie hipotezy $H_0^{\sigma}: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ o jednorodności wariancji w badanych 5 – ciu populacjach typów klientów. Oznacza to, że wariancje można uznać za równe, co pozwala

na przejście do procedur ANOVA dotyczących porównania wartości oczekiwanych (dla tygodniowych wydatków w omawianym przykładzie).

Uzupełnienie. Zilustrujmy test Bartlett'a dla jednorodności wariancji, obliczając z wzoru (16-1.17) w programie Excel statystykę λ dla tego testu. Statystyka testowa ma postać [9]:

$$\lambda = \frac{M}{c},$$

gdzie:

$$M = (n-k)\ln MSE - \sum_{i=1}^{k=5}(n_i - 1)\ln S_i^2,$$

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k=5}\frac{1}{(n_i-1)} - \frac{1}{n_\bullet - k}\right]$$

Statystyka λ ma przy prawdziwości hipotezy o jednorodności wariancji H_0^σ: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$, rozkład chi-kwadrat z liczbą stopni swobody równą k-1=4.

Wyznaczmy licznik i mianownik statystyki λ mają przyjmują w próbce kolejno wartości:

$$M = 537*\ln(189{,}6975) - [83*\ln(189{,}57) + 25*\ln(112{,}71) + 26*\ln(91.15) + 172*\ln(217{,}79) + 232*\ln(187{,}92)] =$$
$$= 537*\ln(189{,}6975) - [435{,}32 + 113{,}40 + 117{,}33 + 925{,}97 + 1214{,}76] =$$
$$= 10{,}03,$$

$$c = 1 + \frac{1}{3*4}\left[\left(\frac{1}{83} + \frac{1}{24} + \frac{1}{26} + \frac{1}{172} + \frac{1}{232}\right) - \frac{1}{542-5}\right] = 1 + \frac{1}{12}*[0{,}10230 - 0{,}00186] = 1{,}0084,$$

skąd:

$$\lambda = \frac{10{,}03}{1{,}0084} = 9{,}9486$$

Wartość krytyczna dla α = 0,05 wynosi $\chi^2_{\alpha,k-1} = 9{,}4877$, zaś dla α = 0,01 wynosi $\chi^2_{\alpha,k-1} = 13{,}2767$. Zatem, wartość statystyki λ otrzymana z obserwacji $\lambda_{obs} = \chi^2_{\alpha,k-1(obs)} = 9{,}9486$ wpada na poziomie istotności α = 0,05 do przedziału krytycznego <9,9486; + ∞), co oznacza, że wynikiem testu byłoby wtedy odrzucenie hipotezy zerowej o jednorodności wariancji. Jeśli jednak poziom istotności α = 0,01, wtedy nie mamy podstaw do odrzucenia hipotezy zerowej o jednorodności wariancji. Dlatego podane w raporcie SAS'a wyniki dwóch innych testów dotyczących jednorodności wariancji są pomocne w podjęciu decyzji. Wyniki tych testów (dla Levene'go p = 0,1827 i dla Brown'a – Forsythe'a, p = 0,1470) są bardziej jednoznaczne i wskazują na brak podstaw do odrzucenia hipotezy zerowej H_0^σ. Taką też podjęliśmy decyzję, która oznacza brak podstaw o odrzuceniu hipotezy o jednorodność wariancji tygodniowych wydatków w pięciu badanych populacjach. Umożliwia to zastosowanie analizy wariancji dla wartości oczekiwanych.

Zatem, poniżej umieszczony jest raport SAS'a dotyczący weryfikacji hipotezy zerowej: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ *o braku istotnego wpływu typu klienta na średnie tygodniowe wydatki na zakupy w hipermarkecie* ABC.

Raport SAS dla omawianego przykładu (analiza ANOVA; do raportu dodano komentarz):

```
                    Przykład_hipermarket ABC 14:40 Sunday, May 16, 2004

                                The ANOVA Procedure
                              Class Level Information
                                Class Levels Values
                               _Source_ 5 D L N NL U

                            Number of observations 1165
   NOTE: Due to missing values, only 542 observations can be used in this analysis.

                    Przykład_hipermarket ABC 14:40 Sunday, May 16, 2004

                                The ANOVA Procedure
                    Dependent Variable: WYDATKI(_Stack_) Stacked Values
```

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ - Weryfikujemy hipotezę zerową

Source	DF(**st.swobody**)	Sum of Squares	Mean Square	F Value	Pr > F
Model	**(*k*−1)=** 4	25525.7236	6381.4309 **(*MSG*)**	33.64	<.0001
Error	**(*n**−*k*)=** 537	101867.5665	189.6975 **(*MSE*)**		
Corrected Total	**(*n**−1)=** 541	127393.2901			

R-Square	Coeff Var	Root MSE	WYDATKI(_Stack_) Mean
0.200369	16.54162	13.77307	83.26315

(siła związku liniowego – słaby)

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Source	4	25525.72364	6381.43091	33.64	<.0001

Z raportu widać, że na każdym poziomie istotności $\alpha \geq p$ (< 0,0001), wartość F w próbie wpada do przedziału krytycznego, co pozwala na podjęcie na tym poziomie istotności decyzji o odrzuceniu hipotezy zerowej na korzyść alternatywnej, tzn. hipotezy H_1 mówiącej o tym, że: *„typ klienta" ma wpływ na ilość wydawanych pieniędzy w powyższym hipermarkecie*.

Oznacza to, że w pobranej próbce, różnica pomiędzy przynajmniej jedną z par średnich $\overline{Y}_1, \overline{Y}_2, \overline{Y}_3, \overline{Y}_4, \overline{Y}_5$ jest istotnie statystycznie różna od zera.

Uzupełnienie. Poniżej przeprowadzono uzupełniające rachunki, które są ilustracją zastosowanych metod oraz wyjaśnieniem raportu SAS'a. Dane wykorzystane do poniższych obliczeń znajdują się w Tabeli 16-3.1. Rachunki wyjaśniają poniżej zamieszczoną postać tablicy ANOVA (Tabela 16-3.2). Korzystając z (16-1.2), (16-1.6), (16-1.7) oraz (16-1.11) i (16-1.12) wyznaczamy (przy $n_\bullet = 542$, k =5):

$$SSG = \sum_{i=1}^{k=5} n_i (\overline{Y}_{i\bullet} - \overline{Y})^2 = \sum_{i=1}^{k=5} (Y_{i\bullet}^2 / n_i) - Y_{\bullet\bullet}{}^2 / n_\bullet =$$

$$= \left([(6456{,}42)^2 / 84] + [(1553{,}87)^2 / 25] + [(2252{,}85)^2 / 27] + [(15757{,}29)^2 / 173] + [(19108{,}20)^2 / 233] \right) - [(45128{,}63)^2 / 542] =$$

$$= 2552{,}572,$$

$$MSG = \frac{SSG}{k-1} = 6381{,}431,$$

$$SSE = \sum_{i=1}^{k=5} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2 = \sum_{i=1}^{k=5} \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k=5} (Y_{i\bullet}^2 / n_i) =$$

$$= (51198{,}875 + 9928{,}51 + 19034{,}523 + 147267{,}472 + 161065{,}120) - \left(\frac{(6456{,}42)^2}{84} + \frac{(1553{,}87)^2}{25} + \frac{(2252{,}85)^2}{27} + \frac{(15757{,}29)^2}{173} + \frac{(19108{,}20)^2}{233} \right) =$$

$$= 10186{,}57,$$

$$MSE = \frac{SSE}{n_\bullet - k} = \frac{1}{542-5}(388494{,}540 - 378307{,}84) = 189{,}6975$$

Statystyka *F*, (16-1.10), przyjmuje w próbce wartość:

$$F = \frac{MSG}{MSE} = \frac{6381{,}431}{189{,}6975} = 33{,}64 \ ,$$

skąd, korzystając z rozkładu F-Snedecora z liczbą stopni swobody licznika $\nu_G = k - 1 = 4$ oraz mianownika $\nu_E = n_\bullet - k = 537$, otrzymujemy w Excel'u empiryczny poziom istotności:

$$p = P(F \geq F_{obs} = 33{,}64) = 4{,}6195 \times 10^{-25} < 0{,}0001,$$

co oznacza, że wartość $F_{obs} = 33{,}64$ jest (wysoce) istotna statystycznie i dlatego hipoteza o równości wartości oczekiwanych wydatków w rozważanych grupach klienckich została odrzucona (na każdym poziomie istotności $\alpha \geq p$).

Np. dla α=0,05 wartość krytyczna statystyki *F* wynosi F_{kr} = 2,38853, stąd zbiór krytyczny jest równy $< 2{,}38853, +\infty)$ i $F_{obs} = 33{,}64 \in < 2{,}38853, +\infty)$. Ta wartość krytyczna F_{kr} testu ogólnego została również podana na początku poniżej podanego raportu SAS'a, przy okazji analizy Scheffe'ego dla kontrastów.

Dodatkowo w celu utworzenia pełnej tablicy ANOVA wyznaczmy całkowitą sumę kwadratów odchyłek zmiennej objaśnianej wydatków (w połączonych populacjach):

$$TSS = SSG + SSE = 2552{,}572 + 10186{,}57 = 12739{,}29.$$

Tablica ANOVA zamieszczona we wcześniejszym raporcie SAS'a ma więc postać:

Tabela 16-3.2. Tablica ANOVA dla jednoczynnikowej analizy wariancji (w przykładzie hipermarket ABC).

Źródła zmienności Y	*df* (stopnie swobody)	*SS*	*MS*	*F*	*p=Pr>F*
Zróżnicowanie międzygrupowe	$\nu_G = k - 1 = 4$	*SSG*=25525,72	*MSG*=6381,431	$F = \frac{MSG}{MSE}$ =33,64	<0,0001
Zróżnicowanie wewnątrzgrupowe	$\nu_E = n_\bullet - k = 537$	*SSE*=101867,57	*MSE*=189,6975		
Ogółem	$\nu = \nu_G + \nu_E = 541$	*TSS*=127393,29			

Ze względu na odrzucenie hipotezy o równości wartości oczekiwanych wydatków wśród k = 5 – ciu grup typów klientów, można zastanowić się nad przyczyną zaistniałej sytuacji i poddać testowi hipotezy zerowe o istnieniu par równych wartości oczekiwanych wydatków (lub ich kombinacji). Służy do tego celu np. omówiony powyżej test Scheffe'ego. Poniżej podano odpowiedni raport SAS'a dla rozważanego przykładu. Hipoteza zerowa dla kontrastów, (16-1.24), ma postać:

$$H_0 : L = \sum_{i=1}^{k=5} c_i \mu_i = 0 \ . \qquad (16\text{-}3.53)$$

```
Przykład_hipermarket ABC 14:40 Sunday, May 16, 2004

The ANOVA Procedure
Scheffe's Test for WYDATKI

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type
II error rate than Tukey's for all pairwise comparisons.

α = Alpha 0.05
n*-k = Error Degrees of Freedom 537
MSE = Error Mean Square 189.6975
Fkr = Critical Value of F 2.38853
```

Powyższa wartości $F_{kr} = F_{k-1,n_\bullet-k,1-\alpha}$ = 2,38853 jest wartością krytyczną dla ogólnego testu o równości wartości oczekiwanych (wyznaczyliśmy ją również powyżej w Excel'u). Występuje ona również jako czynnik w statystyce $S^2 = (k-1)F_{k-1,n_\bullet-k,1-\alpha}$, (16-1.34), wchodzącej w określenie przedziału ufności $\sum_{i=1}^{k} c_i \overline{Y}_i \pm S\sqrt{MSE(\sum_{i=1}^{k}\frac{c_i^2}{n_i})}$, (16-1.32), dla kontrastu L.

Przypomnijmy, że indeksy ***i* = 1,2,3,4,5** odpowiadają kolejno grupom klienckim **L, N, D, NL, U.**

```
Comparisons significant at the 0.05 level are indicated by ***.

                Difference
_Source_        Between                 Simultaneous 95%
Comparison      Means                   Confidence Limits

 NL - D         7.644                   -1.165          16.453
 NL - U         9.073                   4.801           13.346 ***
 NL - L         14.220                  8.559           19.882 ***
 NL - N         28.928                  19.819          38.037 ***
 D - NL         -7.644                  -16.453         1.165
 D - U          1.429                   -7.225          10.084
 D - L          6.577                   -2.841          15.995
 D - N          21.284                  9.468           33.100 ***
 U - NL         -9.073                  -13.346         -4.801 ***
 U - D          -1.429                  -10.084         7.225
 U - L          5.147                   -0.271          10.565
 U - N          19.855                  10.895          28.814 ***
 L - NL         -14.220                 -19.882         -8.559 ***
 L - D          -6.577                  -15.995         2.841
 L - U          -5.147                  -10.565         0.271
 L - N          14.707                  5.008           24.406 ***
 N - NL         -28.928                 -38.037         -19.819 ***
 N - D          -21.284                 -33.100         -9.468 ***
 N - U          -19.855                 -28.814         -10.895 ***
 N - L          -14.707                 -24.406         -5.008 ***
```

*** - oznacza, że średnie różnią się istotnie statystycznie, co widać, gdyż odpowiednie przedziały ufności dla kontrastów nie obejmują wartości zero.

Wniosek. Zatem istotne statystycznie różnice pomiędzy średnimi oznaczone przez *** były przyczyną *odrzucenia* początkowej, ogólnej hipotezy zerowej $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ mówiącej *o braku istotnego wpływu typu klienta na średnie tygodniowe wydatki na zakupy w hipermarkecie* ABC.

Uzupełnienie. Zilustrujmy wyniki metody Scheffé'go dla hipotez szczegółowych dla kontrastów (zawarte w przypadku porównań podwójnych w powyższym raporcie), kilkoma krokami rachunków wykonanymi ręcznie stosując wzór (16-1.35):

$$(\overline{Y}_i - \overline{Y}_j) \pm S\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}\ , \qquad \text{gdzie} \quad S^2 = (k-1)F_{k-1,n_{\bullet}-k,1-\alpha}\ , \qquad (16\text{-}3.54)$$

na przedział ufności dla kontrastu, gdzie odpowiednie wielkości zaczerpniemy z raportów SAS'a lub z Tabel 16-3.1 i 16-3.2.

Rozpatrzmy tylko niektóre z hipotez zerowych, np.:

$$H_{01} : \mu_1 = \mu_2 ; H_{02} : \mu_2 = \mu_3 ; H_{03} : \mu_3 = \mu_4 ; H_{04} : \mu_4 = \mu_5\ , \qquad (16\text{-}3.55)$$

które w języku kontrastów mają postać:

$$H_{01} : L_{(1-2)} = \mu_1 - \mu_2 = 0 ;\ H_{02} : L_{(2-3)} = \mu_2 - \mu_3 = 0 ;\ H_{03} : L_{(3-4)} = \mu_3 - \mu_4 = 0 ; H_{04} : L_{(4-5)} = \mu_4 - \mu_5 = 0 . \qquad (16\text{-}3.56)$$

Hipotezę szczegółową dla układu wartości oczekiwanych odrzucamy, gdy wartość odpowiedniego kontrastu L postawiona w hipotezie zerowej nie wpada do przedziału ufności (16-3.54), tzn. gdy wpada do dopełniającego go zbioru krytycznego wyznaczonego dla poziomu istotności α. Przedziały ufności dla poszczególnych par wartości są zgodnie z (16-3.54) następujące:

dla $L_{(1-2)} = \mu_1 - \mu_2$: $14{,}7073 \pm 9{,}6991 = (5{,}008;\ 24{,}406)^{***}$

dla $L_{(2-3)} = \mu_2 - \mu_3$: $-21{,}284 \pm 11{,}81615 = (-33{,}100; -9{,}468)^{***}$

dla $L_{(3-4)} = \mu_3 - \mu_4$: $-7{,}6437 \pm 8{,}8092 = (-16{,}453; 1{,}166)$

dla $L_{(4-5)} = \mu_4 - \mu_5$: $9{,}07316 \pm 4{,}27256 = (4{,}801; 13{,}346)^{***}$

*** - oznacza przedział ufności, który nie obejmuje zera, co oznacza, że odpowiadająca mu wartości estymatora kontrastu w próbce jest statystycznie istotnie różna od zera. Zatem odpowiednia hipoteza zerowa o tym, że badany kontrast w populacji jest równy zero, zostaje odrzucona. (Indeksy i = 1,2,…,5 odpowiadają kolejno grupom klienckim L, N, D, NL, U).

Powyższe rachunki pokrywają się z wynikami otrzymanymi poprzednio w raporcie SAS'a dla porównań par średnich.

Rozważmy jeszcze kontrast porównujący grupę populacji 1,3,5 z populacją 4. Odpowiednia hipoteza zerowa ma postać:

$$H_{05}: \frac{\mu_1 + \mu_3 + \mu_5}{3} = \mu_4 \tag{16-3.57}$$

lub

$$H_{05}: L_{(135-4)} = \frac{\mu_1 + \mu_3 + \mu_5}{3} - \mu_4 = 0 \,. \tag{16-3.58}$$

W przypadku kontrastu $L_{(135-4)}$ skorzystamy z ogólniejszej postaci przedziału ufności (16-1.32):

$$\sum_{i=1}^{k=5} c_i \overline{Y}_i \pm S\sqrt{MSE(\sum_{i=1}^{k=5} \frac{c_i^2}{n_i})}\,, \tag{16-3.59}$$

gdzie pierwszy składnik w (16-3.59) jest estymatorem kontrastu $L_{(135-4)}$:

$$\hat{L}_{(135-4)} = \sum_{i=1}^{k=5} c_i \overline{Y}_i \,, \qquad \text{gdzie} \quad (c_1 = c_3 = c_5 = \frac{1}{3},\ c_4 = -1,\ \ c_5 = 0), \tag{16-3.60}$$

a realizacją przedziału ufności (16-3.59) w próbce jest:

dla $L_{(135-4)} = \frac{\mu_1 + \mu_3 + \mu_5}{3} - \mu_4$ przedział $-8{,}615 \pm 4{,}604 = (-13{,}219; -4{,}012)^{***}$.

Wynik testu dla hipotezy H_{05} jest istotny statystycznie, co oznacza, że wnioskujemy o tym, że wartość oczekiwana μ_4 tygodniowych wydatków klientów populacji *NL*, różni się od średniej z wartości oczekiwanych $\frac{\mu_1 + \mu_3 + \mu_5}{3}$ tygodniowych wydatków klientów populacji *L*, *D* i *U*.

Wniosek. Na podstawie szczegółowych porównań parami, za równe można uznać wartości oczekiwane (średnie) tygodniowe wydatki wśród klientów populacji: *L* i *U*, następnie *L* i *D*, następnie *U* i *D*, oraz *D* i *NL*. Średnie tygodniowe wydatki wśród klientów populacji *N* (nowi klienci) zasadniczo odbiegają od pozostałych grup, co potwierdziło naszą wcześniejszą analizę „na oko" opartą o wykresy (1)-(3). W miarę jednorodna okazała się grupa klientów pochodząca z populacji *L, U i D* (tworzą one jedną grupę Scheffe'ego), co oznacza, że te trzy populacje można by uznać za jedną. Populacja *NL* odchodzi nieco bardziej od tej trójki.

Tabela 16-3.3. Dane źródłowe (obserwacje Y_{ij}) dla przykładu „hipermarket ABC".

L	N	D	NL	U
94.05	61.18	94.48	102.02	83.98
79.66	53.55	92.44	83.01	80.51
77.06	75.62	80.57	102.49	79.21
70.75	71.26	74.38	100.97	88.65
88.12	84.05	92.23	94.76	70.24
80.72	64.08	90.45	103.85	65.76
76.01	59.08	105.21	95.5	78.33
80.52	53.69	80.25	102.67	89.55
82.97	56.15	68.51	91.49	84.76
73.22	61.6	72.3	80.43	69.04
74.3	68.22	79.65	89.98	65.8
82.77	57.33	68.8	100.36	73.39
79.76	59.64	76.62	85.67	74.72
92.5	81.79	78.04	96.59	72.17

67.32	54.57	89.08	110.51	80.35
79.08	51.8	85.4	64.66	73.99
89.83	59.15	87.42	91.75	85.56
86.61	65.15	87.47	100.88	80.28
62.21	79.72	65.46	83.74	72.85
111.44	39.9	89.09	74.37	89.05
63.99	68.03	98.71	91.84	79.3
105.11	53.19	87.48	77.03	76.26
81.04	48.23	82.09	103.22	86.99
81.93	60.74	75.96	88.73	66.82
88.3	66.15	91.04	96.78	86.16
81.49		79.07	117.23	74.43
70.16		80.65	90.7	94.06
57.35			95.82	95.71
87.47			99.62	80.07
52.29			83.67	98.9
87.83			72.87	78.32
66.57			99.39	90.53
80.56			137.14	73.07
93.47			103.33	64.8
84.12			75.64	75.41
87.01			92.27	83.18
88.03			98.83	110.17
66.26			63.09	73.69
91.81			77.73	67.9
79.27			64.94	75.94
115			90.05	111.67
66.22			81.69	97.35
96.74			97.62	70.7
68.28			113.95	57.35
94.2			89.77	97.26
72.79			88.19	83.28
79.13			72.55	71.06
46.45			83.04	86.35
73.88			93.53	66.35
86.11			100.18	92.55
81.41			98.3	66.68
61.12			95.87	85.06
59.95			51.26	63.37
88.66			108.22	64.58
69.05			82.8	76.36
54.26			102.2	93.66
45.54			85.38	61.54
57.8			83.1	84.63
77.55			80.51	82.38
82.25			103.74	66.3
64.54			86.5	73.76
59.61			91.85	82.43
45.96			97.02	106.11
86.52			89.61	105.66
82.6			75.48	82.23
83.56			105.27	83.09
64.92			101.83	83.13
57.09			84.58	79.47
62.43			81.83	58.15
77.31			81.6	103.77
78.18			87.06	69.35
81.19			81.32	95.76
82.16			90.23	81.39
86.62			85.88	82.69
71.4			118.3	89.13
92.01			86.84	95.55
72.78			100.4	74.64
63.83			89.36	78.85
71.03			60.91	91.92
77.12			106.56	89.11
72.88			105.97	95.18
54.57			111.86	72.49
91.46			84.14	70.74
77.3			72.01	100.72

67.18	86.89
76.64	86.87
96.07	42.13
97.43	75.5
81.76	89.66
84.56	72.7
106.91	111.8
81.53	80.58
68.74	91.1
90	47.48
96.27	91.31
97.65	67.73
94.47	85.36
89.27	90.7
79.72	66.22
115.42	93.88
105.32	72.9
129.11	71.39
68.65	79.54
103	76.69
72.32	79.81
64.78	86.63
82.23	91.87
87.61	73.85
127.27	79.79
65.79	89.06
91.23	86.11
113.53	62.98
81.66	81.64
82.37	91.09
92.89	69.14
108.46	78.25
85.79	72.33
66.74	93.33
94.77	105.61
68.23	57.67
63.03	74.4
91.2	76.53
115.22	69.96
105.42	92.5
85.97	81.12
98.83	73.17
94.55	75.26
82.69	70.01
95.53	91.69
77.06	91.31
88.58	97.33
113.92	83.09
110.13	88.59
104.72	59.5
103.41	72.17
125.32	94.06
88.87	83.6
87.03	129.54
84.56	84.1
103.7	102.09
82.01	70.1
96.72	83.28
103	104.34
108.25	84.87
60.45	97.03
77.93	95.24
84.44	85.49
59.87	80.92
89.28	86.97
92.41	83.22
103.59	71.7
94.17	58.81
116.11	40.24
108.78	71.16
105.07	73.78
90.55	105.44
88.17	63.51
88.65	90.73
103.81	72.22

92.38	82.45
92.91	64.74
89.62	92.37
76.12	83.94
97.07	74.41
104.15	92.15
93.71	77.92
90.15	81.36
56.92	67.82
70.55	116.62
75.75	62.82
80.02	63.53
100.68	96.17
84.96	83.98
	75.66
	88.48
	81.32
	100.92
	69.03
	78.85
	101.48
	81.58
	62.43
	83.27
	99.4
	93.5
	79.5
	61.9
	88.98
	75.96
	76.83
	74.77
	77.56
	91.27
	75.95
	69
	99.45
	96.03
	70.92
	98.21
	89.1
	74.32
	73.35
	89.69
	117.29
	86.94
	67
	106.4
	80.47
	97.44
	94.13
	64.82
	126.26
	105.36
	74.51
	88.22
	70.05
	72.35
	85.82
	72.89
	78.06
	90.49
	88.91
	67.97
	87.99
	87.63
	92.54
	87.54
	90.78
	116.16
	80.37
	76.56
	54.92
	76.98

Rozdział 16-4. Typy czynników; czynnik ustalony i losowy cz.I.

Istnieją dwa typy czynników: czynnik ustalony i losowy. Czynnik ustalony, to taki, którego poziomy są jedynymi spośród rozważanych i wartość poziomu jest znana przed dokonaniem pomiaru na jednostce, np. w powyższym przykładzie, wylosowana osoba pochodzi z góry ustalonej grupy klienckiej supermarketu ABC. Natomiast, czynnik losowy jest czynnikiem, którego poziomy mogą być uważane jako próbka z pewnej obszernej rodziny (populacji) poziomów.

Rozróżnienie tych czynników jest ważne w ANOVA, ponieważ różne testy istotności są wymagane dla różnych konfiguracji losowych i ustalonych czynników. Problem ten będzie bardziej widoczny w dwuczynnikowej analizie wariancji. Do sprawy powrócimy w Rozdziale 17-3.

Czynnik ustalony.

Ogólna zmienność zmiennej objaśnianej jest opisana równaniem (16-1.1):

$$TSS = SSG + SSE\,, \tag{16-4.61}$$

gdzie $SSG = \sum_{i=1}^{k} n_i (\overline{Y_{i\bullet}} - \overline{Y})^2$, (16-1.2), oraz $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2$, (16-1.6).

Wariancja międzygrupowa ma postać (16-1.11):

$$MSG = \frac{1}{k-1} SSG = \frac{1}{k-1}\sum_{i=1}^{k} (\overline{Y}_{i\bullet} - \overline{Y})^2\, n_i = \hat{S}^2(\overline{Y}_{i\bullet})\,, \tag{16-4.62}$$

natomiast *MSE*, (16-1.12), jest średnią wariancją wewnątrzgrupową $MSE = \overline{\hat{S}_i^2(Y)}$. Istotnie:

$$\begin{aligned} MSE &= \frac{1}{n-k} SSE = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2 = \frac{1}{n-k}\sum_{i=1}^{k}[\frac{1}{n_i - 1}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2]\,(n_i - 1) = \\ &= \frac{1}{n-k}\sum_{i=1}^{k} \hat{S}_i^2(Y)\,(n_i - 1) = \overline{\hat{S}_i^2(Y)}, \end{aligned} \tag{16-4.63}$$

gdzie:

$$\hat{S}_i^2(Y) = \frac{1}{n_i - 1}\sum_{i=1}^{k} (Y_{ij} - \overline{Y}_{i\bullet})^2 \tag{16-4.64}$$

jest wariancją wewnątrzgrupową w *i* – tej grupie.

Mianowniki w $MSG = \frac{SSG}{k-1}$ oraz $MSE = \frac{SSE}{n-k}$ są stopniami swobody (*df*) dla sum, kolejno *SSG* oraz *SSE*.

Zatem ponieważ liczba stopni swobody po lewej i prawej stronie równania (16-4.61) musi być równa, zatem odpowiednie równanie dla stopni swobody (*df*) ma postać:

$$\nu = \nu_G + \nu_E \tag{16-4.65}$$

gdzie:

$$df_{TSS} \equiv \nu = n_\bullet - 1,\quad df_{SGG} \equiv \nu_G = k - 1 \quad \text{oraz} \quad df_{SSE} \equiv \nu_E = n_\bullet - k \tag{16-4.66}$$

Rozważmy hipotezę zerową postaci $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, (16-1.8), którą weryfikujemy za pomocą testu $F = \frac{MSG}{MSE}$, (16-1.10), gdzie przy prawdziwości H_0 statystyka F ma rozkład F – Snedecora $F_{k-1;\, n-k}$.

Z powyższych rozważań wynika, że:

$$F = \frac{MSG}{MSE} = \frac{\hat{S}^2(\overline{Y}_{i\bullet})}{\overline{\hat{S}_i^2(Y)}} = \frac{\frac{1}{k-1}\sum_{i=1}^{k}(\overline{Y}_{i\bullet} - \overline{Y})^2\, n_i}{\overline{\hat{S}_i^2(Y)}}\ , \qquad (16\text{-}4.67)$$

to znaczy, że statystyka F jest ilorazem wariancji międzygrupowej i średniej wariancji wewnątrzgrupowej.

Jeśli k populacji ma być jednorodnych pod względem wartości oczekiwanych i wariancji, to istnieje konieczność przeprowadzenia np. testu Levene'go, Brown'a – Forsythe'a lub Bartlett'a przed przystąpieniem do ANOVA dla wartości oczekiwanych. Wariancja wewnątrzgrupowa $\hat{S}_i^2(Y)$ w i – tej grupie, jest estymatorem wariancji σ_i^2 składnika losowego w i – tej populacji, z której pobrano elementy do i – tej (pod)próbki. Estymatory $\hat{S}_i^2(Y)$, (16-4.64), są nieobciążonymi estymatorami σ_i^2, tzn.:

$$E(\hat{S}_i^2(Y)) = \sigma_i^2 . \qquad (16\text{-}4.68)$$

Zatem jeśli $E(\hat{S}_i^2(Y)) = \sigma_i^2 \equiv \sigma_E^2$, dla każdej i – tej grupy pobranej z i – tej populacji, czyli gdy prawdziwa jest hipoteza o jednorodności wariancji w i-tych populacjach, wtedy średnia ważona $\overline{\hat{S}_i^2(Y)} \equiv MSE$ estymatorów $\hat{S}_i^2(Y)$ jest również estymatorem nieobciążonym wariancji σ_E^2 (pokazać):

$$\mu_{MSE} \equiv E(MSE) = \sigma_E^2\ . \qquad (16\text{-}4.69)$$

Wniosek 1. Zatem MSE jest nieobciążonym estymatorem wariancji σ_E^2 składnika losowego, tzn. $\mu_{MSE} \equiv E(MSE) = \sigma_E^2$, tylko wtedy, gdy $H_0^{\sigma} : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ jest prawdziwa. (16-4.70)

Estymatorami wartości oczekiwanych μ_i są $\overline{Y}_{i\bullet}$. Natomiast estymatorem wartości oczekiwanej ogólnej (czyli w populacji generalnej powstałej z połączenia wszystkich i – tych populacji) jest średnia arytmetyczna ogólna $\overline{Y}$. Można pokazać, że dla czynnika *ustalonego* zachodzi [1]:

$$\mu_{MSG} \equiv E(MSG) = \sigma_E^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2\ , \qquad (16\text{-}4.71)$$

gdzie:

$$\overline{\mu} = \frac{1}{n}\sum_{i=1}^{k}\mu_i n_i\ , \qquad n_\bullet = \sum_{i=1}^{k} n_i\ . \qquad (16\text{-}4.72)$$

Wniosek 2. Zatem MSG jest nieobciążonym estymatorem wariancji σ_E^2 składnika losowego, tzn. $\mu_{MSE} \equiv E(MSG) = \sigma_E^2$, tylko wtedy, gdy prawdziwa jest hipoteza $H_0 : \mu_1 = \mu_2 = ... = \mu_k$. Ze względu na model regresji dla ANOVA (Rozdział 16-2) oznacza to, że wartości oczekiwane w i- tych populacjach, czyli

warunkowe wartości oczekiwane zmiennej objaśnianej ze względu na poziom czynnika (indeks populacji), są przy prawdziwości H_0 takie same.

Gdyby $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ *nie była prawdziwa*, wtedy $\mu_{MSG} \equiv E(MSG) = \sigma_E^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2$ $> \sigma_E^2$ i statystyka $F = \frac{MSG}{MSE}$, (16-4.7), która jest przybliżeniem ilorazu:

$$\frac{\mu_{MSG}}{\mu_{MSE}} = \frac{\sigma_E^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2}{\sigma_E^2}, \qquad (16\text{-}4.73)$$

miałaby ze względu na $\mu_{MSG} > \mu_{MSE}$ tendencje do przyjmowania wartości istotnie statystycznie większych od 1. Natomiast, gdy H_0 jest prawdziwa, wtedy $\frac{\mu_{MSG}}{\mu_{MSE}} = 1$ i statystyka *F* jako iloraz dwóch nieobciążonych estymatorów wariancji σ_E^2 przyjmuje na ogół wartości bliskie 1. (Zwróćmy uwagę, że przy prawdziwości H_0, iloraz $\frac{\mu_{MSG}}{\mu_{MSE}} = 1$ wtedy, gdy $\mu_{MSG} = \mu_{MSE} = \sigma_E^2$.)

Czynnik losowy.

Model ANOVA dla ustalonego czynnika można przedstawić w języku analizy regresji ze zmiennymi ukrytymi jako (Rozdział 16-2, (16-2.37)):

$$Y = \mu + \sum_{i=1}^{k-1} \alpha_i X_i + E, \text{ gdzie } X_i = \begin{cases} 1 & dla \quad populacji \quad i-tej \\ -1 & dla \quad populacji \quad k-tej \\ 0 & w \quad pozostalych \quad przypadkach \end{cases} \qquad i = 1, 2, ..., k-1 \cdot$$

Pomiędzy współczynnikami regresji $\alpha_1, \alpha_2, ..., \alpha_{k-1}$ i wartościami oczekiwanymi μ_i (i = 1, 2,..., k) w populacjach zachodzą związki (16-2.41):

$$\alpha_1 = \mu_1 - \overline{\mu}^*, \alpha_2 = \mu_2 - \overline{\mu}^*, ..., \alpha_i = \mu_i - \overline{\mu}^*, ..., \alpha_{k-1} = \mu_{k-1} - \overline{\mu}^*$$

gdzie $\overline{\mu}^* = (\mu_1 + \mu_2 + ... + \mu_k)/k$ jest nieważoną średnią wartości oczekiwanych, a parametr przesunięcia μ jest równy $\overline{\mu}^*$, oraz spełniony jest warunek $\alpha_k \equiv -(\alpha_1 + \alpha_2 + ... + \alpha_{k-1}) = \mu_k - \overline{\mu}^*$. Zatem widać, że współczynniki $\alpha_1, \alpha_2, ..., \alpha_{k-1}, \alpha_k$ spełniają *zawsze* warunek:

$$\sum_{i=1}^{k} \alpha_i = 0. \qquad (16\text{-}4.74)$$

Biorąc powyższe pod uwagę widać, że dla ustalonego czynnika równanie regresji (16-2.37) jest równoważne następującemu sformułowaniu modelu ANOVA w populacji:

$$Y_{ij} = \mu + \alpha_i + E_{ij}, \qquad i = 1, 2, ..., k; \quad j = 1, 2, ... \qquad (16\text{-}4.75)$$

i w próbie:

$$Y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{E}_{ij}, \qquad i = 1, 2, ..., k; \quad j = 1, 2, ..., n_i, \tag{16-4.76}$$

gdzie Y_{ij} jest j – tą obserwacją w i – tej populacji, E_{ij} jest składnikiem losowym (błędem), $\hat{E}_{ij}$ jest resztą związaną z j – tą obserwacją w próbce pobranej z i – tej populacji gdzie n_i jest liczebnością i – tej próbki pobranej z i – tej populacji.

Przechodząc do czynnika losowego, równanie (16-4.75) należy zastąpić równaniem:

$$Y_{ij} = \mu + A_i + E_{ij}, \qquad i = 1, 2, ...; \qquad j = 1, 2, ... \tag{16-4.77}$$

a równanie (16-4.76), równaniem:

$$Y_{ij} = \hat{\mu} + A_i + \hat{E}_{ij}, \qquad i = 1, 2, ..., k; \quad j = 1, 2, ..., n_i, \tag{16-4.78}$$

gdzie w (16-4.78) zmienne A_i, i = 1, 2, ..., tworzą rodzinę zmiennych losowych, a w (16-4.78) zmienne A_i, i = 1, 2, ..., k, tworzą losową próbę tych zmiennych pobraną z tej rodziny.

Każda ze zmiennych A_i, i = 1, 2, ... dla całej rodziny możliwych poziomów (które to poziomy w przykładzie „hipermarket ABC" stanowią całą rodzin możliwych do pomyślenia typów klientów) reprezentuje, przez analogię do występującego w (16-4.75) ustalonego $\alpha_i = \mu_i - \mu$, (16-2.41), różnicę typu:

$$A_i = M_i - \mu, \tag{16-4.79}$$

gdzie zmienna losowa M_i pojawiła się w miejsce ustalonej wartości μ_i występującej dla czynnika ustalonego.

Aby wykonać odpowiednią analizę należy przyjąć jakąś postać rozkładu dla zmiennej A_i. Ogólniej, zakłada się, że każda ze zmiennych A_i ma taki sam rozkład. Dodatkowo przyjmujemy, że wszystkie A_i mają standaryzowany rozkład normalny ze średnią równą zero (na wzór równości $\alpha_i = 0$, (16-2.42), dla H_0):

$$A_i: \; N(0, \sigma_A^2), \qquad \text{dla każdego} \quad i = 1, 2, ..., \tag{16-4.20}$$

gdzie σ^2_A jest wariancją zmiennej A_i (taką samą dla każdego i) oraz, że zmienne A_i są niezależne od E_{ij} oraz nawzajem niezależne pomiędzy sobą. Warunek, aby pojedyncza zmienna losowa A_i miała średnią równą zero ma podobny charakter jak dla czynnika ustalonego warunek $\sum_{i=1}^{k} \alpha_i = 0$, (16-4.74), będący średnią po zespole parametrów α_i. Zmienne losowe M_i mają rozkład $N(\mu, \sigma_A^2)$.

Należy również przyjąć, że reszty E_{ij} mają rozkład $N(0, \sigma_E^2)$. Przy założeniu niezależności zmiennych A_i i E_{ij} oraz normalności ich rozkładów, zmienne Y_{ij}, (16-4.77), mają, dla każdego (i, j), rozkład $N(\mu, \sigma_A^2 + \sigma_E^2)$, przy czym składnik σ_A^2 wariancji zmiennej $Y_{ij} = \mu + A_i + E_{ij}$ jest związany ze zmiennością zmiennej A_i natomiast σ_E^2 ze zmiennością składnika losowego E_{ij}.

Uwaga. Dla różnych grup, tzn. gdy $i \neq i'$ (gdzie j oraz j' są dowolne), zmienne Y_{ij} oraz $Y_{i'j'}$ są niezależne . Natomiast ponieważ w konkretnej i – tej grupie ta sama zmienna A_i występuje dla każdej obserwacji Y_{ij}, dlatego w i – tej próbie zmienne Y_{ij} dla obserwacji j – tej i $Y_{ij'}$ dla obserwacji j' są ze sobą skorelowane, gdzie współczynnik korelacji wewnątrzgrupowej (który jest oszacowaniem współczynnika korelacji wewnątrzpopulacyjnej) jest równy [35]:

$$r = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_E^2}, \qquad \text{dla wszystkich } j \neq j' \text{ z ustalonym } i, \tag{16-4.21}$$

gdzie każda zmienna Y_{ij} ma wariancję równą $\hat{S}^2(Y_{ij}) = \hat{\sigma}_A^2 + \hat{\sigma}_E^2$, a $\hat{\sigma}_A^2$ jest oszacowaniem wariancji międzypopulacyjnej (międzygrupowej).

Uwaga. To, że w liczniku (16-4.21) występuje $\hat{\sigma}_A^2$ nie jest samo w sobie oczywiste i wymaga dowodu [35].

Rozważmy hipotezę zerową w przypadku czynnika losowego, dla którego wartość oczekiwana zmiennej A_i jest równa zero (na wzór $\alpha_i = 0$ dla H_0). Ponieważ przyjęliśmy, że wpływ poziomów uśrednia się do zera, (16-4.20), zatem *jedyna możliwa zmienność pomiędzy wpływami różnych poziomów czynnika na warunkową wartość oczekiwaną* μ_i *zmiennej objaśnianej Y* (*warunkową – gdyż zależy ona od* A_i), mogłaby pochodzić od niezerowej wartości wariancji $\sigma^2{}_A$ zmiennej A_i (czy zmiennej M_i). Jeśli nie ma takiej zmienności, to należy postawić następującą hipotezę zerową:

$$H_0\text{: } \sigma^2{}_A = 0, \tag{16-4.22}$$

wobec hipotezy alternatywnej:

$$H_1\text{: } \sigma^2{}_A > 0. \tag{16-4.23}$$

Na przykład, gdyby typ klienta w przykładzie „hipermarket ABC" był czynnikiem losowym, to powyższa hipoteza zerowa oznaczałaby *brak rozproszenia* warunkowych wartości oczekiwanych wydatków μ_i dla każdej z grup klienckich.

Można pokazać, że dla tak postawionej hipotezy zerowej (16-4.22), statystyka $F = \frac{MSG}{MSE}$ przybliża w próbce następujący stosunek wartości oczekiwanych [1]:

$$\frac{\mu_{MSG}}{\mu_{MSE}} \equiv \frac{E(MSG)}{E(MSE)} = \frac{\sigma_E^2 + n_0\,\sigma_A^2}{\sigma_E^2}, \tag{16-4.24}$$

gdzie:

$$n_0 = \frac{\sum_{i=1}^{k} n_i - \left(\sum_{i=1}^{k} n_i^2 \Big/ \sum_{i=1}^{k} n_i \right)}{k-1}, \tag{16-4.25}$$

spełnia rolę średniej liczby obserwacji w próbkach pobranych z populacji. Gdy liczba obserwacji w próbkach jest taka sama i wynosi $n_i = n$, i=1,2,...,k, wtedy $n_0 = n$.

Gdy hipoteza zerowa H_0: $\sigma^2_A = 0$, (16-4.22), jest prawdziwa wtedy stosunek μ_{MSG} / μ_{MSE}, (16-4.24), jest równy 1.

Stąd statystyka testowa F dla hipotezy H_0: $\sigma^2_A = 0$ ma w przypadku analizy jednoczynnikowej dla czynnika losowego postać:

$$F = \frac{MSG}{MSE}, \qquad (16\text{-}4.26)$$

która jest taka sama jak poprzednio dla czynnika ustalonego (16-4.7). Sytuacja taka nie ma już miejsca w przypadku analizy wieloczynnikowej (Rozdział 17).

Poniższa tabelka podsumowuje rozważania odnośnie hipotez zerowych i testów dla czynnika ustalonego i losowego [1].

Tabela 16-4.1 Tablica wartości oczekiwanych statystyk modelu jednoczynnikowej ANOVA dla czynnika ustalonego i losowego, hipotez zerowych oraz postać testu F [1].

Źródła zmienności Y	df	MS	F	Wartości oczekiwane średnich kwadratów, $E(.)$		$\frac{E(MSG)}{E(MSE)}$
				Czynnik ustalony	Czynnik losowy	
Zróżnicowanie międzygrupowe	$\nu_G = k-1$	MSG	$\frac{MSG}{MSE}$	$E(MSG) = \sigma^2_E + \frac{1}{k-1}\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2$	$E(MSG) = \sigma^2_E + n_0\ \sigma^2_A$	
Zróżnicowanie wewnątrzgrupowe	$\nu_E = n_\bullet - k$	MSE		$E(MSE) = \sigma^2_E$	$E(MSE) = \sigma^2_E$	
Ogółem	$\nu = n_\bullet - 1$					
Hipoteza zerowa			na ogół ~ 1	$H_0 : \mu_1 = \mu_2 = ... = \mu_k$	H_0: $\sigma^2_A = 0$	1

W powyższych rozważaniach, podsumowanych w Tabeli 16-4.1, zwraca uwagę związek postaci ilorazu $\frac{E(MSG)}{E(MSE)}$ z postacią hipotezy zerowej H_0 w ANOVA oraz jego wpływ tak na postać statystyki testowej F jak i jej wartość w przypadku prawdziwości H_0.

C. Rozdział 17. Wieloczynnikowa analiza wariancji – ANOVA (dwuczynnikowa).

Dwuczynnikowa analiza wariancji ANOVA pozwala w jednym eksperymencie ocenić wpływ (efekt) dwóch czynników oraz wpływ oddziaływania (interakcji) między tymi czynnikami na zmienną objaśnianą. Tablica danych dwuczynnikowej analizy wariancji ANOVA charakteryzuje dwa czynniki istniejące w prowadzonym badaniu. Pierwszy z nich (*czynnik* 1) to czynnik „rzędowy" (*R*), który posiada *r* poziomów, którym odpowiada *r* wierszy tablicy. Drugi z nich (*czynnik* 2) to czynnik „kolumnowy" (*C*), który posiada *c* poziomów, którym odpowiada *c* kolumn tablicy. Zmienna objaśniana (odpowiedź) Y jest w danej tablicy reprezentowana przez indywidualne obserwacje tej zmiennej na wszystkich jednostkach zbiorowości statystycznej. Liczebność jednostek, na których dokonano obserwacji wartości zmiennej Y jest w komórce dla *i* – tego poziomu *czynnika R* i *j* – tego poziomu *czynnika C* równa n_{ij}. Końcowa suma dla *i* – tych „rzędów" jest wyrażona przez $n_{i\bullet}$, a końcowa suma dla *j* – tych kolumn przez $n_{j\bullet}$. Suma wszystkich obserwacji (czyli liczba obserwacji w próbie ogólnej) jest równa $n_{\bullet\bullet}$. Poniższa tablica (jeszcze bez *Y*) wygląda więc jak zwykła tablica korelacyjna dla rozkładu dwuwymiarowego zmiennych *R* i *C*.

Tabela 17.1 Rozkład liczebności w tablicy dwuczynnikowej analizy ANOVA [1].

	Czynnik 2 „kolumnowy" *C*				
Czynnik 1 „rzędowy" *R*	1	2	…	*c*	Suma liczebności dla wierszy
1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\bullet}$
Suma liczebności dla kolumn	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet c}$	$n_{\bullet\bullet}$

Schematy liczebności. Sposoby modelowania danych w dwuczynnikowej analizie wariancji.

Główne rozróżnienie schematów liczebności (częstości) w komórkach wynika z podziału na schematy *zrównoważone* i *niezrównoważone*. W schemacie zrównoważonym mamy jednakową liczebność obserwacji w każdej komórce, podczas gdy w niezrównoważonym tak nie jest [1]. Kolejnym podziałem jest podział na schemat *kompletny*, który charakteryzuje się przynajmniej jedną obserwację w komórce, oraz schemat *niekompletny*, który ma zero obserwacji w jednej lub większej liczbie komórek. Wszystkie niekompletne schematy są jednocześnie niezrównoważone. Jednakże niektóre niezrównoważone schematy posiadają własność proporcjonalności częstości w komórkach, co ułatwia analizę ANOVA, upodabniając ją do analizy dla równej liczebności w komórkach [1]. Rozważymy następujące schematy liczebności modelowania danych w komórkach:

1. Pierwszy z typów modelowania dotyczy sytuacji, gdy każda komórka posiada tylko jedną obserwację, tzn. $n_{ij} = 1$ dla wszystkich *i, j*. Sposób ten wyraża model losowego dobierania bloków. Blok *j*-ty obejmuje grupę jednostek w kolumnie *j*-tej i jest on *jednorodny* względem poziomu zmiennej *C* natomiast *różnorodny* ze względu na poziom czynnika *R*. Czynnik kolumnowy *C* można by nazywać *czynnikiem blokowym*.

2. Drugi typ modelowania występuje, gdy liczba obserwacji w każdej komórce jest jednakowa i większa niż jeden.

3. Trzeci typ wiąże się już z tym, że w poszczególnych komórkach liczba obserwacji jest różna, przy czym można rozważać np. następujące schematy dla liczebności w komórkach:

(a) W najprostszym przypadku, komórki w tych samych kolumnach mają tą samą liczbę obserwacji, zaś komórki znajdujące się w tych samych rzędach występują w określonym stałym stosunku, tzn. dla ustalonego *j* = 1,...,*c*, zachodzi warunek [1]:

$$n_{ij} = \frac{n_{\bullet j}}{r}, \text{ gdzie } n_{\bullet 1} = n_{\bullet 2} = \ldots = n_{\bullet c} \equiv n_c, \tag{17.1}$$

gdzie *i* = 1,...,*r*.

(b) W bardziej złożonym przypadku (obejmującym dla szczególnego założenia $n_{\bullet\bullet} = n_c r$ przypadek (a)), mamy [1]:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}} \quad \text{lub} \quad \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \qquad j = 1,\ldots,c; \qquad i = 1,\ldots,r. \tag{17.2}$$

Warunek (17.2) oznacza założenie istnienia niezależności stochastycznej pomiędzy zmienną *R* i *C* [9].

(c) W najogólniejszym przypadku występuje brak jakiegokolwiek schematu dla liczebności w komórkach. Poza Rozdziałem 17-1, kilka uwag na temat tego schematu zostało zamieszczonych w Rozdziale 18. Zainteresowanego czytelnika odsyłamy do pozycji [1].

Układy danych o równej ilości komórek wspomniane w punkcie 1 i 2 rzadko pojawiają się w badaniach przypadkowo, lecz często zdarza się, że są one tworzone przez badacza, który musi również określić typ czynników i ustalić liczbę ich poziomów. Sposób tworzenia układu danych zależy oczywiście od rodzaju badania.

Układy danych mogą być tworzone na trzy sposoby:

a) Poprzez blokowanie, dzięki któremu występuje w każdym bloku zawsze *taka sama* liczba niewielu obserwacji dla każdego poziomu czynnika „głównego” *R*.
b) Poprzez układanie warstwami zgodnie z poziomami dwóch rozważanych czynników *R* i *C*, i dopiero wtedy pobieranie próbek z populacji w warstwie na przecięciu ustalonych poziomów obu czynników.
c) Kształtowanie kombinacji poziomów czynników (zatem i kształtowanie komórek) i dopiero wtedy przypisywanie tych kombinacji każdej wylosowanej jednostce.

W kolejnych rozdziałach zostaną omówione powyższe schematy 1, 2 i 3 liczebności w komórkach.

Rozdział 17-1. Wstępne rozważania dwuczynnikowej ANOVA z dowolną liczebnością komórek.

Przypadek gdy pobrane próbki mają różną liczebność w komórkach pojawia się w praktyce analiz statystycznych dość często i może mieć on miejsce w np. wtedy gdy:

1. Nie wszystkie interesujące nas zmienne zostały sklasyfikowane przed podbraniem danych.
2. Zostały uwzględniane nowe zmienne po tym jak dane zostały już zebrane.
3. Wszystkie zmienne są osobno sklasyfikowane i niepraktyczne albo nawet niemożliwe jest, by z góry sprawdzić jak ich różne grupy połączyć, by utworzyć interesujący nas związek.

Przypadek komórki z różną liczebnością może się też pojawiać w empirycznych badaniach, wtedy, jeśli model opierał się na podstawowych zmiennych, domagających się równej liczebności komórki, natomiast a posteriori został zadany warunek związany z innymi zmiennymi, niż te podstawowe, które nas interesują. Ponadto, różna liczebność komórki pojawia się na ogół zawsze wtedy, kiedy jest brak danych, który może się zdarzyć na przykład z powodu zaniku zapisu części badań.

Rozdział 17-1-1. Tablica danych dla ANOVA.

Poniższa Tabela przedstawia ogólny układ danych z dowolną liczebnością komórek dla przypadku dwuczynnikowej ANOVA z dwoma czynnikami *R* i *C*. Niech liczba obserwacji Y_{ijk} w komórce w *i*-tym rzędzie i *j*-tej kolumnie jest równa $n_{ij,}$ tzn. k=1, 2,..., n_{ij} .

Tabela 17-1-1.1. Rozkład danych dla przypadku nierównej liczebności komórek w dwu-czynnikowej ANOVA.

Czynnik wiersza *R*	Czynnik kolumny *C*				Wierszowe średnie i liczebności brzegowe
	1	2	…	*c*	
1	$Y_{111}, Y_{112}, \ldots Y_{11n_{11}}$ wielkość próbki = n_{11} średnia komórki = $\overline{Y}_{11\cdot}$	$Y_{121}, Y_{122}, \ldots, Y_{12n_{12}}$ wielkość próbki = n_{12} średnia komórki = $\overline{Y}_{12\cdot}$	…	$Y_{1c1}, Y_{1c2}, \ldots, Y_{1cn_{1c}}$ wielkość próbki = n_{1c} średnia komórki = $\overline{Y}_{1c\cdot}$	$n_{1\cdot}$, $\overline{Y}_{1\cdot\cdot}$
2	$Y_{211}, Y_{212}, \ldots, Y_{21n_{21}}$ wielkość próbki = n_{21} średnia komórki = $\overline{Y}_{21\cdot}$	$Y_{221}, Y_{222}, \ldots, Y_{22n_{22}}$ wielkość próbki = n_{22} średnia komórki = $\overline{Y}_{22\cdot}$	…	$Y_{2c1}, Y_{2c2}, \ldots, Y_{2cn_{2c}}$ wielkość próbki = n_{2c} średnia komórki = $\overline{Y}_{2c\cdot}$	$n_{2\cdot}$, $\overline{Y}_{2\cdot\cdot}$
⋮	⋮	⋮	…	⋮	⋮
r	$Y_{r11}, Y_{r12}, \ldots, Y_{r1n_{r1}}$ wielkość próbki = n_{r1} średnia komórki = $\overline{Y}_{r1\cdot}$	$Y_{r21}, Y_{r22}, \ldots, Y_{r2n_{r2}}$ wielkość próbki = n_{r2} średnia komórki = $\overline{Y}_{r2\cdot}$	…	$Y_{rc1}, Y_{rc2}, \ldots, Y_{rcn_{rc}}$ wielkość próbki = n_{rc} średnia komórki = $\overline{Y}_{rc\cdot}$	$n_{r\cdot}$, $\overline{Y}_{r\cdot\cdot}$
Kolumnowe średnie i liczebności brzegowe	$n_{\cdot 1}$, $\overline{Y}_{\cdot 1\cdot}$	$n_{\cdot 2}$, $\overline{Y}_{\cdot 2\cdot}$	…	$n_{\cdot c}$, $\overline{Y}_{\cdot c\cdot}$	$n_{\cdot\cdot}$, $\overline{Y}_{\cdot\cdot\cdot}$

Średnie w komórkach, rzędach, wierszach oraz średnia ogólna są kolejno równe:

$$\begin{aligned}
&\bar{Y}_{ij}. = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} && gdzie \quad i = 1,2,\ldots,r; \quad j = 1,2,\ldots,c \\
&\bar{Y}_{i}.. = \frac{1}{n_{i}.} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} Y_{ijk} && gdzie \quad n_{i}. = \sum_{j=1}^{c} n_{ij} \\
&\bar{Y}_{.j}. = \frac{1}{n_{.j}} \sum_{i=1}^{r} \sum_{k=1}^{n_{ij}} Y_{ijk} && gdzie \quad n_{.j} = \sum_{i=1}^{r} n_{ij} \\
&\bar{Y}... = \frac{1}{n..} \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} Y_{ijk} && gdzie \quad n.. = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}
\end{aligned} \tag{17-1-1.3}$$

Rozdział 17-1-2. Różna liczebność komórek i problem nieortogonalności sum kwadratów.

Kluczowa statystyczna koncepcja wiążąca się ze szczególnymi analitycznymi problemami spotykanymi w przypadku komórek z różną liczebnością w tablicy danych dwuczynnikowej ANOVA odnosi się do „nieortogonalnych" sum kwadratów stosowanych w rozkładzie całkowitej sumy kwadratów odchyłek zmiennej objaśnianej *TSS*, będącej licznikiem ogólnej wariancji zmiennej *Y*. Aby wyjaśnić, jakie jest znaczenie ortogonalności, podajmy ogólne wzory dla tych sum kwadratów, w ogólnym przypadku komórek z różną liczebnością:

$$\begin{aligned}
&SSR = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (\bar{Y}_{i}.. - \bar{Y}...)^2 \\
&SSC = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (\bar{Y}_{.j}. - \bar{Y}...)^2 \\
&SSRC = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (\bar{Y}_{ij}. - \bar{Y}_{i}.. - \bar{Y}_{.j}. + \bar{Y}...)^2 \\
&SSE = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij}.)^2 \\
&TSS = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}...)^2
\end{aligned} \tag{17-1-2.4}$$

Trzy pierwsze z nich, tzn. *SSR, SSC* i *SSRC* są nazywane *bezwarunkowymi sumami kwadratów* dla kolejno czynnika *rzędowego, kolumnowego* i *wzajemnego oddziaływania*.

Każda z bezwarunkowych sum kwadratów może zostać określona osobno, wychodząc z podstawowych zasad, które pozwalają opisać zmienność zmiennej opisywanej *Y* związaną z oszacowaniem wpływów pochodzących od czynnika rzędowego *R*, $(\bar{Y}_i.. - \bar{Y}\ldots)$, kolumowego *C*, $(\bar{Y}_{.j}. - \bar{Y}\ldots)$, oraz od oddziaływania *RC*, $(\bar{Y}_{ij}. - \bar{Y}_i.. - \bar{Y}_{.j}. + \bar{Y}...)$.

Można pokazać, że jeśli zbiór bezwarunkowych sum kwadratów (17-1-2.4) jest „ortogonalny", to spełniona jest równość [1]:

$$SSR + SSC + SSRC + SSE = TSS \tag{17-1-2.5}$$

co oznacza, że składniki po lewej stronie dzielą całkowitą sumę kwadratów *TSS* na nie przekrywające się, ortogonalne źródła zmienności (to znaczy źródła wariancji) zmiennej *Y*. To podstawowe równanie ANOVA obowiązuje jedynie dla przypadku komórek z równą liczebnością.

Niestety, w przypadku istnienia *komórek z różną liczebnością*, bezwarunkowe sumy kwadratów nie odpowiadają ortogonalnym źródłom wariancji, i wtedy:

$$SSR + SSC + SSRC + SSE \neq TSS \qquad (17\text{-}1\text{-}2.6)$$

Aby zobaczyć, co jest powodem powyższego zachowania się rozkładu dla *TSS*, rozważmy ogólne sformułowanie regresji dla dwuczynnikowej ANOVA dla przypadku komórek z różną liczebnością (który oczywiście obejmuje przypadek równej liczebności komórek).

Rozdział 17-1-3. Ogólne sformułowanie regresji dla dwuczynnikowej ANOVA. Fundamentalne równanie analizy regresji.

Ogólne równanie regresji w ANOVA ma postać:

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij} X_i Z_j + E \qquad (17\text{-}1\text{-}3.7)$$

gdzie μ, α_i, β_j i γ_{ij} są współczynnikami regresji, X_i i Z_j są odpowiednio zdefiniowanymi czynnikami (zmiennymi ukrytymi, wskazującymi), natomiast zmienna *E* jest składnikiem losowym. Zmienne *X* nazwijmy *zmiennymi rzędowymi*, natomiast zmienne *Z*, *zmiennymi kolumnowymi*. Podobnie jak w (16-2.38) dla jednoczynnikowej ANOVA zmienne te wskazują poziomy (warianty) zmiennej rzedowej *R* bądź kolumnowej *C*.

W (17-1-1.7) zmienne ukryte X_i, $i = 1, 2, ..., r-1$, oraz Z_j, $j = 1, 2, ..., c-1$, są opisane następującym kodowaniem (17-1-1.30):

$$X_i = \begin{cases} 1 & dla \quad poziomu \quad i-tego \quad czynnika \quad R \quad , \quad i = 1, 2, ..., r-1 \\ -1 & dla \quad poziomu \quad r \quad czynnika \quad R \\ 0 & w \quad pozostalych \quad przypadkach \end{cases}$$

$$Z_j = \begin{cases} 1 & dla \quad poziomu \quad j-tego \quad czynnika \quad C \quad , \quad j = 1, 2, ..., c-1 \\ -1 & dla \quad poziomu \quad c \quad czynnika \quad C \\ 0 & w \quad pozostalych \quad przypadkach \quad . \end{cases} \qquad (17\text{-}1\text{-}3.8)$$

Podstawowe (fundamentalne) równanie dla sum kwadratów w modelu regresji, ma zawsze w ogólności postać [1]:

$$TSS = SS\text{Reg} + SSE \qquad (17\text{-}1\text{-}3.9)$$

gdzie:

$$\begin{aligned} TSS &= \sum_s (Y_s - \bar{Y})^2 \\ SS\text{Reg} &= \sum_s (\hat{Y}_s - \bar{Y})^2 = \text{Regresja}SS(X_1, X_2, ..., X_{r-1};\ Z_1, Z_2, ..., Z_{c-1};\ X_1Z_1, X_1Z_2, ..., X_{r-1}Z_{c-1}), \\ SSE &= \sum_s (Y_s - \hat{Y}_s)^2 \end{aligned} \qquad (17\text{-}1\text{-}3.10)$$

a sumowanie $\sum_s$ przebiega po wszystkich n.. obserwacjach. Równanie (17-1-3.9) bywa nazywane *fundamentalnym równaniem analizy regresji*. Mówi ono, że *„całkowita niewyjaśniona zmienność zmiennej objaśnianej = zmienność wyjaśniona regresją + niewyjaśniona zmienność spowodowana resztami”.*

Wprowadzając *warunkową sumę kwadratów dla zmiennej* X_{p+1} *dodanej na końcu* (porównaj (5-8)):

$$SS(X_{p+1}|X_1, X_2, ..., X_p) = Regresja\ SS(X_1, X_2, ..., X_p, X_{p+1}) - Regresja\ SS(X_1, X_2, ..., X_p) \qquad (17\text{-}1\text{-}3.11)$$

możemy podzielić sumę kwadratów regresji na kilka sposobów, tak aby podkreślić wkład z powodu dodania na końcu całej grupy zmiennych do modelu regresji, który zawiera już w sobie inne grupy zmiennych. Podstawowe równanie regresji można więc zapisać następująco:

$$\begin{aligned} TSS = &\ \text{Regresja } SS(X_1, X_2, ...,X_{r-1}) \\ &+ \text{Regresja } SS(Z_1, Z_2, ...,Z_{c-1} \mid X_1, X_2, ..., X_{r-1}) \\ &+ \text{Regresja } SS(X_1Z_1, X_1Z_2, ..., X_{r-1}Z_{c-1} \mid X_1, X_2, ..., X_{r-1}, Z_1, Z_2, ..., Z_{c-1}) + SSE = \\ &= SS\text{Reg} + SSE \end{aligned} \qquad (17\text{-}1\text{-}3.12)$$

gdzie podziału *TSS* dokonano uwzględniając wpierw wpływ zmiennych rzędowych *X*, potem wpływ zmiennych kolumnowych *Z*, a dopiero na końcu wpływ ich oddziaływania *XZ*.

Natomiast, jeśli chcemy uwzględnić wpływ zmiennych rzędowych w modelu, w którym już są zmienne kolumnowe, to podstawowe równanie regresji ma postać:

$$\begin{aligned} TSS = &\ Regresja\ SS(Z_1, Z_2, ...,Z_{c-1}) + Regresja\ SS(X_1, X_2, ...,X_{r-1} \mid Z_1, Z_2, ...,Z_{c-1}) \\ &+ Regresja\ SS(X_1Z_1, X_1Z_2, ..., X_{r-1}Z_{c-1} \mid X_1, X_2, ..., X_{r-1}, Z_1, Z_2, ..., Z_{c-1}) + SSE = \\ &= SSReg + SSE\ . \end{aligned} \qquad (17\text{-}1\text{-}3.13)$$

Wprowadźmy oznaczenia:

$$\begin{aligned} &Regresja\ SS(X_1, X_2, ...,X_{r-1}) \equiv SSR \\ &Regresja\ SS(Z_1, Z_2, ...,Z_{c-1}) \equiv SSC \\ &Regresja\ SS(X_1Z_1, X_1Z_2, ..., X_{r-1}Z_{c-1}) \equiv SSRC\ , \end{aligned} \qquad (17\text{-}1\text{-}3.14)$$

gdzie *SSR*, *SSC* oraz *SSRC* są bezwarunkowymi sumami kwadratów. Równania (17-1-3.12) i (17-1-3.13) można teraz zapisać następująco:

$$SSR + SS(C|R) + SS(RC|R, C) + SSE = TSS \qquad (17\text{-}1\text{-}3.15a)$$

oraz

$$SSC + SS(R|C) + SS(RC|R, C) + SSE = TSS\ . \qquad (17\text{-}1\text{-}3.15b)$$

Każde z tych równań można zapisać w postaci:

$$SS\text{Reg} + SSE = TSS \qquad (17\text{-}1\text{-}3.16)$$

gdzie:

$$SS\text{Reg} = SSR + SS(C|R) + SS(RC|R, C) = SSC + SS(R|C) + SS(RC|R, C)\ . \qquad (17\text{-}1\text{-}3.17)$$

Jak widać, obok bezwarunkowych sum kwadratów *SSR* oraz *SSC*, człon regresji *SS*Reg równania (3.9) zawiera *warunkowe sumy kwadratów*.

Szczególnym przypadkiem jest schemat z *równą liczebnością komórek*. Można wtedy pokazać, że zachodzą równości:

$$SSR = SS(R|C)$$
$$SSC = SS(C|R) \qquad \text{dla komórek z równą liczebnością .} \qquad (17\text{-}1\text{-}3.18)$$
$$SSRC = SS(RC|R, C)$$

Zatem , gdy wszystkie liczebności próby są w komórkach takie same, wtedy warunkowe sumy kwadratów nie są zależą od zmiennych będących już uprzednio w modelu i fundamentalne równanie regresji przyjmuje znaną postać (17-1-2.5):

$$SSR + SSC + SSRC + SSE = TSS\ . \qquad (17\text{-}1\text{-}3.19)$$

W przypadku komórek z *różną* liczebnością, sytuacja wygląda następująco:

$$SSR \neq SS(R|C)$$
$$SSC \neq SS(C|R) \qquad \text{dla komórek z różną liczebnością ,} \qquad (17\text{-}1\text{-}3.20)$$
$$SSRC \neq SS(RC|R, C)$$

skąd wynika, że w przypadku tym, równanie (17-1-3.19) nie jest poprawne i podstawowe równanie regresji musi mieć postać (17-1-3.12) lub (17-1-3.13). Z równań tych widać znaczenie kolejności, w której czynniki (dające wpływy główne) są wprowadzane do modelu. Wyjątek stanowi przypadek, gdy zachodzi związek (17.2) dla liczebności w komórkach i na brzegach tablicy danych (warunek proporcjonalność częstości w komórkach, oznaczający stochastyczną niezależność czynników *R* i *C*):

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \quad . \qquad (17.2')$$

Gdy warunek (17.2) jest spełniony, wtedy okazuje się, że zachodzą następujące równości:

$$\begin{cases} SSR = SS(R\,|\,C) \\ SSC = SS(C\,|\,R) \\ SSRC \neq SS(RC\,|\,R,C) \end{cases} \qquad (17\text{-}1\text{-}3.21)$$

Widzimy, że chociaż i w tym przypadku równanie (17-1-3.19) nie jest spełnione, to równania (17-1-3.12) i (17-1-3.13) upraszczają się, sprowadzając do pojedynczego równania:

$$SSR + SSC + SS(RC|R,C) + SSE = TSS \qquad (17\text{-}1\text{-}3.22)$$

W równaniu (17-1-1.22) tylko wyraz *SS(RC|R,C)* różni go od równania (17-1-3.19). Ponieważ suma kwadratów *SS(RC|R,C)* może być otrzymana przez wynikające z (17-1-3.22) odejmowanie sum bezwzględnych od *TSS*, zatem w przypadku gdy zachodzi warunek proporcjonalności częstości w komórkach (17.2), standardowe obliczenia ANOVA prowadzone dla komórek z równą liczebnością mogą być wykonane, bez potrzeby stosowania metody analizy regresji.

Poniższy schemat podsumowuje sposób postępowania przy wyborze metody analizy statystycznej w dwuczynnikowej ANOVA. Dolny prostokąt diagramu zostanie omówiony w Rozdziale 18.

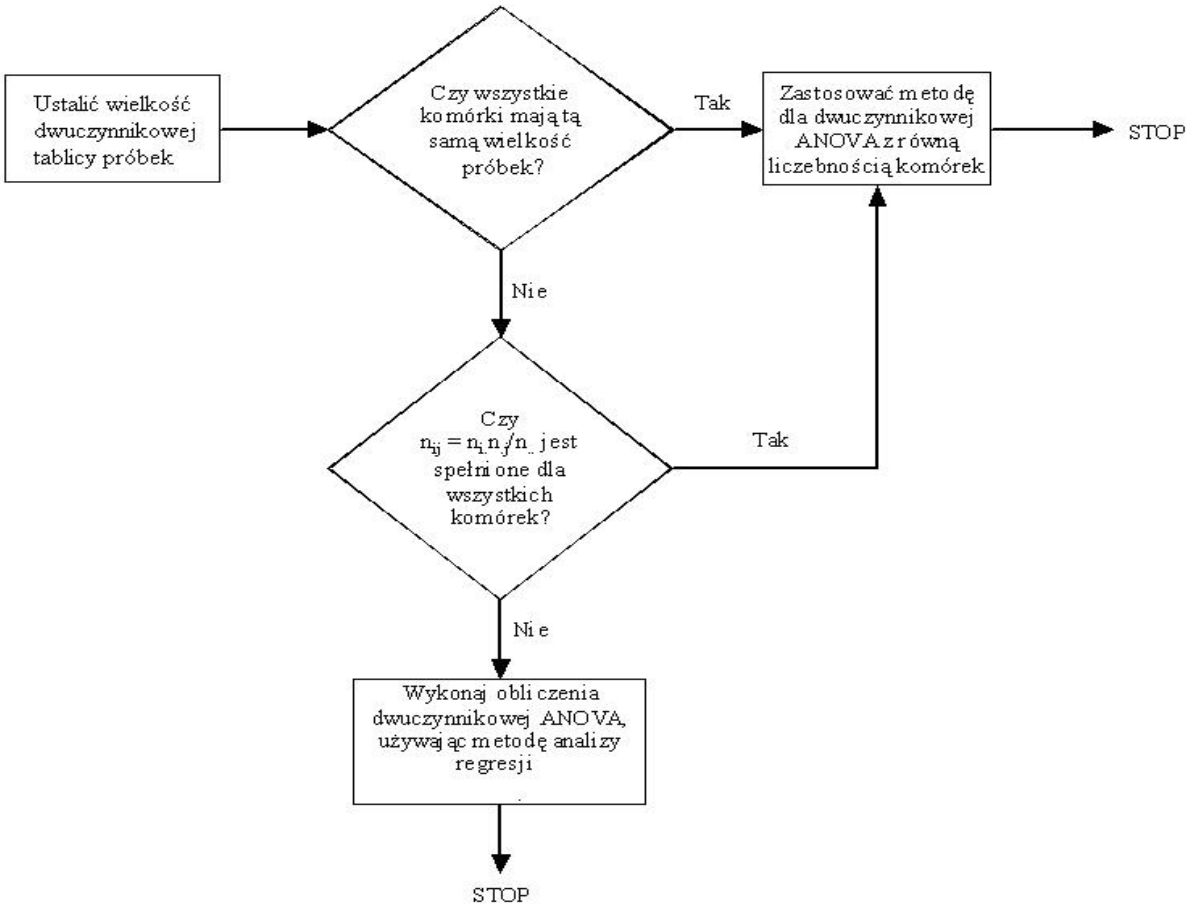


Diagram 17-1-3.1. Diagram postępowania przy wyborze metody analizy w dwuczynnikowej ANOVA.

Rozdział 17-2. Czynnik ustalony i losowy cz.II (równa i większa od 1 liczebność w komórkach).

W Rozdziale 16.4 zostały wprowadzone pojęcia czynnika ustalonego i losowego wraz z uzasadnieniem postaci stawianych hipotez oraz testów. W jednoczynnikowej ANOVA test statystyczny ma taką samą postać dla czynnika ustalonego i losowego, chociaż stawiane hipotezy mają postać różną. Tzn., podczas gdy dla czynnika ustalonego hipoteza zerowa związana była z równością wartości oczekiwanych zmiennej objaśnianej dla różnych poziomów tego czynnika, to dla czynnika losowego dotyczyła ona braku rozproszenia owych wartości oczekiwanych. Podsumowanie sytuacji dla jednoczynnikowej ANOVA zostało zawarte w Tabeli 16-4.1. Również w obecnym rozdziale rozważamy model dwuczynnikowej ANOVA, który realizuje schemat (17.2) dla liczebności w komórkach, oznaczający w ogólności niezależność stochastyczną czynników *R* i *C*. Dodatkowo zakładamy, że **liczba obserwacji w komórkach jest taka sama i nie mniejsza od 2**. Przypadek z jednostkowa liczebnością komórek zostanie omówiony w Rozdziale 17-3.

Układ danych dla dwuczynnikowej ANOVA z równą liczebnościa komórek.

Pierwszym krokiem, który należy wykonać, aby zbadać dwuczynnikowy układ danych, jest stworzenie tablicy składającej się ze średnich z obserwacji w każdej komórce. Mamy *r* poziomów czynnika „rzędowego” *R* oraz *c* poziomów czynnika „kolumnowego” *C*, oraz w każdej z *rc* komórek po *n* obserwacji, tzn. wszystkie komórki zawierają tę samą liczbę obserwacji. Tabela danych dla dwuczynnikowej ANOVA ma poniższą postać.

Tabela. 17-2.1 Dane, średnie dla próbek i populacji, oraz sumy dla dwuczynnikowej ANOVA [1].

	Czynnik „kolumnowy” *C*				
Czynnik „rzędowy” *R*	1	2	…	*c*	Sumy oraz średnie dla rzędów
1	$(Y_{111}, Y_{112}, ..., Y_{11n})$ $Y_{11\bullet}, \overline{Y}_{11\bullet}$, μ_{11}	$(Y_{121}, Y_{122}, ..., Y_{12n})$ $Y_{12\bullet}$, $\overline{Y}_{12\bullet}$, μ_{12}	…	$(Y_{1c1}, Y_{1c2}, ..., Y_{1cn})$ $Y_{1c\bullet}$, $\overline{Y}_{1c\bullet}$, μ_{1c}	$Y_{1\bullet\bullet}$, $\overline{Y}_{1\bullet\bullet}$, $\mu_{1\bullet}$
2	$(Y_{211}, Y_{212}, ..., Y_{21n})$ $Y_{21\bullet}$, $\overline{Y}_{21\bullet}$, μ_{21}	$(Y_{221}, Y_{222}, ..., Y_{22n})$ $Y_{22\bullet}$, $\overline{Y}_{22\bullet}$, μ_{22}	…	$(Y_{2c1}, Y_{2c2}, ..., Y_{2cn})$ $Y_{2c\bullet}$, $\overline{Y}_{2c\bullet}$, μ_{2c}	$Y_{2\bullet\bullet}$, $\overline{Y}_{2\bullet\bullet}$, $\mu_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮
r	$(Y_{r11}, Y_{r12}, ..., Y_{r1n})$ $Y_{r1\bullet}$, $\overline{Y}_{r1\bullet}$, μ_{r1}	$(Y_{r21}, Y_{r22}, ..., Y_{r2n})$ $Y_{r2\bullet}$, $\overline{Y}_{r2\bullet}$, μ_{r2}	…	$(Y_{rc1}, Y_{rc2}, ..., Y_{rcn})$ $Y_{rc\bullet}$, $\overline{Y}_{rc\bullet}$, μ_{rc}	$Y_{r\bullet\bullet}$, $\overline{Y}_{r\bullet\bullet}$, $\mu_{r\bullet}$
Sumy oraz średnie dla kolumn	$Y_{\bullet 1\bullet}$, $\overline{Y}_{\bullet 1\bullet}$, $\mu_{\bullet 1}$	$Y_{\bullet 2\bullet}$, $\overline{Y}_{\bullet 2\bullet}$, $\mu_{\bullet 2}$	…	$Y_{\bullet c\bullet}$, $\overline{Y}_{\bullet c\bullet}$, $\mu_{\bullet c}$	$Y_{\bullet\bullet\bullet}$, $\overline{Y}_{\bullet\bullet\bullet}$, $\mu_{\bullet\bullet}$

Oznaczenia w powyższej tabeli są następujące. Y_{ijk} oznacza *k* – tą obserwację (daną) w komórce (*i*, *j*). Tabela zawiera sumy oraz średnie dla danych, z których pobrane są próbki, jak również odpowiednie wartości oczekiwane μ_{ij} w populacjach (po jednej populacji i jednej próbce, dla każdej komórki).

Łączną sumę w komórce (*i*, *j*) oznaczamy jako $Y_{ij\bullet}$, natomiast sumę w całym *i* – tym rzędzie jako $Y_{i\bullet\bullet}$, a sumę w całej *j* – tej kolumnie jako $Y_{\bullet j\bullet}$, zaś łączną sumę wszystkich obserwacji jako $Y_{\bullet\bullet\bullet}$, zgodnie z zależnościami:

$$Y_{i\bullet\bullet} = \sum_{j=1}^{c}\sum_{k=1}^{n} Y_{ijk} \ , \ Y_{\bullet j\bullet} = \sum_{i=1}^{r}\sum_{k=1}^{n} Y_{ijk} \ , \ Y_{\bullet\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n} Y_{ijk} \ . \qquad (17\text{-}2.23)$$

Średnie w próbach są wyznaczone następująco (ze względu na założoną równą liczebność w komórkach):

$$\overline{Y}_{ij\bullet} = \frac{1}{n}\sum_{k=1}^{n} Y_{ijk} \, , \qquad (17\text{-}2.24)$$

$$\overline{Y}_{i\bullet\bullet} = \frac{Y_{i\bullet\bullet}}{cn}, \qquad \text{dla } i = 1,2,\ldots,r; \tag{17-2.25}$$

$$\overline{Y}_{\bullet j\bullet} = \frac{Y_{\bullet j\bullet}}{rn}, \qquad \text{dla } j = 1,2,\ldots,c; \tag{17-2.26}$$

$$\overline{Y}_{\bullet\bullet\bullet} = \frac{Y_{\bullet\bullet\bullet}}{crn}. \tag{17-2.27}$$

Średnie $\mu_{i\bullet}$ (lub $\mu_{\bullet j}$) są wartościami oczekiwanymi dla populacji utworzonych z połączenia populacji dla wszystkich kolumn (lub wierszy). Średnia $\mu_{\bullet\bullet}$ odpowiada wartości oczekiwanej w populacji generalnej, powstałej z połączenia populacji związanych z wszystkimi poziomami czynnika rzędowego *R* i kolumnowego *C*. Średnia $\overline{Y}_{ij\bullet}$ jest estymatorem wartości oczekiwane μ_{ij}. Średnie (17-2.25)- (17-2.27) są estymatorami wartości oczekiwanych w populacjach, które (ze względu na założoną równą liczebność w komórkach) mają kolejno postać:

$$\mu_{i\bullet} = \frac{1}{c}\sum_{j=1}^{c}\mu_{ij} \qquad \text{dla } i = 1,2,\ldots,r; \tag{17-2.28}$$

$$\mu_{\bullet j} = \frac{1}{r}\sum_{i=1}^{r}\mu_{ij} \qquad \text{dla } j = 1,2,\ldots,c; \tag{17-2.29}$$

$$\mu_{\bullet\bullet} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\mu_{ij}. \tag{17-2.30}$$

Całkowitą zmienność wartości zmiennej objaśnianej w dwuczynnikowej ANOVA, wyrażoną ogólną sumą kwadratów odchyleń (*TSS*), można rozłożyć na zmienność wyjaśnioną zmianą poziomu czynnika rzędowego *R*, następnie, wyjaśnioną zmianą poziomu czynnika kolumnowego *C*, następnie, wyjaśnioną interakcją czynnika rzędowego z kolumnowym *R* x *C* i w końcu wyjaśnioną zmiennością wewnątrz komórek (wewnątrzgrupową) ujętą losowym składnikiem błędu *E*:

$$TSS \equiv \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{\bullet\bullet\bullet})^2 =$$

$$= cn\sum_{i=1}^{r}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2 + rn\sum_{j=1}^{n}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2 + n\sum_{i=1}^{r}\sum_{j=1}^{c}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2 =$$

$$= SSR(\textit{rzędów}) + SSC(\textit{kolumn}) + SSRC(\textit{oddziaływanie } R \text{ x } C) + SSE(\textit{błąd}), \tag{17-2.31}$$

gdzie:

$$SSR = cn\sum_{i=1}^{r}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2, \tag{17-2.32}$$

$$SSC = rn\sum_{j=1}^{n}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2, \tag{17-2.33}$$

$$SSRC = n\sum_{i=1}^{r}\sum_{j=1}^{c}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2 , \qquad (17\text{-}2.34)$$

$$SSE = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2 . \qquad (17\text{-}2.35)$$

Liczba stopni swobody dla każdego z wpływów głównych ujętych w *SSR* oraz *SSC* (dla każdej wariancji międzygrupowej) to liczba poziomów czynnika pomniejszona o 1. Jest ona równa:

$\nu_r = r - 1$, gdzie *r* to liczba poziomów zmiennej w rzędach, (17-2.36)

dla czynnika *R* oraz:

$\nu_c = c - 1$, gdzie *c* to liczba poziomów zmiennej w kolumnach, (17-2.37)

dla czynnika *C*.

Liczbą stopni swobody dla wariancji międzygrupowej przy efekcie interakcji jest pomnożona liczba kolumn i wierszy po odjęciu jednego poziomu z każdego czynnika:

$$\nu_{rc} = (r-1)(c-1) . \qquad (17\text{-}2.38)$$

Wewnątrzgrupowe stopnie swobody, tak jak w przypadku jednoczynnikowej analizy wariancji, to suma stopni swobody dla wszystkich grup:

$$\nu_E = r\,c\,(n-1) . \qquad (17\text{-}2.39)$$

Suma powyższych stopni swobody jest równa ogólnej liczbie stopni swobody *ν* dla *TSS* tzn.:

$$\nu = k + \nu_E = (\nu_r + \nu_c + \nu_{rc}) + \nu_E = r\,c\,n - 1 . \qquad (17\text{-}2.40)$$

Taki *podział liczby stopni* swobody wynika z podanego w Rozdziale 17-1 dowolnego rozkładu (17.12) lub (17.13) (porównaj (5-9) w Rozdziale 5-1-2) ogólnej sumy kwadratów odchyleń na sumy częściowe *TSS* = *SS*Reg + *SSE* (17.16). Podział ten jest również widoczny z przedstawionego modelu regresji (17.16) dla dwuczynnikowej ANOVA, z którego widać, że *liczba stopni swobody dla SS*Reg, która jest związana z modelem regresji ze zmiennymi ukrytymi *X* i *Z* wraz z ich interakcją, jest równa liczbie współczynników kierunkowych , α_i ,β_j *i* γ_{ij} modelu regresji:

$$k = \nu_r + \nu_c + \nu_{rc} . \qquad (17\text{-}2.41)$$

Hipotezy zerowe dla dwuczynnikowej ANOVA.

W dwuczynnikowej ANOVA rozważane są hipotezy zerowe o (a) braku ogólnej zależności korelacyjnej oraz (b) o równości wartości oczekiwanych w rozkładach brzegowych dla czynnika *C* oraz *R* i o niewystępowaniu interakcji pomiędzy tymi czynnikami.

(a) Analiza ogólnej zależności korelacyjnej.

Z powyższego raportu widać, że statystyka testowa *F* dla tesowania hipotezy o braku ogólnej zależności korelacyjnej pomiędzy zmienną objaśnianą *Y* i wszystkimi czynnikami (czyli *R* oraz *C* z uwzględnieniem ich interakcji), czyli statystyka służąca to testowania *łącznie* trzech hipotez:

$$\begin{cases} H_0(R): \mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{r\bullet}\ , \\ H_0(C): \mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet c}\ , \\ H_0(RC): \gamma_{ij} \equiv \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} = 0 \quad \text{dla} \quad i = 1,2,\ldots,\ r\ ;\ j = 1,2,\ldots,\ c, \end{cases} \tag{17-2.42}$$

ma następującą postać [1]:

$$F = \frac{MS\text{Reg}}{MSE} \tag{17-2.43}$$

przy czym:

$$MS\text{Reg} = \frac{SS\text{Reg}}{k} \tag{17-2.44}$$

$$MSE = \frac{SSE}{\nu_E} \tag{17-2.45}$$

gdzie *MS*Reg jest średnią sumą kwadratów dla regresji, natomiast *MSE* jest średnią wariancją wewnątrzgrupową, a odpowiednie liczby stopni swobody wynoszą: dla modelu regresji $k = \nu_r + \nu_c + \nu_{rc}$, (17-2.41), a dla błędu $\nu_E = n_{\bullet\bullet} - 1 - k$. Przy prawdziwości hipotezy zerowej o jednorodności wariancji w elementarnych populacjach (czyli w komórkach dla wszystkich zestawów wartości (17.8) układów zmiennych wskazujących *X* oraz *Z*) oraz przy prawdziwości hipotezy zerowej (17-2.42) o braku ogólnej zależności korelacyjnej zmiennej *Y* od wszystkich czynników, statystyka *F* (17-2.43) ma rozkład F-Snedecora z liczbą stopni swobody licznika *k* oraz mianownika ν_E.

W przypadku, gdy hipoteza zerowa o braku ogólnej zależności korelacyjnej *Y* od czynników zostanie odrzucona, przystępujemy do testowania osobno jej hipotez składowych dla wpływów głównych oraz dla interakcji, szukając przyczyny odrzucenia hipotezy o braku ogólnej zależności korelacyjnej. Analiza ta jest przedstawiona poniżej.

(b) Analiza hipotez zerowych o równości wartości oczekiwanych dla czynników rzędowego *R* i kolumnowego *C* w ich w rozkładach brzegowych oraz o niewystępowaniu członu interakcji $R \times C$. Hipotezy te składają się na hipotezę łączną (17-2.42).

i) Hipoteza zerowa o braku głównego wpływu (efektu) ustalonego czynnika rzędowego *R*, ma postać:

$$H_0(R):\ \mu_{1\bullet} = \mu_{2\bullet} = ... = \mu_{r\bullet}. \qquad (17\text{-}2.46)$$

Hipoteza ta oznacza, że nie ma różnic pomiędzy wartościami oczekiwanymi zmiennej objaśnianej dla różnych poziomów czynnika głównego *R*.

ii) Hipoteza zerowa o braku głównego wpływu (efektu) ustalonego czynnika kolumnowego *C*, ma postać:

$$H_0(C):\ \mu_{\bullet 1} = \mu_{\bullet 2} = ... = \mu_{\bullet c}. \qquad (17\text{-}2.47)$$

Hipoteza ta oznacza, że nie ma różnic pomiędzy wartościami oczekiwanymi zmiennej objaśnianej dla różnych poziomów czynnika głównego *C*.

iii) Hipoteza zerowa o braku interakcji (oddziaływania) pomiędzy wierszami i kolumnami oznacz, że wpływ poziomu czynnika rzędowego *R* wewnątrz jakiejkolwiek kolumny jest taki sam (tzn. nie zależy od kolumny) oraz, że wpływ poziomu czynnika „kolumnowego" *C* jest taki sam wewnątrz jakiegokolwiek wiersza (tzn. nie zależy od wiersza). Hipoteza ta ma postać:

$$H_0(RC): \gamma_{ij} \equiv \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} = 0\ , \qquad \text{dla} \quad i = 1,2,\ldots, r\ ;\ j = 1,2,\ldots, c, \qquad (17\text{-}2.48)$$

co oznacza, że wszystkie wyrażenia $\gamma_{ij} \equiv \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet}$ wewnątrz sumy $\sum_{i=1}^{r}\sum_{j=1}^{c}(\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet})^2$, (której estymatorem jest *SSRC*, (17-2.34)), są równe zero.

Aby określić test statystyczny dla dwuczynnikowej analizy wariancji, musimy ocenić, czy każdy z dwóch czynników jest ustalony czy losowy. Klasyfikacja czynników zależy często od spojrzenia badacza na zagadnienie. Należy rozważyć trzy przypadki:

P1. Przypadek, gdy oba czynniki są ustalone.

P2. Przypadek, gdy oba czynniki są losowe.

P3. Przypadek czynników mieszanych, gdzie jeden czynnik jest ustalony, a drugi losowy.

P1. Oba czynniki ustalone.

a) Uogólnieniem modelu regresji (17-1-1.29) dwuczynnikowej ANOVA do modelu z interakcją czynników ustalonych ma następującą postać:

$$Y = \mu + \sum_{i=1}^{r-1}\alpha_i X_i + \sum_{j=1}^{c-1}\beta_j Z_j + \sum_{i=1}^{r-1}\sum_{j=1}^{c-1}\gamma_{ij} X_i Z_j + E\ . \qquad (17\text{-}2.49)$$

Zmienne ukryte (wskazujące) X_i, $i = 1, 2, ..., r-1$, oraz Z_j, $j = 1, 2, ..., c-1$, są opisane następującym kodowaniem (17-1-1.30):

$$X_i = \begin{cases} 1 & dla \quad poziomu \quad i-tego \quad czynnika \quad R \quad , \quad i = 1,2,...,r-1 \\ -1 & dla \quad poziomu \quad r \quad czynnika \quad R \\ 0 & w \quad pozostalych \quad przypadkach \end{cases}$$

$$Z_j = \begin{cases} 1 & dla \quad poziomu \quad j-tego \quad czynnika \quad C \quad , \quad j = 1,2,...,c-1 \\ -1 & dla \quad poziomu \quad c \quad czynnika \quad C \\ 0 & w \quad pozostalych \quad przypadkach \quad . \end{cases} \qquad (17\text{-}1\text{-}3.8')$$

Wykorzystując kodowanie (17.8), model ANOVA dla obu czynników ustalonych można zapisać następująco:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk} \ , \qquad (17\text{-}2.50)$$

gdzie zmienna losowa E_{ijk} jest błędem (resztą) związanym z k-tą obserwacją w komórce (i, j).

Parametry modelu regresji są w następujący sposób powiązane z wartościami oczekiwanymi w (pod)populacjach (porównaj (17-1-1.31)- (17-1-1.35)):

$$\mu = \mu_{\bullet\bullet} \ ; \quad \alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet} \quad \text{dla} \quad i = 1,2,...,r; \qquad \beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet} \quad \text{dla} \quad j = 1,2,\ldots,c \ ;$$

$$\gamma_{ij} = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} \quad \text{dla} \quad i = 1,2,\ldots,r \ ; \quad j = 1,2,\ldots,c \ , \qquad (17\text{-}2.51)$$

gdzie: $\mu_{i\bullet} = \frac{1}{c}\sum_{j=1}^{c}\mu_{ij}$, (17-2.28), $\mu_{\bullet j} = \frac{1}{r}\sum_{i=1}^{r}\mu_{ij}$, (17-2.29), $\mu_{\bullet\bullet} = \frac{1}{cr}\sum_{i=1}^{r}\sum_{j=1}^{c}\mu_{ij}$, (17-2.30), przy czym stałe α_i, β_j, oraz γ_{ij} spełniają zależności:

$$\sum_{i=1}^{r}\alpha_i = 0, \ \sum_{j=1}^{c}\beta_j = 0, \ \sum_{i=1}^{r}\gamma_{ij} = 0, \ \sum_{j=1}^{c}\gamma_{ij} = 0 \ . \qquad (17\text{-}2.52)$$

Widzimy, że tak jak dla jednokierunkowej analizy wariancji, model ANOVA może być zapisany bądź w postaci zawierającej składnik równania regresji (17-2.49) ze zmiennymi ukrytymi, bądź w postaci sumy różnych kombinacji parametrów strukturalnych tego modelu regresji (17-2.50), która jest bezpośrednio równa (zgodnie z (17-2.51)) wartości oczekiwanej zmiennej objaśnianej Y w populacji oznaczonej parą wskaźników (i, j):

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \ . \qquad (17\text{-}2.53)$$

Zatem model ANOVA dla obu czynników ustalonych ma postać (17-2.53). Poniżej przedstawione są modele ANOVA zapisane w postaci kombinacji sumy parametrów strukturalnych (lub odpowiadających im zmiennych losowych dla poziomów) dla przypadków obu czynników losowych i dla czynników mieszanych.

P2. Oba czynniki losowe.

Model ANOVA ma postać:

$$Y_{ijk} = \mu + A_i + B_j + C_{ij} + E_{ijk} \ , \qquad i = 1,2,\ldots, r \ ; \ j = 1,2,\ldots, c \ ; \ k = 1, 2, \ldots, n, \qquad (17\text{-}2.54)$$

gdzie zmienne losowe A_i, B_j, C_{ij} , oraz E_{ijk} są wzajemnie niezależne i mają rozkłady:

$$A_i\text{: N}(0, \sigma_R^2), \quad B_j\text{: N}(0, \sigma_C^2), \quad C_{ij}\text{: N}(0, \sigma_{RC}^2) \ , \quad E_{ijk}\text{: N}(0, \sigma_E^2). \qquad (17\text{-}2.55)$$

Wariancje σ_R^2, σ_C^2, σ_{RC}^2, σ_{RC}^2 są właściwymi miarami rozproszeń odpowiednich *warunkowych* wartości oczekiwanych zmiennej zależnej *Y* (porównaj tekst przed (16-4.22)), związanymi kolejno z wpływami głównymi czynnika rzędowego *R* i czynnika kolumnowego *C* oraz wpływem oddziaływania *RC*.

P3. Czynniki mieszane.

a) Model ANOVA ma postać:

$$Y_{ijk} = \mu + \alpha_i + B_j + C_{ij} + E_{ijk}\ , \qquad i = 1,2,\ldots, r\ ;\ j = 1,2,\ldots, c\ ;\ k = 1, 2, \ldots, n\ , \qquad (17\text{-}2.56)$$

gdzie czynniki są mieszane, tzn. czynnik rzędowy *R* jest ustalony, a czynnik kolumnowy *C*, losowy.
Zatem ze względu na ustalony czynnik *R* zachodzi warunek:

$$\sum_{i=1}^{r} \alpha_i = 0, \qquad (17\text{-}2.57)$$

natomiast zmienne losowe B_j, C_{ij} oraz E_{ijk} są wzajemnie niezależne i mają rozkłady:

$$B_j\text{: N}(0,\ \sigma_C^2),\ C_{ij}\text{: N}(0,\ \sigma_{RC}^2)\ ,\ E_{ijk}\text{: N}(0,\ \sigma_E^2)\ . \qquad (17\text{-}2.58)$$

b) Model ANOVA ma postać:

$$Y_{ijk} = \mu + A_i + \beta_j + C_{ij} + E_{ijk}\ , \qquad i = 1,2,\ldots, r\ ;\ j = 1,2,\ldots, c\ ;\ k = 1, 2, \ldots, n\ , \qquad (17\text{-}2.59)$$

gdzie tym razem czynniki są mieszane, tzn. czynnik rzędowy *R* jest losowy, a czynnik kolumnowy *C*, ustalony. Ze względu na ustalony czynnik *C* zachodzi warunek:

$$\sum_{j=1}^{c} \beta_j = 0, \qquad (17\text{-}2.60)$$

natomiast zmienne losowe A_i, C_{ij}, oraz E_{ijk} są wzajemnie niezależne i mają rozkłady:

$$A_i\text{: N}(0,\ \sigma_R^2),\ C_{ij}\text{: N}(0,\ \sigma_{RC}^2)\ ,\ E_{ijk}\text{: N}(0,\ \sigma_E^2)\ . \qquad (17\text{-}2.61)$$

Postać hipotez zerowych.

Odpowiednie hipotezy zerowe dla przypadków P1, P2 i P3 zebrano w Tabeli 17-2.2.

Tabela 17-2.2. Tablica hipotez zerowych dla dwuczynnikowej analizy wariancji [1].

Źródła Zmienności *Y*	**P1.** czynniki ustalone H_0:	***P2.*** *czynniki losowe* H_0:	***P3.*** *modele mieszane*	
			R ustalony, *C* losowy H_0:	*R* losowy, *C* ustalony H_0:
Czynnik rzędowy *R*	$\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ lub $\mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{r\bullet}$	$\sigma_R^2 = 0$	$\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ lub $\mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{r\bullet}$	$\sigma_R^2 = 0$
Czynnik kolumnowy *C*	$\beta_1 = \beta_2 = \ldots = \beta_c = 0$ lub $\mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet c}$	$\sigma_C^2 = 0$	$\sigma_C^2 = 0$	$\beta_1 = \beta_2 = \ldots = \beta_c = 0$ lub $\mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet c}$
Oddziaływanie *RC*	$\gamma_{ij} = 0$, dla wszystkich *i*, *j*	$\sigma_{RC}^2 = 0$	$\sigma_{RC}^2 = 0$	$\sigma_{RC}^2 = 0$

Testy statystyczne do weryfikacji hipotez zerowych (17-2.46)-(17-2.48).

W porównaniu z jednokierunkową ANOVA, w przypadku dwuczynnikowej i wieloczynnikowej ANOVA występuje różnica w postaci hipotez statystycznych, które mają, np. dla czynników ustalonych postać $H_0(R)$: $\mu_{1\bullet} = \mu_{2\bullet} = ... = \mu_{r\bullet}$, $H_0(C)$: $\mu_{\bullet 1} = \mu_{\bullet 2} = ... = \mu_{\bullet c}$ oraz $H_0(RC)$: $\gamma_{ij} \equiv \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} = 0$. Różne są też postacie stosowanych statystyk testowych. Jednak rozumowanie, które doprowadziło w jednoczynnikowej ANOVA *od* zależności (16-4.24):

$$\frac{\mu_{MSG}}{\mu_{MSE}} \equiv \frac{E(MSG)}{E(MSE)} = \frac{\sigma_E^2 + n_0\,\sigma_A^2}{\sigma_E^2} \qquad \text{(16-4.24')}$$

do postaci statystyki testowej $F = MSG / MSE$, (16-4.24), można uogólnić na przypadek dwuczynnikowej ANOVA. Poniżej podamy jego najistotniejsze rezultaty.

W celu lepszego uchwycenia istoty hipotez zerowych oraz odpowiednich testów statystycznych *F* (zebranych dalej w Tablicy 17-2-4), warto porównać Tabele 17-2-3 i 17-2-4 z Tabelą 16-4.1 dla jednoczynnikowej ANOVA oraz z (dalszą) Tabelą 17-3.2 dla ANOVA z losowo dobieranymi blokami. W Tabeli 17-2-3 podano wartości oczekiwane liczników i mianowników (średnich sum kwadratów) odpowiednich statystyk testowych *F* dla dwuczynnikowej ANOVA.

Tabela 17-2.3. Wartości oczekiwane średnich sum kwadratów dla dwuczynnikowej ANOVA [1].

Źródła Zmienności *Y*	*Wartości oczekiwane średnich kwadratów E(MS) dla liczników i mianowników statystyk* F			
	P1. Czynnik ustalony	**P2.** Czynnik losowy	**P3.** *R*-ustalony, *C*-losowy	**P3.** *R*-losowy, *C*-ustalony
Czynnik rzędowy *R*	$E(MSR) = \sigma_E^2 + \frac{c\,n}{r-1}\sum_{i=1}^{r}(\mu_{i\bullet} - \mu_{\bullet\bullet})^2 = \sigma_E^2 + \frac{c\,n}{r-1}\sum_{i=1}^{r}\alpha_i^2$	$\sigma_E^2 + n\,\sigma_{RC}^2 + c\,n\,\sigma_R^2$	$\sigma_E^2 + n\,\sigma_{RC}^2 + \frac{c\,n}{r-1}\sum_{i=1}^{r}\alpha_i^2$	$\sigma_E^2 + n\,\sigma_{RC}^2 + c\,n\,\sigma_R^2$
Czynnik kolumnowy *C*	$E(MSC) = \sigma_E^2 + \frac{r\,n}{c-1}\sum_{j=1}^{c}(\mu_{\bullet j} - \mu_{\bullet\bullet})^2 = \sigma_E^2 + \frac{r\,n}{c-1}\sum_{j=1}^{c}\beta_j^2$	$\sigma_E^2 + n\,\sigma_{RC}^2 + r\,n\,\sigma_C^2$	$\sigma_E^2 + n\,\sigma_{RC}^2 + r\,n\,\sigma_C^2$	$\sigma_E^2 + n\,\sigma_{RC}^2 + \frac{r\,n}{c-1}\sum_{j=1}^{c}\beta_j^2$
Oddziaływanie	$E(MSRC) = \sigma_E^2 + \frac{n}{(r-1)(c-1)}\sum_{i=1}^{r}\sum_{j=1}^{c}\gamma_{ij}^2$	$\sigma_E^2 + n\,\sigma_{RC}^2$	$\sigma_E^2 + n\,\sigma_{RC}^2$	$\sigma_E^2 + n\,\sigma_{RC}^2$
Błąd *E*	σ_E^2	σ_E^2	σ_E^2	σ_E^2

Analizując Tabelę 17-2.3 oraz hipotezy zerowe zebrane w Tabeli 17-2.2, można dostrzec postać odpowiednich testów *F* podanych poniżej.

Założenia dla testów F.

Aby użyć testu F muszą być spełnione następujące założenia:

a. Dla modeli z ustalonymi czynnikami obserwacje Y_{ijk} są statystycznie niezależne jedne od drugich. (Założenie to nie jest spełnione wtedy, gdy w model dwuczynnikowej ANOVA z równą liczbą obserwacji w komórkach zostają włączone czynniki losowe, gdyż obserwacje Y_{ijk} są wtedy wzajemnie zależne.)
b. Każda obserwacja pochodzi z populacji o rozkładzie normalnym.
c. Każda populacja związana z konkretną komórką ma taką samą wariancję (czyli zakładamy, że wariancja jest jednorodna).

W weryfikacji hipotez zerowych w ANOVA wykorzystuje się test *F*:

a. dla wpływu głównego czynnika *R* (poziomy w rzędach),
b. dla wpływu głównego czynnika *C* (poziomy w kolumnach),
c. dla interakcji „*RC*" między czynnikami.

W liczniku każdej z tych statystyk testowych *F* występuje wariancja międzygrupowa, która odnosi się do porównań między średnimi dla danego wpływu głównego lub interakcji. W mianowniku znajduję się wariancja wewnątrzgrupowa, która we wszystkich testach *F* z czynnikiem ustalonym jest średnią sumą kwadratów odchyleń wszystkich wyników od średnich w odpowiadających im grupach (czyli od średnich wewnątrzgrupowych).
Jednakże, w przypadku czynnika losowego, wariancja wewnątrzgrupowa jest mianownikiem statystyki *F* jedynie w przypadku testowania interakcji. Jak o tym powiemy poniżej, w przypadku testowania wpływów głównych w mianowniku statystyki *F* występuje średnia suma kwadratów dla interakcji.

Wpływy główne w dwuczynnikowej ANOVA obliczamy poprzez ustalenie stosunku *F* w kolumnach bądź w wierszach. Jak wspomnieliśmy, z postaci wartości oczekiwanych statystyk podanych w Tabeli 17-2.3 i wchodzących w skład rozkładu *TSS*, można podać postać statystyki testowej *F*, która zależy od tego czy czynnik jest losowy czy ustalony. Możemy więc rozważyć następujące sytuacje:

1. Gdy hipotezy zerowe są jak w przypadku P1 (kolumna pierwsza w Tabeli 17-2.2). Zarówno czynnik rzędowy *R* jak i kolumnowy *C* jest ustalony. Testy *F* obliczamy według następujących wzorów (podano też odpowiednie liczby stopni swobody licznika i mianownika statystyki *F* (17-2.36)-(17-2.39)):

$$F(R) = \frac{MSR}{MSE}\ , \qquad \text{z } \nu_r = r-1 \text{ licznika i } \nu_E = r\,c\,(n-1) \text{ mianownika,} \qquad (17\text{-}2.62a)$$

$$F(C) = \frac{MSC}{MSE}, \qquad \text{z } \nu_c = c-1 \text{ licznika i } \nu_E = r\,c\,(n-1) \text{ mianownika,} \qquad (17\text{-}2.62b)$$

$$F(RC) = \frac{MSRC}{MSE} \qquad \text{z } \nu_{rc} = (r-1)(c-1) \text{ licznika i } \nu_E = r\,c\,(n-1) \text{ mianownika.} \qquad (17\text{-}2.62c)$$

2. Gdy hipotezy zerowe są jak w przypadku P2 lub P3 (kolumna druga lub trzecia w Tabeli 17-2.2). Zarówno czynniki rzędowy R jak i kolumnowy C jest losowy lub zachodzi przypadek mieszany, tzn. jeden z czynników jest losowy a drugi ustalony. Testy F obliczamy według następujących wzorów:

$$F(R) = \frac{MSR}{MSRC}, \qquad \text{z } \nu_r = r-1 \text{ licznika i } \nu_{rc} = (r-1)(c-1) \text{ mianownika,} \qquad (17\text{-}2.63a)$$

$$F(C) = \frac{MSC}{MSRC}, \qquad \text{z } \nu_c = c-1 \text{ licznika i } \nu_{rc} = (r-1)(c-1) \text{ mianownika,} \qquad (17\text{-}2.63b)$$

$$F(RC) = \frac{MSRC}{MSE}, \qquad \text{z } \nu_{rc} = (r-1)(c-1) \text{ licznika i } \nu_E = r\,c\,(n-1) \text{ mianownika.} \qquad (17\text{-}2.63c)$$

Jak widać, statystyka do testowania hipotezy o braku interakcji jest taka sama w obu przypadkach.

Tabela 17-2.4. Tablica dwuczynnikowej analizy ANOVA [1].

Źródła zmienności Y	Liczba stopni swobody	Suma kwadratów odchyleń	Średni kwadrat odchylenia	Postać statystyki testowej F	
				P1. Dla wpływu ustalonego	**P2 i P3.** Dla wpływu losowego i mieszanego
Czynnik rzędowy R (wpływ główny)	$\nu_r = r-1$	SSR	$MSR = \frac{SSR}{(r-1)}$	$MSR/\,MSE$	$MSR/\,MSRC$
Czynnik kolumnowy C (wpływ główny)	$\nu_c = c-1$	SSC	$MSC = \frac{SSC}{(c-1)}$	$MSC/\,MSE$	$MSC/\,MSRC$
Oddziaływanie R x C	$\nu_{rc} = (r-1)(c-1)$	$SSRC$	$MSRC = \frac{SSRC}{(r-1)(c-1)}$	$MSRC/\,MSE$	$MSRC/\,MSE$
Błąd E	$\nu_E = r\,c\,(n-1)$	SSE	$MSE = \frac{SSE}{rc(n-1)}$		
Razem	$\nu = r\,c\,n-1$	TSS			

Podsumujmy. Podobnie jak pod koniec Rozdziału 16-4 dla jednoczynnikowej ANOVA tak i teraz, zauważyć można, że w powyższych rozważaniach (podsumowanych w Tabelach 17-2.4 i 17-2.3) zwraca uwagę związek postaci ilorazów wartości oczekiwanych średnich kwadratów odchyłek z postacią hipotez zerowych H_0 (Tabela 17-2.2) i ich wpływ tak na postać statystyk testowych F jak i wartości jakie, w przypadku prawdziwości H_0, statystyki F na ogół przyjmują.

Rozdział 17-2-1. Przykład: „wydolność płuc”

Pewne przedsiębiorstwo postanowiło sprawdzić jaka jest wydolność płuc (WP) pracowników w nim pracujących, którzy są poddani wpływowi *jednej z trzech* możliwych różnych substancji toksycznych (c = 3). Tak się złożyło, że pracownicy tegoż przedsiębiorstwa mieszkali w trzech różnych dzielnicach, a w każdej z nich zasadzono inny gatunek rośliny (trawy). W związku z tym, rozważano także wpływ na wydolność płuc pracowników tych trzech (r = 3) gatunków roślin.
Bardzo niska wydolność płuc pracowników (mała wartość WP) świadczy o zaburzeniach w oddychaniu, zaś wysoka wydolność płuc (duża wartość WP) oznacza brak problemów z oddychaniem.
Dla każdego gatunku rośliny (trawy) i dla każdego rodzaju substancji toksycznej pobrano próbkę n = 12 osób (tzn. jest 3x3 = 9 populacji, z których poprano próbki 12 osobowe).

Zadanie. Przeprowadzić analizę przykładu w SAS, odpowiadając na następujące pytania:

1) Czy rośliny mają istotny wpływ na wydolność płuc pracowników?
2) Czy toksyczne substancje mają istotny wpływ na wydolność płuc pracowników?
3) Czy taka sama jest zmiana wydolności płuc podczas zmiany substancji toksycznej w ramach wpływu konkretnego gatunku roślin. Występowanie tych różnic w zmianie średniego poziomu wydolności płuc byłyby świadectwem występowania interakcji. (Zbadać wpływ interakcji gatunków roślin i typu substancji toksycznej na wydolność płuc.)

Tabela. 17-2-1.1. Dane dla przykładu „wydolność płuc” [1] (Wyliczono również sumy i średnie).

	Rodzaj toksycznej substancji *C*			
Gatunek rośliny *R*	a	b	c	Sumy i średnie w rzędach
1	4,64 5,92 5,25 6,17 4,20 5,90 5,07 4,13 4,07 5,30 4,37 3,76 $Y_{11\bullet} = 58{,}78$ $\bar{Y}_{11\bullet} = 4{,}90$	3,21 3,17 3,88 3,50 2,47 4,12 3,51 3,85 4,22 3,07 3,62 2,95 $Y_{12\bullet} = 41{,}57$ $\bar{Y}_{12\bullet} = 3{,}46$	3,75 2,50 2,65 2,84 3,09 2,90 2,62 2,75 3,10 1,99 2,42 2,37 $Y_{13\bullet} = 32{,}98$ $\bar{Y}_{13\bullet} = 2{,}75$	$Y_{1\bullet\bullet} = 133{,}33$ $\bar{Y}_{1\bullet\bullet} = 3{,}70$
2	5,12 6,10 4,85 4,72 5,36 5,41 5,31 4,78 5,08 4,97 5,85 5,26 $Y_{21\bullet} = 62{,}81$ $\bar{Y}_{21\bullet} = 5{,}23$	3,92 3,75 4,01 4,64 3,63 3,46 4,01 3,39 3,78 3,51 3,19 4,04 $Y_{22\bullet} = 45{,}33$ $\bar{Y}_{22\bullet} = 3{,}78$	2,95 3,21 3,15 3,25 2,30 2,76 3,01 2,31 2,50 2,02 2,64 2,27 $Y_{23\bullet} = 32{,}37$ $\bar{Y}_{23\bullet} = 2{,}70$	$Y_{2\bullet\bullet} = 140{,}51$ $\bar{Y}_{2\bullet\bullet} = 3{,}90$
3	4,64 4,32 4,13 5,17 3,77 3,85 4,12 5,07 3,25 3,49 3,65 4,10 $Y_{31\bullet} = 49{,}56$ $\bar{Y}_{31\bullet} = 4{,}13$	4,95 5,22 5,16 5,35 4,35 4,89 5,61 4,98 5,77 5,23 4,86 5,15 $Y_{32\bullet} = 61{,}52$ $\bar{Y}_{32\bullet} = 5{,}13$	2,95 2,80 3,63 3,85 2,19 3,32 2,68 3,35 3,12 4,11 2,90 2,75 $Y_{33\bullet} = 37{,}65$ $\bar{Y}_{33\bullet} = 3{,}14$	$Y_{3\bullet\bullet} = 148{,}73$ $\bar{Y}_{3\bullet\bullet} = 4{,}13$
Sumy i średnie w kolumnach	$Y_{\bullet 1\bullet} = 171{,}15$ $\bar{Y}_{\bullet 1\bullet} = 4{,}75$	$Y_{\bullet 2\bullet} = 148{,}42$ $\bar{Y}_{\bullet 2\bullet} = 4{,}12$	$Y_{\bullet 3\bullet} = 103{,}00$ $\bar{Y}_{\bullet 3\bullet} = 2{,}86$	$Y_{\bullet\bullet\bullet} = 422{,}57$ $\bar{Y}_{\bullet\bullet\bullet} = 3{,}91$

Rozwiązanie.

Mamy dwa czynniki główne. Różne gatunki rozważanych roślin (rzędy), oznaczające trzy różne poziomy czynnika „głównego” *R*, oznaczymy numerami 1, 2, 3. Różne rodzaje substancji toksycznych oznaczające trzy różne poziomy czynnika „głównego” *C*, oznaczymy jako a, b, c. Wszystkie obserwowane wartości WP zostały zawarte powyżej w **Tabeli 17-2-1.1**.

Ad1) Hipoteza zerowa dla „rzędowych” (brzegowych) wartości oczekiwanych czynnika rzędowego *R*, którym jest gatunek rośliny, ma postać:

$$H_0(R):\ \mu_{1\bullet} = \mu_{2\bullet} = \mu_{3\bullet}$$

Oznacza ona brak głównego wpływu od czynnika *R*, czyli gatunku rośliny na WP pracowników.

Ad 2) Hipoteza zerowa dla „kolumnowych” (brzegowych) wartości oczekiwanych czynnika kolumnowego *C*, którym jest rodzaju substancji toksycznej, ma postać:

$$H_0(C):\ \mu_{\bullet a} = \mu_{\bullet b} = \mu_{\bullet c}$$

Oznacza ona brak głównego wpływu od czynnika *C*, czyli rodzaju toksycznej substancji na WP pracowników.

Ad3) Hipoteza zerowa o braku interakcji pomiędzy czynnikami ma postać:

$$H_0(R\times C): \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} = 0, \quad \text{dla każdej pary } i, j,$$

co można zapisać również następująco:

$$H_0(R\times C): \mu_{ij} - \mu_{i\bullet} = \mu_{\bullet j} - \mu_{\bullet\bullet}, \quad \text{dla każdej pary } i, j.$$

Hipoteza ta oznacza brak zmiany WP na skutek zmiany substancji toksycznej (określonej indeksem *j*) dla wpływu konkretnego gatunku rośliny *i*, przy zmianie gatunku rośliny *i*. Istnienie takiej zmiany średniego poziomu WP jest świadectwem występowania interakcji.

Gdy zamienimy rolami czynnik *R* oraz *C*, wtedy hipotezę tą można zapisać następująco:

$$H_0(R\times C): \mu_{ij} - \mu_{\bullet j} = \mu_{i\bullet} - \mu_{\bullet\bullet}, \quad \text{dla każdej pary } i, j,$$

co oznacza brak zmiany WP na skutek zmiany gatunku roślin (określonej indeksem *i*) w ramach wpływu konkretnej substancji toksycznej *j*, wraz ze zmianą substancji toksycznej *j*.

Polecenia w pakiecie SAS'a.

1) Tworzymy tabelę danych na podstawie, których przeprowadzamy dwuczynnikową ANOVA. Tworzenie tabeli z danymi:

Tools → Table Editor.

2) Zapisujemy tabelę z danymi w wybranej bibliotece. Następnie rozpoczynamy analizę, wybierając z paska Menu polecenia:

Solutions → Analysis → Analyst.

3) Otwieramy wcześniej zapisane dane stosując polecenia:

Edit → Open By SAS Name→ wskazujemy bibliotekę w której zostały zapisane dane → wybieramy dane dla przykładu.

4) Dokonujemy dwuczynnikowej analizy wariancji:

Statistics → ANOVA → Factorial ANOVA → WP (Dependent); Gat.roślin, Rodz.Subst (Independent)

a) Model → Standard Model → Effects up to 2 – way interaction.

b) Tests: Error → Gat.roślin * Rodz.substancji

Add → Gat.roślin; Rodz.substancji

c) Statistics → Type I, III

d) Means → Comparison method → Scheffe's multiple – comparison procedure

Breakdown → zaznaczamy: Mean; Std.dev.; Variance; Num.Obs; Minimum; Maximum

e) Plots: Means → Plot dependent means for main effect; → Plot dependent means for two - way effect

Residual → Plot residual vs variable; Ordinary; Predicted Y; Independents

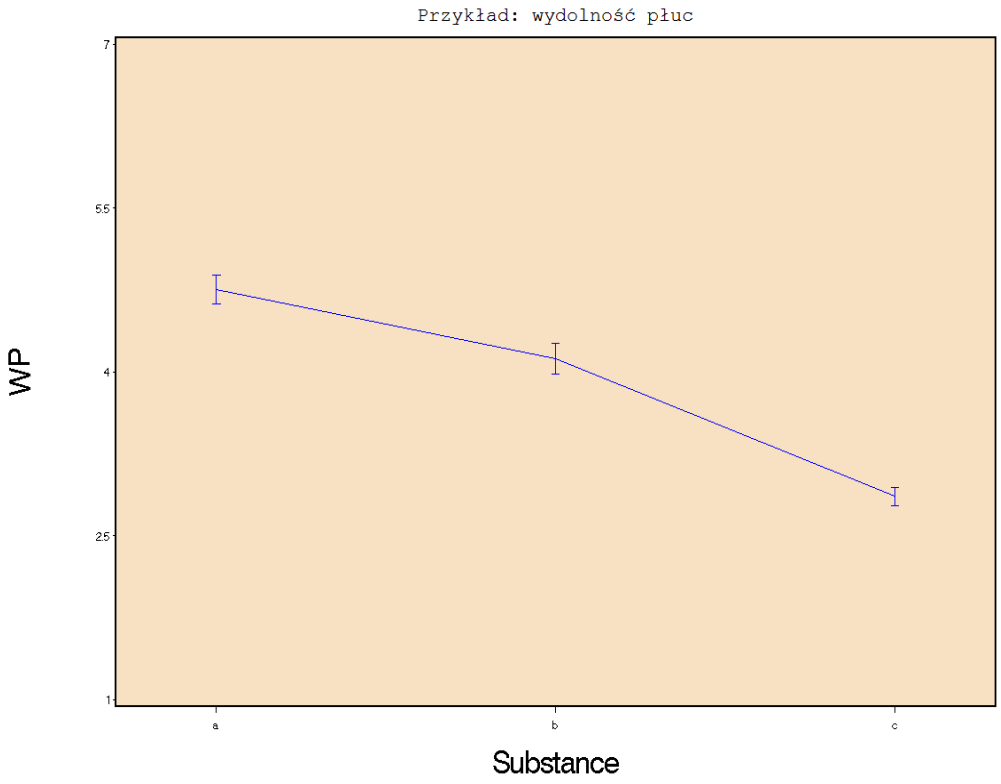
Influence → Plot influence statistics vs variable; Deffits; Predicted Y; Independents

f) Save Data → Create and save diagnostics data; PREDICTED Predicted Values

Wykresy i ich omówienie.

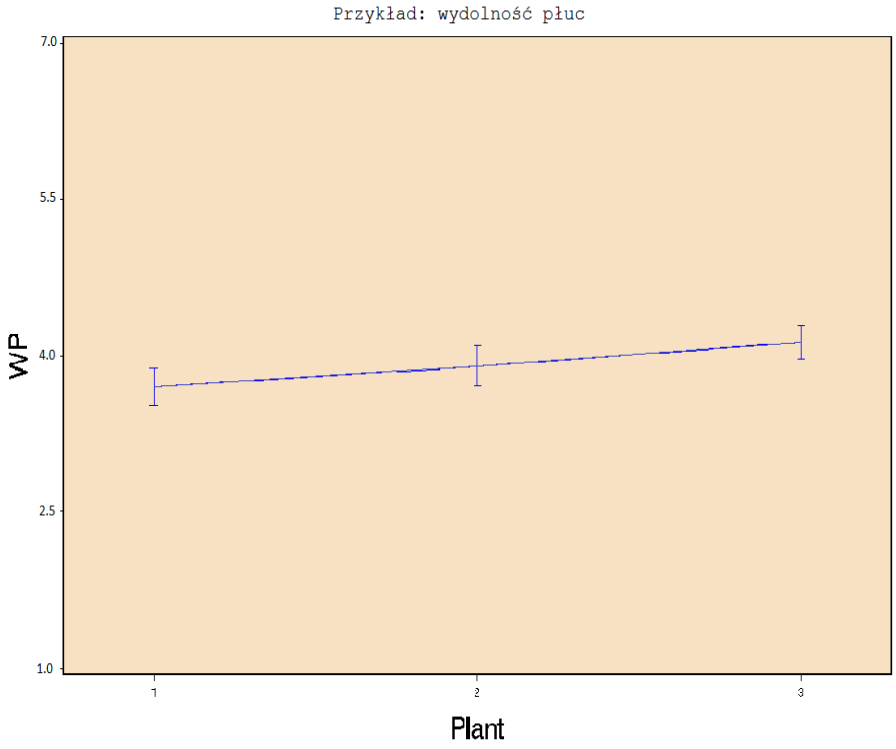
Zanim przystąpimy do analizy numerycznych raportów SAS,a przyjrzyjmy się wykresom pomagającym zorientować się co do związków pomiędzy średnimi. We wstępnym omówieniu sytuacji w pobranej próbce (zgrupowanej w dziewięciu komórkach dla trzech gatunków roślin (1,2,3) i trzech substancji toksycznych a,b,c), odwołamy się do wykresów wygenerowanych przy pomocy pakietu Analyst.

Wykres 1. Poniższy wykres przedstawia empiryczną linię regresji, która wyraża zależności średniej wydolności płuc (WP) od poziomu czynnika głównego, którym jest rodzaj substancji toksycznej (a,b,c) (czynnik *C*). Na wykresie zaznaczono również odchylenia standardowe od wartości średnich „kolumnowych" (brzegowych) dla poziomu a,b,c, wpływu głównego substancji.



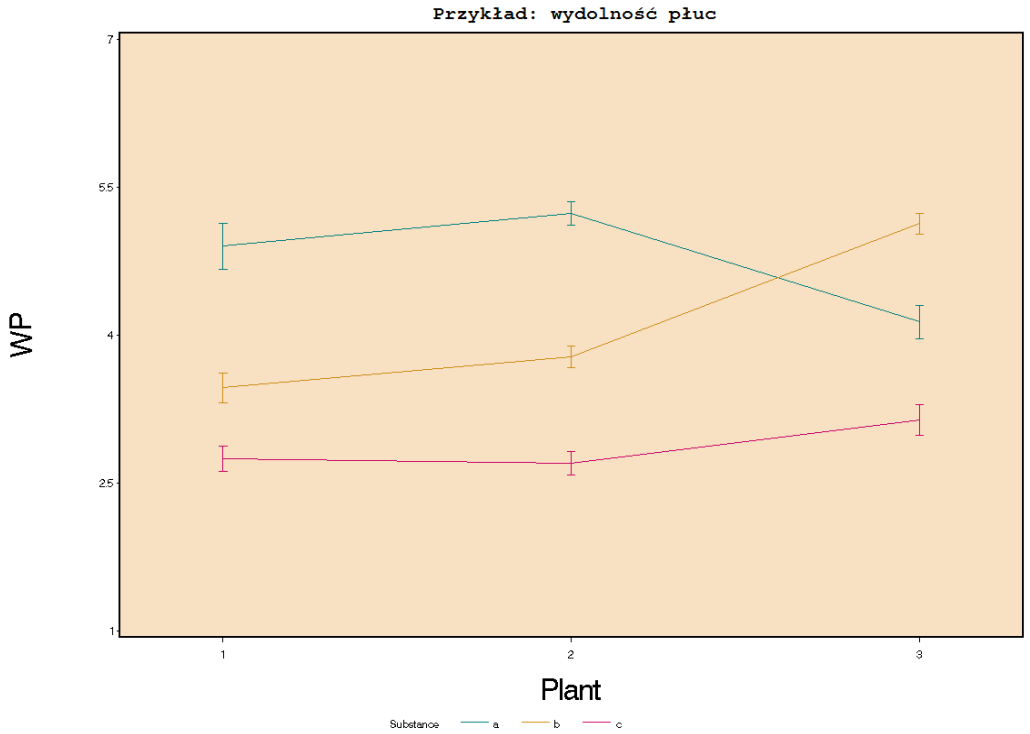
Na podstawie powyższego wykresu można by wyciągnąć wniosek, że zmiana średniej wartości WP wraz z typem substancji toksycznej jest istotna statystycznie. Wniosek ten zgadza się z dalszą analizą numeryczną.

Wykres 2. Poniższy wykres przedstawia empiryczną linię regresji, dla zależności średnich „rzędowych” (brzegowych) WP od poziomu czynnika głównego, którym tym razem jest gatunek rośliny (1,2,3) (czynnik *R*). Na wykresie zaznaczono również odchylenia standardowe od wartości średnich „rzędowych” dla poziomu 1,2,3, wpływu głównego gatunku rośliny.



Na podstawie powyższego wykresu można by wyciągnąć wniosek, że zmiana średniej wartości WP wraz z gatunkiem rośliny jest również istotna statystycznie. Jednakże, chociaż wniosek ten zgadza się z dalszą analizą numeryczną, to nie jest on już tak wyraźny jak w przypadku wpływu substancji toksycznej.

Wykres 3. Poniższy wykres przedstawia empiryczne linie regresji dla rozkładu zmiennej WP wewnątrz tabeli danych. Linie wyrażają zależności średniej WP od poziomu dwóch czynników, którymi są: rodzaj substancji toksycznej (a,b,c, oznaczonej kolorem) oraz gatunek rośliny (1,2,3). Na wykresie zaznaczono odchylenia standardowe od wartości średnich w komórkach dla poziomów a,b,c oraz 1,2,3.



Można zauważyć, mniejszy wpływ gatunku rośliny na średnią wartość WP w przypadku substancji toksycznej c (niż dla pozostałych dwóch substancji a i b, dla których wpływ ten jest również różny). Sugeruje to występowanie interakcji pomiędzy czynnikami (potwierdzonej poniżej w teście statystycznym).

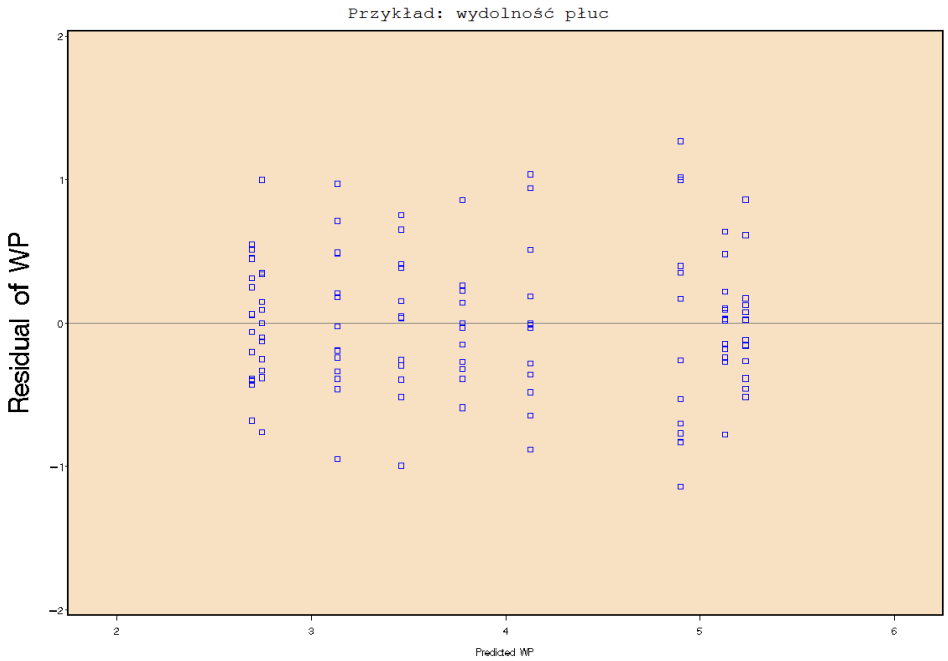
Wykres 4.**Test jednorodności wariancji w komórkach**.

ANOVA zakłada konieczność jednorodności wariancji w populacjach. W celu sprawdzenia tej hipotezy dla $r \times c = 9$ rozważanych populacji (po jednej dla każdej komórki wewnątrz tablicy danych , Tabela. 17-2-1.1.), przeprowadzamy odpowiednie testy statystyczne. Korzystając z procedur dla *jednokierunkowej* ANOVA, skonstruowanej tak, aby każda z $r \times c = 9$ - ciu komórek tablicy danych odpowiadała jednemu z 9-ciu możliwych poziomów nowego czynnika o poziomach numerowanych kolejnymi komórkami, stwierdzamy w oparciu o test Bartlett'a, że nie ma podstaw do odrzucenia hipotezy o jednorodności wariancji.

Bartlett's Test for Homogeneity of WP Variance

Source	DF	Chi-Square	Pr > ChiSq
Plant_Subst_komorki	8	12.3778	0.1351

Dla $\alpha = 0{,}01$ wniosek płynący z testu Brown'a-Forsythe'a (dla którego p = 0.0194) byłby taki sam jak dla testu Bartlett'a, natomiast test Levene'ego (dla którego p = 0.0052) daje inny wniosek. Uznajemy, na podstawie testu Bartlett'a, że nie ma podstaw do odrzucenia hipotezy zerowej o jednorodności wariancji w populacjach dla komórek. Wniosek ten wydaje się być zgodny z poniższym rysunkiem dla rozproszeń w 9 komórkach próby.



Nie mając (jednoznacznych) podstaw do odrzucenia hipotezy zerowej o jednorodności wariancji, przystępujemy do ANOVA dla weryfikacji hipotez o równości brzegowych wartości oczekiwanych oraz o braku interakcji.

Analiza numeryczna podsumowana jest poniższymi raportami SAS'a. W trakcie czytania tego raportu wstawiono odpowiednie komentarze dotyczące weryfikowanych hipotez statystycznych.

Raport SAS'a.

W poniższej części raportu zamieszczono dodatkowe informacje dotyczące wartości charakterystyk opisowych w pobranej próbie pracowników. Ze względu na jego oczywistość pominiemy komentarz.

```
Przyklad_"wydolność płóc"                                    15:49 Saturday, April 17, 2004

                    Breakdown of Means and Other Descriptive Statistics

---------------------------------------- Effect=Overall ----------------------------------------

                                        Std.                    Number
Rodzaj_     Gatunek_       Mean        Dev.     Variance     Non-missing     Minimum     Maximum
substancji  rosliny       of WP      of WP       of WP         of WP          of WP       of WP

                         3.912685  1.064615  1.13340        108          1.99        6.17
```

---------- Effect=GATUNEK_ROSLINY ----------

Rodzaj_ substancji	Gatunek_ rosliny	Mean of WP	Std. Dev. of WP	Variance of WP	Number Non-missing of WP	Minimum of WP	Maximum of WP
	1	3.703611	1.085629	1.17859	36	1.99	6.17
	2	3.903056	1.125806	1.26744	36	2.02	6.1
	3	4.131389	0.961612	0.92470	36	2.19	5.77

---------- Effect=RODZAJ_SUBSTANCJI ----------

Rodzaj_ substancji	Gatunek_ rosliny	Mean of WP	Std. Dev. of WP	Variance of WP	Number Non-missing of WP	Minimum of WP	Maximum of WP
a		4.754167	0.772112	0.59616	36	3.25	6.17
b		4.122778	0.841297	0.70778	36	2.47	5.77
c		2.861111	0.499919	0.24992	36	1.99	4.11

---------- Effect=RODZAJ_SUBSTANCJI*GATUNEK_ROSLINY ----------

Rodzaj_ substancji	Gatunek_ rosliny	Mean of WP	Std. Dev. of WP	Variance of WP	Number Non-missing of WP	Minimum of WP	Maximum of WP
a	1	4.898333	0.819344	0.67132	12	3.76	6.17
a	2	5.234167	0.416358	0.17335	12	4.72	6.1
a	3	4.13	0.594276	0.35316	12	3.25	5.17
b	1	3.464167	0.514348	0.26455	12	2.47	4.22
b	2	3.7775	0.385112	0.14831	12	3.19	4.64
b	3	5.126667	0.369455	0.13650	12	4.35	5.77
c	1	2.748333	0.446091	0.19900	12	1.99	3.75
c	2	2.6975	0.418897	0.17547	12	2.02	3.25
c	3	3.1375	0.542505	0.29431	12	2.19	4.11

Poniższa część raportu dotyczy weryfikacji hipotez zerowych o a) braku ogólnej zależności korelacyjnej b) o równości wartości oczekiwanych w rozkładach brzegowych dla czynnika *C* oraz *R* oraz o niewystępowaniu interakcji pomiędzy czynnikami.

Rozważmy wyniki zawarte w poniższym raporcie SAS'a.

Przyklad_"wydolność płóc" 15:49 Saturday, April 17, 2004

The GLM Procedure
Class Level Information

Class	Levels	Values
Rodzaj_substancji	3	a b c
Gatunek_rosliny	3	1 2 3

Number of observations 108

The GLM Procedure
Dependent Variable: WP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**$(k-1)=$** 8	94.6984630	11.8373079 **(*MSR*)**	44.10	<.0001
Error	**$(n_{**}-k)=$** 99	26.5758583	**0.2684430 (*MSE*)**		
Corrected Total	**$(n_{**}-1)=$** 107	121.2743213		**gdzie:** $k=\nu_r+\nu_c+\nu_{rc}$	

R-Square	Coeff Var	Root MSE	WP Mean
0.780862	13.24193	0.518115	3.912685

(R^2=0.78 **więc siła związku liniowego jest stosunkowo duża)**

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gatunek_rosliny (*R*)	2	3.29889630	**1.64944815**	6.14	0.0031
Rodzaj_substancji (*C*)	2	66.88936852	**33.44468426**	124.59	<.0001
Rodzaj_su*Gatunek_ro (*RC*)	4	24.51019815	**6.12754954**	22.83	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gatunek_rosliny	2	3.29889630	1.64944815	6.14	0.0031
Rodzaj_substancji	2	66.88936852	33.44468426	124.59	<.0001
Rodzaj_su*Gatunek_ro	4	24.51019815	6.12754954	22.83	<.0001

a) Analiza ogólnej zależności korelacyjnej.

Łączna hipoteza o braku ogólnej zalezności korelacyjnej pomiędzy zmienną objaśnianą *Y* (czyli WP) i wszystkimi czynnikami (czyli *R* oraz *C* z uwzględnieniem ich interakcji), ma postać:

$$H_0(R):\ \mu_{1\bullet} = \mu_{2\bullet} = ... = \mu_{r\bullet}. \qquad (17\text{-}2.46')$$

$$H_0(C):\ \mu_{\bullet 1} = \mu_{\bullet 2} = ... = \mu_{\bullet c}. \qquad (17\text{-}2.47')$$

$$H_0(RC): \gamma_{ij} \equiv \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} = 0\ , \quad \text{dla} \quad i = 1,2,\ldots, r;\ j = 1,2,\ldots, c, \qquad (17\text{-}2.48')$$

Statystyka testowa *F* dla tesowania tej hipotezy następującą postać (17-2.43):

$$F = \frac{MS\text{Reg}}{MSE} \qquad (17\text{-}2.43')$$

przy czym $MS\text{Reg} = \frac{SS\text{Reg}}{k}$ oraz $MSE = \frac{SSE}{\nu_E}$, (17-2.44), (17-2.45), gdzie liczba stopni swobody dla modelu regresji wynosi $k = \nu_r + \nu_c + \nu_{rc}$, (17-2.41), a dla błędu $\nu_E = n_{\bullet\bullet} - 1 - k$. Z powyższego raportu widać, że wartość $F = \frac{MS\text{Reg}}{MSE} = 44.10$, (17-2.43'), jest istotna statystycznie (p<0,0001), zatem odrzucamy hipotezę zerową o braku ogólnej zależności korelacyjnej WP od rodzaju substancji toksycznej oraz gatunku rośliny. W takiej sytuacji, przystępujemy do testowania osobno hipotez (17-2.46') oraz (17-2.47') dla wpływów głównych oraz (17-2.48') dla interakcji, szukając przyczyny odrzucenia hipotezy o braku ogólnej zależności korelacyjnej.

b) Analiza dla wartości oczekiwanych w rozkładach brzegowych dla czynnika *C* oraz *R* oraz dla interakcji pomiędzy czynnikami.

Analiza dla czynników ustalonych.

Z powyższego raportu dla sum „Type I SS" oraz wcześniejszej części raportu dla testu ogólnej zależności korelacyjnej widać, że:

$$F(R) = \frac{MSR}{MSE} = \frac{1{,}649}{0{,}268} = 6{,}14\ ,$$

która to wartość ze względu na wartość empirycznego poziomu istotności p = P($F \geq 6{,}14$) = 0,0031 jest istotna statystycznie na każdym poziomie istotności $\alpha \geq p$ = 0.0031, np. α = 0,01 lub 0,05.

Podobnie jest dla:

$$F(C) = \frac{MSC}{MSE} = \frac{33{,}445}{0{,}268} = 124{,}59,$$

z wartością p = P($F \geq 124{,}59$) <0.0001 oraz dla:

$$F(R \times C) = \frac{MSRC}{MSE} = \frac{6{,}128}{0{,}268} = 22{,}83,$$

z p = P($F \geq 22{,}83$) <0.0001.

Powyższe trzy testy, związane z weryfikacją postawionych powyżej hipotez Ad1)- Ad3), oznaczają, że:

1) Hipoteza $H_0(R)$ jest odrzucona, na każdym poziomie istotności $\alpha \geq p$ = 0.0031, np. α = 0,01 lub 0,05. Oznacza to, że przyjmujemy istnienie istnego statystycznie głównego wpływu czynnika R „gatunek rośliny" na wydolność płuc (WP) pracowników.

2) Hipoteza $H_0(C)$ jest odrzucona, na każdym poziomie istotności $\alpha \geq p$ (p < 0,0001), co oznacza, że przyjmujemy istnienie istotnie statystycznego głównego wpływu czynnika „typ substancji" na WP.

3) Hipoteza $H_0(RC)$ jest odrzucona, na każdym poziomie istotności $\alpha \geq p$ (p < 0,0001). Zatem interakcja jest istotna statystycznie, co oznacza istotność wpływu poziomu jednego czynnika na wielkość zmiany średniej WP pracowników pod wpływem zmiany czynnika drugiego. Np. istotnie statystycznie różny jest wpływ zmiany substancji toksycznej na WP pracowników dla różnych gatunków roślin. Przyjmujemy istnienie interakcji pomiędzy czynnikami w populacji.

Poniżej metodą Scheffe'ego porównań szczegółowych szukamy przyczyny odrzucenia hipotezy zerowej $H_0(C)$: $\mu_{\bullet a} = \mu_{\bullet b} = \mu_{\bullet c}$.

```
Przyklad_"wydolność płóc"                                  15:49 Saturday, April 17, 2004

                                  The GLM Procedure
                               Least Squares Means   (LSMean)
                     Adjustment for Multiple Comparisons: Scheffe

                        Rodzaj_                          LSMEAN
                        substancji       WP LSMEAN       Number
                        a               4.75416667           1
                        b               4.12277778           2
                        c               2.86111111           3

                   Least Squares Means for effect Rodzaj_substancji (czynnik C)
                         Pr > |t| for H0: LSMean(i)=LSMean(j)

                               Dependent Variable: WP   (wartości p)

                    i/j              1             2             3
                       1                      <.0001        <.0001
                       2        <.0001                      <.0001
                       3        <.0001        <.0001
```

Widać, że metoda Scheffe'ego porównań szczegółowych (a-b, a-c, b-c) dla wpływu głównego czynnika *C* („typ substancji") wskazała, że na każdym poziomie istotności $\alpha \geq p$ (p = 0,0001), wartości średniej WP pracowników różnią się parami (dla par rodzajów substancji a(nr 1 LSMEAN), b(nr 2 LSMEAN) i c(nr 3 LSMEAN)), w sposób istotny statystycznie. Przyjmujemy więc wszystkie hipotezy alternatywne o tym, że wartości oczekiwane wydolności płuc $\mu_{\bullet j}$ (*j* = a,b,c) pracowników przedsiębiorstwa, poddanych wpływowi różnych substancji toksycznych, są parami różne. Wynika to również z poniższej części raportu, w której utworzone są grupy Scheffe'ego dla czynnika *C*:

```
Przyklad_"wydolność płóc"                                    15:49 Saturday, April 17, 2004

                                  The GLM Procedure
                                 Scheffe's Test for WP

              NOTE: This test controls the Type I experimentwise error rate.

                         Alpha                              0.05
                         Error Degrees of Freedom             99
                         Error Mean Square              0.268443
                         Critical Value of F             3.08824
                         Minimum Significant Difference   0.3035

                 Means with the same letter are not significantly different.
        Scheffe Grouping        Mean      N   substancji (czynnik C)

                          A       4.7542     36    a
                          B       4.1228     36    b
                          C       2.8611     36    c
```

Podana powyżej **wartość krytyczna *F* = 3.08824** dotyczy testu dla wpływu głównego czynnika *C* („typ substancji") dla szczegółowych porównań wartości średnich metodą Scheffe'ego. Statystka testowa ma rozkład F-Snedecora z liczbą stopni swobody licznika (*SSC*) równą $\nu_c = c - 1 = 3 - 1 = 2$ oraz mianownika (*SSE*) równą $\nu_E = rc(n-1) = 3\times3\times(12-1) = 99$. Zgodnie z (16-1.32)-(16-1.33) wyznaczono metodą Scheffe'ego minimalną istotną statystycznie różnicę (Minimum Significant Difference) dla $(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet j'\bullet})$ na równą $S\sqrt{MSE(\frac{1}{n_{\bullet j}} + \frac{1}{n_{\bullet j'}})} = 0{,}3035$, gdzie $S = \sqrt{(c-1)F_{c-1,n_{\bullet\bullet}-c,1-\alpha}}$, $n_{\bullet j} = n_{\bullet j'} = 36$, $n_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = 108$, $c = 3$, oraz $\alpha = 0{,}05$. Stąd porównanie średnich parami, wykazało, że wszystkie one różnią się pomiędzy sobą istotnie statystycznie. Dlatego dla trzech substancji a,b,c, SAS utworzył powyżej trzy grupy Scheffe'ego nazwane A, B oraz C.

Poniżej metodą Scheffe'ego porównań szczegółowych szukamy przyczyny odrzucenia hipotezy zerowej $H_0(R)$: $\mu_{1\bullet} = \mu_{2\bullet} = \mu_{3\bullet}$.

```
Przyklad_"wydolność płóc"                                    15:49 Saturday, April 17, 2004
                                  The GLM Procedure
                                 Least Squares Means
                     Adjustment for Multiple Comparisons: Scheffe

                         Gatunek_                      LSMEAN
                         rosliny        WP LSMEAN      Number
                         1             3.70361111           1
                         2             3.90305556           2
                         3             4.13138889           3
```

```
Least Squares Means for effect Gatunek_rosliny    (czynnik R)
     Pr > |t| for H0: LSMean(i)=LSMean(j)

            Dependent Variable: WP   (wartości p)

i/j             1            2            3
   1                     0.2682       0.0031
   2        0.2682                    0.1795
   3        0.0031       0.1795
```

Zatem, metoda Scheffe'ego szczegółowych porównań (1-2, 1-3, 2-3) dla wpływu głównego czynnika *R* („gatunek rośliny") wskazała, że wartości średniej WP pracowników charakterystyczne dla gatunku roślin 1 i 3 różnią się istotnie na każdym poziomie istotności $\alpha \geq$ p (p = 0,0031), np. na poziomie α = 0,01 lub 0,05. Natomiast pozostałe porównania (1-2, 2-3) nie wykazały statystycznie istotnych różnic pomiędzy średnimi. Wynika to również z poniższej części raportu, w której utworzone są grupy Scheffe'ego dla czynnika *R*:

```
Przyklad_"wydolność płóc"                                  15:49 Saturday, April 17, 2004

                              The GLM Procedure

                             Scheffe's Test for WP

          NOTE: This test controls the Type I experimentwise error rate.

                Alpha                              0.05
                Error Degrees of Freedom             99
                Error Mean Square              0.268443
                Critical Value of F             3.08824
                Minimum Significant Difference   0.3035
          Means with the same letter are not significantly different.
                                                Gatunek_
            Scheffe Grouping          Mean      N   rosliny  (czynnik R)
                          A        4.1314     36   3
                          A
                    B     A        3.9031     36   2
                    B
                    B             3.7036     36   1
```

Podana w powyższym raporcie wartość krytyczna ***F* = 3.08824** dotyczy testu z wpływu głównego czynnika *R* („gatunek rośliny"), dla szczegółowych porównań wartości średnich metodą Scheffe'ego. Statystka testowa ma rozkład F-Snedecora z liczbą stopni swobody licznika (*SSR*) równą $\nu_r = r - 1 = 3 - 1 = 2$ oraz mianownika (*SSE*) równą $\nu_E = rc(n-1) = 3\times3\times(12-1) = 99$. Zgodnie z (16-1.32)- (16-1.33) wyznaczono metodą Scheffe'ego minimalną istotną statystycznie różnicę (`Minimum Significant Difference`) dla $(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{i'\bullet\bullet})$ na równą $S\sqrt{MSE(\frac{1}{n_{i\bullet}} + \frac{1}{n_{i'\bullet}})} = 0{,}3035$, gdzie $S = \sqrt{(r-1)F_{r-1,n_{**}-r,1-\alpha}}$, $n_{i\bullet} = n_{i'\bullet} = 36$, $n_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = 108$, r=3, oraz $\alpha = 0{,}05$. Stąd wniosek płynący z powyższej części raportu jest zgodny z wcześniejszą częścią raportu dla szczegółowych porównań. Istotnie, porównanie metodą Scheffe'ego średnich parami wykazało, że dla trzech roślin 1, 2 oraz 3, można utworzyć dwie grupy Scheffe'ego, grupę A oraz B. Oznacza to, że średnia WP pracowników z punktu widzenia głównego wpływu gatunku rośliny pozwala zgrupować populacje dla rzędów 1 i 2 w jedną populację obejmującą je w ramach grupy Scheffe'ego oznaczonej jako B. Podobnie populacje dla rzędów 2 i 3 można traktować jako jedną z grupę Scheffe'ego oznaczoną jako A.

Poniżej, metodą Scheffe'ego porównań szczegółowych wykonano porównania dla *r*c* średnich w poszczególnych komórkach wewnątrz tablicy danych.

```
Przyklad_"wydolność płóc"                                    15:49 Saturday, April 17, 2004

                                  The GLM Procedure
                                 Least Squares Means
                     Adjustment for Multiple Comparisons: Scheffe
                  Rodzaj_       Gatunek_                      LSMEAN
                  substancji    rosliny        WP LSMEAN      Number
                  a             1             4.89833333           1
                  a             2             5.23416667           2
                  a             3             4.13000000           3
                  b             1             3.46416667           4
                  b             2             3.77750000           5
                  b             3             5.12666667           6
                  c             1             2.74833333           7
                  c             2             2.69750000           8
                  c             3             3.13750000           9

                   Least Squares Means for effect Rodzaj_su*Gatunek_ro
                         Pr > |t| for H0: LSMean(i)=LSMean(j)
                       Dependent Variable: WP (wartości p)
```

i/j	1	2	3	4	5	6	7	8	9
1		0.9587	0.1206	<.0001	0.0013	0.9967	<.0001	<.0001	<.0001
2	0.9587		0.0017	<.0001	<.0001	1.0000	<.0001	<.0001	<.0001
3	0.1206	0.0017		0.2848	0.9451	0.0082	<.0001	<.0001	0.0087
4	<.0001	<.0001	0.2848		0.9729	<.0001	0.1928	0.1225	0.9650
5	0.0013	<.0001	0.9451	0.9729		<.0001	0.0052	0.0025	0.3408
6	0.9967	1.0000	0.0082	<.0001	<.0001		<.0001	<.0001	<.0001
7	<.0001	<.0001	<.0001	0.1928	0.0052	<.0001		1.0000	0.9047
8	<.0001	<.0001	<.0001	0.1225	0.0025	<.0001	1.0000		0.8231
9	<.0001	<.0001	0.0087	0.9650	0.3408	<.0001	0.9047	0.8231	

Wszystkie powyższe porównania z $p \le \alpha = 0{,}05$ lub 0,01 można uznać za statystycznie istotne.

Analiza dla czynników losowych lub mieszanych.

```
Przyklad_"wydolność płóc"                                    15:49 Saturday, April 17, 2004

                                  The GLM Procedure

                              Dependent Variable: WP

Tests of Hypotheses Using the Type III MS for Rodzaj_su*Gatunek_ro as an Error Term

Source                      DF     Type III SS     Mean Square    F Value    Pr > F

Rodzaj_substancji            2     66.88936852     33.44468426       5.46    0.0719
Gatunek_rosliny              2      3.29889630      1.64944815       0.27    0.7768
```

Widać, że statystyki testowe dla obu czynników losowych lub czynników mieszanych (gdy jeden jest losowy, a drugi ustalony), dają:

$$F(R) = \frac{MSR}{MSRC} = \frac{1{,}649}{6{,}128} = 0{,}27,$$

z wartością p = P($F \ge 0{,}27$) = 0,7768

$$F(C) = \frac{MSC}{MSRC} = \frac{33{,}445}{6{,}128} = 5{,}46,$$

z wartością p = P($F \ge 5{,}46$) <0.0719

oraz

$$F(R\times C)=\frac{MSRC}{MSE}=\frac{6{,}128}{0{,}268}=22{,}83,$$

z wartością p = P($F \geq 22{,}83$) <0.0001 .

Zatem w przypadku czynników losowych lub mieszanych, nie ma podstaw do odrzucenia hipotezy zerowej (dla α < p) o nieistotności głównego wpływy czynników (osobno „gatunku rośliny” *R* i osobno „rodzaju toksyny” *C*) na średnią wydolność płuc pracowników. W konsekwencji zarządca przedsiębiorstwa mógłby np. uznać, że nie ma podstaw *statystycznych* do zamknięcia jednego z działów i przesunięcia pracowników do innego działu (np. z działu z substancją „a”, w którym średnia WP pracowników jest równa 4.75 do działu z substancją „c”, w którym średnia ta jest równa 2.86). Widać, że w tej części analizy wnioski płynące z pojawienia się czynnika losowego różnią się zasadniczo od wniosków dla obu czynników ustalonych.

Natomiast interakcja obu czynników jest statystycznie istotna, tak jak poprzednio, co wynika z takiej samej postaci statystyki testowej $F(R\times C)=MSRC/MSE$ jak w przypadku obu czynników ustalonych.

Rozdział 17-3. ANOVA z losowo dobieranymi blokami (jedna obserwacja w komórkach).

Rozpatrzmy przypadek, gdy czynnik „główny” *R* (rzędowy) jest ustalony i przyjmijmy chwilowo, że czynnik blokowy *B* (kolumnowy) przyjął konkretne poziomy. Pozwala to skonstruować poniższą tabelę (za chwilę stwierdzimy wyraźnie, że czynnik blokowy *B* jest *losowy*). W Tabeli 17-3.1 przedstawiono ogólny układ danych dla modelu z blokami dla *k* poziomów czynnika głównego *R* i dla *b* bloków czynnika *B*.

Tabela 17-3.1 Komórki ze średnimi w próbach i populacjach dla układu losowego dobierania bloków [1].

	Czynnik blokowy *B*, *j*					
Czynnik główny *R*, *i*	1	2	…	*b*	Razem: $Y_{i\bullet}=\sum_{j=1}^{b}Y_{ij}$	Średnia
1	Y_{11}, μ_{11}	Y_{12}, μ_{12}	···	Y_{1b}, μ_{1b}	$Y_{1\bullet}$	$\overline{Y}_{1\bullet}$, $\mu_{1\bullet}$
2	Y_{21}, μ_{21}	Y_{22}, μ_{22}	···	Y_{2b}, μ_{2b}	$Y_{2\bullet}$	$\overline{Y}_{2\bullet}$, $\mu_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	Y_{k1}, μ_{k1}	Y_{k2}, μ_{k2}	···	Y_{kb}, μ_{kb}	$Y_{k\bullet}$	$\overline{Y}_{k\bullet}$, $\mu_{k\bullet}$
Razem $Y_{\bullet j}=\sum_{i=1}^{k}Y_{ij}$	$Y_{\bullet 1}$	$Y_{\bullet 2}$	…	$Y_{\bullet b}$	$Y_{\bullet\bullet}=\sum_{i=1}^{k}\sum_{j=1}^{b}Y_{ij}$	
Średnia	$\overline{Y}_{\bullet 1}$, $\mu_{\bullet 1}$	$\overline{Y}_{\bullet 2}$, $\mu_{\bullet 2}$	…	$\overline{Y}_{\bullet b}$, $\mu_{\bullet b}$		$\overline{Y}_{\bullet\bullet}$, $\mu_{\bullet\bullet}$

W powyższej tabeli Y_{ij} oznacza wartość obserwacji zmiennej objaśnianej *Y* w *i* – tym poziomie czynnika głównego *R* i w *j* – tym bloku.

Suma dla czynnika głównego R (w rzędach) jest oznaczona jako:

$$Y_{i\bullet} = \sum_{j=1}^{b} Y_{ij} \ , \qquad i = 1,2,\ldots, k \ , \tag{17-3.64}$$

zaś dla bloków (w kolumnach) ma postać:

$$Y_{\bullet j} = \sum_{i=1}^{k} Y_{ij} \ , \qquad j = 1,2,\ldots, b \ . \tag{17-3.65}$$

Ogólna suma dla wszystkich $b \times k$ obserwacji (po jednej obserwacji w komórce) wynosi:

$$Y_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{b} Y_{ij} \ . \tag{17-3.66}$$

Średnia dla i – tego poziomu czynnika głównego R wynosi:

$$\overline{Y}_{i\bullet} = \frac{Y_{i\bullet}}{b} \ , \qquad i = 1,2,\ldots, k \ , \tag{17-3.67}$$

Średnia w j – tym bloku (kolumnie) czynnika B wynosi:

$$\overline{Y}_{\bullet j} = \frac{Y_{\bullet j}}{k} \ , \qquad j = 1,2,\ldots, b \ , \tag{17-3.68}$$

Średnia ogólna jest równa:

$$\overline{Y}_{\bullet\bullet} = \frac{1}{bk} \sum_{i=1}^{k} \sum_{j=1}^{b} Y_{ij} \ . \tag{17-3.69}$$

Zakładając, że populacje są równoliczne i że μ_{ij} jest wartością oczekiwaną w populacji oznaczonej parą indeksów (i, j) otrzymujemy, że wartość oczekiwana $\mu_{i\bullet}$ w populacji powstałej z połączenia wszystkich (pod)populacji w i-tym rzędzie, następnie, wartość oczekiwana $\mu_{\bullet j}$ w populacji powstałej z połączenia (pod)populacji w j-tym bloku (j-tej kolumnie), oraz wartość oczekiwana $\mu_{\bullet\bullet}$ w ogólnej (generalnej) populacji powstałej z połączenia wszystkich populacji, spełniają kolejno związki:

$$\mu_{i\bullet} = \frac{1}{b} \sum_{j=1}^{b} \mu_{ij} \ , \qquad i = 1,2,\ldots, k \ , \tag{17-3.70}$$

$$\mu_{\bullet j} = \frac{1}{k} \sum_{i=1}^{k} \mu_{ij} \ , \qquad j = 1,2,\ldots, b \ , \tag{17-3.71}$$

oraz:

$$\mu_{\bullet\bullet} = \frac{1}{bk} \sum_{i=1}^{k} \sum_{j=1}^{b} \mu_{ij} \ . \tag{17-3.72}$$

Założenia w metodzie ANOVA z losowo dobieranymi blokami są następujące:

a) obserwacje są statystycznie niezależne jedne od drugich,

b) każda obserwacja jest wybierana z populacji o rozkładzie normalnym,

c) każda obserwacja jest wybierana z populacji, w której wariancja składnika losowego jest taka sama,

d) w modelu regresji dla ANOVA z losowo dobranymi blokami nie ma oddziaływania pomiędzy czynnikiem blokowym a czynnikiem głównym (Rozdział 17-3-1).

ANOVA z losowo dobieranymi blokami realizuje schemat $n_{ij} = n_{i\bullet} n_{\bullet j} / n_{\bullet\bullet}$, (17.2), dla liczebności w komórkach, oznaczający niezależność stochastyczną czynników *R* i *B*, przy czym **liczebność w komórkach jest** taka sama, a w przypadku ANOVA z losowo dobieranymi blokami, jest dodatkowo **jednostkowa**.

W metodzie ANOVA z losowo dobieranymi blokami pomiędzy sumami kwadratów odchyleń zachodzi następujący związek (pokazać):

$$TSS = SSR + SSB + SSRB, \tag{17-3.73}$$

gdzie ogólna suma kwadratów odchyleń obserwacji od średniej ogólnej, będąca miarą wszystkich zmian w danych, ma postać:

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2, \tag{17-3.74}$$

suma kwadratów odchyleń, związana z wpływem czynnika głównego *R*, która jest miarą zmian między średnimi $\overline{Y}_{i\bullet}$ dla *i*-tych poziomów czynnika głównego, ma postać:

$$SSR = b \sum_{i=1}^{k} (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2, \tag{17-3.75}$$

suma kwadratów odchyleń, która jest miarą zmian między średnimi $\overline{Y}_{\bullet j}$ dla bloków, ma postać:

$$SSB = k \sum_{j=1}^{b} (\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2 \ , \tag{17-3.76}$$

suma kwadratów „błędów”, ma postać:

$$SSRB = \sum_{i=1}^{k} \sum_{j=1}^{b} (Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet})^2 \ . \tag{17-3.77}$$

Wyrażenie pod znakiem sumy $(Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet})$ można zapisać jako $(Y_{ij} - \overline{Y}_{i\bullet}) - (\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})$, co oznacza, że jest to składnik związany z oddziaływaniem pomiędzy blokami (z którymi związane są różnice $(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})$) i czynnikiem „głównym” (z którym w *i*-tym wierszu czynnika głównego związane są *j*-te odchyłki $(Y_{ij} - \overline{Y}_{i\bullet})$ w blokach). Suma *SSRB* ma dużą wartość wtedy, gdy wpływ czynnika głównego (na zakres zmian zmiennej *Y*) zmienia się od bloku do bloku.

W metodzie losowego doboru bloków zakłada się jednak, że pojawienie się składnika *SSRB* jako skutku interakcji (tak jak to jest w ogólniejszym przypadku dwuczynnikowej ANOVA opisanym poprzednio) jest wtórne, a jego pierwotnym pochodzeniem jest błąd pomiarowy związany z blokami. Z ogólniejszego przypadku analizy dwuczynnikowej wiemy, że *takie traktowanie błędu w testach F jest związane z pojawieniem się czynnika losowego*. Stąd bierze się nazwa analizy: „*ANOVA z losowo dobieranymi blokami*”.

Uwaga. *Nie należy mylić powyższego SSRB,* (17-3.77)*, z błędem występującym w raportach SAS'a pod nazwą* (`Error`) *i związanym z rozproszeniem wewnątrz komórek w modelu regresji, zadanym przez składnik losowy E*

w (17-1-1.29)*, który w tym przypadku* (*ze względu na występowanie w* ANOVA *z losowo dobieranymi blokami tylko jednego pomiaru w komórce*) *jest równy zero*, jak to zobaczymy również w poniższym raporcie rozważanego przykładu. (Czasami *SSRB* w ANOVA z losowo dobieranymi blokami jest oznaczane jako *SSE* [1].)

A. Podstawowa (główna) hipoteza zerowa o równości wartości oczekiwanych w wariantach (poziomach) czynnika głównego *R*, ma postać:

$$H_0 : \mu_{1\bullet} = \mu_{2\bullet} = ... = \mu_{k\bullet} \tag{17-3.78}$$

Test statystyczny *F* dotyczący weryfikacji hipotezy (17-3.78) ma postać:

$$F = \frac{MSR}{MSRB}, \tag{17-3.79}$$

gdzie średni kwadrat odchyleń w rzędach wynosi:

$$MSR = \frac{SSR}{k-1}, \tag{17-3.80}$$

a średni kwadrat „błędów", ma postać:

$$MSRB = \frac{SSRB}{(k-1)(b-1)}. \tag{17-3.81}$$

Przy prawdziwości hipotezy zerowej (17-3.78) statystyka *F*, (17-3.70), ma rozkład F-Snedecora z liczbą stopni swobody licznika równą $\nu_r = k-1$ i mianownika równą $\nu_E = (k-1)\,(b-1)$. Na poziomie istotności α odrzucamy hipotezę zerową na korzyść alternatywnej, gdy $F \geq F_{k-1,\,(k-1)(b-1),\,1-\alpha}$. Odrzucenie hipotezy głównej oznacza, że zmiana numeru populacji, związana ze zmianą poziomu czynnika głównego *R*, ma wpływ na poziom wartości oczekiwanej $\mu_{i\bullet}$ zmiennej *Y*.

B. Dodatkowo przeprowadzanym testem jest test dotyczący weryfikacji hipotezy zerowej o równości wartości oczekiwanych w blokach:

$$H_0^b : \mu_{\bullet 1} = \mu_{\bullet 2} = ... = \mu_{\bullet b}. \tag{17-3.82}$$

Test ten przeprowadzany jest po fakcie, jako test sprawdzający założenie, że blokowanie zostało przeprowadzone prawidłowo, tzn. że badacz tak dobrał bloki, że przy przejściu od jednego do drugiego, następuje istotna zmiana poziomu średnich $\mu_{\bullet j}$ (co oznacza odrzucenie hipotezy H_0^b).

Statystyka dla tego testu ma postać:

$$F = \frac{MSB}{MSRB}, \tag{17-3.83}$$

gdzie:

$$MSB = \frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}\,(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2 = \frac{1}{b-1}\left(\frac{1}{k}\sum_{j=1}^{b} Y_{\bullet j}^2 - \frac{Y_{\bullet\bullet}^2}{bk}\right), \tag{17-3.84}$$

przy czym, w każdym bloku $n_{\bullet j} = k$, a *MSRB* ma postać określoną w (17-3.81).

Przy prawdziwości hipotezy zerowej (17-3.82) statystyka F, (17-3.83), ma rozkład F-Snedecora z liczbą stopni swobody licznika równą $\nu_b = b-1$ i mianownika równą $\nu_{RB} = (k-1)(b-1)$.

Podsumowując, tablica ANOVA dla doświadczenia losowego dobierania bloków z k poziomami czynnik głównego R i b blokami czynnika blokowego B ma poniższą postać.

Tabela 17-3.2. Tablica ANOVA dla losowego dobierania bloków [1].

Źródła zmienności Y	Liczba stopni swobody	Sumy kwadratów Odchyleń	Średnie kwadraty odchyleń	Statystyka testowa F
Czynnik „główny” R	$k-1$	SSR	$MSR = \dfrac{SSR}{k-1}$	$\dfrac{MSR}{MSRB}$
Bloki (czynnik B)	$b-1$	SSB	$MSB = \dfrac{SSB}{b-1}$	$\dfrac{MSB}{MSRB}$
„Błąd” związany z losowością bloków	$(k-1)(b-1)$	$SSRB$	$MSRB = \dfrac{SSRB}{(k-1)(b-1)}$	
Razem	$kb-1$	TSS		

Rozdział 17-3-1. Model regresji dla ANOVA z losowym doborem bloków.

Model regresji zmiennej objaśnianej Y względem czynnika głównego R o k poziomach i czynnika blokowego B z liczbą kolumn równą b, można sformułować analogicznie jak to było dla ANOVA jednoczynnikowej w Rozdziale 16-2 poprzez wprowadzenie $(k-1)$ zmiennych ukrytych (wskazujących) dla czynnika R oraz $(b-1)$ zmiennych ukrytych dla czynnika B.

Model ten ma następują postać:

$$Y = \mu + \sum_{i=1}^{k-1} \alpha_i X_i + \sum_{j=1}^{b-1} \beta_j Z_j + E, \qquad (17\text{-}3\text{-}1.85)$$

z następującym kodowaniem zmiennych wskazujących:

$$X_i = \begin{cases} 1 & dla \quad poziomu \quad i-tego \quad czynnika \quad R \quad , \quad i = 1,2,\ldots,k-1 \\ -1 & dla \quad poziomu \; k \quad czynnika \quad R \\ 0 & w \quad pozostalych \quad przypadkach \end{cases}$$

(17-3-1.86)

$$Z_j = \begin{cases} 1 & dla \quad poziomu \quad j-tego \quad czynnika \quad kolumnowego \; B \quad , \quad j = 1,2,\ldots,b-1 \\ -1 & dla \quad poziomu \quad b \quad czynnika \quad kolumnowego \; B \\ 0 & w \quad pozostalych \quad przypadkach \quad . \end{cases}$$

Przesunięcie w równaniu regresji (17-3.76) spełnia warunek (pokazać):

$$\mu = \mu_{\bullet\bullet}, \qquad (17\text{-}3\text{-}1.87)$$

gdzie $\mu_{\bullet\bullet} = \dfrac{1}{bk} \sum_{i=1}^{k} \sum_{j=1}^{b} \mu_{ij}$ jest określone w (17-3.72), a współczynniki α_i są postaci:

$$\alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet} \qquad \text{dla} \qquad i = 1,2,\ldots, k-1, \qquad (17\text{-}3\text{-}1.88)$$

gdzie $\mu_{i\bullet} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$, (17-3.70), $\mu_{\bullet j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}$, (17-3.71), przy czym zachodzi związek:

$$\alpha_k \equiv -\sum_{i=1}^{k-1}\alpha_i = \mu_{k\bullet} - \mu_{\bullet\bullet} \quad . \tag{17-3-1.89}$$

Natomiast współczynniki β_j są równe:

$$\beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet} \qquad \text{dla} \qquad j = 1,2,\ldots, b-1, \tag{17-3-1.90}$$

przy czym:

$$\beta_b \equiv -\sum_{j=1}^{b-1}\beta_j = \mu_{\bullet b} - \mu_{\bullet\bullet} \quad . \tag{17-3-1.91}$$

Rozdział 17-3-2. Przykład „samopoczucie" dla ANOVA z losowym doborem bloków.

Przebadano wpływ działania pewnego leku na samopoczucie osób poddanych leczeniu. W tym celu wydzielono dwie grupy osób. Jedną grupę spośród osób poddanych leczeniu (formalnie jest to próbka wylosowana z całej populacji osób leczonych) i drugą grupę spośród osób, którym podano placebo (formalnie jest to próbka wylosowana z całej populacji osób, którym podano placebo). W praktyce grupy są na ogół losowane przed podaniem leku czy placebo.

Zmienną objaśnianą *Y* jest *zmiana samopoczucia* pacjenta po zastosowaniu leku. Niech zmienna ta przyjmuje wartości np. z zakresu od 1 do 20.

Rozważmy *czynnik R* (*czynnik* 1), który przyjmuje dwa poziomy, tzn. podawanie leku może nastąpić na dwóch poziomach wyznaczających dwie grupy dla czynnika *R*:

poziom 1 oznacza, że zastosowano lek

poziom 2 oznacza, że podano placebo

Bloki (różne bloki są w różnych kolumnach) są *jednorodne* ze względu na wiek i płeć. Dysponujemy b = 15 parami indywidualnych dopasowań w jednorodnych wewnętrznie blokach (tzn. kolumnach), które różnią się pomiędzy sobą pod względem wieku oraz płci. Zatem w jednym bloku mogą występować wyłącznie osoby (po jednej, z każdej z dwóch grup wyznaczonych przez czynnik *R*), które mają tą samą płeć i ten sam wiek.

Tablica z danymi została zamieszczona poniżej.

Tabela 17-3-2.1 Przykład „samopoczucie" [1]. W jednoelementowych komórkach podana jest zmierzona *zmiana samopoczucia Y* w *i*-tej grupie (*i*=1,2) oraz w *j*-tym bloku (*j*=1,2,...,15), jednorodnym ze względu na płeć i wiek.

	Bloki (tworzone przez pary)																
Grupa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Razem	Średnia
Leczona *L*	10	12	8	8	13	11	15	16	4	13	2	15	5	6	8	146	9,73
Kontrolna *K*	6	5	7	9	10	12	9	8	3	14	6	10	1	2	1	103	6,87
Razem	16	17	15	17	23	23	24	24	7	27	8	25	6	8	9	249	8,30

Sformułowanie problemu: Zagadnie dotyczy zbadania czy średnia zmiana samopoczucia w grupie pobranej z populacji *L*, w której zastosowano lek, różni się istotnie statystycznie od średniej zmiany samopoczucia w grupie kontrolnej pobranej z populacji *K*, której podano placebo.

Rozważmy hipotezę zerową:

$$H_0 : \mu_L = \mu_K \qquad (17\text{-}3\text{-}2.92)$$

wobec alternatywnej:

$$H_1 : \mu_L \neq \mu_K . \qquad (17\text{-}3\text{-}2.93)$$

gdzie *L* oznacza grupę, której zastosowano lek, a *K* grupę kontrolną.
Hipoteza zerowa oznacza, że nie ma różnicy w wartości oczekiwanej *zmiany samopoczucia* w dwóch badanych populacjach osób. Hipoteza alternatywna jest dwustronna, zatem poza tym, że w populacjach *L* oraz *K* jest różna zmiana samopoczucia, to nie zakładamy z góry, w której populacji zmiana samopoczucia jest większa. Jeśli jednak np. badacz uważa, że grupa pobrana z populacji, w której podano lek będzie odczuwała większą poprawę samopoczucia, wtedy hipoteza alternatywna miałaby postać $H_1 : \mu_L > \mu_K$.

Hipotezę zerową można zweryfikować posługując się testem F dla ANOVA. W teście tym są wymagane trzy składniki odpowiedzialne za zmienność cechy Y:

(a) czynnik główny *R* z dwoma poziomami, $r = 2$,

(b) czynnik blokowy *B*, gdzie w przykładzie bloki są tworzone przez pary, b=15,

(c) oddziaływanie *RB*, pełniące w testach F w przypadku losowych bloków rolę „błędu”, który ma źródło w losowym doborze bloków w oddziaływaniu.

Przypomnijmy, że jednoczynnikowej analizie wariancji występują tylko dwa składniki, składnik międzygrupowy i wewnątrzgrupowy. Zatem, o ile w jednoczynnikowej ANOVA całkowita suma kwadratów odchyleń była równa sumie kwadratów odchyleń międzygrupowej i sumie kwadratów odchyleń wewnątrzgrupowej (16-1.7), tak w tym przypadku całkowita suma kwadratów odchyleń (*TSS*) jest równa sumie sum kwadratów odchyleń trzech składników (a), (b) i (c), (17-3.73):

TSS(całkowita) = *SSR* (czynnik główny) + *SSB*(bloki) + *SSRB*(„błąd”),

gdzie uwzględniono fakt, że *SSE* = 0.

Stopnie swobody, towarzyszące sumie kwadratów odchyleń dla trzech składników w (17-3.73) są następujące:

a. Ponieważ są dwa poziomy ($r = 2$) czynnika głównego *R*, zatem dla *SSR* $\nu_r = r - 1 = 1$.
b. Ponieważ liczba bloków $b = 15$ czynnika *B*, zatem dla *SSB* $\nu_b = b - 1 = 14$.
c. Ponieważ $r = 2$ i $b = 15$, zatem dla interakcji SS*RB*(„błąd”) mamy $\nu_E = (r - 1)(b - 1) = \nu_r \nu_b = 14$.

Statystyka testowa dla hipotezy (17-3-2.92) $H_0 : \mu_L = \mu_K$ ma postać (17-3.79):

$$F = \frac{MSR}{MSRB},$$

gdzie *MSR* dotyczy czynnika głównego *R*, a *MSRB* (17-3.81) dotyczy „błędu”.

Rozważmy hipotezę zerową (17-3.82):

$$H_0^b : \mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet b=15}, \qquad (17\text{-}3\text{-}2.94)$$

o nieistotności podziału na bloki i mówiącą o ty, że ze zmianą bloków nie zmienia się, średnio rzecz biorąc, poziom zmiennej Y, czyli, że nie ulega zmianie *zmiana samopoczucia* przy przejściu od bloku do bloku. Statystyka testowa ma wtedy postać (17-3.84):

$$F = \frac{MSB}{MSRB},$$

gdzie MSB, (17-3.84), jest średnią sumą kwadratów odchyleń dla bloków. Jak wspomnieliśmy poniżej (17-3.82), bloki są tak dobrane, aby hipoteza H_0^b została odrzucona.

Raport SAS’a dla przykładu „samopoczucie”

```
samopoczucie                                        14:32 Sunday, May 16, 2004
                    The GLM Procedure
                 Class Level Information

              Class         Levels    Values
               Blok            15    1 10 11 12 13 14 15 2 3 4 5 6 7 8 9
              _Source_          2    Kontrolna Leczona

              Number of observations    30

samopoczucie                                        14:32 Sunday, May 16, 2004

                     The GLM Procedure
              Dependent Variable: samopocz   Stacked Values
                                          Sum of
   Source                DF             Squares     Mean Square    F Value    Pr > F
   Model                 29          42.3000000      18.7000000        .         .
   Error                  0      SSE = 0.0000000        .
   Corrected Total       29     542.3000000

              R-Square     Coeff Var      Root MSE    samopocz Mean
             1.000000           .              .         8.300000
```

Podana w raporcie liczba stopni swobody modelu jest równa $\nu_r + \nu_b + \nu_E = 29$. Stąd typowy składnik błędu (Error) ma liczbę stopni swobody równą 0 oraz ze względu na występowanie tylko jednego pomiaru w każdej komórce, sumę *SSE* równą zeru; *porównaj* **Uwagę** *nieco poniżej wzoru* (17-3.77). Średnia *MSRB* (17-3.81), pełniąca w ANOVA z losowym doborem bloków rolę składnika „błędu”, ma wartość *MSRB*=6,3476 podaną w poniższym raporcie w wierszu dla źródła o nazwie Blok*_Source_.

Poniższa część raportu związana jest z testowaniem hipotez $H_0 : \mu_L = \mu_K$ (17-3-2.92) oraz H_0^b (17-3-2.94).

```
Tests of Hypotheses Using the Type III MS for Blok*_Source_ as an Error Term

   Source                DF     Type III SS      Mean Square              F Value    Pr > F

   Blok     (B)          14     391.8000000      MSB =   27.9857143        4.41       0.0044
   _Source_(R)            1      61.6333333      MSR =   61.6333333        9.71       0.0076
Blok*_Source_ (RB)       14      88.8666667      MSRB =   6.3476190        .          .
```

Z powyższego raportu można odczytać sumy kwadratów odchyleń (*SS* typu III):

SSR (z poziomami leczenia *L* i *K*)	= 61,63
SSB (bloki)	= 391,80
SSRB („błąd”)	= 88,87
TSS (całkowita)	= 542,30

Widać, że spełniona jest równość (17-3.73):

TSS = *SST* + *SSB* + *SSRB*.

Hipotezę $H_0 : \mu_L = \mu_K$ wobec alternatywnej $H_1 : \mu_L \neq \mu_K$ testujemy za pomocą statystyki $F = \frac{MSR}{MSRB}$, która w pobranej próbce przyjmuje wartość $F = \frac{MSR}{MSRB} = \frac{61{,}63}{6{,}35} = 9{,}71$. Odpowiednie wartości dla średnich *MS* są podane w powyższym raporcie. Przy prawdziwości hipotezy zerowej H_0, statystyka *F* ma rozkład F-Snedecora z ν_r =1 stopniem swobody licznika oraz ν_{RB} =14 stopniami swobody mianownika. W raporcie podana jest również odpowiednia wartość empirycznego poziomu istotności p = P($F \geq F_{obs}$ =9,71) = 0,0076. Zatem na każdym poziomie istotności większym lub równym niż p=0,0076 (np. na poziomie α = 0,01) odrzucamy hipotezę zerową *o braku wpływu podawanego leku na zmianę samopoczucia osób*. Natomiast na każdym poziomie istotności mniejszym od p=0,0076 (np. dla α = 0,001) nie mamy podstaw aby tą hipotezę zerową odrzucić.

Na koniec należy przeprowadzić test dla hipotezy zerowej związanej z blokami, $H_0^b : \mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet b=15}$, (17-3-2.94). Oznacza ona, że *nie ma istotnej różnicy pomiędzy blokami* jeśli chodzi o zmianę *zmiany samopoczucia Y* przy przejściu od bloku do bloku. Statystyka testowa ma teraz postać: *F* = *MSB/MSRB* i (przy prawdziwości hipotezy zerowej H_0^b) ma rozkład F-Snedecora z ν_b =14 stopniami swobody licznika i ν_{RB} =14 stopniami swobody mianownika. W próbce przyjmuje ona wartość *F* = 4,41, a odpowiadający jej empiryczny poziom istotności p= P($F \geq F_{obs}$ =4,41)=0,0044. Zatem, na każdym poziomie istotności $\alpha \geq p = 0{,}0044$ odrzucamy hipotezę zerową, mówiącą o tym, że zmiana samopoczucia osób nie ulega zmianie przy zmianie bloków. Ponieważ *p* jest stosunkowo małe, więc można przyjąć, że bloki zostały dobrane właściwie.

C. Rozdział 18. Podsumowanie ANOVA.

Metoda ANOVA jest wykorzystywana do testowania hipotez o równości wartości oczekiwanych w populacjach. W skrypcie omówiono szczegółowo zastosowanie testów F-Snedecora dla jednoczynnikowej i dwuczynnikowej ANOVA w przypadku równej liczebności w komórkach, przy czym zwrócono uwagę na wyjątkową wagę problemu określenia typu czynnika (ustalony czy losowy) w wyborze rodzaju testu, otrzymaniu określonego wyniku testu statystycznego i w konsekwencji jego wpływ na decyzję statystyczną podejmowaną przez badacza. W ramach dwuczynnikowej ANOVA omówiono testy służące do badania hipotezy o braku głównych wpływów uwzględnionych w modelu czynników na wartości oczekiwane zmiennej objasnianej, jak i hipotezy o braku oddziaływania pomiędzy czynnikami przy badaniu zakresu zmienności wartości oczekiwanych zmiennej objaśnianej. W końcu, w przypadku odrzucenia badanej hipotezy zerowej o równości wartości oczekiwanych, zwrócono uwagę na poszukiwanie przyczyn jej odrzucenia w oparciu o metodę wielokrotnych, szczegółowych porównań wartości oczekiwanych, przedstawiając metodę Scheffe'go, wielokrotnych szczegółowych porównań wartości oczekiwanych odpowiedzi dla różnych poziomów czynnika. Metoda Scheffe'go jest przykładem zastosowania nierówności Bonferroni'ego, omówionej w Rozdziale 9 (o którą opiera się poprawne zrozumienie poziomów istotności dla testów szczegółowych), w analizie wariancji dla wyznaczenia przedziałów ufności dla kontrastów.

Podane w skrypcie sformułowanie analizy regresji, wskazuje wyraźnie na związek pomiędzy analizą statystyczną stosowaną przy badaniu zależności korelacyjnej pomiędzy zmienną objaśnianą a czynnikami, a podstawowym przedmiotem badań analizy wariancji, którym jest testowanie hipotez o równości wartości oczekiwanych zmiennej objasnianej dla różnych poziomów czynników. Określone kombinacje tych (warunkowych) wartości oczekiwanych mogą być widziane jako parametry strukturalne modelu regresji z tzw. zmiennymi kierunkowymi, co oznacza, że każdy problem dwuczynnikowej ANOVA, tak z równą jak i różną liczebnością komórki, może być rozpatrywany poprzez analizę modelu regresji (17-1-3.7):

$$Y = \mu + \sum_{i=1}^{r-1} \alpha_i X_i + \sum_{j=1}^{c-1} \beta_j Z_j + \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \gamma_{ij} X_i Z_j + E \qquad (18.1)$$

gdzie X_i (i=1,2,...,r) oraz Z_j (j=1,2,.., c) są grupami zmiennych kierunkowych X oraz Z wskazujących kolejno r poziomów czynnika wierszowego R i c poziomów czynnika kolumnowego C.

Analiza regresji wykorzystuje warunkowe sumy kwadratów opisujące zmienność zmiennej objaśnianej. Stąd wynika bliskość podejść charakterystycznych dla ANOVA i dla analizy regresji, a sama ANOVA jest czasami widziana jako szczególny przypadek tej drugiej. Podejście to jest szczególnie pomocne wtedy gdy tradycyjna analiza wariancji nie może się posłużyć „ortogonalnym" rozkładem „fundamentalnego równania dla sum kwadratów", co ma miejsce, gdy liczba obserwacji w komórkach tablicy danych nie jest równa (Rozdział 17-1).

Wykorzystując związki wartości oczekiwanych w populacjach z parametrami modelu regresji (18.1), można hipotezy zerowe dla ANOVA zapisać jako hipotezy zerowe tego modelu regresji:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_{r-1} = 0$, brak głównego wpływu czynnika *R* (wierszowego) (18.2a)

H_0: $\beta_1 = \beta_2 = \ldots = \beta_{c-1} = 0$, brak głównego wpływu czynnika *C* (kolumnowego) (18.2b)

H_0: $\gamma_{ij} = 0$ $\quad i = 1, 2, \ldots, r-1; \quad j = 1, 2, \ldots, c-1$, brak interakcji *RC* (18.2c)

Hipotezy te mówią formalnie o niewystępowaniu w równaniu regresji odpowiednich parametrów stojących przy zmiennych kierunkowych, których wartości wskazują warianty czynników. Ich sednem jest założenie braku zależności korelacyjnej zmiennej *Y* od kolejno, grupy zmiennych *X*, *Z*, lub członu interakcji *X**Z*. Można je też ująć jako założenie o nieistotności rozszerzenia modelu o wspomniane grupy zmiennych. Gdy liczebności w komórkach są różne, kolejność, w jakiej testuje się hipotezy (18.2) staje się istotna (Rozdział 17-1-3) i nieostrożna decyzja podjęta w tej kwestii może dać niewłaściwy wynik. W tej sytuacji procedura, którą zalecają autorzy pracy [1] jest odwróconym rodzajem algorytmu, w którym wzajemne oddziaływanie jest uwzględniane przed wpływami głównymi. Jest więc to typ procedury opartej o eliminację wstecz, tyle że eliminowane jako nieistotne statystycznie są całe grupy zmiennych *X*, *Z* lub *X**Z*. Jeśli chodzi o interakcję, to stwierdzenie to może być naruszone w tym sensie, że nie trzeba wyeliminować z modelu jednocześnie wszystkich członów interakcji [1]. Zatem kolejne kroki selekcji właściwego modelu regresji należy przeprowadzić umiejętnie, wykonując serię częściowych testów F dla hipotez o nie występowaniu braku dopasowania. Zawierają one sumy warunkowe a nie bezwzględne, jak to ma miejsce w metodzie ANOVA ze zrównoważonym układem danych. Przykłady tej ostatniej zostały omówione w powyższych Rozdziałach.

Prawie wszystkie analizy danych, a w szczególności te bardziej złożone są obecnie przeprowadzane za pomocą programów komputerowych. Wynika to z faktu, że statystyczne procedury estymacyjne, oraz procedury służące do testowania hipotez, są rachunkowo żmudne. Statystyczne pakiety zawierają zwykle programy do analizy ANOVA, które znajdują się bądź w miejscach specjalnie wydzielonych, bądź w miejscach poświęconych analizie regresji. Tak też jest z pakietem SAS, w którym analiza ANOVA może być przeprowadzana bądź w ramach procedury GLM w Factorial ANOVA wywoływanej w opcji ANOVA pakietu Analyst, bądź wywoływanej w ramach procedury REG wywoływanej w opcji Regression pakietu Analyst. Ich działanie omówiono jako ilustarcję rozważań teoretycznych. Omówiono trzy przykłady. Jeden z nich dotyczył jednoczynnikowa ANOVA i był związany z typowymi badaniami przeprowadzanymi przez firmy próbujące określić swoją skuteczność na rynku. Inne, dotyczyły dwuczynnikowej ANOVA i były ilustracją rzeczywistych problemów powstałych przy badaniu średniego wpływu stosowanych leków bądź trucizn na zdrowie badanych populacji osób. W końcu, w analizie zwrócono uwagę na sprawdzanie założeń (np. jednorodności wariancji składnika losowego), przy których analiza wariancji może być stosowana.

Część II. Metoda największej wiarygodności w analizie regresji Poissona, regresji logistycznej i w szeregach czasowych.

A. Rozdział 1. Wprowadzenie do metody największej wiarygodności.

Z powodu możliwości zastosowania *metody największej wiarygodności* (MNW) do rozwiązania wielu, bardzo różnych problemów estymacyjnych, stała się ona obecnie zarówno metodą podstawową jak również punktem wyjścia dla różnych metod analizy statystycznej. Jej wszechstronność związana jest, po pierwsze z możliwością przeprowadzenia analizy statystycznej dla małej próbki, opisu zjawisk nieliniowych oraz zastosowania zmiennych losowych posiadających zasadniczo dowolny *rozkład prawdopodobieństwa* [1], oraz po drugie, szczególnymi własnościami otrzymywanych przez nią estymatorów, które okazują się być zgodne, asymptotycznie nieobciążone, efektywne oraz dostateczne [36]. MNW zasadza się na intuicyjnie jasnym postulacie przyjęcia za prawdziwe takich wartości parametrów rozkładu prawdopodobieństwa zmiennej losowej, które maksymalizują funkcję wiarygodności realizacji konkretnej próbki.

Rozdział 1-1. Podstawowe pojęcia MNW.

Rozważmy zmienną losową Y [36], [37], która przyjmuje wartości y zgodnie z rozkładem prawdopodobieństwa $p(y \mid \theta)$, gdzie $\theta = (\vartheta_1, \vartheta_2, ..., \vartheta_k)^T \equiv (\vartheta_s)_{s=1}^k$, jest zbiorem k parametrów tego rozkładu (T oznacza transpozycję). Zbiór wszystkich możliwych wartości y zmiennej Y oznaczmy przez $\mathcal{Y}$.

Gdy $k > 1$ wtedy θ nazywamy parametrem *wektorowym*. W szczególnym przypadku $k = 1$ mamy $\theta = \vartheta$. Mówimy wtedy, że parametr θ jest parametrem *skalarnym*.

Pojęcie próby i próbki: Rozważmy *zbiór danych* $y_1, y_2, ..., y_N$ otrzymanych w N obserwacjach zmiennej losowej Y.

Każda z danych y_n, $n = 1,2,...,N$, jest generowana z rozkładu $p_n(y_n \mid \theta_n)$ zmiennej losowej Y w populacji, którą charakteryzuje wartość parametru wektorowego $\theta_n = (\vartheta_1, \vartheta_2, ..., \vartheta_k)_n^T \equiv ((\vartheta_s)_{s=1}^k)_n$, $n = 1,2,...,N$. Stąd zmienną Y w n-tej populacji oznaczymy Y_n. Zbiór zmiennych losowych $\tilde{Y} = (Y_1, Y_2, ..., Y_N) \equiv (Y_n)_{n=1}^N$ nazywamy N-wymiarową *próbą*.

Konkretną realizację $y = (y_1, y_2, ..., y_N) \equiv (y_n)_{n=1}^N$ próby $\tilde{Y}$ nazywamy *próbką*. Zbiór wszystkich możliwych realizacji y próby $\tilde{Y}$ tworzy przestrzeń próby (układu) oznaczaną jako B.

Określenie: Ze względu na to, że n jest indeksem konkretnego punktu pomiarowego próby, rozkład $p_n(y_n \mid \theta_n)$ będziemy nazywali rozkładem „punktowym" (czego nie należy mylić z np. rozkładem dyskretnym).

Określenie funkcji wiarygodności: Centralnym pojęciem MNW jest *funkcja wiarygodności* $L(y;\Theta)$ (pojawienia się) próbki $y = (y_n)_{n=1}^N$, nazywana też *wiarygodnością próbki*. Jest ona funkcją parametru Θ.

Przez wzgląd na zapis stosowany w fizyce, będziemy stosowali oznaczenie $P(y|\,\Theta) \equiv L(y;\Theta)$, które podkreśla, że formalnie *funkcja wiarygodności jest łącznym rozkładem prawdopodobieństwa* [38] pojawienia się realizacji $y \equiv (y_n)_{n=1}^N$ próby $\tilde{Y} \equiv (Y_n)_{n=1}^N$, to znaczy:

$$P(\Theta) \equiv P(y \mid \Theta) = \prod_{n=1}^{N} p_n(y_n \mid \theta_n). \tag{1-1.1}$$

Zwrócenie uwagi w (1-1.1) na występowanie y w argumencie funkcji wiarygodności oznacza, że może być ona rozumiana jako statystyka $P(\tilde{Y} \mid \Theta)$. Z kolei skrócone oznaczenie $P(\Theta)$ podkreśla, że centralną sprawą w MNW jest fakt, że funkcja wiarygodności jest funkcją nieznanych parametrów:

$$\Theta = (\theta_1, \theta_2, ..., \theta_N)^T \equiv (\theta_n)_{n=1}^N \quad \text{przy czym} \quad \theta_n = (\vartheta_{1n}, \vartheta_{2n}, ..., \vartheta_{kn})^T \equiv ((\vartheta_s)_{s=1}^k)_n\,, \tag{1-1.2}$$

gdzie θ_n jest wektorowym parametrem populacji określonej przez indeks próby n. W toku analizy chcemy oszacować wektorowy parametr Θ.

Zbiór wartości parametrów $\Theta = (\theta_n)_{n=1}^N$ tworzy współrzędne rozkładu prawdopodobieństwa rozumianego jako punkt w $d = k \times N$ - wymiarowej (podprzestrzeni) przestrzeni statystycznej S [39], [5].

Uwaga o postaci rozkładów punktowych: Tak jak w [5], zakładamy, że "*punktowe*" rozkłady $p_n(y_n \mid \theta_n)$ dla poszczególnych pomiarów n w N elementowej próbie są *niezależne*[11]. W ogólności [5], rozkłady punktowe $p_n(y_n \mid \theta_n)$ zmiennych Y_n chociaż są *tego samego typu*, jednak nie spełniają warunku $p_n(y_n \mid \theta_n) = p(y \mid \theta)$, charakterystycznego dla próby prostej. Taka ogólna sytuacja ma np. miejsce w analizie regresji Poissona (Rozdział 2).

[11] W przypadku analizy jednej zmiennej losowej Y, rozkłady te obok niezależności spełniają dodatkowo warunek:

$$p_n(y_n \mid \theta_n) = p(y \mid \theta), \tag{1-1.3}$$

co oznacza, że próba jest *prosta*.

Pojęcie estymatora parametru: Załóżmy, że dane $y = (y_n)_{n=1}^N$ są generowane losowo z punktowych rozkładów prawdopodobieństwa $p_n(y_n \mid \theta_n)$, $n = 1,2,...,N$, które chociaż nie są znane, to jednak założono o nich, że dla każdego n należą do określonej, tej samej klasy modeli. Zatem funkcja wiarygodności (1-1.1) należy do określonej, $d = k \times N$ - wymiarowej, przestrzeni statystycznej S.

Celem analizy jest oszacowanie nieznanych parametrów Θ, (1-1.2), poprzez funkcję:

$$\hat{\Theta} \equiv \hat{\Theta}(\tilde{Y}) = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N)^T \equiv (\hat{\theta}_n)_{n=1}^N \quad gdzie \quad \hat{\theta}_n = (\hat{\vartheta}_{1n}, \hat{\vartheta}_{2n}, ..., \hat{\vartheta}_{kn})^T \equiv ((\hat{\vartheta}_s)_{s=1}^k)_n \,, \quad (1\text{-}1.4)$$

mającą $d = k \times N$ składowych.

Każda z funkcji $\hat{\vartheta}_{kn} \equiv \hat{\vartheta}_{kn}(\tilde{Y})$ jako funkcja próby jest *statystyką*, którą przez wzgląd na to, że służy do oszacowywania wartości parametru ϑ_{kn} nazywamy estymatorem tego parametru. *Estymator parametru nie może zależeć od parametru, który oszacowuje*[12].

Podsumowując, odwzorowanie:

$$\hat{\Theta} : B \to \mathbf{R}^d \,, \qquad (1\text{-}1.5)$$

gdzie B jest przestrzenią próby, jest estymatorem parametru (wektorowego) Θ.

Określenie funkcji wynikowej: Funkcję $S(\Theta)$ będącą gradientem logarytmu funkcji wiarygodności:

$$S(\Theta) \equiv \frac{\partial}{\partial \Theta} \ln P(y|\,\Theta) = \begin{pmatrix} \dfrac{\partial \ln P(y \mid \Theta)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ln P(y \mid \Theta)}{\partial \theta_N} \end{pmatrix} \quad gdzie \quad \frac{\partial \ln P(y \mid \Theta)}{\partial \theta_n} = \begin{pmatrix} \dfrac{\partial \ln P(y \mid \Theta)}{\partial \vartheta_{1n}} \\ \vdots \\ \dfrac{\partial \ln P(y \mid \Theta)}{\partial \vartheta_{kn}} \end{pmatrix}, \quad (1\text{-}1.6)$$

nazywamy *funkcją wynikową*.

Równania wiarygodności: Będąc funkcją $\Theta = (\theta_n)_{n=1}^N$, funkcja wiarygodności służy do konstrukcji estymatorów $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N)^T \equiv (\hat{\theta}_n)_{n=1}^N$ parametrów $\Theta \equiv (\theta_n)_{n=1}^N$. Procedura polega na wyborze takich $(\hat{\theta}_n)_{n=1}^N$, dla których funkcja wiarygodności przyjmuje maksymalną wartość, skąd statystyki te nazywamy estymatorami MNW.

Zatem, wprowadzony przez Fishera, warunek konieczny otrzymania estymatorów $\hat{\Theta}$ MNW sprowadza się do znalezienia rozwiązania układu $d = k \times N$ tzw. *równań wiarygodności* [38]:

[12] Natomiast rozkład estymatora oszacowywanego parametru, zależy od tego parametru.

$$S(\Theta)_{|\Theta=\hat{\Theta}} \equiv \frac{\partial}{\partial\Theta} \ln P(y|\Theta)_{|\Theta=\hat{\Theta}} = 0\,, \tag{1-1.7}$$

gdzie zagadnienie maksymalizacji funkcji wiarygodności $P(y|\Theta)$ sprowadzono do (na ogół) analitycznie równoważnego mu problemu maksymalizacji jej logarytmu $\ln P(y|\Theta)$.

Ponieważ w Rozdziałach dotyczących regresji Poissona i regresji logistycznej ograniczymy się do rozkładów punktowych, dla których parametr jest skalarny, dlatego poniżej pominiemy indeks wewnętrzny k parametru.

Niech $\hat{\Theta} \equiv (\hat{\theta}_n)_{n=1}^N$ jest estymatorem MNW wektorowego parametru $\Theta \equiv (\theta_n)_{n=1}^N$ otrzymanym po rozwiązaniu układu równań wiarygodności (1-1.7). Aby zagwarantować maksimum funkcji wiarygodności, warunek (1-1.7) musi być uzupełniony warunkiem ujemności określoności poniższej formy kwadratowej w punkcie będącym rozwiązaniem równania (1-1.7) [38], [5]:

$$\sum_{n=0}^{N} \sum_{n'=0}^{N} \left. \frac{\partial^2 P(\tilde{Y},\Theta)}{\partial\theta_n \partial\theta_{n'}} \right|_{\Theta=\hat{\Theta}} \Delta\theta_n \Delta\theta_{n'} < 0 \tag{1-1.8}$$

gdzie rzeczywiste przyrosty $\Delta\theta_n$ $\Delta\theta_{n'}$ nie zerują się jednocześnie.

Macierz,

$$\mathbf{iF}(\hat{\Theta}) \equiv \left(-\left. \frac{\partial^2 P(\tilde{Y}\,|\,\Theta)}{\partial\theta_n \partial\theta_{n'}} \right|_{\Theta=\hat{\Theta}} \right) \tag{1-1.9}$$

jest tak zwaną obserwowaną w próbie informacją Fishera [38] dla parametru Θ, a warunek (1-1.8) oznacza żądanie jej dodatniej określoności. Poprzez macierz obserwowanej informacji Fishera można zdefiniować macierz kowariancji estymatorów parametrów modelu:

$$\hat{\mathbf{V}}(\hat{\Theta}) = \begin{bmatrix} \hat{\sigma}^2(\hat{\theta}_1) & \hat{Cov}(\hat{\theta}_1,\hat{\theta}_2) & \hat{Cov}(\hat{\theta}_1,\hat{\theta}_3) & \\ \hat{Cov}(\hat{\theta}_2,\hat{\theta}_1) & \hat{\sigma}^2(\hat{\theta}_2) & \hat{Cov}(\hat{\theta}_2,\hat{\theta}_3) & \\ \hat{Cov}(\hat{\theta}_3,\hat{\theta}_1) & \hat{Cov}(\hat{\theta}_3,\hat{\theta}_2) & \hat{\sigma}^2(\hat{\theta}_3) & \\ & & & \ddots \end{bmatrix} := \mathbf{iF}^{-1}(\hat{\Theta}) \tag{1-1.10}$$

Powyższa macierz kowariancji posiada na głównej przekątnej wariancję kolejnych estymatorów parametrów, a w miejscach poza przekątną, kowariancje między kolejnymi estymatorami parametrami. Występujące na przekątnej wariancje estymatorów, zostaną wykorzystane w budowie przedziałów ufności dla odpowiadających im parametrów.

Po otrzymaniu (wektora) estymatorów $\hat{\Theta}$, *zmaksymalizowaną* wartość funkcji wiarygodności definiujemy jako numeryczną wartość funkcji wiarygodności powstałą przez podstawienie do $P(y|\Theta)$ wartości oszacowanej $\hat{\Theta}$ w miejsce parametru Θ.

Przykład: Rozważmy problem estymacji skalarnego parametru, tzn. $\Theta = \theta$ (tzn. $k = 1$ oraz $N = 1$), dla zmiennej losowej Y opisanej rozkładem dwumianowym (Bernoulliego):

$$P(y \mid \theta) = \binom{m}{y} \theta^{y} (1-\theta)^{m-y} . \tag{1-1.11}$$

Estymacji parametru θ dokonamy na podstawie *pojedynczej* obserwacji (długość próby $N = 1$) zmiennej Y bądź Y/m. Parametr m charakteryzuje rozkład zmiennej Bernoulliego Y (i nie ma związku z długością N próby).

Zatem ponieważ $y \equiv (y_1)$, więc $P(y \mid \theta)$ jest funkcją wiarygodności dla $N = 1$ wymiarowej próby. Jej logarytm wynosi:

$$\ln P(y \mid \theta) = \ln \binom{m}{y} + y \ln \theta + (m-y) \ln(1-\theta). \tag{1-1.12}$$

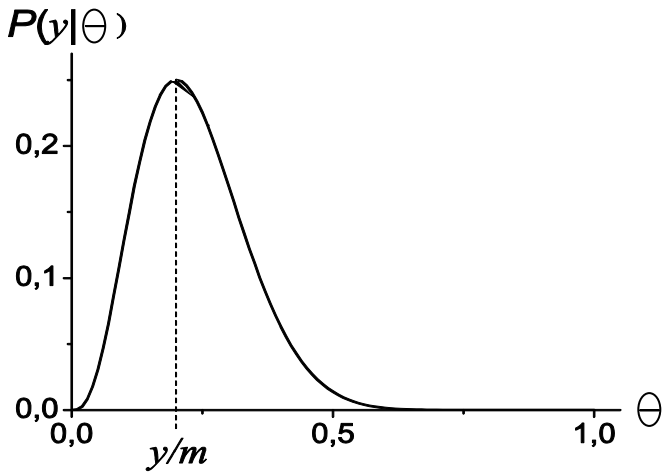
W rozważanym przypadku otrzymujemy jedno równanie wiarygodności (1-1.7):

$$S(\theta) = \frac{1}{\theta} y - \frac{1}{1-\theta} (m-y) \Big|_{\theta=\hat{\theta}} = 0 \tag{1-1.13}$$

a jego rozwiązanie daje estymator MNW parametru θ rozkładu dwumianowego, równy:

$$\hat{\theta} = \frac{y}{m} \tag{1-1.14}$$

Ilustracją powyższej procedury znajdowania wartości estymatora parametru θ jest Rysunek 1.1 (gdzie przyjęto $m = 5$). Na skutek pomiaru zaobserwowano wartość Y równą $y = 1$.



Rysunek 1.1. Graficzna ilustracja metody największej wiarygodności dla $P(y \mid \theta)$ określonego wzorem (1-1.11) dla rozkładu dwumianowego. Przyjęto wartość parametru $m = 5$. W pomiarze zaobserwowano wartość $Y = y = 1$.

Maksimum $P(y \mid \theta)$ przypada na wartość θ równą punktowemu oszacowaniu $\hat{\theta} = y/m = 1/5$ tego parametru. Maksymalizowana wartość funkcji wiarygodności wynosi $P(y \mid \hat{\theta})$.

Rozdział 1-2. Wnioskowanie w MNW.

Z powyższych rozważań wynika, że konstrukcja punktowego oszacowania parametru w MNW oparta jest o postulat maksymalizacji funkcji wiarygodności przedstawiony powyżej. Jest on wstępem do statystycznej procedury wnioskowania. Kolejnym krokiem jest konstrukcja przedziału wiarygodności. Jest on odpowiednikiem przedziału ufności, otrzymywanego w częstotliwościowym podejściu statystyki klasycznej do procedury estymacyjnej. Do jego konstrukcji niezbędna jest znajomość rozkładu prawdopodobieństwa estymatora parametru, co (dzięki ”porządnym”' granicznym własnościom stosowanych estymatorów) jest możliwe niejednokrotnie jedynie asymptotycznie, tzn. dla wielkości próby dążącej do nieskończoności. Znajomość rozkładu estymatora jest też niezbędna we wnioskowaniu statystycznym odnoszącym się do weryfikacji hipotez.

W sytuacji, gdy nie dysponujemy wystarczającą ilością danych, potrzebnych do przeprowadzenia skutecznego częstotliwościowego wnioskowania, Fisher [38] zaproponował do określenia niepewności dotyczącej parametru Θ wykorzystanie maksymalizowanej wartość funkcji wiarygodności.

Przedział wiarygodności jest zdefiniowany jako zbiór wartości parametru Θ, dla których funkcja wiarygodności osiąga (umownie) wystarczająco wysoką wartość, tzn.:

$$\left\{\Theta, \frac{P(y \mid \Theta)}{P(y \mid \hat{\Theta})} > c\right\}, \tag{1-1.15}$$

dla pewnego *parametru obcięcia* c, nazywanego *poziomem wiarygodności*.

Iloraz wiarygodności:

$$\frac{P(y \mid \Theta)}{P(y \mid \hat{\Theta})} \tag{1-1.16}$$

reprezentuje pewien typ unormowanej wiarygodności i jako taki jest wielkością skalarną. Jednak z powodu niejasnego znaczenia określonej wartości parametru obcięcia c pojęcie to wydaje się być na pierwszy rzut oka za słabe, aby dostarczyć taką precyzję wypowiedzi jaką daje analiza częstotliwościowa.

Istotnie, wartość c nie odnosi się do żadnej wielkości obserwowanej, tzn. na przykład 1%-we ($c = 0{,}01$) obcięcie nie ma ścisłego probabilistycznego znaczenia. Inaczej ma się sprawa dla częstotliwościowych przedziałów ufności. W tym przypadku wartość współczynnika $\alpha = 0{,}01$ oznacza, że gdybyśmy rozważyli realizację przedziału ufności na poziomie ufności $1-\alpha = 0{,}99$, to przy pobraniu nieskończonej (w praktyce wystarczająco dużej) liczby próbek, 99% wszystkich wyznaczonych przedziałów ufności pokryłoby prawdziwą (teoretyczną) wartość parametru Θ w populacji generalnej (składającej się z N podpopulacji). Pomimo tej słabości MNW, rozbudowanie analizy stosunku wiarygodności okazuje się być istotne we wnioskowaniu statystycznym analizy doboru modeli i to aż po konstrukcję równań teorii pola [5].

Rozdział 1-2-1. Wiarygodnościowy przedział ufności.

Przykład rozkładu normalnego z jednym estymowanym parametrem: Istnieje przypadek pozwalający na prostą *interpretację przedziału wiarygodnościowego jako przedziału ufności*. Dotyczy on zmiennej Y posiadającej rozkład Gaussa oraz sytuacji gdy (dla próby prostej) interesuje nas estymacja skalarnego parametru θ będącego wartością oczekiwaną $E(Y)$ zmiennej Y. Przypadek ten omówimy poniżej. W ogólności, przedział wiarygodności posiadający określony poziom ufności jest nazywany przedziałem ufności.

Częstotliwościowe wnioskowanie o nieznanym parametrze θ wymaga określenia rozkładu jego estymatora, co jest zazwyczaj możliwe jedynie granicznie [38]. Podobnie w MNW, o ile to możliwe, korzystamy przy dużych próbkach z twierdzeń granicznych dotyczących rozkładu ilorazu wiarygodności [38]. W przypadku rozkładu normalnego i parametru skalarnego okazuje się, że możliwa jest konstrukcja skończenie wymiarowa. Niech więc zmienna Y ma rozkład normalny $N(\theta,\sigma^2)$:

$$p(y\,|\,\theta,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(y-\theta)^2}{2\sigma^2}\right). \tag{1-1.17}$$

Rozważmy próbkę $y\equiv(y_1,\ldots,y_N)$, która jest realizacją próby prostej $\tilde{Y}$ i załóżmy, że *wariancja* σ^2 *jest znana*. Logarytm funkcji wiarygodności dla $N(\theta,\sigma^2)$ ma postać:

$$\ln P(y\,|\,\theta)=-\frac{N}{2}\ln 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\theta)^2\ , \tag{1-1.18}$$

gdzie ze względu na próbę prostą, w argumencie funkcji wiarygodności wpisano w miejsce $\Theta\equiv(\theta)_{n=1}^{N}$ parametr θ, jedyny który podlega estymacji.

Z postaci funkcji wiarygodności (1-1.18) oraz związku $\sum_{n=1}^{N}\left(y_n-\hat{\theta}\right)^2 = \sum_{n=1}^{N}\left((y_n-\theta)+(\theta-\hat{\theta})\right)^2$, otrzymujemy[13]:

---

[13] **Postać estymatora parametru skalarnego θ rozkładu $N(\theta,\sigma^2)$**: Korzystając z równania wiarygodności (1-1.7) dla przypadku skalarnego parametru θ, otrzymujemy:

$$S(\theta)\big|_{\theta=\hat{\theta}}\equiv\frac{\partial}{\partial\theta}\ln P(y|\,\theta)\big|_{\theta=\hat{\theta}}=0\,, \tag{1-1.20}$$

skąd dla log funkcji wiarygodności (1-1.18), otrzymujemy:

$$\hat{\theta}=\bar{y}=\frac{1}{N}\sum_{n=1}^{N}y_n\ . \tag{1-1.21}$$

Zatem estymatorem parametru θ jest średnia arytmetyczna:

$$\hat{\theta}=\bar{Y}=\frac{1}{N}\sum_{n=1}^{N}Y_n\ . \tag{1-1.22}$$

Estymator i jego realizowaną wartość będziemy oznaczali tak samo, tzn. $\hat{\theta}$ dla przypadku skalarnego i $\hat{\Theta}$ dla wektorowego.

$$\ln \frac{P(y \mid \theta)}{P(y \mid \hat{\theta})} = -\frac{N}{2\sigma^2}\left(\hat{\theta} - \theta\right)^2 , \tag{1-1.19}$$

gdzie $\hat{\theta} = \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$ jest estymatorem MNW parametru θ.

Statystyka Wilka: Widać, że po prawej stronie (1-1.19) otrzymaliśmy wyrażenie kwadratowe. Ponieważ $\bar{Y}$ jest nieobciążonym estymatorem parametru θ, co oznacza, że wartość oczekiwana $E(\bar{Y}) = \theta$, zatem (dla rozkładu $Y \sim N(\theta, \sigma^2)$) średnia arytmetyczna $\bar{Y}$ ma rozkład normalny $N\left(\theta, \frac{\sigma^2}{N}\right)$. Z normalności rozkładu $\bar{Y}$ wynika, że tzw. *statystyka ilorazu wiarygodności Wilka*:

$$W \equiv 2\ln \frac{P(\tilde{Y} \mid \hat{\theta})}{P(\tilde{Y} \mid \theta)} \sim \chi_1^2 , \tag{1-1.23}$$

ma rozkład χ^2, w tym przypadku z jednym stopniem swobody [38], [1].

Wyskalowanie statystyki Wilka w przypadku normalnym: Wykorzystując (1-1.23) możemy wykonać wyskalowanie wiarygodności oparte o możliwość powiązania przedziału wiarygodności z jego częstotliwościowym odpowiednikiem.

Mianowicie z (1-1.23) otrzymujemy, że dla ustalonego (chociaż nieznanego) parametru θ prawdopodobieństwo, że iloraz wiarygodności znajduje się w wyznaczonym dla parametru obcięcia c, wiarygodnościowym przedziale ufności, wynosi:

$$P\left(\frac{P(\tilde{Y} \mid \theta)}{P(\tilde{Y} \mid \hat{\theta})} > c\right) = P\left(2\ln \frac{P(\tilde{Y} \mid \hat{\theta})}{P(\tilde{Y} \mid \theta)} < -2\ln c\right) = P\left(\chi_1^2 < -2\ln c\right). \tag{1-1.24}$$

Zatem jeśli dla jakiegoś $0 < (1-\alpha) < 1$ wybierzemy parametr obcięcia:

$$c = e^{-\frac{1}{2}\chi^2_{1,(1-\alpha)}} , \tag{1-1.25}$$

gdzie $\chi^2_{1,(1-\alpha)}$ jest kwantylem rzędu $100(1-\alpha)\%$ rozkładu χ-kwadrat, to spełnienie przez θ związku:

$$P\left(\frac{P(\tilde{Y} \mid \theta)}{P(\tilde{Y} \mid \hat{\theta})} > c\right) = P\left(\chi_1^2 < \chi^2_{1,(1-\alpha)}\right) = 1-\alpha \tag{1-1.26}$$

oznacza, że przyjęcie wartości c zgodnej z (1-1.25) daje zbiór możliwych wartości parametru θ:

$$\left\{\theta, \frac{P(\tilde{Y} \mid \theta)}{P(\tilde{Y} \mid \hat{\theta})} > c\right\}, \tag{1-1.27}$$

nazywany $100(1-\alpha)\%$-owym (*wiarygodnościowym*) *przedziałem ufności*. Jest on odpowiednikiem wyznaczonego na poziomie ufności $(1-\alpha)$ częstotliowściowego przedziału ufności dla θ. Dla analizowanego przypadku rozkładu normalnego z estymacją skalarnego parametru θ oczekiwanego poziomu zjawiska, otrzymujemy po skorzystaniu z wzoru (1-1.25) wartość parametru obcięcia równego $c = 0.15$ lub $c = 0.04$ dla odpowiednio 95%-owego $(1-\alpha = 0.95)$ bądź 99%-owego $(1-\alpha = 0.99)$ przedziału ufności. Tak więc w przypadku, *gdy przedział wiarygodności da się wyskalować rozkładem prawdopodobieństwa, parametr obcięcia* c *posiada własność wielkości obserwowanej, interpretowanej częstotliwościowo poprzez związek z poziomem ufności*.

Zwróćmy uwagę, że chociaż konstrukcje częstotliwościowego i wiarygodnościowego przedziału ufności są różne, to *ich losowość wynika* w obu przypadkach *z rozkładu prawdopodobieństwa estymatora* $\hat{\theta}$.

Ćwiczenie: W oparciu o powyższe rozważania wyznaczyć, korzystając z (1-1.19) ogólną postać przedziału wiarygodności dla skalarnego parametru θ rozkładu normalnego.

Rozdział 1-2-2. Rozkłady regularne.

Dla zmiennych o innym rozkładzie niż rozkład normalny, statystyka Wilka W ma w ogólności inny rozkład niż χ^2 [38]. Jeśli więc zmienne nie mają dokładnie rozkładu normalnego lub dysponujemy za małą próbką by móc odwoływać się do (wynikających z twierdzeń granicznych) rozkładów granicznych dla estymatorów parametrów, wtedy związek (1-1.23) (więc i (1-1.25)) daje jedynie przybliżone wyskalowanie przedziału wiarygodności rozkładem χ^2.

Jednakże w przypadkach wystarczająco *regularnych rozkładów*, zdefiniowanych jako takie, w których możemy zastosować przybliżenie kwadratowe:

$$\ln \frac{P(y \mid \theta)}{P(y \mid \hat{\theta})} \approx -\frac{1}{2}\mathbf{iF}(\hat{\theta})(\hat{\theta}-\theta)^2 , \qquad (1\text{-}1.28)$$

powyższe rozumowanie oparte o wyskalowanie wiarygodności rozkładem χ_1^2 jest w przybliżeniu słuszne. Wielkość $\mathbf{iF}(\hat{\theta})$, która pojawiła się powyżej jest *obserwowaną* informacją Fishera, a powyższa formuła stanowi poważne narzędzie w analizie doboru modeli [5], [38].

Przykład: Rozważmy przypadek parametru skalarnego θ w jednym eksperymencie ($N = 1$) ze zmienną Y posiadającą rozkład Bernoulliego z $m = 15$. W wyniku pomiaru zaobserwowaliśmy wartość $Y = \boldsymbol{y} = 3$. Prosta analiza pozwala wyznaczyć wiarygodnościowy przedział ufności dla parametru θ. Ponieważ przestrzeń V_θ parametru θ wynosi $V_\theta = (0,1)$, zatem łatwo pokazać, że dla $c = 0{,}01$, $c = 0{,}1$ oraz $c = 0{,}5$ miałby on realizację odpowiednio $(0{,}019;0{,}583)$, $(0{,}046;0{,}465)$ oraz $(0{,}098;0{,}337)$. Widać, że wraz ze

wzrostem wartości c, przedział wiarygodności zacieśnia się wokół wartości oszacowania punktowego $\hat{\theta} = y/m = 1/5$ parametru θ i nic dziwnego, bo wzrost wartości c oznacza akceptowanie jako możliwych do przyjęcia tylko takich *modelowych wartości parametru* θ, które gwarantują wystarczająco wysoką wiarygodność próbki.

Powyższy przykład pozwala nabyć pewnej intuicji co do sensu stosowania ilorazu funkcji wiarygodności. Mianowicie po otrzymaniu w pomiarze określonej wartości y/m oszacowującej parametr θ, jesteśmy skłonni preferować model z taką wartością parametru θ, która daje większą wartość (logarytmu) ilorazu wiarygodności $P(y\,|\,\theta)/P(y\,|\,\hat{\theta})$. Zgodnie z podejściem statystyki klasycznej *nie oznacza to jednak*, że uważamy, że parametr θ ma jakiś rozkład. Jedynie wobec niewiedzy co do modelowej (populacyjnej) wartość parametru θ preferujemy ten model, który daje większą wartość ilorazu wiarygodności w próbce.

Rozdział 1-2-3. Weryfikacja hipotez z wykorzystaniem ilorazu wiarygodności.

Powyżej wykorzystaliśmy funkcję wiarygodności do *estymacji wartości parametru* Θ. Funkcję wiarygodności można również wykorzystać w drugim typie wnioskowania statystycznego, tzn. w *weryfikacji hipotez statystycznych*.

Rozważmy prostą hipotezę zerową $H_0 : \Theta = \Theta_0$ wobec złożonej hipotezy alternatywnej $H_1 : \Theta \neq \Theta_0$. W celu przeprowadzenia *testu statystycznego* wprowadźmy unormowaną funkcję wiarygodności:

$$\frac{P(y\,|\,\Theta_0)}{P(y\,|\,\hat{\Theta})}, \qquad (1\text{-}1.29)$$

skonstruowaną przy założeniu prawdziwości hipotezy zerowej. Hipotezę zerowa H_0 odrzucamy na rzecz hipotezy alternatywnej, jeśli jej wiarygodność $P(y\,|\,\Theta_0)$ jest ”za mała”. Sugerowałoby to, że złożona hipoteza alternatywna H_1 zawiera pewną hipotezę prostą, która jest lepiej poparta przez dane otrzymane w próbce, niż hipoteza zerowa.

Jak o tym wspomnieliśmy powyżej, np. 5%-owe obcięcie c w zagadnieniu estymacyjnym, samo w sobie nie mówi nic o frakcji liczby przedziałów wiarygodności pokrywających nieznaną wartość szacowanego parametru. Potrzebne jest wyskalowanie ilorazu wiarygodności. Również dla weryfikacji hipotez skalowanie wiarygodności jest istotne. Stwierdziliśmy, że takie skalowanie jest możliwe wtedy gdy mamy do czynienia z jednoparametrowym przypadkiem rozkładu Gaussa, a przynajmniej z przypadkiem wystarczająco regularnym.

Empiryczny poziom istotności: W przypadku jednoparametrowego, regularnego problemu z ($\Theta \equiv (\theta)_{n=1}^{N}$) jak w Przykładzie z Rozdziału 1-2-1, skalowanie poprzez wykorzystanie statystki Wilka służy otrzymaniu empirycznego poziomu istotności p. Ze związku (1-1.23) otrzymujemy wtedy przybliżony (a dokładny dla rozkładu normalnego) *empiryczny poziom istotności*:

$$\begin{aligned} p &\approx P\left(\frac{P(\tilde{Y} \mid \hat{\theta})}{P(\tilde{Y} \mid \theta_0)} \geq \frac{P(y \mid \hat{\theta}_{obs})}{P(y \mid \theta_0)}\right) = P\left(2\ln\frac{P(\tilde{Y} \mid \hat{\theta})}{P(\tilde{Y} \mid \theta_0)} \geq -2\ln c_{obs}\right) \\ &= P(\chi_1^2 \geq -2\ln c_{obs}), \quad gdzie \quad c_{obs} \equiv \frac{P(y \mid \theta_0)}{P(y \mid \hat{\theta}_{obs})}, \end{aligned} \tag{1-1.30}$$

przy czym $\hat{\theta}_{obs}$ jest wartością estymatora MNW $\hat{\theta}$ wyznaczoną w obserwowanej (obs) próbce y. Powyższe określenie empirycznego poziomu istotności p oznacza, że w przypadku wystarczająco regularnego problemu [38], istnieje typowy związek pomiędzy prawdopodobieństwem (1-1.26), a empirycznym poziomem istotności p, podobny do związku jaki istnieje pomiędzy poziomem ufności $1-\alpha$, a poziomem istotności α w analizie częstotliwościowej. I tak, np. w przypadku jednoparametrowego rozkładu normalnego możemy wykorzystać wartość empirycznego poziomu istotności p do stwierdzenia, że gdy $p \leq \alpha$ to hipotezę H_0 odrzucamy na rzecz hipotezy H_1, a w przypadku $p > \alpha$ nie mamy podstawy do odrzucenia H_0.

Problem błędu pierwszego i drugiego rodzaju: Jednakże podobne skalowanie ilorazu wiarygodności okazuje się być znacznie trudniejsze już chociażby tylko w przypadku dwuparametrowego rozkładu normalnego, gdy obok θ estymujemy σ^2 [38]. Wtedy określenie co oznacza sformułowanie „zbyt mała" wartość c jest dość dowolne i zależy od rozważanego problemu lub wcześniejszej wiedzy wynikającej z innych źródeł niż prowadzone statystyczne wnioskowanie. Wybór dużego parametru obcięcia c spowoduje, że istnieje większe prawdopodobieństwo popełnienia *błędu pierwszego rodzaju* polegającego na odrzuceniu hipotezy zerowej w przypadku, gdy jest ona prawdziwa. Wybór małego c spowoduje zwiększenie prawdopodobieństwa popełnienia *błędu drugiego rodzaju*, tzn. przyjęcia hipotezy zerowej w sytuacji, gdy jest ona błędna.

Rozdział 1-3. MNW w analizie regresji.

W metodzie regresji klasycznej, estymatory parametrów strukturalnych modelu regresji są otrzymane arytmetyczną metodą najmniejszych kwadratów (MNK). Zmienne objaśniające $X_n = x_n$, $n = 1,..., N$, nie mają wtedy charakteru stochastycznego, co oznacza, że eksperyment jest ze względu na nie kontrolowany.

MNK polega na minimalizacji sumy kwadratów odchyleń obserwowanych wartości zmiennej objaśnianej (tzw. odpowiedzi) od ich wartości teoretycznych spełniających równanie regresji. MNK ma znaczenie probabilistyczne tylko w przypadku analizy standardowej, gdy zmienna objaśniana Y ma rozkład normalny. Jej estymatory pokrywają się wtedy z estymatorami MNW. Pokażemy, że tak się sprawy mają.

Załóżmy, że zmienne $Y_1, Y_2,..., Y_N$ odpowiadające kolejnym wartościom zmiennej objaśniającej, x_1, $x_2,..., x_N$, są względem siebie niezależne i mają rozkład normalny ze średnią $\mu_n = E(Y|x_n) = E(Y_n)$ zależną od wariantu zmiennej objaśniającej x_n, oraz taką samą wariancję $\sigma^2(Y_n) = \sigma^2(Y)$.

Funkcja wiarygodności próbki $(y_1, y_2,..., y_N)$ dla normalnego klasycznego modelu regresji z parametrem $\Theta = \mu \equiv (\mu_n)_{n=1}^N$, ma postać:

$$\begin{aligned} P(\mu) \equiv P(y \mid \mu) &= \prod_{n=1}^{N} f(y_n \mid \mu_n) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mu_n)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mu_n)^2\right\}, \end{aligned} \tag{1-1.31}$$

gdzie $f(y_n \mid \mu_n)$, $n = 1,2..., N$, są punktowymi rozkładami gestości prawdopodobieństwa Gaussa. Widać, że maksymalizacja $P(\mu)$ ze względu na $(\mu_n)_{n=1}^N$ pociąga za sobą minimalizację sumy kwadratów reszt[14] (SKR):

$$SKR = \sum_{n=1}^{N}(y_n - \mu_n)^2 , \tag{1-1.32}$$

gdzie $\mu_n = E(Y|x_n)$ jest *postulowanym modelem regresji*. Zatem w standardowej, klasycznej analizie regresji, estymatory MNW pokrywają się z estymatorami MNK. Widać, że procedura minimalizacji dla SKR prowadzi do liniowej w Y_n postaci estymatorów $\hat{\mu}_n$ parametrów μ_n.

[14] SSE w literaturze anglojęzycznej (w tym w raportach SAS'a).

Problem z nieliniowym układem równań wiarygodności: Jednak rozwiązanie układu równań wiarygodności (1-1.7) jest zazwyczaj nietrywialne. Jest tak, gdy otrzymany w wyniku ekstremizacji układ algebraicznych równań wiarygodności dla estymatorów jest nieliniowy, co w konsekwencji oznacza, że możemy nie otrzymać ich w zwartej analitycznej postaci. Przykładem może być analiza regresji Poissona, w której do rozwiązania równań wiarygodności wykorzystujemy metody iteracyjne. W takich sytuacjach wykorzystujemy na ogół jakiś program komputerowy do analizy statystycznej, np. zawarty w pakiecie SAS. Po podaniu postaci funkcji wiarygodności, program komputerowy dokonuje jej maksymalizacji rozwiązując układ (1-1.7) np. metodą Newton-Raphson'a [38], wyznaczając numerycznie wartości estymatorów parametrów modelu.

Testy statystyczne: Logarytm ilorazu wiarygodności jest również wykorzystywany w analizie regresji do przeprowadzania testów statystycznych przy weryfikacji hipotez o nie występowaniu braku dopasowania modelu mniej złożonego, tzw. ”niższego”, o mniejszej liczbie parametrów, w stosunku do bardziej złożonego modelu ”wyższego”, posiadającego większą liczbę parametrów. Tzw. *statystyka ilorazu wiarygodności* (likelihood ratio) wykorzystywana do tego typu testów ma postać [1], [38]:

$$LR_{\Theta_1/\Theta_2} = -2\ln\frac{P\left(\tilde{Y} \mid \hat{\Theta}_1\right)}{P\left(\tilde{Y} \mid \hat{\Theta}_2\right)} \tag{1-1.33}$$

gdzie $P\left(\tilde{Y} \mid \hat{\Theta}_1\right)$ jest maksymalizowaną wartością funkcji wiarygodności dla modelu mniej złożonego, a $P\left(\tilde{Y} \mid \hat{\Theta}_2\right)$ dla modelu bardziej złożonego. Przy prawdziwości hipotezy zerowej H_0 o braku konieczności rozszerzania modelu niższego do wyższego, statystyka (1-1.33) ma asymptotycznie rozkład χ^2 z liczbą stopni swobody równą różnicy liczby parametrów modelu wyższego i niższego.

Ponieważ $P\left(\tilde{Y} \mid \hat{\Theta}_2\right) \geq P\left(\tilde{Y} \mid \hat{\Theta}_1\right)$ zatem:

$$0 \leq LR_{\Theta_1/\Theta_2} < +\infty \tag{1-1.34}$$

Analogia współczynnika determinacji: Maksymalizowana wartość funkcji wiarygodności zachowuje się podobnie jak *współczynnik determinacji* R^2 [1], tzn. rośnie wraz ze wzrostem liczby parametrów w modelu, zatem wielkość pod logarytmem należy do przedziału $(0,1)$ i statystyka (1-1.33) przyjmuje wartości z przedziału $(0,+\infty)$. Stąd (asymptotycznie) zbiór krytyczny dla H_0 jest prawostronny. Im lepiej więc model wyższy dopasowuje się do danych empirycznych w stosunku do modelu niższego, tym większa jest wartość statystyki ilorazu wiarygodności (1-1.33) i większa szansa, że wpadnie ona w przedział odrzuceń hipotezy zerowej H_0, który leży w prawym ogonie wspomnianego rozkładu χ^2 [1].

Rozdział 1-4. Test statystyczny dla doboru modelu.

Przyczyna nielosowej zmiany wartości zmiennej objaśnianej: Rozważmy model regresji dla zmiennej objaśnianej Y posiadającej określony rozkład (np. Poissona albo dychotomiczny). Zmienne Y_n, $n = 1,2,..., N$ posiadają więc również ten sam rozkład. Zakładamy, że są one *parami wzajemnie niezależne*. Niech X jest zmienną objaśniającą (tzw. czynnikiem) kontrolowanego eksperymentu, w którym X nie jest zmienną losową, ale *jej zmiana*, jest rozważana jako możliwa przyczyna warunkująca *nielosową zmianę wartości zmiennej* Y.

Gdy czynników $X_1, X_2,...X_k$ jest więcej, wtedy dla każdego punktu n próby podane są wszystkie ich wartości:

$$x_{1n}, x_{2n},...x_{kn}\ , \quad gdzie \quad n = 1,2,..., N\ , \tag{1-1.35}$$

gdzie pierwszy indeks w x_{in}, $i = 1,2,..., k$, numeruje zmienną objaśniającą.

Brak możliwości eksperymentalnej separacji podstawowego kanału n: Niech $x_n = (x_{1n}, x_{2n},..., \ x_{kn})$ oznacza zbiór wartości jednego wariantu zmiennych $(X_1, X_2,..., X_k)$, tzn. dla jednej konkretnej podgrupy n. Zwróćmy uwagę, że *indeks próby* n numeruje podgrupę, co oznacza, że w pomiarze wartości Y_n nie ma możliwości eksperymentalnego sięgnięcia ”w głąb” indeksu n - tego kanału, tzn. do rozróżnienia wpływów na wartość y_n płynących z różnych ”pod-kanałów” i, gdzie $i = 1,2,..., k$.

Rozdział 1-4-1. Model podstawowy.

Zakładając brak zależności zmiennej Y od czynników $X_1, X_2,...X_k$, rozważa się tzw. *model podstawowy*. Dla próby $\tilde{Y} \equiv (Y_n)_{n=1}^N$, funkcja wiarygodności próby z parametrem $\Theta = \mu \equiv (\mu_n)_{n=1}^N$:

$$P(\mu) \equiv P(\tilde{Y} \mid \mu) = \prod_{n=1}^{N} p(y_n \mid \mu_n) \tag{1-1.36}$$

jest funkcją wektorowego parametru $\mu \equiv (\mu_n)_{n=1}^N$. W poniższych rozdziałach rozważymy regresję dla Poissona oraz dychotomicznego, dla których każdy z parametrów $\mu_n = E(Y_n)$ jest parametrem skalarnym. N jest równocześnie liczebnością zbioru danych, która może być liczbą podgrup, komórek lub kategorii, oraz liczbą parametrów modelu podstawowego występującą w wiarygodności (2-1-2.5).

Układ równań MNW ma postać:

$$\frac{\partial}{\partial \mu_n}\left[\ln P(\tilde{Y} \mid \mu)\right]_{|\mu=\hat{\mu}} = 0\ , \quad n = 1,2,...,N\ . \tag{1-1.37}$$

Rozwiązaniem powyższego układu równań wiarygodności są estymatory $\hat{\mu} \equiv (\hat{\mu}_1, \hat{\mu}_2, ...\hat{\mu}_N)$ wektorowego parametru $\mu \equiv (\mu_n)_{n=1}^{N}$. Na funkcję wiarygodności modelu podstawowego *nie narzuca się żadnych ograniczeń na postać* μ_n, zatem w żadnym modelu regresji nie można otrzymać większej wartości funkcji wiarygodności w próbce niż w modelu zmaksymalizowanym modelu podstawowym.

Niech $\mu_n \equiv E(Y_n) = \mu(x_n, \beta), \quad n = 1,2,...,N,$ jest funkcją regresji badanego modelu z parametrami $\beta = (\beta_0, \beta_1, ..., \beta_k)$. Funkcja wiarygodności próby przy rozważanym modelu regresji ma postać:

$$P(\tilde{Y} \mid \beta) \;=\; \prod_{n=1}^{N} p(Y_n \mid \beta) \tag{1-1.38}$$

Estymatory MNW, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, parametrów $\beta_0, \beta_1, ..., \beta_k$ otrzymuje się rozwiązując $k+1$ równań wiarygodności:

$$\frac{\partial}{\partial \beta_j} \ln P(\tilde{Y} \mid \beta)\,|_{\beta=\hat{\beta}} = 0, \quad j = 0,1,2,...,k. \tag{1-1.39}$$

Rozwiązaniem powyższego układu równań wiarygodności są estymatory $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ wektorowego parametru $\beta = (\beta_0, \beta_1, ..., \beta_k)$.

Hipotezę zerową o *nie występowaniu braku dopasowania* badanego modelu w porównaniu z modelem podstawowym można zapisać w postaci proponującej postać tego modelu:

$$\mathrm{H}_0 : \mu_n = \mu(x_n, \beta), \quad n = 1,2,...,N. \tag{1-1.40}$$

Stawiamy ją wobec hipotezy alternatywnej:

$$\mathrm{H}_A : \; \mu_n \text{ nie ma ograniczonej postaci}, \; n = 1,2,...,N. \tag{1-1.41}$$

która odpowiada wyborowi modelu podstawowego zawierającego tyle parametrów μ_n ile jest punktów pomiarowych, tzn. *N*, z funkcją wiarygodności (1-1.36).

Niech więc $P(\tilde{Y} \mid \hat{\beta})$ jest maksymalną wartością funkcji wiarygodności określoną jak w (1-1.38). Oznacza to, że w miejsce parametrów $\beta = (\beta_0, \beta_1, ..., \beta_k)$ podstawiono ich estymatory $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ wyznaczone przez MNW, jako te które maksymalizują funkcję wiarygodności (1-1.38). Podobnie rozumiemy funkcję wiarygodności $P(\tilde{Y} \mid \hat{\mu})$ modelu podstawowego.

Ponieważ celem każdej analizy jest otrzymanie możliwie najprostszego opisu danych, model $\mu_n = \mu(x_n, \beta)$ zawierający $k+1$ parametrów β, będzie uznany za dobry, jeśli maksymalna wartość funkcji wiarygodności wyznaczona dla niego, będzie prawie tak duża, jak funkcji wiarygodności dla nie niosącego żadnej informacji modelu podstawowego z liczbą parametrów μ_n równą licznie punktów pomiarowych N. Sformułowanie ”prawie tak duża” oznacza, że wartość funkcji wiarygodności $P(y \mid \hat{\beta})$ nie może być istotnie statystycznie

mniejsza od $P(y \mid \hat{\mu})$. Zasadniczo powinno to oznaczać, że musimy podać miary pozwalające na określenie statystycznej istotności przy posługiwaniu się intuicyjnym parametrem obcięcia c (Rozdział 1-2-1). Okazuje się, że dla dużej próby, miary typu (1-1.42), podane poniżej, uzyskują cechy pozwalające na budownie wiarygodnościowych obszarów krytycznych nabywających charakteru standardowego (częstotliwościowego).

Określenie dewiancji: Wprowadźmy *statystykę typu ilorazu wiarygodności*:

$$D(\hat{\beta}) = -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta})}{P(\tilde{Y} \mid \hat{\mu})}\right] \tag{1-1.42}$$

nazywaną *dewiancją* (deviance) dla modelu regresji. Jej przykład zostanie omówiony dalej na przykładzie modelu regresji Poissona. Służy ona do badania dobroci dopasowania modelu z zadaną postacią $\mu_n = \mu(x_n, \beta)$ w stosunku do modelu podstawowego, bez narzuconej postaci na μ_n, tzn. do stwierdzenia, czy $P(y \mid \hat{\beta})$ jest istotnie *mniejsza* od $P(y \mid \hat{\mu})$, co sugerowałoby istotny statystycznie brak dopasowania badanego modelu $\mu_n = \mu(x_n, \beta)$, do danych empirycznych. Jak pokażemy poniżej dewiancja może być rozumiana jako *miara zmienności reszt wokół linii regresji* (tzn. odchylenia wartości obserwowanych w próbie od wartości szacowanych przez model), na której leżą wartości przewidywane $\hat{y}_j$ przez model [1].

A. Rozdział 2. Analiza doboru modelu regresji Poissona.

Celem obecnego Rozdziału jest praktyczne wyjaśnienie działania metody największej wiarygodności (MNW) oparte o przykład analizy doboru modelu dla regresji Poissona, z wykorzystaniem możliwości procedur zawartych w pakiecie SAS (system analiz statystycznych). Podstawy teoretyczne MNW oraz aparatu matematycznego związanego z zastosowaniem informacji Fishera może czytelnik znaleźć między innymi w pozycji [5].

MNW jest ogólną statystyczną metodą otrzymywania estymatorów parametrów populacyjnych modelu statystycznego. Estymatory MNW mają dla dużej próbki optymalne właściwości statystyczne [36]. Dla małej próbki skorzystanie z pełni praktycznych zalet MNW możliwe jest dopiero po odwołaniu się do formalizmu geometrii różniczkowej na przestrzeni statystycznej modeli statystycznych [39], [5].

Zaletą MNW w estymacji parametrów jest to, że można ją zastosować w rozmaitych sytuacjach. Jej ważną cechą jest to, że ogólne zasady i procedury mogą być używane do przeprowadzania wnioskowania statystycznego dla modeli regresji ze zmienną objaśnianą o dowolnym rozkładzie. Stąd to samo wnioskowanie statystyczne MNW może być (z dokładnością do różnic modelowych) zastosowane w analizie regresji np. klasycznego modelu normalnego, jak i w analizie regresji Poissona.

Gdy model wielorakiej regresji liniowej jest dopasowany do danych empirycznych zmiennej objaśnianej posiadającej rozkład normalny, wtedy estymatory współczynników regresji metody najmniejszych kwadratów

(MNK) są identyczne jak estymatory otrzymane w MNW [38], [1]. Estymacja MNW parametrów modelu umożliwia również analizę modeli nieliniowych, takich jak np. model regresji logistycznej [1] oraz rozważany w niniejszej części skryptu model regresji Poissona. Zrozumienie działania MNW w estymacji parametrów i umiejętność dokonywania wyboru modelu w oparciu o odpowiednie testy statystyczne jest niezbędną umiejętnością współczesnych analiz statystycznych w wielu dziedzinach nauk empirycznych.

Analiza regresji Poissona jest stosowana w modelowaniu zależności pomiędzy zmiennymi w przypadku, gdy zależna zmienna losowa (nazywana też zmienną opisywaną lub odpowiedzią) przyjmuje z natury tej zmiennej realizacje w postaci zbioru dyskretnych danych. Na przykład zmienna objaśniana może być liczbą zliczeń przypadków interesującego nas zdarzenia, np. liczbą przypadków awarii, które pojawiają się w ustalonym czasie badania.

Dla typowego modelu regresji Poissona naturalną miarą estymowanego defektu jest ryzyko względne, związane z określonym, interesującym nas czynnikiem.

Celem obecnego Rozdziału jest wyjaśnienie jak postulować i badać postać modelu regresji Poissona oraz jak wykorzystywać kluczowe cechy modelu do estymacji parametru ryzyka względnego, kontrastującego porównywane zbiorowości ze względu na warianty czynników ryzyka. Wykorzystamy pojęcia statystyki ilorazu wiarygodności oraz dewiancji [1], stosując je do analizy selekcji modelu właściwego dla przykładowych danych (których realizacja jest możliwa), co do których uznamy, że pochodzą z rozkładu [2] Poissona. Przedstawiony zostanie typowy model regresji Poissona, który wyraża w postaci logarytmicznej tempo porażki (np. awarii) jako liniowej funkcji zbioru czynników. Metoda regresji Poissona, może być również zastosowana w bardziej skomplikowanych nieliniowych modelach. Zainteresowanego czytelnika odsyłamy do [1].

Rozdział 2-1. Analiza doboru modelu regresji dla rozkładu Poissona.

Rozdział 2-1-1. Dewiancja jako miara dobroci dopasowania. Rozkład Poissona.

Rozważmy zmienną losową Y posiadającą rozkład Poissona. Rozkład ten jest wykorzystywany do modelowania zjawisk związanych z rzadko zachodzącymi zdarzeniami, jak na przykład z liczbą rozpadających się niestabilnych jąder w czasie t. Ma on postać:

$$p(Y = y \mid \mu) = \frac{\mu^y e^{-\mu}}{y!}, \quad oraz \quad y = 0,1,...,\infty, \tag{2-1-1.1}$$

gdzie μ jest parametrem rozkładu. Zmienna losowa podlegająca rozkładowi Poissona może przyjąć tylko nieujemną wartość całkowitą. Rozkład ten można wyprowadzić z rozkładu dwumianowego, bądź wykorzystując rozkłady Erlanga i wykładniczy [36].

Na przykład, zgodnie z (2-1-1.1) prawdopodobieństwo, że Y przyjmuje wartość $y = 7$ wynosi:

$$p(y = 7 \mid \mu) = \frac{\mu^7 e^{-\mu}}{7!} = \frac{\mu^7 e^{-\mu}}{5040} \ .$$

Widać, że prawdopodobieństwo to zmienia się jako funkcja wartości parametru μ. Jak już wiemy w MNW koncentrujemy się na badaniu zależności rozkładu prawdopodobieństwa zmiennej objaśnianej, od parametrów tego rozkładu.

Związek wariancji z wartością oczekiwaną rozkład Poissona: Rozkład Poissona posiada pewną interesującą właściwość statystyczną, mianowicie jego wartość oczekiwana, wariancja i trzeci moment centralny są równe parametrowi rozkładu μ:

$$E(Y) = \sigma^2(Y) = \mu_3 = \mu \, . \qquad (2\text{-}1\text{-}1.2)$$

Aby pokazać dwie pierwsze równości w (2-1-1.2) skorzystajmy bezpośrednio z definicji odpowiednich momentów, otrzymując:

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} y \cdot p(Y = y \mid \mu) = \sum_{y=0}^{\infty} y \cdot \frac{\mu^y e^{-\mu}}{y!} = e^{-\mu} \sum_{y=1}^{\infty} \frac{\mu^y}{(y-1)!} \\ &= e^{-\mu} \mu \sum_{y=1}^{\infty} \frac{\mu^{y-1}}{(y-1)!} = e^{-\mu} \mu \sum_{l=0}^{\infty} \frac{\mu^l}{l!} = e^{-\mu} \mu \, e^{\mu} = \mu \, , \end{aligned} \qquad (2\text{-}1\text{-}1.3)$$

oraz, korzystając z (2-1-1.3):

$$\begin{aligned} \sigma^2(Y) &= E(Y^2) - [E(Y)]^2 = E(Y^2) - \mu^2 = \sum_{y=0}^{\infty} y^2 \cdot p(Y = y \mid \mu) - \mu^2 \\ &= \sum_{y=0}^{\infty} y^2 \cdot \frac{\mu^y e^{-\mu}}{y!} - \mu^2 = e^{-\mu} \sum_{y=1}^{\infty} y \frac{\mu^y}{(y-1)!} - \mu^2 = e^{-\mu} \mu \sum_{l=0}^{\infty} (l+1) \frac{\mu^l}{l!} - \mu^2 \\ &= e^{-\mu} \mu \left[\sum_{l=0}^{\infty} l \frac{\mu^l}{l!} + e^{\mu} \right] - \mu^2 = e^{-\mu} \mu \left[e^{\mu} \mu + e^{\mu} \right] - \mu^2 = (\mu^2 + \mu) - \mu^2 = \mu \, . \end{aligned} \qquad (2\text{-}1\text{-}1.4)$$

Uwaga: *Zatem otrzymaliśmy ważną własność rozkładu Poissona*, która mówi, że stosunek dyspersji σ do wartości oczekiwanej $E(Y)$ maleje pierwiastkowo wraz ze wzrostem poziomu zmiennej Y opisanej tym rozkładem:

$$\frac{\sigma}{E(Y)} = \frac{1}{\sqrt{\mu}} \ . \qquad (2\text{-}1\text{-}1.5)$$

Fakt ten oznacza z założenia *inne zachowanie się odchylenia standardowego* w modelu regresji Poissona niż w klasycznym modelu regresji normalnej (w którym zakładamy jednorodność wariancji zmiennej objaśnianej w różnych wariantach zmiennej objaśniającej).

Ćwiczenie: Pokazać (2-1-1.2) dla trzeciego momentu.

Rozdział 2-1-2. Model podstawowy.

Zakładając brak zależności zmiennej Y od czynników $X_1, X_2, ... X_k$, rozważa się tzw. *model podstawowy*. Dla rozkładu (2-1-1.1) i próby $\widetilde{Y} \equiv (Y_n)_{n=1}^N$, funkcja wiarygodności przy parametrze $\Theta = \mu \equiv (\mu_n)_{n=1}^N$, ma postać:

$$P\left(\widetilde{Y} \mid \mu\right) = \prod_{n=1}^{N} \frac{\mu_n^{Y_n} e^{-\mu_n}}{Y_n!} = \frac{\left(\prod_{n=1}^{N} \mu_n^{Y_n}\right) \exp\left(-\sum_{n=1}^{N} \mu_n\right)}{\prod_{n=1}^{N} Y_n!}, \qquad (2\text{-}1\text{-}2.5)$$

jest więc wyrażona jako funkcja wektorowego parametru $\mu \equiv (\mu_n)_{n=1}^N$, gdzie każdy z parametrów $\mu_n = E(Y_n)$ jest parametrem skalarnym. N jest równocześnie liczebnością zbioru danych, która może być liczbą podgrup, komórek lub kategorii, oraz liczbą parametrów modelu podstawowego występującą w wiarygodności (2-1-2.5).

Rozważmy układ równań MNW:

$$\frac{\partial}{\partial \mu_n}\left[\ln P\left(\widetilde{Y} \mid \mu\right)\right] = 0, \quad n = 1,2,..., N. \qquad (2\text{-}1\text{-}2.6)$$

Dla funkcji wiarygodności (2-1-2.5) otrzymujemy:

$$\ln P\left(\widetilde{Y} \mid \mu\right) = \sum_{n=1}^{N} Y_n \ln \mu_n - \sum_{n=1}^{N} \mu_n - \sum_{n=1}^{N} \ln Y_n!. \qquad (2\text{-}1\text{-}2.7)$$

Zatem rozwiązanie układu (2-1-2.6) daje:

$$\mu_n = \hat{\mu}_n = Y_n, \quad n = 1,2,..., N, \qquad (2\text{-}1\text{-}2.8)$$

jako estymatory modelu postawowego. Zatem funkcja wiarygodności (2-1-2.5) modelu podstawowego przyjmuje w punkcie μ zadanym przez estymatory (2-1-2.8) wartość maksymalną:

$$P\left(\widetilde{Y} \mid \hat{\mu}\right) = \frac{\left(\prod_{n=1}^{N} Y_n^{Y_n}\right) \exp\left(-\sum_{n=1}^{N} Y_n\right)}{\prod_{n=1}^{N} Y_n!}, \qquad (2\text{-}1\text{-}2.9)$$

gdzie zastosowano oznaczenie $\hat{\mu} = \left(\hat{\mu}_1, \hat{\mu}_2, ... \hat{\mu}_N\right)$.

Rozdział 2-1-3. Analiza regresji Poissona.

Niech zmienna zależna Y reprezentuje liczbę zliczeń badanego zjawiska (np. przypadków awarii określonego zakupionego sprzętu), otrzymaną dla każdej z N podgrup (np. klienckich). Każda z tych podgrup wyznaczona jest przez komplet wartości zmiennych objaśniających $X \equiv \left(X_1, X_2, ..., X_k\right) = x \equiv \left(x_1, x_2, ..., x_k\right)$

(np. wiek, poziom wykształcenia, cel nabycia sprzętu). Zmienna Y_n określa liczbę zliczeń zjawiska w n-tej podgrupie, $n = 1,2,...,N$. W konkretnej próbce $(Y_n)_{n=1}^N = (y_n)_{n=1}^N$.

Określenie modelu regresji Poissona: Rozważmy następujący model regresji Poissona:

$$\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta), \quad n = 1,2,...,N, \tag{2-1-3.10}$$

opisujący zmianę wartości oczekiwanej liczby zdarzeń Y_n (dla rozkładu Poissona) wraz ze zmianą *wariantu* $x_n = (x_{1n}, x_{2n},..., x_{kn})$.

Funkcja regresji po prawej stronie (2-1-3.10) ma dwa czynniki. Czynnik funkcyjny funkcji regresji, $r(x_n, \beta)$, opisuje *tempo zdarzeń* określanych mianem porażek (np. awarii) w n-tej podgrupie (tzn. jest *częstością* tego zjawiska), skąd $r(x_n, \beta) > 0$, gdzie $\beta \equiv (\beta_0, \beta_1,..., \beta_k)$ jest zbiorem nieznanych parametrów tego modelu regresji. Natomiast czynnik ℓ_n jest współczynnikiem określającym *dla każdej* n*-tej podgrupy* (np. klientów) *skumulowany czas prowadzenia badań kontrolnych dla wszystkich jednostek tej podgrupy*.

Ponieważ funkcja regresji[15] $r(x_n, \beta)$ przedstawia typową liczbę porażek na jednostkę czasu, zatem nazywamy ją *ryzykiem*.

Uwaga o postaci funkcji regresji: Funkcję $r(x_n, \beta)$ można zamodelować na różne sposoby [38]. Wprowadźmy oznaczenie:

$$\lambda_n^* \equiv \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn}. \tag{2-1-3.11}$$

Funkcja regresji $r(x_n, \beta)$ ma różną postać w zależności od typu danych. Może mieć ona postać charakterystyczną dla regresji liniowej (wielokrotnej), $r(x_n, \beta) = \lambda_n^*$, którą stosujemy szczególnie wtedy gdy zmienna Y ma *rozkład normalny*. Postać $r(x_n \beta) = 1/\lambda_n^*$ jest stosowana w analizie z danymi pochodzącymi z *rozkładu eksponentialnego*, natomiast $r(x_n, \beta) = 1/(1 + \exp(-\lambda_n^*))$ w modelowaniu regresji logistycznej dla opisu zmiennej *dychotomicznej* [1], [38].

Postać funkcji regresji użyteczna w regresji Poissona jest następująca:

$$r(x_n, \beta) = \exp(\lambda_n^*), \quad \lambda_n^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn}. \tag{2-1-3.12}$$

Ogólniej mówiąc analiza regresji odnosi się do modelowania wartości oczekiwanej zmiennej zależnej (objaśnianej) jako funkcji pewnych czynników. Postać funkcji wiarygodności stosowanej do estymacji

[15] Czynnik $r(x_n, \beta)$ nazywany dalej funkcją regresji, chociaż właściwie nazwa ta odnosi się do całej $E(Y_n)$.

współczynników regresji β odpowiada założeniom dotyczącym rozkładu zmiennej zależnej. Tzn. zastosowanie konkretnej funkcji regresji $r(x_n,\beta)$, np. jak w (2-1-3.12), wymaga określenia postaci funkcji częstości $r(x_n,\beta)$, zgodnie z jej postacią dobraną do charakteru losowej zmiennej Y przy której generowane są dane w badanym zjawisku. Na ogół przy konstrukcji $r(x_n,\beta)$ pomocna jest uprzednia wiedza dotyczącą relacji między rozważanymi zmiennymi.

Funkcja wiarygodności dla analizy regresji Poissona: Ponieważ Y_n ma rozkład Poissona (2-1-1.1) ze średnią μ_n, $p(Y_n \mid \mu_n) = \frac{\mu_n^{Y_n}}{Y_n!} e^{-\mu_n}$, $n = 1,2,...,N$, zatem dane $Y_n = 0,1,...,\infty$ dla określonego $n = 1,2,...,N$ są generowane z rozkładów warunkowych:

$$p(Y_n \mid \beta) = \frac{\left[\ell_n r(x_n,\beta)\right]^{Y_n}}{Y_n!} e^{-\ell_n r(x_n,\beta)}, \qquad (2\text{-}1\text{-}3.13)$$

wokół funkcji regresji, (2-1-3.10), $\mu_n = \ell_n r(x_n,\beta)$, dla $n = 1,2,...,N$. (*Bezwarunkowa*) *funkcja wiarygodności*[16] dla analizy regresji Poissona ma więc postać:

$$\begin{aligned} P(\tilde{Y} \mid \beta) &= \prod_{n=1}^{N} p(Y_n \mid \beta) = \prod_{n=1}^{N} \frac{(\ell_n r(x_n,\beta))^{Y_n} e^{-\ell_n r(x_n,\beta)}}{Y_n!} \\ &= \frac{\prod_{n=1}^{N} (\ell_n r(x_n,\beta))^{Y_n} \exp\left[-\sum_{n=1}^{N} \ell_n r(x_n,\beta)\right]}{\prod_{n=1}^{N} Y_n!}. \end{aligned} \qquad (2\text{-}1\text{-}3.14)$$

Aby w praktyce posłużyć się funkcją regresji $r(x_n,\beta)$ będącą określoną funkcją zmiennej $\lambda_n^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn}$, parametry $\beta_0, \beta_1, ..., \beta_k$ muszą być oszacowane. Estymatory MNW, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, tych parametrów otrzymuje się rozwiązując $k+1$ równań wiarygodności:

$$\frac{\partial}{\partial \beta_j} \ln P(\tilde{Y} \mid \beta) = 0, \quad j = 0,1,2,...,k. \qquad (2\text{-}1\text{-}3.15)$$

W przypadku regresji Poissona $P(\tilde{Y} \mid \beta)$ jest określona zgodnie z (2-1-3.14).

---

[16] Termin *bezwarunkowa funkcja wiarygodności* odnosi się do bezwarunkowych (tzn. łącznych) rozkładów prawdopodobieństwa otrzymania określonego zbioru danych. *Bezwarunkowa funkcja wiarygodności* jest łącznym rozkładem prawdopodobieństwa pojawienia się określonych dyskretnych danych (określonej próbki), a w przypadku zmiennej typu ciągłego jest ona łącznym rozkładem gęstości prawdopodobieństwa dla danych w próbce generowanych z rozkładów warunkowych tej zmiennej. Na ogół będziemy pomijali słowo „bezwarunkowa".

Algorytmy IRLS: Zauważmy, że dla rozkładu Poissona zachodzi zgodnie z (2-1-1.2) oraz (2-1-3.10), $\sigma^2(Y_n) = E(Y_n) = \ell_n r(x_n, \beta)$, *co oznacza, że wariancja* $\sigma^2(Y_n)$ *zmiennej objaśnianej nie jest stała lecz zmienia się jako funkcja* ℓ_n *oraz* x_n*, wchodząc w analizę z różnymi wagami wraz ze zmianą* n. Na fakt ten zwróciliśmy już uwagę przy okazji związku (2-1-1.5). Ponieważ układ równań wiarygodności (2-1-3.15) jest na ogół rozwiązywany iteracyjnymi metodami numerycznymi [1], a wariancja $\sigma^2(Y_n)$ jest również funkcją β, zatem na każdym kroku procesu iteracyjnego *wagi* te zmieniają się jako funkcja zmieniających się składowych estymatora $\hat{\beta}$. Algorytmy takiej analizy określa się ogólnym mianem *algorytmów najmniejszych kwadratów*[17] *iteracyjnie ważonych* (IRLS[18]) [38], [1]. Nazwa ta pozostała jedynie z powodu „pierwszeństwa" MNK, ale ogólnie nie odnosi się do MNK, która ma probabilistyczne znaczenie tylko gdy zmienne Y_j mają rozkład normalny.

Uwaga o programach: Różne programy do analiz statystycznych, w tym SAS wykorzystujący procedurę PROC GENMOD, mogą być użyte do znajdowania estymatorów $\hat{\beta}$ MNW dla funkcji wiarygodności (2-1-3.14). Również *obserwowana macierz kowariancji estymatorów*[19] oraz miary dobroci dopasowania modelu, takie jak omówiona dalej dewiancja, mogą być otrzymane przy użyciu powyżej wspomnianych programów.

Rozdział 2-1-4. Test statystyczny dla doboru modelu w regresji Poissona.

Uwaga o większej wiarygodności modelu podstawowego: Maksymalna wartość funkcji wiarygodności $P(y \mid \mu)$ wyznaczona w oparciu o (2-1-2.9) będzie, dla każdego zbioru danych i dla liczby parametrów $k+1 < N$, większa niż otrzymana przez maksymalizację funkcji wiarygodności (2-1-3.14). Jest tak, ponieważ w wyrażeniu (2-1-2.9) na funkcję wiarygodności modelu podstawowego *nie narzuca się żadnych ograniczeń na postać* μ_n, natomiast (2-1-3.14) wymaga aby $\mu_n = \ell_n r(x_n, \beta)$.

Pomyśl o tym tak: Model podstawowy dopasowuje się do danych, w każdym punkcie z osobna, leżąc zgodnie z (2-1-2.8) maksymalnie blisko tych danych, natomiast MNW dla modelu regresji $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$, $n = 1,2,...,N$, (2-1-3.10), wyznacza krzywą regresji przechodzącą pomiędzy punkami pomiarowymi.

---

[17] Należy jednak pamiętać, że zwrotu "najmniejszych kwadratów" nie należy tu brać dosłownie, gdyż metoda najmniejszych kwadratów ma sens jedynie wtedy, gdy rozkład zmiennej Y jest normalny (por. Rozdział W1.3).

[18] *iteratively reweighted least squares*

[19] Obserwowana macierz (wariancji-) kowariancji $\hat{V}(\hat{\beta})$ estymatorów $\hat{\beta}$ MNW jest zdefiniowana jako odwrotność macierzy obserwowanej informacji Fishera [5,1] (por. (2-2-8.45)):

$$\hat{V}(\hat{\beta}) := \mathbf{iF}^{-1}(\hat{\beta}) . \qquad (2\text{-}1\text{-}4.16)$$

Hipoteza zerowa o nie występowaniu braku dopasowania w modelu niższym: Zgodnie z powyższym zdaniem, analizę doboru modelu regresji można rozpocząć od postawienia hipotezy zerowej wobec alternatywnej. W hipotezie zerowej wyróżnimy proponowany model regresji. Wybór modelu badanego oznacza wybór funkcji wiarygodności (2-1-3.14) z nim związanej.

Stawiamy więc hipotezę zerową:

$$\mathrm{H}_0 : \mu_n = \ell_n r(x_n, \beta), \quad n = 1,2,...,N, \tag{2-1-4.17}$$

która odpowiada wyborowi modelu z funkcją wiarygodności (2-1-3.14), wobec hipotezy alternatywnej:

$$\mathrm{H}_A : \ \mu_n \ \textit{nie ma ograniczonej postaci}, n = 1,2,...,N\ , \tag{2-1-4.18}$$

która odpowiada wyborowi modelu podstawowego zawierającego tyle parametrów μ_n ile jest punktów pomiarowych, tzn. *N*, z funkcją wiarygodności (2-1-2.9).

Niech więc $P(\tilde{Y} \mid \hat{\beta})$ jest maksymalną wartością funkcji wiarygodności określoną jak w (2-1-3.14). Oznacza to, że w miejsce parametrów $\beta = (\beta_0, \beta_1,..., \beta_k)$ podstawiono ich estymatory $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1,..., \hat{\beta}_k)$ wyznaczone przez MNW, jako te które maksymalizują funkcję wiarygodności (2-1-3.14). Podobnie rozumiemy funkcję wiarygodności $P(\tilde{Y} \mid \hat{\mu})$ modelu podstawowego.

Ponieważ celem każdej analizy jest otrzymanie możliwie najprostszego opisu danych, model $\mu_n = \ell_n r(x_n, \beta)$ zawierający $k+1$ parametrów β, będzie uznany za dobry, jeśli maksymalna wartość funkcji wiarygodności wyznaczona dla niego, będzie prawie tak duża, jak funkcji wiarygodności dla nie niosącego żadnej informacji modelu podstawowego z liczbą parametrów μ_n równą licznie punktów pomiarowych N. Sformułowanie ”prawie tak duża” oznacza, że wartość funkcji wiarygodności $P(y \mid \hat{\beta})$ nie może być istotnie statystycznie mniejsza od $P(y \mid \hat{\mu})$. Zasadniczo powinno to oznaczać, że musimy podać miary pozwalające na określenie statystycznej istotności przy posługiwaniu się intuicyjnym parametrem obcięcia c (Rozdział 1-2-1). Okazuje się, że dla dużej próby, miary typu (1-1.42), podane poniżej, uzyskują cechy pozwalające na budownie wiarygodnościowych obszarów krytycznych nabywających charakteru standardowego (częstotliwościowego).

Określenie dewiancji: Wprowadźmy *statystykę typu ilorazu wiarygodności*:

$$D(\hat{\beta}) = -2 \ln \left[\frac{P(\tilde{Y} \mid \hat{\beta})}{P(\tilde{Y} \mid \hat{\mu})} \right] \tag{1-1.42'}$$

nazywaną *dewiancją* (deviance) dla modelu regresji, w tym przypadku dla modelu Poissona z określoną postacią $\mu_n = \ell_n r(x_n, \beta)$. Służy ona do badania dobroci dopasowania modelu z zadaną postacią

$\mu_n = \ell_n r(x_n, \beta)$ w stosunku do modelu podstawowego, bez narzuconej postaci na μ_n, tzn. do stwierdzenia, czy $P(y \mid \hat{\beta})$ jest istotnie *mniejsza* od $P(y \mid \hat{\mu})$, co sugerowałoby istotny statystycznie brak dopasowania badanego modelu $\mu_n = \ell_n r(x_n, \beta)$, do danych empirycznych. Jak pokażemy poniżej dewiancja może być rozumiana jako *miara zmienności reszt* (tzn. odchylenia wartości obserwowanych w próbie od wartości szacowanych przez model) *wokół linii regresji,* na której leżą wartości przewidywane $\hat{y}_j$ przez model [1].

Przy prawdziwości hipotezy $H_0 : \mu_n = \ell_n r(x_n, \beta)$, rozkład dewiancji $D(\hat{\beta})$ dla regresji Poissona, można asymptotycznie przybliżyć rozkładem chi-kwadrat (por. dyskusja w [1], [38]) z $N-k-1$ stopniami swobody.

Wyznaczenie liczby stopni swobody dewiancji: Podana liczba stopni swobody dewiancji $D(\hat{\beta})$ wynika z następującego rozumowania. Zapiszmy (1-1.42) w postaci:

$$D(\hat{\beta}) + 2\ln P(\tilde{Y} \mid \hat{\beta}) = 2\ln P(\tilde{Y} \mid \hat{\mu}), \qquad (2\text{-}1\text{-}4.19)$$

co po skorzystaniu z (2-1-3.14) dla $\beta = \hat{\beta}$ ma postać:

$$D(\hat{\beta}) - 2\sum_{n=1}^{N} \ell_n r(x_n, \hat{\beta}) = 2\ln P(\tilde{Y} \mid \hat{\mu}) + 2\ln\left(\prod_{n=1}^{N} Y_n!\right) - 2\ln\left(\prod_{n=1}^{N} \left(\ell_n r(x_n, \hat{\beta})\right)^{Y_n}\right). \qquad (2\text{-}1\text{-}4.20)$$

Można zauważyć, że prawa strona tego równania ma N -stopni swobody. Istotnie, ze względu na (2-1-2.8)[20], $\hat{\mu} \equiv (\hat{\mu}_n) = (Y_n)$, $n = 1,2,...,N$, liczba niezależnych zmiennych po prawej strony powyższego równania, których wartości trzeba określić z eksperymentu, wynosi N. Natomiast drugi składnik po lewej stronie ma liczbę stopni swobody równą $k+1$, co jest liczbą estymatorów parametrów strukturalnych $\hat{\beta}$ modelu regresji, których wartości trzeba określić z eksperymentu. Ponieważ liczba stopni swobody po prawej i lewej stronie równania musi być taka sama, zatem liczba stopni swobody dewiancji $D(\hat{\beta})$ wynosi $N-k-1$.

Decyzja testu statystycznego: Z powyższego wynika, że dla *bardzo dużej* próbki dewiancja $D(\hat{\beta})$ posiada, przy prawdziwości hipotezy $H_0 : \mu_n = \ell_n r(x_n, \beta)$ (2-1-4.17) w przybliżeniu rozkład chi-kwadrat z $N-k-1$ stopniami swobody. Zatem, przybliżony statystyczny test dobroci dopasowania (tzn. niewystępowania braku dopasowania) modelu $\mu_n = \ell_n r(x_n, \beta)$ do danych w stosunku do modelu podstawowego, może zostać wykonany przez sprawdzenie czy w zaobserwowanej (*obs*) próbce $\tilde{Y} = y$, wartości estymatorów MNW

[20] W przyjętym przedstawieniu danych jak dla diagramu punktowego, N jest ogólnie liczbą punktów pomiarowych (równą liczbie wariantów czy komórek). Tylko dla modelu podstawowego jest N również liczbą parametrów.

$\hat{\beta} \equiv \hat{\beta}_{obs}$ modelu regresji (2-1-3.10) oraz $\hat{\mu} \equiv \hat{\mu}_{obs} = y$ modelu podstawowego (2-1-2.8), dają wartość dewiancji $D(\hat{\beta}) = D(\hat{\beta}_{obs})$:

$$D(\hat{\beta}_{obs}) = -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta}_{obs})}{P(\tilde{Y} \mid \hat{\mu}_{obs})}\right], \qquad (2\text{-}1\text{-}4.21)$$

która jest nie mniejsza niż wartość krytyczna w prawym ogonie rozkładu chi-kwadrat z $N-k-1$ stopniami swobody [1]. Przyjęcie przez $D(\hat{\beta}_{obs})$ wartości równej lub większej od krytycznej skutkuje odrzuceniem hipotezy zerowej. Alternatywnie, mając wartości $D(\hat{\beta}_{obs})$, można policzyć empiryczny poziom istotności $p = \Pr(\chi^2_{N-k-1} \geq D(\hat{\beta}_{obs}))$ i porównać jego wartość z przyjętą (w dziedzinie badań) wartością poziomu istotności α [9]. Gdy $p \leq \alpha$ wtedy odrzucamy hipotezę zerową H_0, która mówi o nie występowaniu braku dopasowania w badanym modelu regresji w porównaniu z modelem podstawowym i decydujemy się na statystycznie uzasadnioną rozbudowę modelu, o dalsze parametry strukturalne. Gdy $p > \alpha$ wtedy nie mamy podstaw do odrzucenia hipotezy zerowej H_0.

Uwaga dotycząca zapisu indeksu *obs*: W dalszej części będziemy pomijać indeks '*obs'* w indeksie wartości estymatora w próbce, za wyjątkiem sytuacji, gdy rozróżnienie pomiędzy estymatorem jako statystyką, a jego realizacją w próbce, nie wynika jasno z kontekstu.

Rozdział 2-1-4-1. Testy ilorazu wiarygodności.

Dewiancje dla hierarchicznych klas modeli mogą służyć do budowy testów stosunku wiarygodności. Zwróćmy szczególnie uwagę na funkcję wiarygodności (2-1-3.14) zawierającą zbiór parametrów $\beta = (\beta_0, \beta_1, ..., \beta_k)$ z *dewiancją* $D(\hat{\beta})$ daną wyrażeniem (1-1.42). Przypuśćmy, że chcemy zweryfikować hipotezę o tym, że $k-r$ (gdzie $0 < r < k$) ostatnich parametrów będących składowymi wektora β jest równych *zeru*.

Hipoteza zerowa, o nieistotności rozszerzenia modelu niższego do wyższego, ma wtedy postać:

$$\mathrm{H}_0 : \beta_{r+1} = \beta_{r+2} = ... = \beta_k = 0, \qquad (2\text{-}1\text{-}4\text{-}1.22)$$

Hipoteza alternatywna H_A mówi, że przynajmniej jeden z parametrów strukturalnych $\beta_{r+1}, \beta_{r+2}, ..., \beta_k$ jest różny od *zera*.

Funkcja wiarygodności przy prawdziwości hipotezy zerowej H_0, (2-1-4.20), ma postać taką jak w (2-1-3.14), tyle, że zastąpiono w niej parametr β parametrem $\beta_{(r)}$:

$$\beta_{(r)} \equiv (\beta_0, \beta_1, ..., \beta_r; 0,0,...,0) \quad \textit{gdzie liczba zer wynosi } k-r\,. \qquad (2\text{-}1\text{-}4\text{-}1.23)$$

Oznaczmy funkcje wiarygodności tego modelu jako $P(\tilde{Y} \mid \beta_{(r)})$, a $\hat{\beta}_{(r)}$ niech będzie estymatorem MNW wektorowego parametru $\beta_{(r)}$, wyznaczonym przez rozwiązanie odpowiadającego mu układu równań wiarygodności (oczywiście dla niezerowych parametrów $\beta_0, \beta_1, ..., \beta_r$). Estymator $\hat{\beta}_{(r)} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_r;$ $0,0,...,0)$ maksymalizuje funkcję wiarygodności $P(\tilde{Y} \mid \beta_{(r)})$.

Test ilorazu wiarygodności dla weryfikacji hipotezy H_0 przeprowadzamy posługując się *statystyką ilorazu wiarygodności*:

$$LR_{r/k} = -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta}_{(r)})}{P(\tilde{Y} \mid \hat{\beta})}\right], \qquad (2\text{-}1\text{-}4\text{-}1.24)$$

która przy prawdziwości hipotezy zerowej ma asymptotycznie rozkład chi-kwadrat z $k-r$ stopniami swobody [1], co widać, gdy zapiszemy (2-1-4-1.24) jako różnicę dewiancji:

$$-2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta}_{(r)})}{P(\tilde{Y} \mid \hat{\beta})}\right] = -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta}_{(r)})}{P(\tilde{Y} \mid \hat{\mu})}\right] + 2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta})}{P(\tilde{Y} \mid \hat{\mu})}\right] = D(\hat{\beta}_{(r)}) - D(\hat{\beta}), \qquad (2\text{-}1\text{-}4\text{-}1.25)$$

oraz skorzystamy z podobnej analizy jak dla (2-1-4.20).

Zatem, przy prawdziwości hipotezy zerowej (2-1-4-1.22), którą można zapisać jako $\mathrm{H}_0 : \beta_{r+1} = \beta_{r+2} = ... = \beta_k = 0$, różnica $D(\hat{\beta}_{(r)}) - D(\hat{\beta})$ ma dla dużej próby w przybliżeniu rozkład chi-kwadrat z $k-r$ stopniami swobody. Natomiast **decyzja testu statystycznego** ma w przypadku posługiwania się statystyką testową (2-1-4-1.24) analogiczny przebieg jak dla omówionej poprzednio przypadku dewiancji.

Wniosek: Jeśli używamy regresji Poissona do analizowania danych empirycznych, modele tworzące hierarchiczne klasy mogą być porównywane miedzy sobą poprzez wyznaczenie statystyki ilorazu wiarygodności (2-1-4-1.24), lub co na jedno wychodzi, poprzez wyznaczenie różnicy (2-1-4-1.25) między parami dewiancji dla tych modeli. Należy przy tym pamiętać o wniosku jaki już znamy z analizy dewiancji, że *im model gorzej dopasowuje się do danych empirycznych tym jego dewiancja jest większa.*

Rozdział 2-1-5. Podobieństwo dewiancji do SKR analizy częstotliwościowej.

Warunkowe wartości oczekiwane $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$, $n = 1,2,..., N$, (2-1-3.10), są w analizie regresji przyjmowane jako teoretyczne przewidywania modelu regresji dla wartości zmiennej objaśnianej Y_n, zwanej odpowiedzią (układu).

W próbie oszacowania $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$ oznaczamy jako $\hat{Y}_n$. W n-tej komórce jest ono następujące:

$$\hat{Y}_n = \ell_n r(x_n, \hat{\beta}), \quad n = 1,2,..., N, \tag{2-1-5.26}$$

zgodnie z wyestymowaną postacią modelu regresji. Wykorzystując (2-1-5.26) w (2-1-3.14) możemy zapisać dewiancję modelu (1-1.42) następująco:

$$\begin{aligned} D(\hat{\beta}) &= -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta})}{P(\tilde{Y} \mid \hat{\mu})}\right] = -2\ln\left[\frac{\prod_{n=1}^{N} \hat{Y}_n^{Y_n} \exp\left(-\sum_{n=1}^{N} \hat{Y}_n\right)}{\prod_{n=1}^{N} Y_n^{Y_n} \exp\left(-\sum_{n=1}^{N} Y_n\right)}\right] \\ &= -2\left[\sum_{n=1}^{N} Y_n \ln \hat{Y}_n - \sum_{n=1}^{N} \hat{Y}_n - \sum_{n=1}^{N} Y_n \ln Y_n + \sum_{n=1}^{N} Y_n\right] \end{aligned} \tag{2-1-5.27}$$

tzn:

$$D(\hat{\beta}) = 2\sum_{n=1}^{N}\left[Y_n \ln\left(\frac{Y_n}{\hat{Y}_n}\right) - \left(Y_n - \hat{Y}_n\right)\right]. \tag{2-1-5.28}$$

Podobieństwo D do SKR: Powyższa postać dewiancji oznacza, że $D(\hat{\beta})$ zachowuje się w poniższym sensie jak suma kwadratów reszt $SKR = \sum_{n=1}^{N}(Y_n - \hat{Y}_n)^2$ w standardowej wielorakiej regresji liniowej. Otóż, gdy dopasowywany model dokładnie przewiduje obserwowane wartości, tzn. $\hat{Y}_n = Y_n, n = 1,2,.., N$ wtedy, jak SKR w analizie standardowej [1], tak $D(\hat{\beta})$ w analizie wiarygodnościowej jest równe zeru [1]. Z drugiej strony wartość $D(\hat{\beta})$ jest tym większa im większa jest różnica między wartościami obserwowanymi Y_n i wartościami przewidywanymi $\hat{Y}_n$ przez oszacowany model.

Asymptotyczna postać D: W analizowanym modelu Y_n, $n = 1,2,..., N$ są niezależnymi zmiennymi Poissona (np. zmiennymi częstości), natomiast wartości $\hat{Y}_n$ są ich przewidywaniami. Nietrudno przekonać się, że gdy wartości przewidywane mają rozsądną wartość[21], np. $\hat{Y}_n > 3$ oraz $(Y_n - \hat{Y}_n) << Y_n$, $n = 1,2,.., N$ tak,

[21] Zauważmy, że statystyka (2-1-5.28) może mieć myląco dużą wartość gdy wielkości $\hat{Y}_n$ są bardzo małe.

że $(Y_n - \hat{Y}_n)/Y_n << 1$, wtedy wyrażenie w nawiasie kwadratowym w (2-1-5.28) można przybliżyć przez $(Y_n - \hat{Y}_n)^2/(2Y_n)$, a statystykę (2-1-5.28) można przybliżyć statystyką o postaci (pokazać):

$$\chi^2 = \sum_{n=1}^{N} \frac{\left(Y_n - \hat{Y}_n\right)^2}{\hat{Y}_n}, \qquad (2\text{-}1\text{-}5.29)$$

która (dla dużej próby) ma rozkład chi-kwadrat z $N-k-1$ stopniami swobody [1].

Rozdział 2-2. Przykład analizy doboru modelu w regresji Poissona.

Poniżej przedstawimy na przykładzie działanie MNW w estymacji parametrów modelu regresji Poissona [1], [38] oraz pokażemy jak połączyć wnioskowanie statystyczne z tworzeniem odpowiedniej procedury pakietu SAS.

Przyjmijmy więc, że zmienna objaśniana jest zmienną losową *Y* przyjmującą wartości *y* zgodnie z rozkładem Poissona (2-1-1.1), $p(y \mid \mu) = \mu^y e^{-\mu} / y!$, $y = 0,1,...,\infty$. Jak wspomnieliśmy we Wprowadzeniu, rozkład Poissona (2-1-1.1) jest często używany do modelowania pojawiania się rzadkich zdarzeń, takich jak np. nowych przypadków awarii w pewnej populacji w pewnym okresie czasu albo zajścia określonej liczby wypadków samochodowych w pewnym określonym miejscu w ciągu roku.

Analizę regresji Poissona stosuje się w modelowaniu zachowania się zmiennej objaśnianej przyjmującej, z natury tej zmiennej, dyskretne realizacje widoczne w danych i powstałe np. ze zliczeń modyfikowanych zmiennymi objaśniającymi (nazywanych czynnikiami). Po pierwsze, wyjaśnimy jak konstruować postać modelu regresji Poissona dla tzw. ryzyka względnego i zastosujemy przedstawioną we Wprowadzeniu estymację MNW parametrów modelu. Zastosowanie wnioskowania związanego z weryfikacją hipotez o braku dopasowania w modelu niższym przedstawimy w drugiej kolejności.

Rozdział 2-2-1. Przykład danych dla regresji Poissona.

Aby zilustrować działanie MNW w analizie regresji Poissona rozważmy dane przedstawiające awarię urządzenia określonego typu (pomijając awarię niszczącą całkowicie urządzenie). Tego typu analiza została zastosowana ze sporym sukcesem w badaniach medycznych [1].

Poniższa Tabela 2-2-5. 1 przedstawia dwie *przykładowe* próbki pobrane z populacji silników serwisowanych samochodów pewnej firmy (nazwijmy ją „Auto”) i jej modelu typu „Model”, które uległy niedestrukcyjnej awarii, tzn. takiej, po której silnik można jeszcze naprawić nie zmniejszając tym samym wielkości populacji, z których dokonujemy losowania.

Próbki powstały na skutek losowania pewnej liczby aglomeracji miejskich i takiej samej liczby aglomeracji wiejskich na całym obszarze ziemi, na którym firma „Auto” ma swój serwis. Próbki w obszarach Miejskim i Wiejskim zostały uporządkowane wg wariantów wieku (miesięcy używania).

W przykładzie, zmienna zależna Y jest zmienną zliczeń przypadków awarii silnika. Generalne populacje dwóch obszarów używania samochodów zakwalikowano do ośmiu wariantów wiekowych. Stąd zmienną Y indeksujemy podwójnym indeksem grupowym, tzn. Y_{ij} oznacza liczbę zliczeń dla i–tego wariantu wiekowego i j–tego obszaru, gdzie i zmienia się od 1 do 8, natomiast $j = 0$ dla obszaru „Miasta” oraz $j = 1$ dla obszaru „Wsie”. Oznaczmy przez ℓ_{ij} rozmiar (pod)populacji dla i–tego wariantu wieku samochodu i j–tego obszaru.

Celem analizy jest ustalenie, czy ryzyko awarii silnika samochodu, przy dopasowaniu ze względu na wiek, jest wyższe w pierwszym badanym obszarze czy w drugim.

Rozdział 2-2-2. Rola kowarianta.

„Wiek” jest wspólną *zmienną poboczną* dla obu rozważanych populacji, tzw. *kowariantem* zmiennej „obszar”. Należy wprowadzić go do analizy bądź w *członach interakcji* ze zmienną „obszar” lub jako *zaburzenie* wpływu głównego, którym jest zmienna „obszar” [1]. Wprowadzenie „wieku” do analizy oznacza, że zmienna ta jest pod kontrolą oraz, że oszacowany parametr, którym w naszym przykładzie okaże się być ryzyko względne, jest estymowany w sytuacji dopasowania zmiennych i estymatorów parametrów modelu ze względu na zmienną „wiek” samochodu. Pominięcie kowarianta oznaczałoby wyznaczanie *surowych* estymatorów *parametrów*.

Rozdział 2-2-3. Pojęcie ryzyka.

Termin ryzyko w rozważanym przykładzie odnosi się do (rozwijającego się z wiekiem) prawdopodobieństwa zajścia wady silnika. **Przez r_{ij} będziemy oznaczać rzeczywiste populacyjne ryzyko w grupie** (i, j).

Rozdział 2-2-3-1. Analogia ryzyka awarii i prawdopodobieństwa zajścia porażki na jednostkę czasu. Estymowane tempo defektu.

Rozważmy rozkład dwumianowy z parametrem prawdopodobieństwa p oraz parametrem liczby losowań m. Związek określający oczekiwaną liczbę sukcesów w m losowaniach Bernoulliego $\mu = m\,p$, można zapisać następująco:

$$\mu = (m \cdot \Delta t)\frac{p}{\Delta t} \quad , \qquad (2\text{-}2\text{-}3\text{-}1.30)$$

skąd widać, że jeśli Δt jest czasem prowadzonego badania, wtedy $l \equiv m \cdot \Delta t$ jest zakumulowaną w tym czasie liczbą „samochodo–lat”, a

$$r \equiv \frac{p}{\Delta t} \tag{2-2-3-1.31}$$

jest tzw. **intensywnością,** *czyli prawdopodobieństwem zajścia zdarzenia na jednostkę czasu, nazywanym ryzykiem*.

Pojęcie ryzyka: Ze względu na to, że μ jest liczebnością, związek $\mu = (m \cdot \Delta t)\frac{p}{\Delta t}$ ma postać analogiczną do stosowanej w analizie regresji Poissona postaci funkcji regresji (2-1-3.10) $\mu_{ij} = \ell_{ij} r_{ij}$ [1], dla wartości oczekiwanej μ_{ij} liczby zliczeń zdarzeń awarii w grupie (*i*, *j*), gdzie ℓ_{ij}, który jest odpowiednikiem $(m \cdot \Delta t)$, jest parametrem określającym liczbę wszystkich wyników zakumulowanych w czasie badania.

R**yzyko** w grupie (*i*, *j*) jest zdefiniowane jako:

$$r_{ij} = \frac{\mu_{ij}}{\ell_{ij}} \quad . \tag{2-2-3-1.32}$$

Jest ono *analogiem* **intensywności** (awarii) $r = \frac{\mu}{(m \cdot \Delta t)}$.

Oszacowane ryzyko, nazywane **tempem defektu** rozumianego jako porażka, jest definiowane jako:

$$\hat{r}_{ij} = \frac{Y_{ij}}{\ell_{ij}} \quad , \tag{2-2-3-1.34}$$

gdzie Y_{ij} jest ilością zaobserwowanych zliczeń defektów silnika dla podgrupy (*i*, *j*), a ℓ_{ij} oznacza zakumulowaną (tzn. sumaryczną) długość czasu wolnego od defektu dla wszystkich samochodów w tej podgrupie. Zatem $\hat{r}_{ij}$ mierzy liczbę defektów w stosunku do całkowitej zakumulowanej liczby wszystkich samochodów poddanych serwisowaniu w danej podgrupie na ustaloną jednostkę czasu (np. roku). Zwróćmy uwagę, że występująca w liczniku (2-2-3-1.34) zmienna Y_{ij} nie jest w ogólności estymatorem MNW parametru μ_{ij} dla modelu regresji Poissona, chociaż jest tak dla modelu podstawowego. Dla modelu regresji zamiast empirycznego związku (2-2-3-1.34) pojawia się zgodnie z (2-1-5.26) estymator $r(x_{ij}, \hat{\beta}) = \frac{\hat{Y}_{ij}}{\ell_{ij}}$.

Rozdział 2-2-3-2. Ryzyko względne.

Stosunek:

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} \qquad (2\text{-}2\text{-}3\text{-}2.35)$$

jest parametrem nazywamym *ryzykiem względnym* lub *ilorazem ryzyk*, który w tym przypadku jest stosunkiem r_{i1} dla populacji „Wiejskiej" w i –tym wariancie wiekowym do ryzyka r_{i0} dla populacji „Miejskiej", również w i –tym wariancie wiekowym.

Jeżeli $R_{Wi} = 1$, to ryzyka populacyjne są takie same w obu i –tych wariantach wiekowych, jeżeli $R_{Wi} > 1$, to ryzyko dla Wsi jest wyższe niż dla Miast w danym wariancie wieku samochodu.

Alternatywne nazwy ryzyka względnego.

Innymi używanymi określeniami (wskaźnika) ryzyka względnego $R_{Wi} = r_{i1} / r_{i0}$ są: stosunek temp, stosunek intensywności (IDR), iloraz zapadalności, stosunek częstości, iloraz prawdopodobieństw lub po prostu, stosunek ryzyk.

Rozdział 2-2-4. Uwaga o ogólnym indeksowaniu podgrup populacji.

W ogólnych rozważaniach, każda wartość indeksu grupowego j=1,2,...,N, wskazuje j-tą (generalną) populację, w której (nielosowe) czynniki X_i, i=1,2,...,p, przyjmują ustalone, im właściwe wartości. W ten sposób liczba wszystkich (pod)populacji wskazanych indeksem j oraz wartościami zmiennych X_i, i=1,2,...,p wynosi $N \times z_1 \times z_2 \times ... \times z_p$, gdzie z_i jest liczbą dyskretnych wartości, które przyjmuje zmienna X_i. W każdej z tych podpopulacji zmienną losową Y oznaczamy jako $Y_{l_1,...,l_p,j}$, gdzie l_i = 1,...,z_i dla i=1,2,...,p, a jej zmierzoną wartość jako $y_{l_1,...,l_p,j}$. Zbiór wszystkich $Y_{l_1,...,l_p,j}$ tworzy próbę oznaczaną tak jak poprzednio przez $\tilde{Y}$.

Rozdział 2-2-5. Dane dla przykładu.

Tabela 2-2-5.1 danych dla przykładu: Porównanie wystąpienia awarii silnika samochodów „Model" firmy „Auto" użytkowanych przez mieszkańców obszarów Miejskich oraz Wiejskich na całym obszarze dostępnym przez serwis tej firmy. Liczebności występujące w tabeli w są sumarycznymi liczebnościami dla próbki powstałej z wszystkich wylosowanych aglomeracji Miejskich (lub Wiejskich).

Wiek grupy samochodów (w miesiącach)	Obszary Miejskie		Obszary Wiejskie		Estymowany wskaźnik ryzyka, gdzie obszar Miast jest grupą referencyjną
	Ilość przypad-ków	Rozmiar próbek serwisowanych samochodów	Ilość przypadków	Rozmiar próbek serwisowanych samochodów	
0 – 12	1	172675	4	181343	3,81
13 – 24	16	123065	38	146207	2,00
25 – 36	30	96216	119	121374	3,14
37 – 48	71	92051	221	111353	2,57
49 – 60	102	72159	259	83004	2,21
61 – 72	130	54722	310	55932	2,33
73 – 84	133	32185	226	29007	1,89
85 +	40	8328	65	7538	1,80

Uwaga: Dla danych w Tabeli 2-2-5.1 dotyczących jednego wariantu wielu badań serwisowych w populacji „Miejskiej" (j=0) lub „Wiejskiej" (j=1), jedna liczba w kolumnach 3 lub 5 podająca rozmiar próbki, jest rozumiana jako liczba samochodo–lat w czasie prowadzonego badania dla określonego j-tego obszaru i i-tego wariantu wieku podgrupy (i, j), gdzie i =1,2,...,8.

Rozdział 2-2-6. Cel badań.

W ostatniej kolumnie Tabeli 2-2-5.1 podano *estymowane* z pobranych próbek ryzyka względne, w każdym z wariantów wiekowych. W każdym wariancie wieku, ryzyka wyniosły więcej niż 1, co jasno sugeruje, że obszar Wiejski ma wyższy ogólny wskaźnik awaryjności niż Miejski.

Rozdział 2-2-6-1. Uzasadnienie zastosowania rozkładu Poissona w analizie.

Fakt, że rozkład Poissona jest użyteczny dla modelowania pewnych typów zliczeń zdarzeń dla danych serwisowych, jest oparty na tym, że rozkład Poissona jest przybliżeniem rozkładu dwumianowego B [36], [2]. Ściśle rzecz biorąc, rozkład dwumianowy $B(m, p)$ przechodzi w rozkład Poissona $Poisson(\mu = m\,p)$ zmiennej Y tylko granicznie wtedy, gdy przy liczbie pomiarów m dążącym do nieskończoności i dwumianowym parametrze prawdopodobieństwa p bardzo małym, wartości oczekiwana liczby zdarzeń $\mu = E(Y) = m\,p$ pozostaje ustalona na wartości oczekiwanej rozkładu dwumianowego [36]. W granicy tej

oczekiwana dwumianowa liczba zliczeń „sukcesów” (wartość oczekiwana μ) jest względnie mała w porównaniu z liczbą wszystkich wyników, a rozkład Poissona daje dobre przybliżenie rozkładu dwumianowego dla rzadkich przypadków awarii (które są tu „sukcesami”). Dlatego zastosowanie modelu Poissona jest sugerowane, gdy otrzymujemy dużą liczbę wszystkich wyników dla próbki pobranej z populacji, w której bada się rozwój awaryjności, np. rozwój rzadkiej awarii silnika, tak że wielkość zakumulowanego (samochodo-) czasu jest duża, a jednocześnie tempo r_{ij} pojawiania się interesujących nas zdarzeń jest małe.

Dane w Tabeli 1 satysfakcjonująco spełniają to założenie, gdyż w każdej kategorii wiekowej występuje stosunkowo mały udział względny przypadków awarii w porównaniu do rozmiaru, tzn. liczby wszystkich wyników w odpowiedniej próbce pobranej z podpopulacji. Jednak pełna analiza powinna obejmować test nieparametrycznej hipotezy o typie rozkładu [2], z którego generowane są dane. Przemyślenie tego faktu pozostawiamy czytelnikowi jako ćwiczenie.

Rozdział 2-2-6-2. Przykład fizycznego odpowiednika danych w przykładzie.

Pojęcie „serwisowego” ryzyka względnego nie jest niepodobne do żadnej wielkości pojawiającej się np. w modelach fizycznych. Jej fizycznym odpowiednikiem jest iloraz przekrojów czynnych stosowany do opisu zajścia badanego procesu, który jest typem kontrastu różnych możliwych kanałów zachodzącej reakcji. W przypadku, gdy zmienna objaśniana ma pewien rozkład z wartością oczekiwaną zmieniającą się w zależności od wariantów zmiennej głównej oraz zmiennych pobocznych (kowariantów), wtedy w przypadku braku interakcji zmiennej głównej ze wspomnianymi kowariantami, zastosowanie stosunku temp może być przyczyną „zniknięcia” wpływu tych drugich na wartość ilorazu. W przypadku braku interakcji sytuacja ta byłaby więc podobna do omówionej w Rozdziale 2-2-7-2.

Rozdział 2-2-7. Równanie regresji Poissona ze zmiennymi ukrytymi.

Zmienną objaśnianą Y jest liczba zliczeń defektów (silników) otrzymanych dla każdej podgrupy, której wartości są wyjaśniane w regresji przez ustaloną liczbę czynników $X_1, X_2, ..., X_k$. Analizę dla regresji Poissona, omówimy na przykładzie danych z Tabeli 1 opisujących liczbę niedestrukcyjnych awarii silnika dla samochodów sklasyfikowanych wg wariantów wiekowych w Miastach i Wsiach.

Jedyną modelową różnicą pomiędzy regresją Poissona a standardową regresją wieloraką jest to, że pierwsza zakłada zastosowanie rozkładu Poissona, podczas gdy druga zakłada zastosowanie rozkładu normalnego, co oczywiście wpływa na postać równania regresji zgodnie z uwagami zawartymi pomiędzy (2-1-3.11) a (2-1-3.12). Równanie (2-1-3.10) jest treścią równania regresji pierwszego rodzaju. Jego współczynniki musimy oszacować na podstawie pobranej próbki odwołując się do MNW. Równanie regresji z oszacowanymi współczynnikami nazywamy *równaniem regresji drugiego rodzaju*. Funkcja wiarygodności dla analizy regresji Poissona ma ogólną postać (2-1-3.14) [1].

Rozdział 2-2-7-1. Indeksowanie grup w przykładzie.

W rozważanym przykładzie występują dwa czynniki, czynnik „obszar” serwisowania oraz „wiek” samochodu. Czynnik „obszar” przyjmuje 2 wartości. Zgodnie z już wcześniej wprowadzonymi oznaczeniami, ponieważ mamy dwie populacje „Miast” i „Wsi”, liczba generalnych populacji N=2 skąd, ze względu na poniżej wprowadzone kodowanie zmiennych ukrytych (tzn. kierunkowych), przyjmujemy j=0,1. Natomiast zgodnie z danymi z Tabeli 1, (kategoryczny) czynnik „wieku” samochodu przyjmuje z = 8 wartości. Stąd liczba wszystkich podpopulacji wynosi $N \times z = 2 \times 8 = 16$, a każdą z podpopulacji (podgrup) wskazuje para indeksów grupowych (i, j). Zmienne losowe Y oznaczamy jako Y_{ij}, gdzie i = 1,...,8 , a j=0,1. Indeksowanie dla populacji i podpopulacji przenosi się automatycznie na indeksowanie pobranych z tych populacji próbek.

Zbudowanie modelu regresji Poissona dla powyższej sytuacji oznacza opisanie oczekiwanej liczby przypadków awarii silnika, $E(Y_{ij})$, poprzez wprowadzone do modelu zmienne objaśniające. Liczba zliczeń Y_{ij} jest zmienną losową Poissona (teoretycznie zmienną losową dwumianową) z wartością oczekiwaną równą $\mu_{ij} = \ell_{ij} r_{ij}$. Równanie to, dla określonej postaci zależności r_{ij} od czynników, wyraża treść funkcji regresji pierwszego rodzaju, tzn. postulowaną jej postać w całej generalnej populacji[22].
Analiza regresji Poissona ma ustalić, czy widoczny „na oko” wzorzec danych w Tabeli 1 jest statystycznie istotny oraz otrzymać estymator ogólnego ryzyka względnego, który byłby dopasowany ze względu na wiek samochodu (tzn. wiek samochodu jest zmienną pod kontrolą).

W rozważanym przykładzie występują więc k=2 czynniki, czynnik wpływu głównego X_1, którym jest „obszar” serwisowania oraz czynnik poboczny X_2 „wieku” samochodu. Ponieważ „wiek” będzie klasyfikowany w ośmiu kategoriach, użyjemy do ich wskazania (indeksowania) siedmiu zmiennych ukrytych [1]. Zmienna „obszar”, która zawiera dwa warianty, wymaga tylko jednej zmiennej kierunkowej.

Ogólna postać modelu regresji, czyli funkcji opisującej zmianę wartości oczekiwanej liczby awarii (silnika) wraz ze zmianą grupy (i, j), może być zapisana zgodnie z (2-1-3.10) [1] następująco:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij} r_{ij} \quad , \qquad i = 1, 2, \ldots, 8 ; \quad j = 0, 1 . \qquad (2\text{-}2\text{-}7\text{-}1.36)$$

[22] We wstępnych rozważaniach indeksowaliśmy grupy jednym indeksem n. Np. dla grupy n, gdzie n = 1,2,...,N, zmienną obserwowanej ilości defektów oznaczaliśmy przez Y_n, a całkowitą wielkość zakumulowanego czasu dla wszystkich samochodów w n-tej podgrupie przez ℓ_n. Równanie regresji miało wtedy postać (2-1-3.10) $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta), \quad n = 1,2,\ldots,N$.

Wspomniane **zmienne ukryte** (wskazujące, kierunkowe) U_k oraz M wskazującą w następujący sposób [1] odpowiednio wariant „wieku” oraz „obszaru”:

$$U_k = \begin{cases} 1 & \text{jeśli} \quad k = i, \qquad \text{gdzie} \quad i = 1, 2, ..., 7 \\ 0 & \text{w przeciwnym wypadku} \end{cases} \tag{2-2-7-1.37}$$

$$M = \begin{cases} 1 & \text{jeśli} \quad j = 1 \qquad \text{(Wsie)} \\ 0 & \text{jeśli} \quad j = 0 \qquad \text{(Miasta)} \quad . \end{cases} \tag{2-2-7-1.38}$$

Podstawowa dla wielu analiz regresji Poissona, logarytmiczna postać funkcji ryzyka [1], [38], która pojawi się w (2-2-7-1.36) i korzystająca z kodowania (2-2-7-1.37) oraz (2-2-7-1.38) ma w przypadku **bez interakcji** następującą postać:

Model 1: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M \quad . \tag{2-2-7-1.39}$$

Korzystając z kodowania (2-2-7-1.37) i (2-2-7-1.38), możemy w powyższym „Modelu 1” ryzyka, wyrazić r_{ij} poprzez parametry α_i i β w następujący sposób:

$$\ln r_{i0} = \alpha + \alpha_i \qquad \text{oraz} \qquad \ln r_{i1} = \alpha + \alpha_i + \beta \ , \qquad i = 1, 2, ..., 7, \tag{2-2-7-1.40}$$

oraz

$$\ln r_{80} = \alpha \qquad \text{oraz} \qquad \ln r_{81} = \alpha + \beta \ , \qquad \text{dla } i = 8 \ , \tag{2-2-7-1.41}$$

co wynikało z tego, że $U_k = 1$ dla $k = i = 1, 2, ..., 7$ oraz $U_k = 0$ dla $i = 8$.

Powyższy przykład modelowania jest wykorzystywany w estymacji ryzyka rozwijania się uszkodzenia (silnika samochodu) z wiekiem. Bardziej ogólne i popularne zastosowania regresji Poissona dotyczą modelowania tempa defektów, czyli tzw. *intensywności* procesu, dla różnych interesujących nas podgrup.

Wniosek: Ze związków (2-2-7-1.40) i (2-2-7-1.41) widzimy, że w traktowanych osobno obszarach „Miejskim” i „Wiejskim” ryzyko (tempo awarii) r_{ij} zmienia się z wariantem „wieku”, co z powodu niezerowych oszacowań współczynników α_i będzie widoczne w poniższych raportach SAS.

Uwaga o alternatywnym kodowaniu:
Alternatywnie model może być zdefiniowany poprzez użycie ośmiu zmiennych kierunkowych dla wieku i jednej zmiennej kierunkowej dla obszaru [1]. Gdyby zastosowano osiem zmiennych kierunkowych dla wieku, użycie wyrazu wolnego byłoby błędem.

Rozdział 2-2-7-2. Estymator ogólnego ryzyka względnego w modelu bez interakcji.

Poniżej wyprowadzimy ważny wniosek dotyczący ryzyka względnego w modelu bez interakcji czynnika „obszar" z czynnikiem pobocznym „wieku".

Korzystając z (2-2-7-1.40) i (2-2-7-1.41) otrzymujemy:

$$\ln r_{i1} - \ln r_{i0} = (\alpha + \alpha_i + \beta - \alpha - \alpha_i) = \beta \quad , \qquad i = 1, 2, ..., 7 \qquad (2\text{-}2\text{-}7\text{-}2.42)$$

oraz

$$\ln r_{81} - \ln r_{80} = (\alpha + \beta - \alpha) = \beta \ , \qquad i = 8 \ . \qquad (2\text{-}2\text{-}7\text{-}2.43)$$

Korzystając z (2-2-7-2.42) oraz (2-2-7-2.43) widzimy, że **ryzyko względne** (2-2-3-2.35) dla modelu (2-2-7-1.39) nie zawierającego interakcji jest równe:

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} = \exp\left[\ln\left(\frac{r_{i1}}{r_{i0}}\right)\right] = \exp[\ln r_{i1} - \ln r_{i0}] = \exp[\beta] = e^{\beta} \quad , \qquad i = 1, 2, ..., 8. \qquad (2\text{-}2\text{-}7\text{-}2.44)$$

Powyższy model pozwala na estymację wskaźnika ryzyka względnego dla każdej kategorii wiekowej. Czynimy to stosując MNW do estymacji współczynnika kierunkowego β stojącego obok zmiennej M i w ten sposób dopasowując model do danych, a następnie licząc eksponentę tego estymatora.

Estymator ogólnego ryzyka względnego: Ponieważ estymowany wskaźnik ryzyka względnego e^{β} jest niezależny od i (tzn. od kategorii wiekowej), zatem możemy interpretować $\hat{r}_{Wi} = e^{\hat{\beta}}$ jako *estymator ogólnego ryzyka względnego* R_{Wi}, dopasowanego do wieku, gdzie $\hat{\beta}$ jest estymatorem MNW parametru β.

Wniosek *o postaci ryzyka względnego w modelu bez interakcji*: **Dla modelu** (2-2-7-1.39) **bez interakcji zmiennych „obszar" i „wiek"** (oznaczonego jako Model 1), **ryzyko względne nie zależy od wariantu wiekowego**, tzn. wpływ „obszaru" nie jest modyfikowany przez „wiek".

Rozważany przykład przedstawia model statystyczny przydatny do przeprowadzenia analizy regresji Poissona przy dwóch czynnikach. W ogólności, zamiast dwóch czynników (wiek i obszar), możemy mieć k - czynników: X_1, X_2, ..., X_k. Wtedy ogólna metoda dopasowywania modelu regresji Poissona nie zmienia się i

polega na wykorzystaniu rozkładu Poissona do otrzymania funkcji wiarygodności, która może być później maksymalizowana w celu otrzymania estymatorów parametrów modelu oraz oszacowanych błędów standardowych zmaksymalizowanych statystyk MNW. Ponieważ pakiety programów (zawarte np. w systemie analiz statystycznych SAS) mogą wykonywać takie analizy, zatem użytkownik musi jedynie wyszczególnić trafny model, który ma być dopasowany. Numeryczna analiza dla powyższego przykładu zostanie przeprowadzona w dalszej części.

Rozdział 2-2-8. Macierz kowariancji i obserwowana informacja Fishera.

Dodatkowo procedurami SAS estymowana jest, po pierwsze, obserwowana macierz kowariancji $\hat{V}(\hat{\beta})$ estymatorów parametrów β będąca w metodzie MNW odwrotnością *obserwowanej* informacji Fishera $\mathbf{iF}$:

$$\hat{\mathrm{V}}(\hat{\beta}) = \begin{bmatrix} \hat{\sigma}^2(\hat{\beta}_0) & \hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \hat{Cov}(\hat{\beta}_0, \hat{\beta}_2) & \\ \hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \hat{\sigma}^2(\hat{\beta}_1) & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \\ \hat{Cov}(\hat{\beta}_0, \hat{\beta}_2) & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \hat{\sigma}^2(\hat{\beta}_2) & \\ & & & \ddots \end{bmatrix} := \mathbf{iF}^{-1}(\hat{\beta}) \quad , \qquad (2\text{-}2\text{-}8.45)$$

oraz po drugie, miary dobroci dopasowania rozważanego modelu i pewne statystyki diagnostyczne regresji, użyteczne dla wykrywania obserwacji wpływowych oraz współliniowości [1]. Wszystko to w raportach SASa pojawia się jako część wydruku komputerowego. Więcej na temat *obserwowanej* informacji Fishera $\mathbf{iF}$, jej definicji oraz przykładów, można znaleźć w [38], [5].

Rozdział 2-2-9. Statystyczne kryterium doboru modelu.

Do weryfikacji hipotez o nie występowaniu statystycznie istotnego braku dopasowania w jednym modelu w porównaniu z innym modelem, będącym członkiem tej samej hierarchii modeli, wykorzystamy logarytmiczny ilorazu zmaksymalizowanych wiarygodności tych modeli (2-1-4-1.23) oraz dewiację (1-1.42), jako jego szczególny typ. W Rozdziale 2-1-4-1 przekonaliśmy się, że modele mogą być porównywane poprzez obliczenie różnic pomiędzy parami dewiancji tych modeli.

„**Model podstawowy**” został omówiony w Rozdziale 2-1-2. Wiarygodność próby $\tilde{Y}$ przyjmuje dla modelu podstawowego postać (2-1-2.5) a jej postać zmaksymalizowana jest określona zgodnie z (2-1-2.9).

Powód konstrukcji modelu regresji: Powodem analizowania modelu regresji, a nie trwania przy modelu podstawowym, nie jest sama dokładność dopasowania (która nie może być lepsza niż w modelu podstawowym), lecz próba zrozumienia istoty opisywanego zjawiska oraz mniejsza liczba parametrów, co wpływa na zmniejszenie kosztów oszacowywania parametrów z określoną dokładnością [1].

Rozdział 2-2-9-1. Minimalny oszczędny model opisu danych.

Model podstawowy bez struktury parametrów zawiera tyle parametrów ile jest grup danych pomiarowych, czyli N . Celem analizy regresji jest otrzymanie oszczędnego opisu danych. Model $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta), \quad n = 1,2,...,N$, zawierający $k+1$ parametrów, uznamy za oszczędny, jeśli ma wartość zmaksymalizowaną wiarygodności prawie tak dużą, jak dla modelu podstawowego i jednocześnie najmniejszą liczbę parametrów funkcji regresji w klasie modeli hierarchicznych, do których należy. Dla modelu oszczędnego wartość dewiancji wpadnie w wiarygodnościowy obszar przyjęć hipotezy zerowej.

Rozdział 2-2-10. Analiza regresji dla przykładu: Model 1.

Pierwszy rozważany model regresji Poissona dla oczekiwanej liczby przypadków awarii silnika w podgrupach (i, j) ma postać zadaną przez (2-2-7-1.36) oraz (2-2-7-1.39). Jest więc to uprzednio wprowadzony Model 1:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij} r_{ij} \quad , \qquad i = 1, 2, ... , 8 ; \quad j = 0, 1 , \tag{2-2-10.46}$$

gdzie:

Model 1: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M \quad . \tag{2-2-10.47}$$

Zmienne U_k były „sztucznie" wprowadzonymi zmiennymi kierunkowymi (ukrytymi) (2-2-7-1.37) wskazującymi wariant wiekowy i przyjmującymi wartości 0 lub 1, a zmienna kierunkowa M przyjmowała zgodnie z (2-2-7-1.38) wartości 0 lub 1, wskazując odpowiednio obszar Miejski lub Wiejski.
Dla powyższego modelu ryzyko względne wynosi (2-2-3-2.34):

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} \quad , \tag{2-2-10.48}$$

a zgodnie z (2-2-7-2.44) jego postać redukuje się do:

$$R_{Wi} = e^{\beta} \quad , \tag{2-2-10.49}$$

gdzie e^{β} jest niezależne od i, reprezentując ogólne ryzyko względne dopasowane do „wieku".

Konkretna postać funkcji wiarygodności powyższego modelu jest konsekwencją założenia, że liczba zliczeń Y_{ij} ma rozkład Poissona ze średnią $\mu_{ij} = \ell_{ij} r_{ij}$. Zgodnie z (2-1-3.14) ma ona w próbce postać:

$$P(y \mid (\beta)) = \prod_{i=1}^{8} \left\{ \left[\frac{(\ell_{i0} r_{i0})^{y_{i0}} e^{-\ell_{i0} r_{i0}}}{y_{i0}!} \right] \left[\frac{(\ell_{i1} r_{i1})^{y_{i1}} e^{-\ell_{i1} r_{i1}}}{y_{i1}!} \right] \right\} \quad , \tag{2-2-10.50}$$

gdzie zgodnie z (2-2-7-1.40)-(2-2-7-1.41) mamy $r_{i0} = \exp(\alpha + \alpha_i)$ i $r_{i1} = \exp(\alpha + \alpha_i + \beta)$ dla i = 1, ..., 7 , oraz $r_{80} = \exp(\alpha)$ i $r_{81} = \exp(\alpha + \beta)$ dla $i = 8$.

Użycie pakietu komputerowego dla regresji Poissona będzie maksymalizowało powyższą funkcję wiarygodności, dając 9 estymatorów parametrów badanego modelu:

$$\{\hat{\alpha}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, \hat{\alpha}_7, \hat{\beta}\} \qquad (2\text{-}2\text{-}10.51)$$

oraz oszacowaną 9×9 - wymiarową macierz kowariancji (2-2-8.45).

Rozdział 2-2-11. Analiza numeryczna programem SAS.

W celu wykonania selekcji modelu regresji Poissona dla powyższego przykładu z wykorzystaniem SAS należy utworzyć zbiór danych oraz program wyznaczający oszacowania parametrów modelu zgodnie z procedurą języka programowania 4GL tej aplikacji. Następnie zbiór danych należy umieścić w edytorze systemu SAS (Widok -> Enhanced Editor) i uruchamiając właściwą procedurę, dokonać przeliczenia modelu (Uruchom -> Przekaż) [11]. Język 4GL dzięki swojej budowie umożliwia przetwarzanie oraz pełną obsługę zbiorów danych.
Ramka ogólnej składni wprowadzonej procedury ma postać:

proc nazwa procedury *data* = zbior_danych opcje_procedury;
...
Instrukcje;
...
run;

Niektóre z wykorzystywanych poniżej instrukcji potrzebnych w dalszej analizie (zarówno dla programów wczytujących dane jak i procedur do analizy modeli), podane zostały w **Uzupełnieniu 1** w Rozdziale 2-2-15. Pełny ich wykaz oraz zastosowanie można znaleźć w pomocy pakietu SAS.

Rozdział 2-2-11-1. Dane oraz programy.

W analizie rozważanego przykładu można wykorzystać jeden, poniższy zbiór danych, wprowadzając w zależności od modelu odpowiednie modyfikacje dopiero na poziomie programu analizującego rozważany model. Wyjaśnienie używanych poleceń języka 4GL oraz opis zmiennych od A do O znajdują się w Uzupełnieniu 1 w Rozdziale 2-2-15.

Zbiór danych:

```
data awaria;
input A Y  L  M  U1  U2  U3  U4  U5  U6  U7 U1M U2M U3M U4M U5M U6M U7M O;
ln = log(L);
datalines;
1    1    172675  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2    16   123065  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
3    30   96216   0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
4    71   92051   0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0
5    102  72159   0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
6    130  54722   0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
7    133  32185   0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
8    40   8328    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
1    4    181343  1  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0
2    38   146207  1  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0
3    119  121374  1  0  0  1  0  0  0  0  0  0  1  0  0  0  0  0
4    221  111353  1  0  0  0  1  0  0  0  0  0  0  1  0  0  0  0
5    259  83004   1  0  0  0  0  1  0  0  0  0  0  0  1  0  0  0
6    310  55932   1  0  0  0  0  0  1  0  0  0  0  0  0  1  0  0
7    226  29007   1  0  0  0  0  0  0  1  0  0  0  0  0  0  1  0
8    65   7538    1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
;
run;
```

Określenie funkcji wiążacej (Link Function) **oraz czynnika przesunięcia** (Offset Variable): Funkcja wiążacą $g(\mu_n)$ wiąże wartość oczekiwaną warunkową $\mu_n \equiv E(Y_n)$ z kombinacją liniową zmiennych objaśniających ujętą w postaci funkcji regresji. W przypadku rozkładu Poissona funkcja $g(\mu_n) = \ln(\mu_n)$, a w przypadku rozkładu normalnego $g(\mu_n) = \mu_n$.

W omawianych raportach SAS jest ona skonstruowana jako funkcja $g(r(x_n, \beta))$, zatem dla rozkładu Poissona otrzymujemy $g(r(x_n, \beta)) = \ln(r(x_n, \beta))$, gdzie funkcja ryzyka $r(x_n, \beta)$ ma postać jak w (2-1-3.12). Ponieważ zgodnie z (2-1-3.10), μ_n oraz $r(x_n, \beta)$ różnią się czynnikiem (skumulowanego czasu badania) $\ell_n \equiv \mathrm{L}$, zatem w powyższym programie dla zbioru danych, pojawiła się komenda ln = log(L) służąca do wczytania niezbędnego wyrażenia $\ln(\ell_n)$.

Natomiast na skutek przypisania w *poniższym* programie zmiennej '*offset*' wartości $\ln \equiv \ln(\ell_n)$, wyrażenie to pojawi się w raporcie SAS jako tzw. zmienna przesunięcia (Offset Variable).

Wczytanie programu analizującego model:

Po wczytaniu zbioru danych, przystępujemy do wpisania programu analizującego konkretną postać modelu, który wykorzysta całość lub część powyższego zbioru danych, w zależności od modelu regresji Poissona dla

przykładu awarii silnika. I tak dla Modelu 1 program ten, wykorzystujący procedurę GENMOD, ma następującą postać:

proc ***genmod*** *data* = awaria*;*
model Y = M U1 U2 U3 U4 U5 U6 U7 / covb
dist = poisson
link = log
offset = ln*;*
run;
quit;

Uwaga: Zamiast wpisywać *model* w powyższej postaci można użyć wprowadzonej zmiennej klasującej A i zamienić odpowiednie linie:

ref = 8;
class A;
model Y = M A / corrb
a program SAS odda raport identycznej postaci.
Uwaga: Dodatkowo zamieniono komendę wyznaczenia macierzy kowariancji estymatorów 'covb' na polecenie 'corrb' wyznaczenia ich macierzy korelacyjnej.

Rozdział 2-2-11-2. Wynik analizy numerycznej SAS dla Modelu. 1

Jako rezultat wczytania powyższych danych i uruchomienia procedury GENMOD dla rozważanego aktualnie Modelu 1, otrzymujemy raport systemu SAS, który ma następującą postać:

```
                System SAS
            The GENMOD Procedure

            Informacje o modelu

 Zbiór                        WORK.MODEL1
 Rozkład                          Poisson
 Funkcja wiążąca                      Log
 Zmienna zależna                        Y
 Zmienna przesunięcia                  ln

Liczba obserwacji wczytanych           16
Liczba obserwacji użytych              16

  Informacje o poziomie klasyfikacji

 Klasa     Poziomy    Wartości

 A               8    1 2 3 4 5 6 7 8
```

Informacje o parametrach

Parametr	Efekt
Prm1	Intercept
Prm2	M
Prm3	U1
Prm4	U2
Prm5	U3
Prm6	U4
Prm7	U5
Prm8	U6
Prm9	U7

Kryteria oceny zgodności

Kryterium	St. sw.	Wartość	Wartość/st. sw.
Dewiancja	7	**8.1950**	1.1707
Skalowana dewia	7	8.1950	1.1707
Chi-kwadrat Pearso	7	8.0626	1.1518
Scaled Pearson X2	7	8.0626	1.1518
Log. wiarogodn		7201.8635	

System SAS
The GENMOD Procedure
Algorytm osiągnął zbieżność.

Macierz kowariancji szacunkowych

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.01074	-0.001824	-0.009465	-0.009419	-0.009398
Prm2	-0.001824	**0.002725**	-0.000087	-0.000156	-0.000188
Prm3	-0.009465	-0.000087	0.20953	0.009529	0.009530
Prm4	-0.009419	-0.000156	0.009529	0.02805	0.009535
Prm5	-0.009398	-0.000188	0.009530	0.009535	0.01625
Prm6	-0.009413	-0.000166	0.009529	0.009533	0.009535
Prm7	-0.009431	-0.000138	0.009528	0.009532	0.009533
Prm8	-0.009476	-0.000072	0.009526	0.009528	0.009529
Prm9	-0.009526	2.5943E-6	0.009524	0.009524	0.009524

Macierz kowariancji szacunkowych

	Prm6	Prm7	Prm8	Prm9
Prm1	-0.009413	-0.009431	-0.009476	-0.009526
Prm2	-0.000166	-0.000138	-0.000072	2.5943E-6
Prm3	0.009529	0.009528	0.009526	0.009524
Prm4	0.009533	0.009532	0.009528	0.009524
Prm5	0.009535	0.009533	0.009529	0.009524
Prm6	0.01296	0.009532	0.009528	0.009524
Prm7	0.009532	0.01230	0.009527	0.009524
Prm8	0.009528	0.009527	0.01180	0.009524
Prm9	0.009524	0.009524	0.009524	0.01231

Analiza ocen parametrów

Parametr	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
Intercept	1	-5.4797	0.1037	-5.6828	-5.2765	2794.67	<.0001
M	1	**0.8043**	**0.0522**	0.7020	0.9066	237.34	<.0001
U1	1	-6.1782	0.4577	-7.0753	-5.2810	182.17	<.0001
U2	1	-3.5480	0.1675	-3.8763	-3.2197	448.76	<.0001
U3	1	-2.3308	0.1275	-2.5807	-2.0810	334.36	<.0001
U4	1	-1.5830	0.1138	-1.8061	-1.3599	193.38	<.0001
U5	1	-1.0909	0.1109	-1.3083	-0.8735	96.75	<.0001
U6	1	-0.5328	0.1086	-0.7457	-0.3199	24.06	<.0001
U7	1	-0.1196	0.1109	-0.3371	0.0978	1.16	0.2809
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed. (Przedział ufności Wald'a, Rozdział 7.1, Uzupełnienie 1).

Rozdział 2-2-11-3. Oszacowanie parametru i błąd standardowy oszacowania dla Modelu 1.

Z powyższego raportu możemy odczytać *oszacowanie* $\hat{\beta}$ *MNW parametru* β:

$$\hat{\beta} = 0{,}8043 \qquad (2\text{-}2\text{-}11\text{-}3.52)$$

oraz *błąd standardowy* (*se*) tego oszacowania wyznaczony jako element macierzy (wariancji-) kowariancji (2-2-8.45) [1]:

$$\hat{\sigma}_{\hat{\beta}} \equiv se(\hat{\beta}) = (0{,}002725)^{1/2} = 0{,}0522 \ . \qquad (2\text{-}2\text{-}11\text{-}3.53)$$

Punktowe oszacowanie $\hat{r}_{Wi}^{Model1}$ dopasowanego ze względu na wiek ryzyka względnego R_{Wi} wynosi więc:

$$\hat{r}_{Wi}^{Model1} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513 \ . \qquad (2\text{-}2\text{-}11\text{-}3.54)$$

Natomiast 95%-owy wiarygodnościowy przedział ufności Wald'a (Rozdział 7.1, Uzupełnienie 1) dla e^{β} [1], przy odwołaniu się do faktu, że dla dużej próbki estymator MNW ma *w przybliżeniu* rozkład normalny, ma postać:

$$\exp[\hat{\beta} \pm 1{,}96\ \hat{\sigma}_{\hat{\beta}}] = \exp\left[0{,}8043 \pm 1{,}96\ (0{,}0522)\right] = \exp\left(0{,}8043 \pm 0{,}1023\right) \ , \qquad (2\text{-}2\text{-}11\text{-}3.55)$$

lub

$$\left(e^{0.7020}, e^{0.9066}\right) = \left(2{,}01778;\ 2{,}47589\right) \ . \qquad (2\text{-}2\text{-}11\text{-}3.56)$$

Rozdział 2-2-11-4. Test hipotezy zerowej z wykorzystaniem statystyki Wald'a.

Dla dużej próbki, test hipotezy zerowej:

$$H_0\text{:}\ \beta = 0 \qquad (2\text{-}2\text{-}11\text{-}4.57)$$

o braku zależności korelacyjnej tempa awarii od lokalizacji, wobec hipotezy alternatywnej:

$$H_1\text{:}\ \beta \neq 0 \ , \qquad (2\text{-}2\text{-}11\text{-}4.58)$$

może być przeprowadzony z zastosowaniem statystyki Wald'a [1] (Rozdział 7.1, Uzupełnienie 1):

$$U = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} \ . \qquad (2\text{-}2\text{-}11\text{-}4.59)$$

Przy prawdziwości hipotezy zerowej H_0: $\beta = 0$ statystyka U ma asymptotycznie rozkład normalny N(0,1).

Dla rozważanego przykładu wartość statystyki Wald'a wynosi:

$$U = \frac{0{,}8043 - 0}{0{,}0522} = 15{,}408 \ , \qquad (2\text{-}2\text{-}11\text{-}4.60)$$

natomiast empiryczny poziom istotności [1], [9] ma wartość (wyznaczoną np. w pakiecie kalkulacyjnym Excel):

$$p = \Pr(|U| \geq 15{,}408) < 0{,}0001 \ . \qquad (2\text{-}2\text{-}11\text{-}4.61)$$

Rozdział 2-2-11-5. Wniosek.

Ze względu na $p < 0{,}0001$ przeprowadzona analiza regresji Poissona wskazuje na statystycznie istotny wpływ lokalizacji (tzn. na statystyczną istotność wprowadzenia parametru β przy zmiennej kierunkowej *M* wskazującej lokalizację). Ze względu na wartość oszacowanego ryzyka względnego $\hat{r}_{Wi} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513$ ogólne, dopasowane ze względu na wiek, tempo awarii silników samochodów na Wsiach jest około 2,2 razy większe niż w Miastach. Wyznaczony 95%-owy wiarygodnościowy przedział ufności dla ogólnego dopasowania ryzyka względnego wynosi $(2{,}01776;\ 2{,}47591)$.
Do analizy Modelu 1 wrócimy jeszcze poniżej, aby omówić interakcję czynnika „wiek" ze zmienną „obszar", bądź uwzględnienie „wieku" jako ewentualnego zaburzenia w modelu [1] oraz porównać dobroć dopasowania Modelu 1 z innymi modelami w hierarchii.

Rozdział 2-2-12. Charakter kowarianta „wiek" - interakcja czy zaburzenie.

Głównym wpływem interesującym nas w analizie ryzyka jest zmienna „obszar". W formule (2-2-7-2.44) na ryzyko względne, zmienna wiek okazała się nawet nie występować. Jednak w wyprowadzeniu (2-2-7-2.44) nie braliśmy pod uwagę możliwości występowania zmiennej pobocznej „wiek" jako kowarianta w interakcji ze zmienną „obszar". Przyjrzyjmy się więc bliżej charakterowi zmiennej „wiek" z punktu widzenia sposobu wprowadzenia jej do modelu regresji.

Punkt 1. Zmienna poboczna „wiek" może być wprowadzona do multiplikatywnego **członu interakcji** ze zmienną „obszar". Rozważanie tej możliwości związane jest z odpowiedzią na pytanie o to *czy zmienna* „wiek" *modyfikuje wpływ zmiennej* „obszar", to znaczy, czy wpływ zmiennej „obszar" mierzony ryzykiem względnym, różni się dla różnych wariantów wieku?
Punkt 2. Zmienna „wiek" może być wprowadzona do modelu tylko jako **zaburzenie**. Możliwość ta jest rozważana wtedy, gdy po analizie Punktu 1, okazało się, że wprowadzenie zmiennej „wiek" do modelu w członie interakcji jest nieistotne statystycznie. W takiej sytuacji rozważamy czy zmienna „wiek" jest *zaburzeniem*, tzn. czy powinna znaleźć się w modelu w jakiejkolwiek formie po to, aby dać właściwe określenie jej wpływu na oszacowanie interesującego nas parametru, którym w rozważanym przykładzie jest ryzyko względne?

Jest różnica pomiędzy wprowadzeniem do modelu nowej zmiennej w postaci zaburzenia lub w postaci iloczynowego członu interakcji. **Nie wykonuje się testów statystycznych w przypadku, gdy zmienna ma wejść do modelu w postaci zaburzenia** [1].

Szczegółowe omówienie problemu rozróżnienia pomiędzy interakcją, czyli własnością modyfikacji wpływu głównego zmiennej typu „obszar" przez kowarianta będącego zmienną poboczną typu „wiek", a problemem zaburzenia głównego wpływu zmiennej „obszar" przez zmienną poboczną „wiek", można znaleźć w [1].

Rozdział 2-2-12-1. Analiza interakcji obszaru i wieku. Model 2.

Aby rozstrzygnąć kwestię zawartą w powyższym Punkcie 1, dotyczącą możliwości, że zmienna „wiek" jest kowariantem modyfikującym wpływ zmiennej „obszar", rozszerzmy Model 1 , (2-2-10.47) (porównaj (2-2-7-1.39)), o człon interakcji, otrzymując:

Model 2: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M + \sum_{k=1}^{7} \delta_k \left(M U_k \right) \quad , i = 1, 2, ..., 8; \quad j = 0, 1 . \qquad (2\text{-}2\text{-}12\text{-}1.62)$$

Aby uniknąć osobliwości, tzn. idealnej współliniowości, możemy dodać człony interakcji tylko dla siedmiu (a nie ośmiu) zmiennych kierunkowych U_k .

Istotność interakcji „wieku" z „obszarem" możemy testować weryfikując hipotezę zerową:

$$H_0: \ \delta_1 = \delta_2 = \ldots = \delta_7 = 0 \ , \qquad (2\text{-}2\text{-}12\text{-}1.63)$$

z wykorzystaniem statystyki ilorazu wiarygodności (1-1.42). Ma ona przy prawdziwości hipotezy zerowej H_0 asymptotycznie rozkład χ^2 z 7 stopniami swobody, co jest liczbą nowych parametrów wprowadzonych do wyższego Modelu 2.

Statystyka testowa (1-1.42) pozwala więc na porównanie Modelu 1 (bez interakcji) z Modelem 2, który zawiera siedem iloczynowych członów interakcji $M U_k$.

Rozdział 2-2-12-2. Program SAS dla Modelu 2.

Ponieważ w Modelu 2 chcemy uwzględnić również interakcję „wieku” i „obszaru”, zatem po wczytaniu danych takich samych jak w Punkcie 1.11.1, należy przy korzystaniu z procedury GENMOD (Punkt 1.11.1) zmienić linię *model* na uwzględniający człony interakcji $M\,U_k$, k=1,2,...,7, wczytując program:

proc ***genmod*** *data* = awaria*;*
model Y = M U1 U2 U3 U4 U5 U6 U7 U1M U2M U3M U4M U5M U6M U7M / covb
dist = poisson
link = log
offset = ln*;*
run;
quit;

Rozdział 2-2-12-3. Raport z dopasowania Modelu 2.

W wyniku analizy otrzymujemy następujący komputerowy raport SAS z dopasowywania Modelu 2. Jak to wynika z powyższych rozważań, raport ten dotyczy analizy z uwzględnieniem interakcji zmiennych „wiek” i „obszar”.

```
                    System SAS
                 The GENMOD Procedure

                 Informacje o modelu

 Zbiór                        WORK.MODEL2
 Rozkład                          Poisson
 Funkcja wiążąca                      Log
 Zmienna zależna                        Y
 Zmienna przesunięcia                  ln

Liczba obserwacji wczytanych           16
Liczba obserwacji użytych              16


  Informacje o poziomie klasyfikacji

 Klasa     Poziomy    Wartości

 A               8    1 2 3 4 5 6 7 8

                    System SAS
                 The GENMOD Procedure

              Kryteria oceny zgodności

                        St.                     Wartość/st.
Kryterium               sw.         Wartość             sw.

Dewiancja                 0          0.0000           .
Skalowana dewia           0          0.0000           .
Chi-kwadrat Pearso        0          0.0000           .
Scaled Pearson X2         0          0.0000           .
Log. wiarogodn                     7205.9610

Algorytm osiągnął zbieżność.
```

System SAS
The GENMOD Procedure

Analiza ocen parametrów

Parametr	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
Intercept	1	-5.3385	0.1581	-5.6484	-5.0286	1139.98	<.0001
M	1	0.5852	0.2010	0.1913	0.9790	8.48	0.0036
U1	1	-6.7207	1.0124	-8.7050	-4.7364	44.07	<.0001
U2	1	-3.6094	0.2958	-4.1891	-3.0296	148.89	<.0001
U3	1	-2.7347	0.2415	-3.2080	-2.2613	128.20	<.0001
U4	1	-1.8289	0.1977	-2.2164	-1.4414	85.58	<.0001
U5	1	-1.2232	0.1866	-1.5888	-0.8575	42.99	<.0001
U6	1	-0.7040	0.1808	-1.0584	-0.3496	15.16	<.0001
U7	1	-0.1504	0.1803	-0.5038	0.2030	0.70	0.4042
U1M	1	0.7521	1.1360	-1.4743	2.9786	0.44	**0.5079**
U2M	1	0.1075	0.3594	-0.5970	0.8120	0.09	**0.7649**
U3M	1	0.5605	0.2866	-0.0012	1.1221	3.83	**0.0505**
U4M	1	0.3599	0.2429	-0.1161	0.8360	2.20	**0.1384**
U5M	1	0.2067	0.2325	-0.2490	0.6623	0.79	**0.3740**
U6M	1	0.2620	0.2265	-0.1819	0.7059	1.34	**0.2474**
U7M	1	0.0490	0.2288	-0.3994	0.4973	0.05	**0.8305**
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed.

Z raportu SAS widać, że dewiancja dla Modelu 2 jest dokładnie równa zero:

$$D\left(\hat{\beta}\right)^{Model2} = 0 \ , \qquad (2\text{-}2\text{-}12\text{-}3.64)$$

co oznacza, że model ten dopasowuje się do danych empirycznych idealnie. *Fakt ten jest spowodowany dopasowywaniem modelu z* 16 *parametrami do* $N = 16$ *elementowego zbioru danych.*

Jednak z raportu widać (pogrubienie na końcu linii U1M do U7M), że oszacowania parametrów interakcji $\delta_1, \delta_2, \ldots, \delta_7$ różnią się na poziomie istotności $\alpha = 0{,}05$ statystycznie nieistotnie od zera, co oznacza, że nie ma potrzeby aby wprowadzać interakcję. Sprawdźmy ten wniosek odwołując się do analizy z wykorzystaniem statystyki logarytmu ilorazu wiarygodności (1-1.42) dla Modelu 1 i Modelu 2.

Rozdział 2-2-12-4. Testowanie braku dopasowania w Modelu 1 w porównaniu z Modelem 2.

Rozważmy hipotezę zerową (2-2-12-1.63):

$$\mathrm{H}_0\text{:}\ \delta_1 = \delta_2 = \ldots = \delta_7 = 0 \qquad (2\text{-}2\text{-}12\text{-}1.63')$$

mówiącą o nieistotności rozszerzenia Modelu 1 do Modelu 2, czyli statystycznej nieistotności interakcji.

Okazuje się, że w rozważanym przypadku test statystyczny weryfikujący hipotezę zerową (2-2-12-1.63), można by przeprowadzić zarówno wykorzystując statystykę ilorazu wiarygodności (co jest oczywiste), jak i dewiancję Modelu 1.

Istotnie, po pierwsze w Rozdziałach 2-1-4 oraz 2-1-4-1 zauważyliśmy, że obie te statystyki mają w przybliżeniu rozkład chi-kwadrat [1]. Po drugie, zauważmy, że dewiancja dla Modelu 1 otrzymana w raporcie w D**1.11.2** przyjęła w próbce wartość:

$$D\left(\hat{\beta}_{(r)}\right)^{Model1} = 8{,}195 \quad . \qquad (2\text{-}2\text{-}12\text{-}4.65)$$

Natomiast liczba stopni swobody dewiancji $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ wynosi [1]:

$d.f. =$ [*liczba zmiennych* $\left(Y_{ij}\right)$] – [*liczba parametrów w* Modelu 1] = $N - (r+1)$

$$= 16 - 9 = 7. \qquad (2\text{-}2\text{-}12\text{-}4.66)$$

Statystyka $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ ma więc w przybliżeniu rozkład chi-kwadrat z *d.f.* = 7.

Z kolei statystyka testowa ilorazu wiarygodności (2-1-4-1.24) dla hipotezy zerowej (2-2-12-1.63) jest zgodnie z (2-1-4-1.25) otrzymana przez odjęcie dewiancji dla Modelu 2 (która jest równa zero) od dewiancji dla Modelu 1, tzn.:

$$-2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right] = D\left(\hat{\beta}_{(r)}\right)^{Model1} - D\left(\hat{\beta}\right)^{Model2} = 8{,}195 - 0 = 8{,}195 \quad , \qquad (2\text{-}2\text{-}12\text{-}4.67)$$

zatem jej wartość w próbce jest równa $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ jak w (2-2-12-4.65).

Również liczba stopni swobody statystyki ilorazu wiarygodności (1-1.42), równa [1]:

$d.f. =$ [*liczba parametrów w* Modelu 2] – [*liczba parametrów w* Modelu 1]

$$= 16 - 9 = 7 \ , \qquad (2\text{-}2\text{-}12\text{-}4.68)$$

wynosi tyle ile *d.f.* dewiancji Modelu 1, więc i ona ma w przybliżeniu rozkład chi-kwadrat z *d.f.* = 7.

Zbierzmy informacje zawarte we wzorach od (2-2-12-4.65) do (2-2-12-4.68). Wynika z nich, że skoro zarówno rozkład, jak i wartość liczbowa oraz liczba stopni swobody dewiancji Modelu 1, (2-2-12-4.65), oraz log ilorazu wiarygodności, (2-2-12-4.67), są takie same, zatem równoważnie można weryfikować hipotezę zerową (2-2-12-1.63) korzystając ze statystyki (2-2-12-4.67) bądź (2-2-12-4.65).

Przyjmijmy więc, w tym przypadku, dewiancję $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ dla Modelu 1 jako statystykę testową hipotezy (2-2-12-1.63). Korzystając z (2-2-12-4.65) oraz (2-2-12-4.66) otrzymujemy, wykonując pomocnicze rachunki na przykład w arkuszu kalkulacyjnym Excel, że empiryczny poziom istotności wynosi:

$$p = \Pr\left(\chi_7^2 \geq D\left(\hat{\beta}_{(r)}\right)^{Model1} = 8{,}195\right) = 0.3157 \ . \qquad (2\text{-}2\text{-}12\text{-}4.69)$$

Rozdział 2-2-12-4-1. Wniosek dla analizy interakcji zmiennych „obszar” i „wiek”.

Zatem na żadnym poziomie istotności α mniejszym od jak widać dość dużego p = 0.3157, np. na poziomie $\alpha = 0{,}1$, nie mamy podstaw do odrzucenia hipotezy zerowej o statystycznej nieistotności rozszerzenia Modelu 1 do Modelu 2. Uznajemy więc, że *w* Modelu 1 *nie ma statystycznie istotnego braku dopasowania do danych empirycznych w porównaniu z* Modelem 2. Ponieważ Model 2 oraz model podstawowy dopasowują się do danych pomiarowych tak samo dobrze, zatem widzimy, że w Modelu 1 nie ma istotnego odchylenia obserwowanych wartości Y_{ij} od wartości przewidywanych $\hat{Y}_{ij}$ tym modelem.

Pozostawiamy więc prostszy Model 1 jako wystarczający do *przewidywania oczekiwanej ilości przypadków awarii* silnika, stwierdzając, że dodanie członów interakcji $M\,U_k$ do Modelu 1 skomplikowałoby niepotrzebnie model, nie poprawiając w sposób statystycznie istotny dopasowania do danych empirycznych.

Rozdział 2-2-12-5. Analiza „wieku” jako zaburzenia czynnika głównego.

Rozważenie Punktu 2 w Rozdziale 2-2-12 polega na szukaniu odpowiedzi na pytanie o to, czy „wiek” jest kowariantem zaburzającym główny wpływ czynnika jakim jest „obszar”. Odpowiedz tą otrzymuje się wraz ze zbadaniem czy ryzyko względne $\hat{r}_{Wi} = e^{\hat{\beta}}$ albo równoważnie $\hat{\beta}$ zmienia się znacząco, jeśli zignorujemy zmienną „wiek”. Nie wprowadzenie „wieku” do analizy w Modelu 1 pozostawia tą zmienną poza kontrolą [1].

Rozdział 2-2-12-5-1. Znacząca różnica ekspercka.

Aby przeprowadzić potrzebną analizę należy więc pominąć wyrażenia dla „wieku”, tzn. składnik $\sum_{k=1}^{7} \alpha_k U_k$ z Modelu 1 i zobaczyć, czy otrzymane oszacowanie współczynnika przy *M* różnić się będzie **znacząco** od wartości $\hat{\beta} = 0{,}8043$, (2-2-11-3.52), albo lepiej czy oszacowanie względnego ryzyka (bo to ono ostatecznie interesuje badacza) różni się znacząco od wartości $\hat{r}_{Wi}^{Model1} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513$. *Termin „znacząca różnica” nie odnosi się do testów statystycznych, ale do wiedzy ekspertów w dziedzinie*.

Rozdział 2-2-12-5-2. Analiza SAS dla Modelu 3.

Aby odpowiedzieć na pytanie o ile zmieni oszacowanie współczynnika β przy M, musimy dopasowywać do danych następujący model:

Model 3: $\ln r_{ij} = \alpha + \beta M$, i =1, 2,..., 8, j = 0, 1 (2-2-12-5-2.70)

Zadanie dla Modelu 3. Napisać program korzystający z procedury GENMOD dla Modelu 3, a następnie wykorzystując dane podane w Punkcie 1.11.1 uruchomić go, otrzymując poniższy raport SAS.

Rozdział 2-2-12-5-3. Raport SAS dla Modelu 3.

```
                                   System SAS
                              The GENMOD Procedure

                              Informacje o modelu

                         Zbiór                      WORK.MODEL3
                         Rozkład                        Poisson
                         Funkcja wiążąca                    Log
                         Zmienna zależna                      Y
                         Zmienna przesunięcia                ln

                        Liczba obserwacji wczytanych          17
                        Liczba obserwacji użytych             16
                        Braki danych                           1

                          Informacje o poziomie klasyfikacji

                         Klasa     Poziomy    Wartości

                         A               8    1 2 3 4 5 6 7 8

                               Informacje o parametrach

                               Parametr        Efekt

                               Prm1            Intercept
                               Prm2            M

                                   Kryteria oceny zgodności

                                       St.                     Wartość/st.
                   Kryterium           sw.         Wartość            sw.

                   Dewiancj             14       2569.7700       183.5550
                   Skalowana dewia      14       2569.7700       183.5550
                   Chi-kwadrat Pearso   14       3012.0987       215.1499
                   Scaled Pearson X2    14       3012.0987       215.1499
                   Log. wiarogodn                5921.0760

          Algorytm osiągnął zbieżność.

                                       System SAS
                                  The GENMOD Procedure

                                  Analiza ocen parametrów

                                                   95% granice
              St.                       Błąd     przedziału ufności        Chi-
Parametr      sw.       Ocena    standardowy            Walda           kwadrat    Pr > chi kw..

Intercept      1     -7.1273         0.0437     -7.2130     -7.0416    26567.6          <.0001
M              1      0.7431         0.0521      0.6410      0.8453     203.23          <.0001
Skala          0      1.0000         0.0000      1.0000      1.0000

UWAGA: The scale parameter was held fixed.
```

Rozdział 2-2-12-5-4. Analiza raportu SAS dla Modelu 3.

Z powyższego raportu odczytujemy, że oszacowanie parametru β wynosi $\hat{\beta} = 0{,}7431$, skąd „surowe" oszacowanie (z powodu braku w analizie zmiennej „wiek") względnego ryzyka, wynosi:

$$\hat{r}_W^{Model3} = e^{\hat{\beta}} = e^{0{,}7431} = 2{,}1024 \ . \qquad (2\text{-}2\text{-}12\text{-}5\text{-}4.71)$$

Uwaga: Podkreślmy raz jeszcze, że w przeciwieństwie do różnicy istotnej statystycznie, wypowiedź o znaczącej różnicy, nie jest poparta żadnym statystycznym testem i nie należy testów takich wykonywać. O tym, czy różnica jest znacząca wypowiadają się specjaliści w branży.

Wniosek dotyczący zaburzenia: Porównując wartości Modelu 1 oraz Modelu 3 dla $\hat{\beta}$, które wynoszą odpowiednio 0,8043 oraz 0,7431 lub lepiej dla względnego ryzyka $\hat{r}_W = e^{\hat{\beta}}$, które wynoszą odpowiednio 2,2351 oraz 2,1024, uznajmy (chociaż nie jesteśmy specjalistami z branży samochodowej), że różnią się one znacząco i *zmienną poboczną „wiek" eksploatacji samochodu należy wprowadzić do modelu jako zaburzenie głównego wpływu zmiennej „obszar" eksploatacji samochodu.*

Rozdział 2-2-12-5-5. Analiza rozszerzenia Modelu 3 do wyższego w hierarchii Modelu 1.

Z porównania raportów dla Modelu 3 oraz Modelu 1 widać, że różnica dewiancji tych modeli wynosi: 2569,77 – 8,195 = `2561,58`. Różnica dewiancji tych modeli (2-1-4-1.25), tzn. log ilorazu funkcji wiarygodności, ma w przybliżeniu rozkład chi-kwadrat. Przy różnicy 14-7=7 stopni swobody dewiancji tych modeli, wartość `2561,58` jest wysoce istotna statystycznie, wskazując na istotny brak dopasowania Modelu 3 w stosunku do Modelu 1.

Zadanie: Sformułować postać hipotezy zerowej mówiącej o nie występowaniu braku dopasowania do danych pomiarowych w Modelu 3 w porównaniu z Modelem 1. Wyznaczyć empiryczny poziom istotności dla przeprowadzanego testu tej hipotezy.

Rozdział 2-2-13. Analiza regresji Poissona w SAS dla modelu z przesunięciem.

Dla skompletowania analizy dla wszystkich modeli ze zbioru modeli hierarchicznych rozważymy jeszcze model tylko z wyrazem wolnym, czyli taki w którym występuje brak zależności modelowej od zmiennych objaśniających. Model ten ma postać:

Model 0: $\ln r_{ij} = \alpha$, i =1, 2,..., 8; j = 0, 1 . (2-2-13.71)

Rozdział 2-2-13-1. Dane i program SAS dla Modelu 0.

Aby przeprowadzić analizę z użyciem procedury GENMOD została w danych podanych w Rozdziale 2-2-11-1 wprowadzona dodatkowa zmienna O, przyjmująca zawsze wartość zero.

Ponieważ w Modelu 0 występuje brak zależności modelowej od zmiennych objaśniających, w związku z tym modyfikujemy następująco wiersz *model* poleceń w procedurze GENMOD:

model Y = O / covb

lub

model Y = O / pred covb

Rozdział 2-2-13-2. Raport SAS dla Modelu 0.

Po wczytaniu danych zawartych w Rozdziale 2-2-11-1 oraz uruchomieniu programu procedury GENMOD, otrzymujemy poniższy raport.

```
                    System SAS
              The GENMOD Procedure

              Informacje o modelu

 Zbiór                        WORK.MODEL0
 Rozkład                          Poisson
 Funkcja wiążąca                      Log
 Zmienna zależna                        Y
 Zmienna przesunięcia                  ln

Liczba obserwacji wczytanych           16
Liczba obserwacji użytych              16

  Informacje o poziomie klasyfikacji

 Klasa     Poziomy    Wartości

 A               8    1 2 3 4 5 6 7 8

       Informacje o parametrach

       Parametr        Efekt

       Prm1            Intercept
       Prm2            O

        Kryteria oceny zgodności

                 St.                     Wartość/st.
Kryterium        sw.         Wartość            sw.

Dewiancja         15       2790.3403        186.0227
Skalowana dewia   15       2790.3403        186.0227
Chi-kwadrat Pearso 15      3480.1347        232.0090
Scaled Pearson X2 15       3480.1347        232.0090
Log. wiarogodn             5810.7909

Algorytm osiągnął zbieżność.
```

Analiza ocen parametrów

Parametr	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
Intercept	1	-6.6669	0.0238	-6.7135	-6.6202	78449.1	<.0001
O	0	0.0000	0.0000	0.0000	0.0000	.	.
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed.

Rozdział 2-2-13-3. Wynik analizy dla Modelu 0.

Dewiancja dla Modelu 0, $\ln r_{ij} = \alpha$, wynosi 2790,3403. Jak można się było spodziewać, model posiadający tylko przesunięcie i bez zależności od zmiennych objaśniających wykazuje istotny brak dopasowania w stosunku do Modelu 1, $\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M$, co przejawia się gwałtownym wzrostem dewiancji Modelu 0, (2-2-13.71), w stosunku do dewiancji Modelu 1, (2-2-10.47).

Zadanie: Sformułować postać hipotezy zerowej mówiącej o nie występowaniu braku dopasowania do danych pomiarowych w Modelu 0 w porównaniu z Modelem 1 . Wyznaczyć empiryczny poziom istotności dla przeprowadzanego testu tej hipotezy sprawdzając powyższy wynik analizy dla Modelu 0.

Zadanie: Pokazać, że różnica dewiancji Modelu 0, (2-2-13.71), oraz Modelu 3, (2-2-12-5-2.70), jest również statystycznie istotna, znajdując wartość odpowiedniego empirycznego poziomu istotności.

Rozdział 2-2-14. Podsumowanie analizy regresji doboru modelu Poissona.

Poniższa Tabela 2-2-14.2 podsumowuje przeprowadzoną analizę regresji Poissona dla przykładu zależności liczby awarii silnika w klasie modeli hierarchicznych z uwzględnieniem „obszaru” jako czynnika głównego wpływu, a zmiennej „wiek” jako zmiennej pobocznej.

Tabela 2-2-14.2. Tabela ANOVA dla przykładu awarii silnika ($N = 16$).

	Model dla $\ln r_{ij}$	Liczba parametrów	$D(\beta)$	*d.f.*	Istotna statystycznie różnica w $D(\beta)$
Model 0	α	1	2790,34	15	Istotna $p \approx 0$
Model 3	$\alpha + \beta M$	2	2569,77	14	Istotna $p \approx 0$
Model 1	$\alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M$	9	8,2	7	Nieistotna $p = 0{,}32$
Model 2	$\alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M + + \sum_{k=1}^{7} \delta_k \; MU_k$	16	0	0	
Model podstawowy	μ_j	16	0	0	

Rozdział 2-2-14-1. Wniosek z analizy.

Z przeprowadzonej analizy widać, że dane zawierają wskazanie, że spośród rozważanego zbioru modeli hierarchicznych należałoby wybrać Model 1 jako ten, który nie ma statystycznie istotnego braku dopasowania do danych pomiarowych, a jednocześnie ma prostszą strukturę (9 parametrów) niż model podstawowy lub Model 2 z interakcją (16 parametrów).

Uwaga: Wykroczenie poza klasę modeli hierarchicznych i potraktowanie „wieku” jako zmiennej typu ciągłego mogłoby doprowadzić do wyselekcjonowania modelu z mniejszą liczbą parametrów niż Model 1 [1].

Rozdział 2-2-15. Uzupełnienie.

Rozdział 2-2-15-1. Polecenia języka 4GL procedury GENMOD dla rozważanego przykładu.

Poniżej podane zostały podstawowe komendy programów napisanych w języku 4GL dla celów przeprowadzenia analizy regresji Poissona, w tym rozważanego powyżej przykładu.

data 'awaria' wskazuje nazwę zbioru z danymi;
input wskazuje zmienne, które mają być wczytane do modelu;
ln wskazuje zewnętrzną zmienną funkcyjną (tutaj logarytm);
datalines wskazuje, że poniżej będą się znajdowały linie danych;
run wskazuje na koniec linii danych;

proc oznacza początek odpowiedniej procedury („genmod" dla regresji Poissona);
model wskazuje zmienne użyte w modelu;
pred wskazuje na konieczność wyliczenia wartości prognozowanych;
ref wskazuje referencyjną populację (tzn. linię, w której wszystkie zmienne kierunkowe dla przyjętego systemu kodowania oraz ich interakcje mają wartość 0);
covb wskazuje na wyliczenie macierzy kowariancji estymatorów;
corrb wskazuje na wyliczenie macierzy korelacyjnej estymatorów;
dist informuje o użyciu określonego rozkładu;
link informuje o użyciu wskazanej funkcji linku (logarytmicznej dla regresji Poissona);
offset wskazuje zmienną, znajdującą się poza modelem, w której przechowywana jest funkcja linkująca;
run informuje o uruchomieniu procedury liczącej;
quit powoduje wyjście z programu i wyświetlenie wydruku.

Rozdział 2-2-15-2. Opis zmiennych występujących w zbiorze danych w Rozdziale 2-2-11-1.

Zmienna A jest zmienną jakościową z wariantem wieku serwisowanych samochodów;
Y oznacza zmienną objaśnianą ilości występujących przypadków (zmienna o rozkładzie Poissona);
N oznacza liczebność badanych populacji;
M jest zmienną kierunkową wskazującą na obszar;
U1, U2, U3, U4, U5, U6, U7 są zmiennymi kierunkowymi, wskazującymi na odpowiednią przynależność do klasy wiekowej;
U1M, U2M, U3M, U4M, U5M, U6M, U7M to interakcje zmiennych kierunkowych „wiek" U1, U2, U3, U4, U5, U6, U7 oraz „obszar" M;
O jest zmienną sztucznie wprowadzoną dla celu analizy Modelu 0, która nie jest zmienną objaśniającą.

A. Rozdział 3. Podsumowanie zastosowania MNW w analizie regresji Poissona.

Przedmiotem Rozdziału 2 było przećwiczenie zastosowania metody największej wiarygodności (MNW) w problemach estymacyjnych analizy regresji Poissona. Rozważania zostały poparte przykładami przeliczonymi z wykorzystaniem systemu analiz statystycznych SAS.
Omówiono sposób konstrukcji funkcji wiarygodności wykorzystywany dla celów budowy estymatorów parametrów modelu oraz wynikające z tej metody procedury wnioskowania statystycznego. Procedury dla testowania hipotez i konstruowania przedziałów ufności wykorzystują nie tylko zmaksymalizowane wartości funkcji wiarygodności, ale również oszacowane macierze kowariancji wyznaczane w ramach szerzej rozumianej metody największej wiarygodności odwołującej się do tzw. informacji Fishera zawartej w próbie.
Teoretyczne podstawy MNW wraz ze znaczeniem informacji Fishera dla (estymacji) macierzy kowariancji estymatorów parametrów modelu znajdują się w zacytowanej literaturze.

W omówionych przykładach zmienna losowa objaśniana zawsze była liczbą zliczeń przypadków interesującego nas zdarzenia. Dlatego przy spełnieniu warunku małej liczby defektów w stosunku do wszystkich obserwacji w rozważanych podgrupach próbek pobranych z dwóch porównywanych populacji, wykorzystywana postać funkcji wiarygodności odwoływała się do zmiennej mającej rozkład Poissona. W praktyce, dla typowego modelu regresji Poissona naturalną miarą estymowanego efektu jest tempo awarii (tzn. ryzyko) oraz ryzyko względne, związane z określonym, interesującym nas czynnikiem, którego warianty kontrastują badane populacje.

Przedstawiono metodę selekcji modelu z wykorzystaniem statystyki ilorazu wiarygodności oraz zastosowanie statystyki dewiancji, która jest rodzajem statystyki ilorazu wiarygodności, opisującej dobroć dopasowania badanego modelu względem modelu podstawowego. Ponieważ różnica statystyk dewiancji, otrzymana dla dwóch porównywanych modeli, jest równa statystce logarytmu ilorazu funkcji wiarygodności dla tych modeli, więc testy hipotez o braku dopasowania w modelach niższych w hierarchii, mogą być przeprowadzony z wykorzystaniem różnicy statystyk dewiancji, które pojawiają się raportach SAS.
Zastosowanie MNW w analizie regresji Poissona ma kluczowe znaczenie ze względu na możliwość selekcji modelu, który nie tylko ma estymatory posiadające (asymptotycznie) optymalne własności [36], ale jak na to zwrócono uwagę w analizowanych przykładach, nie wykazuje również statystycznie istotnie gorszego dopasowania do danych empirycznych niż model podstawowy, posiadając przy tym najmniejszą możliwą liczbę parametrów.

Typowy model regresji Poissona, użyty w przykładach, wyraża w postaci logarytmicznej tempo porażki jako liniową funkcję zbioru czynników. Niemniej, estymacja MNW jest szczególnie przydatna w estymacji współczynników regresji w modelach nieliniowych, takich jak model regresji logistycznej czy nieliniowy model regresji Poissona. Ponieważ układ równań wiarygodności nie prowadzi wtedy do liniowych równań algebraicznych na estymatory tych parametrów, dlatego procedury estymacji dla takich modeli wymagają

programu komputerowego stosującego algorytmy z wielokrotnymi iteracjami estymatorów parametrów modelu. Taki pakiet numerycznych procedur komputerowych jest zawarty w systemie SAS.

Podstawową procedurą SAS stosowaną w analizie regresji Poissona w sytuacji, gdy logarytm ryzyka jest liniową kombinacją czynników, jest procedura GENMOD. W bardziej skomplikowanych nieliniowych modelach regresji Poissona, gdy logarytmu ryzyka nie da się przedstawić w postaci liniowej kombinacji czynników, właściwą procedurą, którą można wykorzystać jest procedura NLMIXED [1].

A. Rozdział 4. Analiza doboru modelu w regresji logistycznej.

Obecny rozdział dotyczy zastosowania regresji logistycznej. Posiada ona wyjątkowo szerokie zastosowania w opisie danych empirycznych w przypadku, gdy zmienna objaśniana (odpowiedź) ma charakter dychotomiczny, tzn. gdy odpowiedź może przyjmować jedynie dwa warianty klasowe, poziomy numerowane wartością 0 lub 1. (Uogólnieniem regresji logistycznej jest wielomianowa regresja logistyczna, w której zmienna objaśniana może przyjąć więcej niż dwa warianty.) W tym aspekcie różni się ona od poprzednio omawianych modeli regresji i ANOVA, w których odpowiedź była zmienną ilościową.

Sednem analizy logistycznej jest określenie zmiany prawdopodobieństwa (szansy) pojawienia się sukcesu, na skutek wpływu określonego czynnika. Pozwala ona określić tak siłę, jak i kierunek zależności pomiędzy czynnikiem jakościowym (typu klasowego) lub ilościowym (typu dyskretnego lub ciągłego), a dychotomiczną zmienną objaśnianą Y. Można jej użyć w celu opisania wpływu na dychotomiczną zmienną Y kilku czynników X_1, X_2, ... ,X_k. Np. w badaniach ekonomicznych można dowiedzieć się jaka jest *szansa* spłacenia kredytu przez przedsiębiorstwo działające w określonej branży i przy określonych warunkach (w odniesieniu do osób działających przy innych warunkach, gdzie warunek jest wariantem tego samego czynnika). Również w badaniach medycznych, gdzie podstawowym problemem jest zdrowie pacjenta, odpowiedź w przeprowadzanych badaniach może przyjmować wartości 0 (zdrowy) lub 1 (chory).

Do oszacowania parametrów modelu regresji (w postaci logit'owej) wykorzystuje się bezwarunkową estymację metodą największej wiarygodności. Kolejne rozdziały zostały poświęcone ogólnemu sformułowaniu założeń i metody regresji logistycznej oraz zastosowaniu metody największej wiarygodności do estymacji parametrów modelu logistycznego. Przedstawiono również przykład analizy logistycznej, wykorzystując procedury programu SAS.

Rozdział 4-1. Wprowadzenie teoretyczne.

Rozdział 4-1-1. Zmienne dychotomiczne.

W modelu regresji logistycznej podstawowa zmienna dychotomiczna Y, przyjmuje wartość 1 z prawdopodobieństwem θ oraz 0 z prawdopodobieństwem 1- θ. Zmienną dychotomiczną określa się również mianem zmiennej wskazującej, binarnej czy zero-jedynkowej (0-1). Danej obserwacji zmiennej Y przypisujemy wartość 1 wtedy, gdy interesujący nas poziom zjawiska A został osiągnięty, zjawisko to się wydarzyło lub zaistniał interesujący nas stan, oraz wartość 0, gdy zaszło zdarzenie przeciwne. Zmienna wskazująca pokazuje więc czy pewien warunek (np. w jednym, konkretnym losowaniu do próby, w jednej obserwacji) został spełniony, czy nie został spełniony. Zatem, jako zmienna binarna, która wskazuje poziom A, zmienna Y ma postać:

$$Y = \begin{cases} 1 \text{ , gdy osiągnięto poziom A} \\ 0 \text{ , gdy nie osiągnięto poziomu A} \end{cases} \qquad (4\text{-}1\text{-}1.1)$$

Rozkład prawdopodobieństwa zmiennej dychotomicznej jest więc następujący:

$$p(Y \mid \theta) \equiv p(Y) = \theta^{Y}(1-\theta)^{1-Y}, \quad Y = 0,1 \ , \tag{4-1-1.2}$$

skąd łatwo sprawdzić, że jej wartość oczekiwana wynosi:

$$E(Y) = \sum_{y=0}^{1} y\, p(Y = y \mid \theta) = p(Y = 1 \mid \theta) = \theta \ . \tag{4-1-1.3}$$

W N-wymiarowej próbie (Y_1, Y_2, ..., Y_N), każda ze zmiennych dychotomicznych Y_i (i = 1,2,...,N) jest zmienną losową o rozkładzie zero-jedynkowym:

$$p_i(Y_i \mid \theta_i) = \theta_i^{Y_i}(1-\theta_i)^{1-Y_i}, \quad Y_i = 0,1, \qquad i = 1,2,...,N \ . \tag{4-1-1.4}$$

Chociaż zakładamy, że zmienne dychotomiczne (Y_1, Y_2, ... ,Y_N) są parami niezależne, jednak ponieważ θ_i mogą być różne dla różnych i, więc próba nie musi być prosta. Parametr θ_i jest prawdopodobieństwem, np. że i-ta osoba w losowej próbce N osób wylosowanych z populacji, nie spłaci pożyczki w ustalonym, od pewnej chwili, okresie czasu.

Rozkład zero-jedynkowy (0-1) jest również nazywany *punktowym rozkładem dwumianowym*, co wynika z tego, że dla N = 1 jest on szczególnym przypadkiem rozkładu dwumianowego:

$$\binom{N}{Y} \theta^{Y}(1-\theta)^{N-Y} \tag{4-1-1.5}$$

zmiennej losowej Y (zliczającej np. liczbę sukcesów w N losowaniach, Y=0,1,2,...,N, gdzie w każdym z tych losowań, prawdopodobieństwo sukcesu θ jest takie samo).

Rozdział 4-1-2. Metoda największej wiarygodności w regresji logistycznej.

Funkcja wiarygodności próby $\tilde{Y} \equiv (Y_1, Y_2, \ldots, Y_N)$ dla zmiennych posiadających punktowe rozkłady jak w (4-1-1.3), ma postać:

$$P(\tilde{Y} \mid \Theta) = \prod_{i=1}^{N} p_i(Y_i \mid \theta_i) = \prod_{i=1}^{N} [\theta_i^{Y_i}(1-\theta_i)^{1-Y_i}] \tag{4-1-2.6}$$

gdzie $\Theta = (\theta_1, \theta_2, \ldots, \theta_N)^T$.

Ponieważ nie interesuje nas kolejność zajścia zdarzenia w ciągu N losowań, zatem bez straty ogólności rozważań załóżmy, że N_1 pierwszych jednostek z N elementowej losowej próby, nabyło rozważaną własność (np. nie spłaciło pożyczki w umówionym terminie), tak że $Y_1 = Y_2 = \ldots = Y_{N_1} = 1$. Oznacza to że ($N_2$ = N - N_1) pozostałych jednostek próby nie nabyło tej własności, czyli $Y_{N_1+1} = Y_{N_1+2} = \ldots = Y_N = 0$. Postać funkcji wiarygodności jest zatem następująca:

$$P\left(\tilde{Y} \mid \Theta\right) = \left(\prod_{i=1}^{N_1} \theta_i\right)\left(\prod_{i=N_1+1}^{N} (1-\theta_i)\right) \qquad (4\text{-}1\text{-}2.7)$$

Ponieważ prawdopodobieństwo θ_i (i = 1,2,...,N) jest również warunkową wartością oczekiwaną zmiennej dychotomicznej Y_i, tzn.:

$$\mu_i \equiv E(Y_i) = p(Y_i = 1 \mid \theta_i) = \theta_i\,, \quad i = 1,2,...,N, \qquad (4\text{-}1\text{-}2.8)$$

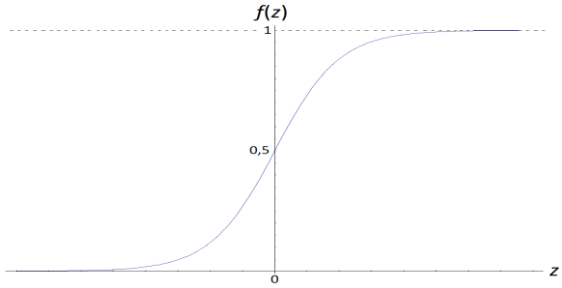
zatem funkcję wiarygodności (4-1-2.7) można wyrazić poprzez współczynniki regresji β_j modelu logistycznego. Logistyczna analiza regresji jest jednym ze sposobów modelowania zależności korelacyjnej, zmiennej zależnej Y dychotomicznej w jej związku z czynnikami X_1, X_2, ... , X_k, które mogą być tak ilościowe jak i jakościowe. Niech $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$ oznacza zbiór wartości k czynników X_j (j = 1,2,...,k) dla określonej i–tej jednostki ($i = 1,2,...,N$) w N-wymiarowej próbie $\tilde{Y}$. W modelu logistycznym warunkowe wartości oczekiwane zmiennej dychotomicznej Y są związane z czynnikami X_j następującą transformacją:

$$E(Y_i) = \theta_i = \frac{1}{1+\exp(-\lambda_i^*)} = \frac{1}{1+\exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\right)\right]}\,, \quad i = 1,2,...,N, \qquad (4\text{-}1\text{-}2.9)$$

gdzie $\lambda_i^* \equiv \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}$, (2-1-3.11), a β_j, (j=0,1,2,...,k) są nieznanymi współczynnikami regresji logistycznej, które trzeba estymować. Wyrażenie stojące po prawej stronie (4-1-2.9) ma postać:

$$f(z) = \frac{1}{1+e^{-z}}\,. \qquad (4\text{-}1\text{-}2.10)$$

Funkcja *f(z)* jest nazywana *funkcją logistyczną.* Nadaje się ona do modelowania zmiany prawdopodobieństwa wraz ze zmianą wartości z, co wynika z tego, że dla $z \in (-\infty; +\infty)$, *f(z)* zmienia się zgodnie z jej sigmoidalnym kształtem w granicach od 0 do 1. W badaniach ekonomicznych można ją wykorzystać np. do opisu (indywidualnego) ryzyka niespłacenia długu przez dłużnika. Sigmoidalny kształt wykresu funkcji logistycznej *f(z)* jest wykorzystywany przez epidemiologów, szczególnie wtedy gdy zmienna z reprezentuje indeks, który podsumowuje działanie kilku czynników ryzyka (X_1, X_2, ... , X_k), tak, że $f(z)$ opisuje ryzyko dla danej wartości z.



Rysunek. 4-1-2.1 Wykres funkcji logistycznej.

Funkcja linku (wiążacą), czyli funkcja $g(\mu_i)$ wiążaca wartość oczekiwaną warunkową $\mu_i \equiv E(Y_i) = \theta_i$ z kombinacją liniową λ_i^* zmiennych objaśniających, ma więc dla regresji logistycznej $\lambda_i^* = g(\mu_i) = \ln(\mu_i / (1 - \mu_i))$ (tzw. postać „Logit").

Wstawiając θ_i z (4-1-2.9) do funkcji wiarygodności (4-1-2.7), otrzymujemy *bezwarunkową funkcję wiarygodności* wykorzystywaną w standardowej analizie logistycznej:

$$P(\tilde{Y} \mid \beta) = \left(\prod_{i=1}^{N_1} \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right]} \right) \times \prod_{i=N_1+1}^{N} \left(\frac{\exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right]}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right]} \right)$$

$$= \frac{\prod_{i=N_1+1}^{N} \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right]}{\prod_{i=1}^{N} \left[1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right] \right]} \quad , \qquad (4\text{-}1\text{-}2.11)$$

gdzie $\beta \equiv (\beta_1, \beta_2, ..., \beta_k)$, co można zapisać następująco:

$$P(\tilde{Y} \mid \beta) = \frac{\prod_{i=1}^{N_1} \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right)}{\prod_{i=1}^{N} \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right]} \quad . \qquad (4\text{-}1\text{-}2.12)$$

W ten sposób otrzymaliśmy ostateczną postać (bezwarunkowej) funkcji wiarygodności dla modelu logistycznego. Jej logarytm ma postać:

$$\ln P(\tilde{Y} \mid \beta) = \ln \left(\frac{\prod_{i=1}^{N_1} \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right)}{\prod_{i=1}^{N} \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right]} \right) = \qquad (4\text{-}1\text{-}2.13)$$

$$= \sum_{i=1}^{N_1} \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) - \sum_{i=1}^{N} \ln \left[1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \right) \right] \quad .$$

Ze względu na konieczność oszacowania k+1 niezależnych parametrów β_j, j=0,1,2,...,k, odpowiedni układ równań wiarygodności, którego rozwiązanie daje oszacowania MNW parametrów modelu, ma postać:

$$\left. \frac{\partial \ln P(\tilde{Y} \mid \beta)}{\partial \beta_j} \right|_{\beta = \hat{\beta}} = 0 \ , \qquad j = 0,1,2,..., k \ . \qquad (4\text{-}1\text{-}2.14)$$

Ponieważ logarytm funkcji wiarygodności (4-1-2.13) jest skomplikowaną nieliniową funkcją parametrów $\beta \equiv (\beta_1, \beta_2, ..., \beta_k)$, zatem jego maksymalizacja (mająca na celu otrzymanie estymatorów $\hat{\beta}$ MNW dla

parametrów β) oznacza na ogół konieczność wykorzystywania właściwie zaimplementowanych komputerowo algorytmów. Analiza regresji logistycznej zawarta w systemie SAS dysponuje odpowiednimi procedurami (LOGISTIC procedure), dając po rozwiązaniu układu równań (4-1-2.14) maksymalizowaną wartość $P(\tilde{Y} \mid \hat{\beta})$ funkcji wiarygodności oraz macierz kowariancji $\hat{\mathbf{V}}(\hat{\beta})$ estymatorów $\hat{\beta}$ rozważanego modelu. Macierz $\hat{\mathbf{V}}(\hat{\beta})$ jest konieczna do przeprowadzenia właściwego statystycznego wnioskowania.
W dalszej części rozdziału, zostanie omówiony przykład zastosowania procedur SAS'a, pozwalających na wyznaczenie zarówno wartości estymatorów MNW, jak i macierzy kowariancji $\hat{\mathbf{V}}(\hat{\beta})$ w regresji logistycznej.

Rozdział 4-1-3. Modelowanie ilorazu szans.

Współczynniki regresji β_j odgrywają ważną rolę w dostarczaniu informacji na temat związku pomiędzy czynnikami X_j ($j=1,...,k$) a odpowiedzią Y. W modelu regresji logistycznej ilościowe określenie tego związku przedstawia „*iloraz szans*" („*odds ratio*").
Iloraz szans rozumiany jest jako częstość wystąpienia danego zdarzenia (np. nie spłacenia długu czy zachorowania) u grupy badanej charakteryzujących się daną własnością (np. kolejno, grupy osób mieszkających w miastach czy osób palących) w stosunku do grupy porównawczej, nie posiadającej danej własności (np. kolejno, grupy osób mieszkających na wsiach czy osób niepalących). *Iloraz szans* jest miernikiem przedstawiającym możliwości poznawcze w ocenie modeli służących do prognozowania zagrożeń, będąc *miarą wpływu* wartości rozważanych czynników na zmienną Y [1].

Zdefiniujmy „szansę" ($odds(G)$) jako stosunek prawdopodobieństwa $pr(G)$ wystąpienia zdarzenia G do $pr(\overline{G}) \equiv pr(nieG) = 1 - pr(G)$, czyli prawdopodobieństwa nie wystąpienia zdarzenia G:

$$odds(G) = \frac{pr(G)}{pr(\overline{G})} = \frac{pr(G)}{1 - pr(G)} \qquad (4\text{-}1\text{-}3.15)$$

Iloraz ten może być interpretowany jako stosunek prawdopodobieństwa wystąpienia zdarzenia do jego nie wystąpienia. Przykładowo, gdy $pr(G)$ = 1/3, wtedy $odds(G) = \frac{1/3}{1-1/3} = \frac{1}{2}$ i mamy szansę 2 do 1, że zdarzenie *nie* zajdzie.

Każdy *iloraz szans* („Odds Ratio", skąd oznaczenie *OR*) jest z definicji *proporcją dwóch szans*, tzn.:

$$OR_{A vs. B} = \frac{odds(G_A)}{odds(G_B)} = \frac{\dfrac{pr(G_A)}{1 - pr(G_A)}}{\dfrac{pr(G_B)}{1 - pr(G_B)}} \qquad (4\text{-}1\text{-}3.16)$$

gdzie indeksy A i B odpowiadają dwóm grupom jednostek, które mają być porównane.

Przykład. Przypuścimy, że A=M oznacza grupę osób mieszkających w (dużych) miastach (*m*), a B=W oznacza grupę osób mieszkających na wsiach (łącznie z małymi miastami) (*w*). Niech G_M opisuje zdarzenie, że osoba mieszkająca w grupie osób mieszkających w miastach nie spłaci długu (kredytu) i odpowiednio, zdarzenie G_W odpowiada grupie osób mieszkających na wsiach, którzy nie spłacili długu. Jeśli $pr(G_M)=1/3$, a $pr(G_W)=1/5$, to iloraz szans porównujący szanse niespłacenia długu przez osoby w miastach w stosunku do osób mieszkających na wsiach, przedstawia się następująco:

$$OR_{M\ vs.W} = \frac{\frac{1/3}{1-1/3}}{\frac{1/5}{1-1/5}} = \frac{\frac{1}{2}}{\frac{1}{4}} = 2 \qquad (4\text{-}1\text{-}3.17)$$

Zatem dla tego (przykładowego) układu danych, szansa niespłacenia długu przez osoby mieszkające w miastach jest dwukrotnie większa niż szansa niespłacenia długu przez osoby mieszkające na wsiach. Iloraz szans równy 1, oznaczałby *równoważność ryzyka* dla porównywanych grup. Taka wartość ilorazu szans mówiłaby, że miejsce zamieszkania nie ma wpływu na spłacenie długu. Wystąpienie wartości ilorazu szans mniejszej niż 1, wskazywałoby na wystąpienie większej szansy niespłacenia długu przez osoby z grupy B=W, niż w grupie A=M.

Istnieje równoważny sposób zapisu regresji logistycznej w tzw. postaci *logitowej* („*logit form*"). Funkcja *logit* z $pr(Y=1)$ jest transformacją zdefiniowaną jako *logarytm naturalny szansy zajścia zdarzenia G* $\equiv\{Y=1\}$:

$$\text{logit}[pr(Y=1)] = \ln[odds(Y=1)] = \ln\left[\frac{pr(Y=1)}{1-pr(Y=1)}\right] \qquad (4\text{-}1\text{-}3.18)$$

lub

$$odds(Y=1) = \frac{pr(Y=1)}{1-pr(Y=1)} = e^{\text{logit}[pr(Y=1)]} \qquad (4\text{-}1\text{-}3.19)$$

Po wstawieniu (4-1-2.8)-(4-1-2.9) do (4-1-3.18) otrzymujemy tzw. *logit'ową postać modelu logistycznego*:

$$\text{logit}[pr(Y=1)] = \beta_0 + \sum_{j=1}^{k} \beta_j X_j\ , \qquad (4\text{-}1\text{-}3.20)$$

w której czynniki X_j pojawiają się liniowo, a zaletą jej jest to, że podobnie jak w liniowej regresji wielorakiej $\text{logit}[pr(Y=1)] \in (-\infty; +\infty)$.

Ogólnie, przy liczeniu ilorazu szans, należałoby podać dwie grupy (*lub jednostki*), które mają zostać porównane poprzez dokonanie wyszczególnienia wartości zbioru czynników $X_1, X_2, ..., X_k$. Niech $x_A=(x_{1A}, x_{2A}, ..., x_{kA})$ i $x_B=(x_{1B}, x_{2B}, ..., x_{kB})$ oznaczają zbiór wartości tych czynników dla grup (lub jednostek), kolejno A i B.

Dzieląc szansę dla grupy (lub jednostki) A przez szansę dla grupy (lub jednostki) B, a następnie podstawiając (4-1-3.19) do (4-1-3.16), otrzymujemy *ogólną*, logit'ową postać ilorazu szans dla dwóch grup (lub jednostek) A i B:

$$OR_{Avs.B} = \frac{odds(G_A)}{odds(G_B)} = \frac{\dfrac{pr(G_A)}{1-pr(G_A)}}{\dfrac{pr(G_B)}{1-pr(G_B)}} = \frac{\exp(\text{logit}[pr(Y_A=1)])}{\exp(\text{logit}[pr(Y_B=1)])} \qquad (4\text{-}1\text{-}3.21)$$

Podstawiając do (4-1-3.21) logit'ową postać modelu logistycznego, (4-1-3.20), otrzymujemy ogólną postać iloraz szans dla dwóch grup (lub jednostek) A i B, wskazanych szczegółowo poziomami czynników $X_1, X_2, \ldots, X_k$, i wyrażoną poprzez parametry modelu logistycznego:

$$OR_{X_A vs. X_B} = \frac{odds\_dla\_X_A}{odds\_dla\_X_B} = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{jA}\right)}{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{jB}\right)} \qquad (4\text{-}1\text{-}3.22)$$

Po przekształceniu (4-1-3.22), otrzymujemy następującą, ogólną postać ilorazu szans:

$$OR_{X_A vs. X_B} = \exp\left(\sum_{j=1}^{k} \beta_j (x_{jA} - x_{jB})\right) \quad . \qquad (4\text{-}1\text{-}3.23)$$

Zauważmy, że wyraz wolny β_0 modelu logistycznego skasował się w ilorazie szans (4-1-3.23), tak że iloraz szans zależy jedynie od eksponenty sumy parametrów β_j (j = 1,2,…,k) modelu pomnożonych przez różnice wartości czynników X_j dla grup (lub jednostek) A i B.

Ponieważ β_j są nieznanymi parametrami populacyjnymi, zatem parametr $OR_{X_A vs. X_B}$ (4-1-3.23) jest *populacyjnym* ilorazem szans. Jego estymator można otrzymać poprzez dopasowanie modelu logistycznego, wykorzystując metodę największej wiarygodności i podstawiając w (4-1-3.23) estymatory $\hat{\beta}_j$ MNW w miejsce β_j (j = 1,2,…,k). Podając wartości $\hat{\beta}_j$ w konkretnej próbce oraz konkretne wartości specyfikacji x_{jA} i x_{jB} czynników, otrzymujemy liczbową wartość tego estymatora.

Przykład 1. Załóżmy, że Y może przyjąć dwie wartości, 1-nie spłacił długu i 0-spłacił dług.

a) Niech jedynym czynnikiem X_1 (k=1) będzie *lokalizacja* (l) przyjmująca poziomy 0,1. Wskazuje on przynależność do grupy osób mieszkających w miastach $(X_1 \equiv l = 1)$ bądź na wsiach $(X_1 \equiv l = 0)$. Model (4-1-3.20) ma zatem postać:

$$\text{logit}[pr(Y=1)] = \beta_0 + \beta_1 \cdot l \quad . \qquad (4\text{-}1\text{-}3.24)$$

Aby otrzymać wyrażenie na iloraz szans w modelu logistycznym, musimy porównać szanse dla dwóch grup jednostek. W omawianym przykładzie należy rozważyć następujące logit'owe transformacje prawdopodobieństw. W grupie A=M osób mieszkających w miastach ma ona postać:

$$\text{logit}[pr(G_A)] \equiv \text{logit}[pr(Y_A=1)] = \ln odds(Y_A=1) = \beta_0 + (\beta_1 \cdot 1) = \beta_0 + \beta_1 \ , \qquad (4\text{-}1\text{-}3.25)$$

a w grupie B=W osób mieszkających na wsiach:

$$\text{logit}[pr(G_B)] \equiv \text{logit}[pr(Y_B=1)] = \ln odds(Y_B=1) = \beta_0 + (\beta_1 \cdot 0) = \beta_0 \ . \qquad (4\text{-}1\text{-}3.26)$$

Zatem postać (4-1-3.22) ilorazu szans jest dla (4-1-3.24) i (4-1-3.25) następująca:

$$OR_{A\ vs.B} = \frac{odds(grupa\,A)}{odds(grupa\,B)} \equiv \frac{odds(G_A)}{odds(G_B)} = \frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}} = e^{\beta_1} \quad . \tag{4-1-3.27}$$

Widać, że wyraz wolny β_0 skasował się w ilorazie szans $OR_{Avs.B}$. Jednocześnie współczynnik β_1 pozostał, co jest spowodowane niejednorodnością czynnika X_1, tzn. przyjmowaniem innego poziomu w grupach A i B. Postać (4-1-3.27) pozwala jedynie na podstawie oszacowania $\hat{\beta}_1$ (np. metodą MNW) wartości parametru β_1 na porównanie szans spłacenia długu (kredytu) dla osób mieszkających w miastach (A=M) z mieszkającymi na wsiach (B=W).

Wprowadźmy obok głównego czynnika *lokalizacji* $X_1 \equiv l$ dwa jego kowarianty, zmienną ciągłą $X_2 \equiv w$ oznaczającą wiek dłużnika oraz zmienną kategoryczną $X_3 \equiv s$ wskazującą płeć (1-kobieta, 0-mężczyzna). Mamy więc k=3 czynniki. Zatem specyfikacja całej grupy (lub jednostki) A i B jest bardziej złożona niż poprzednio, gdzie grupa (jednostka) była wskazana jedynie miejscem zamieszkania. Niech więc $x_A=(x_{1A},x_{2A},x_{3A})$ i $x_B=(x_{1B},x_{2B},x_{3B})$ są dwoma szczególnymi specyfikacjami A i B (wartości) trzech czynników $X = (X_1,X_2,X_3)$, przykładowo:

$$x_A = (x_{1A} \equiv l_A = 1, x_{2A} \equiv w_A, x_{3A} \equiv s_A = 1)\,, \tag{4-1-3.28a}$$

$$x_B = (x_{1B} \equiv l_B = 0, x_{2B} \equiv w_B, x_{3B} \equiv s_B = 1)\,. \tag{4-1-3.28b}$$

Specyfikacja x_A wskazuje grupę kobiet ($s_A = 1$) w wieku w_A lat, mieszkających w miastach ($l_A = 1$), a x_B wskazuje również grupę kobiet ($s_B = 1$) w wieku w_B lat lecz mieszkających na wsiach ($l_B = 0$).

W omawianym przykładzie należy rozważyć więc następującą logit'ową transformację (4-1-3.20) prawdopodobieństwa:

$$\text{logit}[pr(G)] \equiv \text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot l + \beta_2 \cdot w + \beta_3 \cdot s\,. \tag{4-1-3.29}$$

Dla grupy (jednostki) A jak w (4-1-3.28a) ma ona postać:

$$\begin{aligned} &\text{logit}[pr(G_A)] \equiv \text{logit}[pr(Y_A=1)] = \ln odds(Y_A=1) \\ &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot w_A + \beta_3 \cdot s_A = \beta_0 + \beta_1 + \beta_2 \cdot w_A + \beta_3 \cdot s_A\,, \end{aligned} \tag{4-1-3.30}$$

a dla grupy (jednostki) B jak w (4-1-3.28b) ma ona postać:

$$\begin{aligned} &\text{logit}[pr(G_B)] \equiv \text{logit}[pr(Y_B=1)] = \ln odds(Y_B=1) \\ &= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot w_B + \beta_3 \cdot s_B = \beta_0 + \beta_2 \cdot w_B + \beta_3 \cdot s_B\,. \end{aligned} \tag{4-1-3.31}$$

Iloraz szans porównujący stosunek szans spłacenia długu (kredytu) dla osób wskazanych łącznym poziomem A (więc mieszkających w miastach) do osób wskazanych łącznym poziomem B (więc mieszkających na wsiach) ma zgodnie z (4-1-3.22) postać (4-1-3.23):

$$OR_{A(w_A,s_A)vs.B(w_B,s_B)} = \frac{odds(grupa\,A)}{odds(grupa\,B)} \equiv \frac{odds(G_A)}{odds(G_B)} = \frac{e^{(\beta_0+\beta_1+\beta_2\cdot w_A+\beta_3\cdot s_A)}}{e^{(\beta_0+\beta_2\cdot w_B+\beta_3\cdot s_B)}} = e^{\beta_1+\beta_2\cdot(w_A-w_B)+\beta_3\cdot(s_A-s_B)}.$$

(4-1-3.32)

Gdyby w powyższym przykładzie rozważać grupy, w których wartości kowariantów czynnika lokalizacji $X_1 \equiv l$, tzn. wieku $X_2 \equiv w$ i płci $X_3 \equiv s$ były jednorodne w podgrupach kredytobiorców miast i wsi, czyli gdyby $w_A = w_B$ (ten sam wiek) oraz $s_A = s_B$ (ta sama płeć), to (ze względu na brak interakcji z czynnikiem lokalizacja $X_1 \equiv l$) wpływ kowariantów skasowałby się w liczniku i mianowniku $odds(\text{grupa A}) / odds(grupa\,B)$:

$$OR_{Avs.B} = \frac{odds(grupa\,A)}{odds(grupa\,B)} = e^{\beta_1} \qquad (4\text{-}1\text{-}3.33)$$

i otrzymalibyśmy iloraz szans niespłacenia długu w mieście w stosunku do wsi, który nie zależałby od poziomów tych kowariantów. (Podobna sytuacja miała miejsce w Rozdziale ROZDZIAŁ 2-2-7-2, wzór (2-2-7-2.44) dla ryzyka względnego, w przypadku analizy awarii w regresji Poissona, w którym kowariantem lokalizacji był rok serwisowania samochodu).

Podsumujmy. W ogólności, kiedy zmienia się tylko jeden czynnik (np. lokalizacja w powyższym przykładzie), a poziomy pozostałych czynników są ustalone, to mówimy, że iloraz szans porównujący dwa poziomy jednej, zmieniającej się cechy (zamieszkanie w mieście vs. na wsi) jest *dopasowanym*[23] (*skorygowanym*) ilorazem szans[24], który został wyznaczony, gdy pozostałe czynniki są pod kontrolą (tzn. model je uwzględnił) i ich wartości są ustalone w porównywanych grupach. Podlegający zmianie, interesujący nas czynnik jest często nazywany czynnikiem badanym (jawnym), inne czynniki to zmienne kontrolowane. Czynniki jawne są traktowane w analizie jako mające główny wpływ (*main effect*) na zachowania się wartości oczekiwanej zmiennej objaśnianej, stąd są nazywane ogólnym terminem czynników głównych (zasadniczych). Gdy człon interakcji jest nieistotny statystycznie, wtedy wprowadzone do analizy kowarianty są czynnikami zaburzającymi. O sprawie tej mówiliśmy poprzednio w Rozdziale 2-2-12, przy okazji regresji Poissona.

Pod warunkiem, że nie ma żadnych interakcji włączających zmienne główne, dopasowany iloraz szans może zostać wyliczony jako eksponenta ze współczynnika kierunkowego stojącego przy zero-jedynkowej zmiennej głównej, występującej w modelu logistycznym. Warto pamiętać, że gdy oszacowujemy współczynnik β_1 na podstawie pobranej próbki (czyniąc to np. zgodnie z MNW), to chociaż $OR_{Avs.B}$ w (4-1-3.27) i w (4-1-3.33)

[23] Dopasowanym ze względu na wprowadzone czynniki pod kontrolą. Ustalenie wartości kontrolowanych czynników na tych samych poziomach umożliwia (przy jednoczesnym traktowaniu rozkładów tych czynników w (pod)populacjach jako identycznych) przeprowadzenie wspólnej analizy szans w (pod)populacjach połączonych w jedną ogólniejszą populację [1]. Założenie o identyczności rozkładów czynników w (pod)populacjach należałoby jednak najpierw przetestować.

[24] (*adjusted odds ratio*)

wyglądają tak samo, jednak wartość surowego oszacowania $\hat{\beta}_1$ otrzymana w modelu (4-1-3.24) jest inna niż otrzymana w modelu (4-1-3.29), w którym kowarianty wiek oraz płeć zostały uwzględnione.

Do tej pory rozważaliśmy tylko wpływy główne w badanym modelu, np. *lokalizacja* (*l*) i ewentualne kowarianty np. *wiek* (*w*), *płeć* (*s*). (Powyższy przykład mógłby być tak sformułowany, aby wpływem głównym była np. płeć). Nie rozważaliśmy ani członów interakcji np. typu $l \times w$ czy $l \times s$, ani innych czynników głównych jak tylko zero-jedynkowe (0-1). Gdy model zawiera bądź interakcje typu $l \times w$, bądź czynniki główne nie przyjmujące wartości 0 lub 1, wtedy poprzednie wyprowadzenie ilorazu szans w postaci (4-1-3.33) nie będzie na ogół słuszne. Wtedy, w celu wyznaczenia dopasowanego iloraz szans należy skorzystać z ogólnej zależności (4-1-3.21).

Przykład 1 (c.d.) Niech tak jak wcześniej *lokalizacja* (*l*) jest czynnikiem głównym (0-1), a (ciągła zmienna) *wiek* oraz zero-jedynkowa zmienna *płeć* są zmiennymi kontrolowanymi. Model logistyczny w postaci logit'owej ma teraz postać:

$$\begin{aligned}&\text{logit}[pr(G)] \equiv \text{logit}[pr(Y=1)] = \ln odds(Y=1)\\ &= \beta_0 + \beta_1\, l + \beta_2\, w + \beta_3\, s + \beta_4\,(l \times w) + \beta_5\,(l \times s)\quad .\end{aligned} \tag{4-1-3.34}$$

Aby otrzymać iloraz szans dopasowany do wieku i rasy, musimy określić dwa zbiory wartości dla x_A i x_B dla układu czynników X:

X = (*lokalizacja (l), wiek (w), płeć (s), lokalizacja* $\times$ *wiek* ($l \times w$), *lokalizacja*$\times$*płeć* ($l \times s$))

(4-1-3.35)

Ogólna postać dopasowanego ilorazu szans (4-1-3.23) dla modelu (4-1-3.34) jest więc następująca:

$$OR_{A\ vs.\ B} = e^{\beta_1(l_A - l_B) + \beta_2(w_A - w_B) + \beta_3(s_A - s_B) + \beta_4(l_A \times w_A - l_B \times w_B) + \beta_5(l_A \times s_A - l_B \times s_B)} \tag{4-1-3.36}$$

Jeśli jak poprzednio, specyfikacja x_A wskazuje na grupę kobiet ($s_A = 1$) w wieku w_A lat, mieszkających w miastach ($l_A = 1$), a x_B wskazuje na grupę kobiet ($s_B = 1$) w wieku w_B lat i mieszkających na wsiach ($l_B = 0$), wtedy mamy dwa pięcio-elementowe zbiory wartości dla układu czynników X (4-1-3.35):

$$x_A = (l_A = 1, w_A, s_A = 1, l_A \times w_A = w_A, l_A \times s_A = 1)\ , \tag{4-1-3.37a}$$

$$x_B = (l_B = 0, w_B, s_B = 1, l_B \times w_B = 0, l_B \times s_B = 0)\ , \tag{4-1-3.37b}$$

gdzie $w_A = w_B = w$ (ten sam wiek) oraz $s_A = s_B = s$ (ta sama płeć) . (4-1-3.37c)

Dla powyższych specyfikacji poziomów x_A oraz x_B, *dopasowany* iloraz szans (4-1-3.36) dla modelu (4-1-3.34), ma postać:

$$OR_{(A\ vs.\ B|wiek=w,\ p\ell eć=1)} = e^{\beta_1(1-0) + \beta_2(w-w) + \beta_3(s-s) + \beta_4\, w + \beta_5(1-0)} = e^{\beta_1 + \beta_4\, w + \beta_5}\ , \tag{4-1-3.38}$$

gdzie po symbolu | podano warunek wskazujący badaną grupę. Otrzymany dopasowany iloraz szans na spłatę długu ze względu na wpływ lokalizacji, został wyznaczony przy ustalonych poziomach innych czynników. Obok współczynnika β_1 pojawiły się parametry β_4, β_5 włączone do modelu (4-1-3.34) ze względu na ich występowanie w dopasowanym modelu w członach interakcji z czynnikiem lokalizacji. Model (4-1-3.34) oznacza więc, że wartość ilorazu szans dla wpływu *lokalizacji* zmienia się w zależności od poziomu czynników *wiek* i *płeć* oraz zależy od wartości parametrów β_4, β_5, które stoją przy iloczynach tych czynników ze zmienną *lokalizacja*. Aby zobaczyć w jaki sposób kowarianty *wiek* i *płeć* modyfikują wpływ główny zmiennej zasadniczej *lokalizacja*, wystarczy do (4-1-3.36) wstawić różne, ustalone specyfikacje x_A oraz x_B porównywanych grup (lub jednostek) A i B.

Podsumujmy: Dla modelu logistycznego opisującego główny wpływ jawnej zmiennej zero-jedynkowej wraz z jej interakcjami (opisanymi przez iloczyny tej zmiennej głównej ze zmiennymi kontrolowanymi), dopasowany iloraz szans jest wyznaczany jako eksponenta funkcji liniowej współczynników regresji dla zarówno wpływu głównego jak i jego interakcji z kowariantami. Ponadto, ponieważ dopasowywany model zawiera człony interakcji, liczbowa wartość dopasowanego ilorazu szans będzie się zmieniała w zależności od wartości zmiennych pod kontrolą, które jako wchodzące w interakcję ze zmienną główną modyfikują wpływ główny.

Przykład 2. Przypadek czynnika głównego typu ciągłego.

Przypuśćmy, że główny czynnik modelu jest zmienną typu ciągłego, taką jak np. są w praktyce dochody osoby (d) (pomijając jednostkę jednego grosza). Kowariantami są *wiek* (w) oraz *płeć* (s) a zmienną objaśnianą, tak jak w Przykładzie 1, *spłacalność zadłużenia* (Y). Model przyjmuje teraz następującą postać:

$$\text{logit}[\text{pr}(Y=1)] = \beta_0 + \beta_1 d + \beta_2\, w + \beta_3\, s \qquad (4\text{-}1\text{-}3.39)$$

Aby uzyskać skumulowany iloraz szans dla takiego modelu należy wyodrębnić dwie wartości zmiennej d, kontrolując wpływ wieku w i płci s. Dwie wyodrębnione wartości d są potrzebne nawet gdy czynnik jest typu ciągłego, ponieważ iloraz szans z założenia porównuje dwa stany zmiennej.

Ogólna postać ilorazu szans, gdy zmienne wiek w i płeć s są ustalone w grupie (dla jednostki) A i B, jest następująca:

$$OR_{(d_A \ vs.\ d_B | w_A = w_B = w, s_A = s_B = s)} = e^{(d_A - d_B)\beta_1 + (w-w)\beta_2 + (s-s)\beta_3} = e^{(d_A - d_B)\beta_1} \qquad (4\text{-}1\text{-}3.40)$$

Na przykład, jeśli dwa stany zmiennej d to d_A = 3200 PLN oraz d_B = 1200 PLN, to iloraz szans ma postać:

$$OR_{(d_A = 3200 \ vs.\ d_B = 1200 | w_A = w_B = w, s_A = s_B = s)} = e^{(3200-1200)\beta_1} = e^{2000\beta_1} \ . \qquad (4\text{-}1\text{-}3.41)$$

Gdyby d_A - d_B = 1 PLN, wtedy wyrażenie (4-1-3.40) uprościłoby się do e^{β_1}. Przypadek różnicy w *d* równej 1 PLN, jest z ekonomicznego punktu widzenia nieciekawe. Dlatego zmienne typu ciągłego, takie jak *d*, można potraktować następująco. Po zbadaniu rozkładu zmiennej głównej (np. *d*), można dokonać pogrupowania jej wartości w warianty, np. wg kolejnych kwantyli, następnie wyznaczyć w kwantylach średnie lub mediany dla rozkładu tej zmiennej i w końcu wyznaczyć ilorazy szans dla oceny wpływu tej zmiennej głównej na odpowiedź *Y* [1], [40].

Rozdział 4-1-4. Estymacja ilorazu szans oraz weryfikacja hipotez statystycznych.

Właściwie przeprowadzona analiza logistyczna oznacza uchwycenie właściwego związku pomiędzy wybranym czynnikiem głównym, a zmienną objaśnianą (charakteryzującą np. spłacalność zadłużenia). Istotną sprawa jest właściwe wprowadzenie do analizy kowariantów, które należy mieć pod kontrolą i ze względu na które dopasowywany jest iloraz szans *OR*. Nie wprowadzenie kowariantów do analizy oznacza oszacowanie surowego *OR*, co może być uzasadnione, ale jedynie po uprzednim sprawdzeniu nieistotności możliwych interakcji, a następnie stwierdzeniu (eksperckim, bez wykonywania statystycznego testu) nieznaczności kowarianta jako zaburzenia. Niewłaściwe ujęcie wpływu kowarantów skutkuje złym ocenieniem siły związku pomiędzy zmienną jawną a odpowiedzią. Oznacza ono posługiwanie się niedopasowanymi ze względu na kowarianty zmiennymi głównymi i estymatorami parametrów.

Estymatory MNW współczynników regresji β_j, *j*=0,1,2,...,*k*, otrzymujemy za pomocą standardowego zestawu procedur regresji logistycznej w programie SAS. Są one z kolei wykorzystane w budowaniu estymatora ilorazu szans $\hat{O}R$ po wstawieniu w (4-1-3.23) $\hat{\beta}_j$ w miejsce β_j, w celu otrzymania z próbki numerycznej, punktowej wartości oszacowania dopasowanego ilorazu szans.
Wnioskowanie statystyczne dotyczące ilorazu szans może być związane bądź z testowaniem hipotez bądź estymacją przedziałową.

Dla prostego przykładu bez interakcji z kowariantami i przy ich ustalonych wartościach, iloraz szans ma postać $OR = e^{\beta_1}$, (4-1-3.33), a przedział ufności dla dopasowanego ilorazu szans można wyznaczyć licząc przedział ufności dla β_1, a następnie eksponenty z jego dolnej i górnej granicy. W konsekwencji, dla wystarczająco dużej próby, otrzymujemy (1-α) 100%-ową realizację przedziału ufności dla ilorazu szans *OR*:

$$\left(\exp(\hat{\beta}_1 - u_{1-\alpha/2}\ \hat{\sigma}_{\hat{\beta}_1}), \exp(\hat{\beta}_1 + u_{1-\alpha/2}\ \hat{\sigma}_{\hat{\beta}_1})\right), \qquad (4\text{-}1\text{-}4.42)$$

gdzie $\hat{\beta}_1$ jest estymatorem MNW parametru β_1, $\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$ jest oszacowanym błędem standardowym (estymatorem odchylenia standardowego) estymatora $\hat{\beta}_1$ parametru β_1, a $u_{1-\alpha/2}$ jest kwantylem rzędu $1-\alpha/2$ statystyki Wald'a, która ma asymptotycznie rozkład normalny (Rozdział 4). Oszacowanie $\hat{\sigma}^2_{\hat{\beta}_1}$ wariancji $\sigma^2_{\hat{\beta}_1}$ estymatora parametru β_1 leży na przekątnej obserwowanej macierz kowariancji $\hat{V}(\hat{\beta})$ (2-2-8.45) estymatorów parametrów β, która w metodzie MNW jest odwrotnością *obserwowanej* informacji Fishera **iF** (Rozdział 2-2-8) [5]. Zarówno $\hat{V}(\hat{\beta})$ jak i $\hat{\sigma}_{\hat{\beta}_1}$ może być podana w raportach SAS'a.

W przypadku weryfikacji hipotez statystycznych szczególnie ważna jest hipoteza zerowa mówiąca o tym , że iloraz szans wynosi 1. Gdy na przykład dopasowany jest iloraz szans dany prostym wyrażeniem $OR = e^{\beta_1}$, (4-1-3.33), zawierający zależność od jednego współczynnika, wtedy hipoteza zerowa:

$$H_0 : e^{\beta_1} = 1 \,, \qquad (4\text{-}1\text{-}4.43)$$

że iloraz szans wynosi 1 może, ze względu na $e^{\beta_1} = e^0 = 1$, zostać wyrażona następująco:

$$H_0 : \beta_1 = 0 \,. \qquad (4\text{-}1\text{-}4.44)$$

Test tej hipotezy zerowej może zostać przeprowadzony bądź za pomocą statystyki testowej Wald'a (Rozdział 7.1, Uzupełnienie 1, (7-1.6)), bądź z wykorzystaniem statystyki ilorazu wiarygodności (Rozdziały 1-3 i 1-4). Przykład odpowiedniej analizy zostanie zaprezentowany w następnym rozdziale.

Przy analizie wyboru modelu logistycznego wygodną postacią jest jego postać logit'owa (4-1-3.20):

$$\text{logit}[pr(Y=1)] = \beta_0 + \sum_{j=1}^{k} \beta_j X_j \,. \qquad (4\text{-}1\text{-}3.20')$$

Analizę związana z selekcją modelu można oprzeć o badanie istotności (kolejnych w hierarchii) parametrów modeli o postaci (2-1-1.19'). Hipoteza zerowa o nieistotności parametru β_j w powyższym równaniu regresji ma postać:

$$\mathrm{H}_0 : \beta_j = 0 \qquad j = 0,1,..k, \qquad (4\text{-}1\text{-}4.45)$$

gdzie *k* jest liczbą czynników w modelu.

Hipoteza alternatywnej ma postać:

$$H_1 : \beta_j \neq 0 \,. \qquad (4\text{-}1\text{-}4.46)$$

Ponieważ wszystkie parametry modelu logistycznego estymuje się metodą największej wiarygodności, zatem statystyka testowa dla hipotezy zerowej (4-1-4.45):

$$U = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \tag{4-1-4.47}$$

ma asymptotyczne rozkład normalny [36], a statystyka (4-1-4.47) ma asymptotycznie rozkład normalny standaryzowany $N(0,1)$. Oznacza to, że dla wystarczająco dużej próby, do testowania hipotezy (4-1-4.45) wobec hipotezy (4-1-4.46) można wykorzystać *statystykę Wald*'a (Rozdział 7.1, Uzupełnienie 1):

$$U = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}, \tag{4-1-4.48}$$

która przy prawdziwości hipotezy zerowej (4-1-4.45) ma rozkład N(0, 1). Wartość empirycznego poziom istotności:

$$p = P(U \geq u_{obs}) \tag{4-1-4.49}$$

jest wykorzystywana do stwierdzenia, czy wartość statystyki (4-1-4.48) różni się od zera istotnie statystycznie, a więc czy wartość statystyki $\hat{\beta}_j$ różni się od zera w sposób istotny statystycznie.

Zamiast posługiwać się statystyką (4-1-4.48) można posłużyć się związaną z nią statystyką:

$$\chi^2 = U^2. \tag{4-1-4.50}$$

Gdy zmienna U ma rozkład $N(0,1)$, to można łatwo pokazać, że zmienna U^2 ma rozkład chi-kwadrat z jednym stopniem swobody (proszę pokazać ten fakt). Zatem statystyka chi-kwadrat Wald'a:

$$\chi_1^2 = \frac{\hat{\beta}_j^2}{\hat{\sigma}_{\hat{\beta}_j}^2}, \tag{4-1-4.51}$$

ma asymptotycznie i przy prawdziwości hipotezy zerowej (4-1-4.45) rozkład chi-kwadrat z liczbą stopni swobody równą 1.

A. Rozdział 5. Przykład regresji logistycznej.

Przykład dotyczy problemu niespłacenia długu (kredytu). Podane dane (Rozdział 5-3) opisują próbę losową 196 osób wylosowanych z trzech ustalonych dzielnic wybranego, średniej wielkości miasta (40 – 400 tysięcy). Z pośród wylosowanych osób 57 nie spłaciło długu w wyznaczonym terminie.

Celem analizy jest określenie czynników „ryzyka" mających wpływ na niespłacenie długu. Po pierwsze, badanym wpływem głównym jest „Płeć" osoby. Jej kowariantami są: czynnik „*Czas*" lat pracy osoby (włączając w to okres na emeryturze) oraz czynnik „Lokalizacja" (*Lok*), czyli dzielnica miasta. Zmienną objaśnianą Y jest „Dług". Zmienne te mają następujący charakter:

a) zmienną objaśnianą Y jest „Dług"; jest to zmienna dychotomiczna mająca dwa poziomy: 1=dług niespłacony, 0=dług spłacony,
b) czynnik główny „Płeć" osoby przyjmuje dwa poziomy: zmienna nominalna (kategoryczna) przyjmująca warianty 1=kobieta", 0=„mężczyzna",
c) czynnik poboczny (kowariant) „*Czas*" lat pracy osoby (włączając w to okres na emeryturze): zmienna typu ilościowego, interwałowa (czas podany jest w latach),
d) czynnik poboczny „Lokalizacja" (*Lok*) (dzielnica miasta): zmienna kategoryczna przyjmująca warianty 1,2,3.

W analizie statystycznej SAS często wygodnie jest posługiwać się zmiennymi wskazującymi zamiast zmiennymi pierwotnymi. Ponieważ zmienna „Płeć" ma dwa warianty, zatem wystarczy jedna zmienna wskazująca do ich wskazania; nazwijmy ją P_c. Jej kodowanie jest następujące:

$$P_c = \begin{cases} 1 \text{ , gdy osoba jest kobietą} \\ 0 \text{ , gdy osoba jest mężczyzną} \end{cases} \tag{5.1}$$

Zmienna „Lokalizacja" (*Lok*) dostarcza informacji o dzielnicy zamieszkania badanej osoby. Rozważono trzy dzielnice, stąd próbkę podzielono na trzy lokalizację. Zmienna *Lok* ma trzy warianty: 1, 2 i 3, zatem w jej miejsce wprowadzamy dwie zmienne wskazujące (inne kodowanie też jest możliwe):

$$L_1 = \begin{cases} 1 \text{ gdy osoba mieszka w dzielnicy1} \\ 0 \text{ gdy osoba mieszka w dzielnicy2} \\ 0 \text{ gdy osoba mieszka w dzielnicy3} \end{cases} \tag{5.2a}$$

$$L_2 = \begin{cases} 0 \text{ gdy osoba mieszka w dzielnicy1} \\ 1 \text{ gdy osoba mieszka w dzielnicy2} \\ 0 \text{ gdy osoba mieszka w dzielnicy3} \end{cases} \tag{5.2b}$$

Przez podanie wartości zmiennych L_1 i L_2 w sposób jednoznaczny wskazujemy dzielnicę miasta, w której mieszka badana osoba. Ponieważ możliwe są tylko trzy warianty zmiennej Lokalizacja (*Lok*), zatem jeśli zmienna wskazująca L_1 przyjmuje wartość 0, a zmienna L_2 również przyjmuje wartość 0, to badana osoba nie mieszka ani w dzielnicy 1, ani w dzielnicy 2, co oznacza, że mieszka ona w dzielnicy 3, a zmienna pierwotna *Lok* ma wartość 3. Po przygotowaniu zmiennych kierunkowych i wprowadzeniu ich w miejsce zmiennych

pierwotnych do zbioru danych (Rozdział 5-3), korzystamy z analizy SAS'a, traktując wszystkie zmienne jako ilościowe (*quantitative*).

Rozdział 5-1. Analiza bez interakcji głównego wpływu z kowariantami.

W pierwszej kolejności przeprowadzamy analizę bez wprowadzania interakcji. Dla modelu zawierającego kolejno, zmienną *Czas*, zmienną P_c wskazującą płeć i zmienne L_1 oraz L_2 wskazujące lokalizację, forma logit'owa (4-1-3.20) modelu (czyli logarytm naturalny szansy zajścia zdarzenia {*Y*=1}), przybiera postać:

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_2 \cdot P_c + \beta_3 \cdot L_1 + \beta_4 \cdot L_2 \qquad (5\text{-}1.3)$$

Do oszacowania parametrów modelu (5-1.3) wykorzystamy aplikację Analyst pakietu SAS. W Rozdziale 5-3 podano dane pomiarowe. (W nazwach zmiennych w zbiorze danych pomijamy polskie znaki oraz podnosimy dolne indeksy). W pierwszej kolumnie podano numerem porządkowy (Nr) osoby biorącej udział w badaniu. W drugie kolumnie znajdują sie obserwacje dla zmiennej „Dług" *Y* (Dlug), będąca zmienną charakteryzującą stan spłacenia długu. W kolumnie trzeciej podane są wartości zmiennej *Czas*; w czwartej P_c (czyli zmienna kierunkowa płci (5.1)); potem kolumny zmiennych kierunkowych lokalizacji L_1 oraz L_2 (5.2a) i (5.2b), następnie zmienna interakcji czasu i płci $CP_c \equiv Czas \times P_c$ oraz dodatkowo zmienna pierwotna „Lokalizacji" (*Lok*).

Rozdział 5-1-1. Omówienie kolejnych kroków analizy przykładu w programie SAS.

Raport z analizy przeprowadzonej w SAS otrzymujemy po wykonaniu następujących kroków. Po uruchomieniu SAS:

c) Z paska MENU wybieramy Solutions → Analysis → Analyst.

d) Po wczytaniu zbioru danych (File → Open By SAS Name) przechodzimy do regresji logistycznej: Statistics → Regression → Logistic.

e) W prawym górnym okienku „Model Pr{ }" wybieramy wartość 1, gdyż ta wartość zmiennej *Y* („Dług") oznacza zajście zdarzenia niespłacenia długu (co w naszym przypadku jest sukcesem) i modelowanie odpowiadającego mu prawdopodobieństwa. (Wybór wartości 0, odpowiadającej porażce, oznaczałoby modelowanie prawdopodobieństwa związanego ze spłatą długu).

f) Z opcji „Dependent type" wybieramy „Single trial" jako opis eksperymentu pojedynczego losowania, dla każdej jednostki (osoby), wartości zmiennej dychotomicznej *Y*.

g) W okno „Quantitative" prowadzamy zmienne *Czas*, Pc, L1 i L2.

h) W opcji „Model" (na dole okna głównego) wybieramy: z (pod)okna „Model" → Standard Models → Main effects only, czyli uwzględniamy jedynie wpływy główne (którymi *w znaczeniu tego okna wyboru* są wpływy od czynnika głównego i jego kowariantów (tzn. *Czas*, Pc, L1 i L2), ale bez uwzględnienia interakcji). Można by wprowadzić opcje z interakcjami do rzędu 2-giego lub 3-ciego (pozostawiając jedynie interakcje

mające sens), jednakże interakcję rozważymy w dalszej części, wprowadzając zmienną interakcji $CP_c \equiv Czas \times$ Pc bezpośrednio do danych.

i) Ciągle w opcji „Model” wybieramy: z (pod)okna „Selection” opcję „Full Model” (możliwość wyboru procedury eliminacji wstecz „Backward elimination”, omówimy dalej).

j) W opcji „Statistics” (na dole okna głównego) wybieramy: z (pod)okna „Intervals” → Wald limits (For parameter estimates), co umożliwi otrzymanie przedziałów ufności Wald’a dla parametrów β_j.

Po zatwierdzeniu powyższych procedur, otrzymujemy raport SAS’a.

Uwaga: W zbiorze danych oraz w raportach SAS’a: Pc to P_c, L1 to L_1, L2 to L_2 oraz CP_c to CPc.

Raport 1 dla modelu (5-1.3). Interesująca nas część raportu ma postać:

```
Dlug_logit          23:47 Thursday, February 6, 2014   1

                                Procedura LOGISTIC

                                   Informacje

                Zbiór                          DLUG_KIERUNKOWE
                Zmienna objaśniana             Dlug                      Dlug
                Liczba poziomów odpowiedzi     2
                Model                          logit binarny
                Technika optymalizacji         Ocena Fishera

                           Wczytano obserwacji                 196
                           Użyto obserwacji                    196

                                   Profil odpowiedzi

                                 Wartość                   Całkowita
                           uporządkowana         Dlug     liczebność

                                       1            1             57
                                       2            0            139

                        Modelowane prawdopodobieństwo wynosi Dlug=1.

                                     Status zbieżności

                        Kryterium zbieżności (GCONV=1E-8) spełnione.


                                 Statystyki dopasowania

                                           Tylko      Wyraz wolny
                                           wyraz                i
                           Kryterium       wolny     współzmienne

                           AIC           238.329          221.220
                           SC            241.607          237.611
                           -2 log L      236.329          211.220


                       Testowanie globalnej hipotezy zerowej: BETA=0

                                                  St.
                Test                 Chi-kwadrat      sw.     Pr. > chi-kw.

                Iloraz wiarygod.         25.1094        4            <.0001
                Ocena                    24.9977        4            <.0001
                Wald                     21.9895        4            0.0002
```

```
                                        Dlug_logit          23:47 Thursday, February 6, 2014
2

                                    Procedura LOGISTIC

                        Analiza ocen maksymalnej wiarygodności

                 St.                          Błąd    Chi-kwadrat
Parametr         sw.       Ocena    standardowy          Walda    Pr. > chi-kw.

Intercept         1     -2.0405         0.3895        27.4474           <.0001
Czas              1      0.0270        0.00868         9.6796           0.0019
Pc                1      1.2436         0.3523        12.4615           0.0004
L1                1     -0.2534         0.4056         0.3905           0.5320
L2                1     -0.2088         0.4545         0.2111           0.6459


                               Oceny ilorazu szans

                                 Ocena      Przedział ufności
                       Efekt    punktowa          Walda 95%

                       Czas        1.027       1.010       1.045
                       Pc          3.468       1.739       6.918
                       L1          0.776       0.351       1.718
                       L2          0.812       0.333       1.978


                         Przedział ufności Walda dla parametrów

                     Parametr         Ocena     Przedział ufności 95%

                     Intercept      -2.0405      -2.8038      -1.2771
                     Czas            0.0270      0.00999       0.0440
                     Pc              1.2436       0.5531       1.9341
                     L1             -0.2534      -1.0483       0.5414
                     L2             -0.2088      -1.0997       0.6820
```

Odpowiedni kod języka programowania 4GL SAS'a potrzebny do otrzymania powyższego raportu ma postać:

```
proc logistic data= Dlug_kierunkowe DESCEND;
model Dlug = Czas  Pc  L1  L2 / waldcl;
run;
```

Uwaga. Opcja DESCEND powoduje modelowanie prawdopodobieństwa sukcesu Y=1, o czym w powyższym raporcie informuje komunikat „`Modelowane prawdopodobieństwo wynosi Dlug=1`". Jej pominięcie oznaczałoby modelowanie prawdopodobieństwa porażki Y=0. Komenda „waldcl" powoduje wyznaczenie przedziałów ufności Wald'a dla parametrów. Konstrukcja przedziału ufności Wald'a opiera się o asymptotyczną normalność parametrów oszacowanych metodą NW.

Z części powyższego raportu: „`Analiza ocen maksymalnej wiarygodności`", odczytujemy, że estymatory MNW współczynników β_i otrzymane dla dopasowanego modelu przyjmują w pobranej próbce wartości:

$$\hat{\beta}_0 = -2.0405;\ \hat{\beta}_1 = 0.0270;\ \hat{\beta}_2 = 1.2436;\ \hat{\beta}_3 = -0.2534;\ \hat{\beta}_4 = -0.2088. \qquad (5\text{-}1\text{-}1.4)$$

Zatem oszacowana w pobranej próbce forma logit'owa modelu (5-1.3), ma postać:

$$\begin{aligned}&\text{logit}[pr(Y=1)] = \ln odds(Y=1)\\ &= -2.0405 + 0.0270 \cdot Czas + 1.2436 \cdot P_c - 0.2534 \cdot L_1 - 0.2088 \cdot L_2 \quad .\end{aligned} \tag{5-1-1.5}$$

Wychodząc z powyższego oszacowania modelu oraz powyższego raportu SAS'a, wyznaczymy iloraz szans niespłacenia długu kobiet względem mężczyzn, w sytuacji, gdy czas oraz lokalizacja (dzielnica zamieszkania w rozważanym mieście) są pod kontrolą.

Oszacowany iloraz szans ma wtedy (zgodnie z Raportem 1 SAS'a) postać wynikającą z (4-1-3.33):

$$\hat{O}R_{(Pc=1 vs. Pc=0 | Czas, Lok)} = \frac{odds(grupa \text{kobiet})}{odds(grupa \text{mezczyzn})} = e^{\hat{\beta}_2} = e^{1,2436} = 3,4681. \tag{5-1-1.6}$$

Ponieważ 95%-owy przedział ufności Wald'a dla β_2 wynosi (0,5531, 1.9341), zatem 95%-owy przedział ufności (4-1-4.42) dla ilorazu szans wynosi (zgodnie z Raportem 1 SAS'a):

$$\begin{aligned}&\left(\exp(\hat{\beta}_2 - u_{1-\alpha/2}\,\hat{\sigma}_{\hat{\beta}_2}), \exp(\hat{\beta}_2 + u_{1-\alpha/2}\,\hat{\sigma}_{\hat{\beta}_2})\right) = \exp(1,2436 \pm 1,96 \cdot 0,3523)\\ &= \left(\exp(0,5531), \exp(1,9341)\right) = (1,7386,\ 6,9179)\ ,\end{aligned} \tag{5-1-1.7}$$

gdzie $\hat{\sigma}_{\hat{\beta}_2} = 0,3523$ jest oszacowanym standardowym błędem estymatora $\hat{\beta}_2$ parametru β_2.

Rozważmy istotność statystyczną parametru β_2 stojącego przy czynniku głównym P_c (Płeć) modelu regresji (5-1.3):

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_2 \cdot P_c + \beta_3 \cdot L_1 + \beta_4 \cdot L_2 \ . \tag{5-1.3'}$$

Hipoteza zerowa o nieistotności parametru β_2 ma postać:

$$\mathrm{H}_0 : \beta_2 = 0, \tag{5-1-1.8}$$

a hipotezą alternatywną jest:

$$\mathrm{H}_1 : \beta_2 \neq 0. \tag{5-1-1.9}$$

Statystyka chi-kwadrat Wald'a (4-1-4.51):

$$\chi_1^2 = \frac{\hat{\beta}_2^2}{\hat{\sigma}_{\hat{\beta}_2}^2}, \tag{5-1-1.10}$$

ma przy prawdziwości hipotezy zerowej (5-1-1.8) asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody równą 1. Z Raportu 1 SAS'a widać, że ze względu na małą wartość empirycznego poziom istotności $p = P(\chi_1^2 \geq \chi_{1\,obs}^2 = 12,4615) = 0,0004$, można przyjąć, że otrzymana wartość oszacowania $\hat{\beta}_2 = 1,2436$ jest istotnie statystycznie różna od zera. Stąd wnioskujemy, że oszacowana wartość ilorazu szans niespłacenia długu przez kobiety w stosunku do mężczyzn wyniosła w tym mieście $\hat{O}R_{(Pc=1 vs. Pc=0 | Czas, Lok)}$

$= e^{1,2436} = 3,4681$, (5-1-1.6), i wartość ta jest istotnie statystycznie różna od 1. Wskazuje na to również nie pokrycie wartości 0 przez 95%-owy przedział ufności dla *OR*, (5-1-1.7).

Taki sam wniosek otrzymamy przeprowadzając test braku dopasowania do danych empirycznych, modelu niższego w hierarchii:

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_3 \cdot L_1 + \beta_4 \cdot L_2 \qquad (5\text{-}1\text{-}1.11)$$

w stosunku do modelu (5-1.3). Hipoteza zerowa ma również postać $\mathrm{H}_0 : \beta_2 = 0$, (5-1-1.8), ale można ją sformułować następująco:

H_0 : Nie ma istotnego statystycznego braku dopasowania do danych empirycznych modelu niższego w stosunku do modelu wyższego. (5-1-1.12)

Statystyką testową dla powyższej hipotezy jest statystyka ilorazu wiarygodności (2-1-4-1.24):

$$L_{r/k} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right] = -2\ln P(\tilde{Y} \mid \hat{\beta}_{(r)}) - \left(-2\ln P(\tilde{Y} \mid \hat{\beta})\right), \qquad (5\text{-}1\text{-}1.13)$$

gdzie funkcja wiarygodności dla modelu logistycznego ma postać **(4-1-2.13)**, a w miejsce parametrów wprowadzono ich estymatory MNW tak, że $\ln P(\tilde{Y} \mid \hat{\beta})$ jest zmaksymalizowanym logarytmem funkcji wiarygodności modelu wyższego z k + 1 parametrami β, a $\ln P(\tilde{Y} \mid \hat{\beta}_{(r)})$ jest zmaksymalizowanym logarytmem funkcji wiarygodności dla modelu niższego z r + 1< k+1 parametrami $\beta_{(r)}$.

Ze względu na to, że w przypadku powyższych modeli mamy do czynienia z klasą modeli hierarchicznych, tzn. jeden jest zredukowaną formą drugiego, można pokazać, że również w przypadku modelu logistycznego (porównaj Rozdział 2-1-5 dla rozkładu Poissona) i przy prawdziwości hipotezy zerowej (5-1-1.12), statystyka (5-1-1.13) ma asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody (*l.st.sw.*) równą *l.st.sw.*= $k - r$, gdzie k jest liczbą parametrów kierunkowych modelu wyższego, a r jest liczbą parametrów kierunkowych modelu niższego [1], [38]. Uznajemy, że analizowana próba w przykładzie jest duża (n = 196).

W analizowanym przykładzie $\beta_{(r)} = (\beta_0, \beta_1, \beta_3, \beta_4)$, czyli r =3, a $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$, czyli k = 4, zatem liczba stopni swobody statystyki (5-1-1.13) wynosi k-r = 1. Zweryfikujmy więc hipotezę H_0 (5-1-1.12) w rozpatrywanym przykładzie, stosując statystykę (5-1-1.13). W powyższym raporcie SAS'a, dla modelu (5-1.3) z wszystkimi czynnikami, otrzymaliśmy:

$$-2\ln P\left(\tilde{Y} \mid \hat{\beta}\right) = 211{,}22\,. \qquad (5\text{-}1\text{-}1.14)$$

Aby otrzymać analogiczną wartość $-2\ln P(\tilde{Y} \mid \hat{\beta}_{(r=3)})$ dla modelu (5-1-1.11), należy przeprowadzić analizę z pominięciem czynnika głównego P_c (Płeć). (Podobnie jak poprzednio, po wywołaniu ciągu „Analyst →Statistic →Regression →Logistic" w oknie „Quantitative" wprowadzamy zmienne *Czas* oraz *L*1 i *L*2, jednak z pominięciem Pc.)

Raport 2 dla modelu (5-1-1.11). Odpowiedni fragment raportu SAS'a ma postać:

```
Dlug_logit. Model bez Pc (Płeć).                                  1
                                                  16:45 Saturday, February 8, 2014

                          Procedura LOGISTIC

                       Statystyki dopasowania

                                Tylko      Wyraz wolny
                                wyraz                i
Kryterium                       wolny     współzmienne

AIC                           238.329          232.216
SC                            241.607          245.329
-2 log L                      236.329          224.216
```

W powyższym raporcie SAS'a dla modelu (5-1-1.11), otrzymaliśmy:

$$-2\ln P\left(\tilde{Y}|\hat{\beta}_{(r=3)}\right) = 224{,}216 \quad . \tag{5-1-1.15}$$

Wstawiając (5-1-1.14) oraz (5-1-1.15) do statystyki ilorazu wiarygodności (5-1-1.13), otrzymujemy:

$$LR_{3/4} = -2\ln\left[\frac{P\left(\tilde{Y}|\hat{\beta}_{(r=3)}\right)}{P\left(\tilde{Y}|\hat{\beta}\right)}\right] = 224{,}216 - 211{,}22 = 12{,}996, \tag{5-1-1.16}$$

Przy prawdziwości hipotezy zerowej (5-1-1.12) statystyka $LR_{3/4}$ ma asymptotycznie rozkład chi-kwadrat z *l.st.sw.* = $k - r$ = 1. Zatem empiryczny poziom istotności dla (5-1-1.13) wynosi :

$$p = P(LR_{3/4} \sim \chi_1^2 \geq 12{,}996) \approx 0{,}0003. \tag{5-1-1.17}$$

Zatem, na każdym poziomie istotności $\alpha \geq p = 0{,}0003$ model niższy (5-1-1.11) dopasowuje się do danych empirycznych istotnie statystycznie gorzej niż model wyższy (5-1.3). Oznacza to, że nie można pominąć wpływu głównego czynnika „Płeć" na spłatę zadłużenia i uznajemy rozszerzenie modelu (5-1-1.11) do (5-1.3) za istotne statystycznie.

Powróćmy więc do modelu (5-1.3) z wszystkimi czynnikami. Aby zweryfikować hipotezę zerową o nieistotności kowarianta „Lokalizacja" (w zastępstwie którego wprowadzono parę zmiennych kierunkowych L_1 oraz L_2) testujemy hipotezę:

$$\mathrm{H}_0 : \beta_3 = \beta_4 = 0 \tag{5-1-1.18}$$

wobec hipotezy alternatywnej

$$H_1 : \beta_3 \neq 0 \vee \beta_4 \neq 0 \ . \qquad (5\text{-}1\text{-}1.19)$$

Hipoteza H_0 jest równocześnie hipotezą o nie występowaniu braku dopasowania do danych empirycznych w modelu zredukowanym (z r = 2):

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_2 \cdot P_c \qquad (5\text{-}1\text{-}1.20)$$

w porównaniu z modelem (5-1.3).

Statystyka testowa (5-1-1.13) służąca do weryfikacji hipotezy (5-1-1.18) ma postać:

$$L_{2/4} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r=2)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right] = -2\ln P(\tilde{Y} \mid \hat{\beta}_{(r=2)}) - \left(-2\ln P(\tilde{Y} \mid \hat{\beta})\right), \qquad (5\text{-}1\text{-}1.21)$$

gdzie $\ln P(\tilde{Y} \mid \hat{\beta}_{(r=2)})$ jest zmaksymalizowanym logarytmem funkcji wiarygodności zredukowanego modelu (5-1-1.20). Statystyka $L_{2/4}$ ma przy prawdziwości hipotezy zerowej (5-1-1.18) asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody *l.st.sw.* = $k - r$ = 4 – 2 = 2.

Podobnie jak poprzednio, po wywołaniu ciągu „Analyst →Statistic →Regression →Logistic” w oknie „Quantitative” wprowadzamy zmienne *Czas* oraz Pc, jednak z pominięciem L1 i L2, tzn. nie włączamy do modelu zmiennych kierunkowych dla lokalizacji.

Raport 3 dla modelu (5-1-1.20). Odpowiedni fragment raportu SAS'a ma postać:

```
Dlug logit. Model bez L1 i L2 (Lokalizacja).                          1
                                                                      02:53 Sunday, February 9,
2014

                                  Procedura LOGISTIC

                                     Informacje

                 Zbiór                          JACEK.DLUG_KIERUNKOWE
                 Zmienna objaśniana             Dlug                      Dlug
                 Liczba poziomów odpowiedzi     2
                 Model                          logit binarny
                 Technika optymalizacji         Ocena Fishera

                       Wczytano obserwacji                 196
                       Użyto obserwacji                    196

                                 Profil odpowiedzi

                            Wartość                   Całkowita
                        uporządkowana         Dlug     liczebność

                                    1            1             57
                                    2            0            139
                 Modelowane prawdopodobieństwo wynosi Dlug=1.


                                 Status zbieżności

                   Kryterium zbieżności (GCONV=1E-8) spełnione.
```

Statystyki dopasowania

Kryterium	Tylko wyraz wolny	Wyraz wolny i współzmienne
AIC	238.329	217.639
SC	241.607	227.474
-2 log L	236.329	**211.639**

Testowanie globalnej hipotezy zerowej: BETA=0

Test	Chi-kwadrat	St. sw.	Pr. > chi-kw.
Iloraz wiarygod.	24.6901	2	<.0001
Ocena	24.6315	2	<.0001
Wald	21.6714	2	<.0001

Dlug logit. Model bez L1 i L2 (Lokalizacja).
2
02:53 Sunday, February 9,
2014

Procedura LOGISTIC

Analiza ocen maksymalnej wiarygodności

Parametr	St. sw.	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.
Intercept	1	-2.1596	0.3439	39.4357	<.0001
Czas	1	0.0268	0.00865	9.6082	0.0019
Pc	1	1.1817	0.3370	12.2981	0.0005

Oceny ilorazu szans

Efekt	Ocena punktowa	Przedział ufności Walda 95%	
Czas	1.027	1.010	1.045
Pc	3.260	1.684	6.310

Przedział ufności Walda dla parametrów

Parametr	Ocena	Przedział ufności 95%	
Intercept	-2.1596	-2.8337	-1.4856
Czas	0.0268	0.00986	0.0438
Pc	1.1817	0.5212	1.8421

Przedział ufności Walda dla ilorazów szans

Efekt	Jednostka	Ocena	Przedział ufności 95%	
Czas	1.0000	1.027	1.010	1.045
Pc	1.0000	3.260	1.684	6.310

W powyższym raporcie SAS'a dla modelu (5-1-1.20), otrzymaliśmy:

$$-2\ln P\left(\tilde{Y}|\hat{\beta}_{(r=2)}\right)=211{,}639. \qquad (5\text{-}1\text{-}1.22)$$

Wstawiając (5-1-1.14) oraz (5-1-1.22) do statystyki ilorazu wiarygodności (5-1-1.21), otrzymujemy:

$$LR_{2/4} = -2\ln\left[\frac{P\left(\tilde{Y}|\hat{\beta}_{(r=2)}\right)}{P\left(\tilde{Y}|\hat{\beta}\right)}\right] = 211{,}639 - 211{,}22 = 0{,}419\,. \qquad (5\text{-}1\text{-}1.23)$$

Przy prawdziwości hipotezy zerowej (5-1-1.18) statystyka $LR_{2/4}$ ma asymptotycznie rozkład chi-kwadrat z *l.st.sw.* = $k - r$ = 4 – 2 = 2. Zatem empiryczny poziom istotności dla (5-1-1.23) wynosi:

$$p = P(LR_{2/4} \sim \chi_2^2 \geq 0{,}419) \approx 0{,}811. \qquad (5\text{-}1\text{-}1.24)$$

Tak więc, na żadnym poziomie istotności $\alpha < p = 0{,}811$ nie ma podstaw do odrzucenia hipotezy H_0 (5-1-1.18) o nie występowaniu braku dopasowania do danych empirycznych w modelu zredukowanym (5-1-1.20) w porównaniu z modelem (5-1.3).

Wnioskujemy więc, że rozszerzenie modelu o czynnik „Lokalizacja" jest nieistotne statystycznie, czyli na podstawie przeprowadzonego testu uznajemy brak wpływ dzielnicy zamieszkania na niespłacenie długu w całej populacji osób tego miasta. Decydujemy się więc na wybór modelu zredukowanego (5-1-1.20) jako prostszego od modelu (5-1.3) i jednocześnie prawie tak samo dobrze dopasowującego się do danych empirycznych.

W końcu, z Raportu 3 odczytujemy następującą wartość oszacowania parametru β_2:

$$\hat{\beta}_2 = 1{,}1817\ . \qquad (5\text{-}1\text{-}1.25)$$

Oszacowany iloraz szans (4-1-3.33) ma zgodnie z Raportem 3 SAS'a wartość:

$$\hat{O}R_{(Pc=1\,vs.\,Pc=0\,|\,Czas)} = \frac{odds(grupa\ \text{kobiet})}{odds(grupa\ \text{mezczyzn})} = e^{\hat{\beta}_2} = e^{1{,}1817} = 3{,}26\ . \qquad (5\text{-}1\text{-}1.26)$$

Otrzymany iloraz szans nie zależy od wariantu czasu zatrudnienia, jest więc on ogólnym ilorazem szans.

Zatem zgodnie z modelem (5-1-1.20), w populacji osób rozważanego miasta, oszacowana szansa niespłacenia długu przez kobiety jest 3,26 razy większa niż przez mężczyzn.

Ponieważ 95%-owy przedział ufności Wald'a dla β_2 wynosi (0,5212, 1,8421), zatem 95%-owy przedział ufności (4-1-4.42) dla ilorazu szans jest (zgodnie z Raportem 3 SAS'a) równy:

$$\begin{aligned}&\left(\exp(\hat{\beta}_2 - u_{1-\alpha/2}\,\hat{\sigma}_{\hat{\beta}_2}), \exp(\hat{\beta}_2 + u_{1-\alpha/2}\,\hat{\sigma}_{\hat{\beta}_2})\right) = \exp(1{,}1817 \pm 1{,}96 \cdot 0{,}337)\\ &= \left(\exp(0{,}5212), \exp(1{,}8421)\right) = (1{,}684,\ 6{,}310)\quad ,\end{aligned} \qquad (5\text{-}1\text{-}1.27)$$

gdzie $\hat{\sigma}_{\hat{\beta}_2} = 0{,}337$ jest oszacowanym standardowym błędem estymatora $\hat{\beta}_2$ parametru β_2.

Sprawdźmy czy z modelu zredukowanego (5-1-1.20) nie można zrezygnować, przechodząc do modelu bez zmiennych objaśniających *Czas* i „Płeć", a jedynie z wyrazem wolnym β_0 (r = 0):

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0\ . \qquad (5\text{-}1\text{-}1.28)$$

W modelu tym wartość ilorazu szans (4-1-3.23) jest równa:

$$OR = e^0 = 1. \qquad (5\text{-}1\text{-}1.29)$$

Zweryfikujmy więc hipotezę zerową o niewystępowaniu braku dopasowania do danych empirycznych w modelu (5-1-1.28), w stosunku do modelu (5-1-1.20) (z $k = 2$), zapisując ją następująco:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = 0. \qquad (5\text{-}1\text{-}1.30)$$

Hipoteza alternatywna ma postać:

$$\mathrm{H}_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \quad . \qquad (5\text{-}1\text{-}1.31)$$

Statystyka testowa (5-1-1.13) dla hipotezy (5-1-1.18) ma postać:

$$L_{0/2} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_0\right)}{P\left(\tilde{Y} \mid \hat{\beta}_{(k=2)}\right)}\right] = -2\ln P(\tilde{Y} \mid \hat{\beta}_0) - \left(-2\ln P(\tilde{Y} \mid \hat{\beta}_{(k=2)})\right)$$
$$= 236{,}329 - 211{,}639 = 24{,}69 \quad , \qquad (5\text{-}1\text{-}1.32)$$

gdzie w Raporcie 3 wartość $-2\ln P(\tilde{Y} \mid \hat{\beta}_0) = 236{,}329$ jest podana w części raportu „`Statystyki dopasowania`” w kolumnie „`Tylko wyraz wolny`”.

Statystyka $L_{0/2}$ ma przy prawdziwości hipotezy zerowej (5-1-1.30) asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody *l.st.sw.* = $k - r = 2 - 0 = 2$.

Wartość empirycznego poziomu istotności:

$$p = P\left(L_{0/2} \sim \chi_2^2 \geq 24{,}69\right) = 0{,}0000044 \qquad (5\text{-}1\text{-}1.33)$$

jest bardzo mała, zatem odrzucamy hipotezę zerową o nieistotności rozszerzenia modelu (5-1-1.28) do modelu (5-1-1.20), w którym występują czynniki *Czas* zatrudnienia oraz „Płeć”.

Test trendu liniowego wpływu czynnika „Czas”.

Z Raportu 1 dla modelu (5-1.3) można również wyciągnąć wniosek, co do istotności czynnika „Czas”. Rozważmy więc sytuację, w której „*Czas*” jest traktowany jako czynnik główny, a pozostałe zmienne tzn. P_c (Płeć) oraz L_1 i L_2 są pod kontrolą. Rozważmy istotność statystyczną parametru β_1 stojącego przy czynniku *Czas* modelu regresji (5-1.3). Testujemy więc hipotezę zerową o nieistotności parametru β_1:

$$\mathrm{H}_0 : \beta_1 = 0, \qquad (5\text{-}1\text{-}1.34)$$

wobec hipotezy alternatywnej:

$$\mathrm{H}_1 : \beta_1 \neq 0. \qquad (5\text{-}1\text{-}1.35)$$

Statystyka testowa chi-kwadrat Wald'a (4-1-4.51):

$$\chi_1^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2_{\hat{\beta}_1}}, \qquad (5\text{-}1\text{-}1.36)$$

ma przy prawdziwości hipotezy zerowej (5-1-1.34) asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody równą 1. Z Raportu 1 SAS'a widać, że ze względu na stosunkowo małą wartość empirycznego poziom istotności $p = P(\chi_1^2 \geq \chi_{1\,obs}^2 = 9{,}6796) = 0{,}0019$ (tzn. $p = 0{,}0019 \leq \alpha$, np. dla $\alpha = 0{,}01$), można uznać

otrzymaną wartość oszacowania $\hat{\beta}_2 = 0{,}0270 > 0$ za istotnie statystycznie różną od zera. Uznajemy więc rozszerzenie modelu bez czynnika „*Czas*" do modelu wyższego (5-1.3) włączającego ten czynnik, za istotne statystycznie. Jednocześnie dodatnia wartość oszacowania $\hat{\beta}_2 > 0$ sugeruje, że ryzyko niespłacenia długu wzrasta z czasem.

Uwaga. Podana w Raporcie 1 wartość ilorazu szans odnosi się do *zmiany* wartości (ciągłego) czynnika „*Czas*" pomiędzy jednostką *A* i *B* ($\Delta Czas = Czas_A - Czas_B$) o jednostkę:

$$\hat{O}R_{(\Delta Czas = 1 rok | P_c, Lok)} = e^{1 \times \hat{\beta}_1} = e^{0{,}027} = 1{,}027, \qquad (5\text{-}1\text{-}1.37)$$

(gdzie pod kontrolą jest *ustalona* wartość czynnika „Płeć" oraz „Lokalizacja"). Dotyczy ona oszacowania ilorazu szansy niespłacenia długu dla wzrostu czasu (zatrudnienia) o jeden rok, co samo w sobie może nie być interesujące w badaniach ekonomicznych. Jednak w oparciu o to oszacowanie można podać oszacowanie ilorazu szansy niespłacenia długu dla wzrostu czasu (zatrudnienia) ($\Delta Czas$) o np. dziesięć lat. Zgodnie z podobną formułą do (4-1-3.40), to oszacowanie ilorazu szansy wynosi $\hat{O}R_{(\Delta Czas = 10 lat | P_c, Lok)}$ $= e^{10 \times \hat{\beta}_1} = e^{0{,}27} \approx 1{,}31$, a odpowiedni 95%-owy przedział ufności dla $OR_{(\Delta Czas = 10 lat | P_c, Lok)}$ jest równy:

$$\left(\exp(10 \times \hat{\beta}_1 - u_{1-\alpha/2}\, \hat{\sigma}_{10 \times \hat{\beta}_1}), \exp(10 \times \hat{\beta}_1 + u_{1-\alpha/2}\, \hat{\sigma}_{10 \times \hat{\beta}_1})\right) =$$
$$\left(\exp(10 \times \hat{\beta}_1 - u_{1-\alpha/2}\, 10 \times \hat{\sigma}_{\hat{\beta}_1}), \exp(10 \times \hat{\beta}_1 + u_{1-\alpha/2}\, 10 \times \hat{\sigma}_{\hat{\beta}_1})\right) = e^{10 \times (0{,}027 \pm 1{,}96 \times 0{,}00868)}$$
$$= \left(e^{0{,}99872}, e^{0{,}440128}\right) \approx (1{,}105,\ 1{,}553) \quad .$$

Ponieważ przedział ten nie obejmuje wartości 1, dlatego wpływ zmiany czasu o 10 lat jest istotny statystycznie dla wzrostu szansy niespłacenia długu. Dla dowolnej różnicy czynnika „*Czas*" wartość ilorazu szans (gdy pod kontrolą jest *ustalona* wartość czynnika „Płeć" oraz czynnika „Lokalizacja") oraz dolna i górna granica przedziału ufności dla $OR_{(\Delta Czas | Pc, Lok)}$, zmieniają się w kolejności, następująco:

$$\hat{O}R_{(\Delta Czas | Pc, Lok)} = e^{\Delta Czas \times \hat{\beta}_1} = e^{\Delta Czas \times 0{,}027}, \qquad (5\text{-}1\text{-}1.38)$$

oraz

$$e^{(\Delta Czas \times \hat{\beta}_1 \pm u_{1-\alpha/2}\, \hat{\sigma}_{\Delta Czas \times \hat{\beta}_1})} = e^{\Delta Czas \times (\hat{\beta}_1 \pm u_{1-\alpha/2}\, \hat{\sigma}_{\hat{\beta}_1})} = e^{\Delta Czas \times (0{,}027 \pm 1{,}96 \cdot 0{,}00868)} \quad . \qquad (5\text{-}1\text{-}1.39)$$
$$= \left(e^{\Delta Czas \times 0{,}09987},\ e^{\Delta Czas \times 0{,}044013}\right)$$

Numeryczne wartości powyższego ilorazu szans wraz z 95%-owymi przedziałami ufności, zostały podane w poniższej Tabeli.

Tabela 5-1-1.1. Wpływ różnicy czasu zatrudnienia (włączając w to czas na emeryturze) na oszacowaną wartość dopasowanego ilorazu szans niespłacenia długu.

Zmiana czasu zatrudnienia $\Delta Czas$	Oszacowanie punktowe, $\hat{O}R_{(\Delta Czas \mid Pc,Lok)} = e^{\Delta Czas \times \hat{\beta}_1}$	95%-owe przedziały ufnośc $e^{(\Delta Czas \times \hat{\beta}_1 \pm u_{1-\alpha/2}\, \hat{\sigma}_{\Delta Czas \times \hat{\beta}_1})}$ dla $OR_{(\Delta Czas \mid Pc,Lok)}$
10	1,310	(1,10, 1,56)
20	1,716	(1,22, 2,42)
30	2,248	(1,34, 3,77)
40	2,945	(1,48, 5,86)
50	3,857	(1,63, 9,12)
60	5,053	(1,8, 14,19)
70	6,619	(1,99, 22,08)
80	8,671	(2,19, 34,35)
85	9,924	(2,3, 42,85)

Wszystkie otrzymane 95%-owe przedziały ufności nie obejmują jedynki hipotezy zerowej dla $OR_{(\Delta Czas \mid Pc,Lok)}$ (co jest równoznaczne z odrzuceniem hipotezy zerowej $\mathsf{H}_0 : \beta_1 = 0$, (5-1-1.34)). Dla $\Delta Czas > 40$ lat szansa niespłacenia kredytu wzrasta ponad trzykrotnie. Jednakże dla $\Delta Czas > 40$ lat, dokładność oszacowania mocno spada wraz ze wzrostem różnicy czasu zatrudnienia.

Zadanie. Pokazać, że taki sam wniosek dla hipotezy (5-1-1.34) otrzymalibyśmy przeprowadzając test ilorazu wiarygodności odpowiednich modeli (chociaż numerycznie otrzymany empiryczny poziom istotności $p = 0{,}0015$ różniłyby się nieco od $p = 0{,}0019$).

Wniosek. W powiązaniu z wcześniejszą analizą, test dla hipotezy (5-1-1.34) utwierdza nas w przekonaniu o słuszności wyboru modelu zredukowanego (5-1-1.20) z czynnikiem głównym „Płeć" oraz kowariantem „*Czas*". Jednak wniosek ten *nie* uwzględnia np. ewentualnego, istotnie statystycznego wpływu interakcji.
Istotność statystyczna wprowadzonego kowarianta „*Czas*" oznacza (przy braku czynników typu „$Czas^2$" oraz interakcji „$Czas \times$ Płeć", włączających „*Czas*"), że liniowe wprowadzenie czynnika „*Czas*" jest (w świetle analizowanych danych) bardziej możliwe do przyjęcia, niż pominięcie go (które wynikałoby z nieodrzucenia hipotezy $\mathsf{H}_0 : \beta_1 = 0$,(5-1-1.34)). Dlatego też test odnoszący się do hipotezy $\mathsf{H}_0 : \beta_1 = 0$,(5-1-1.34), jest nazywany *testem trendu liniowego wpływu* czynnika „*Czas*".

Rozdział 5-2. Analiza interakcji głównego wpływu z kowariantami.

Włączenie interakcji.

Rozważmy dodanie do modelu (5-1.3) interakcji czynnika „*Czas*" i czynnika P_c (Płeć) w postaci iloczynu pomiędzy tymi zmiennymi:

$$CP_c = Czas \times P_c \qquad (5\text{-}2.40)$$

Zatem, forma logit'owa (4-1-3.20) modelu zawiera kolejno, zmienną *Czas*, zmienną P_c, zmienne L_1 i L_2 (wskazujące lokalizację), oraz człon interakcji CP_c:

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_2 \cdot P_c + \beta_3 \cdot L_1 + \beta_4 \cdot L_2 + \beta_5 \cdot CP_c \qquad (5\text{-}2.41)$$

Raport 4 dla modelu (5-2.41). Poniższy raport SAS'a dotyczy analizy powyższego modelu.

```
     Dlug_logit z interakcją CzasxPc                                    1
                                                                  21:37 Monday, February 10,
2014
                                     Procedura LOGISTIC

                                         Informacje

                    Zbiór                          DLUG_KIERUNKOWE
                    Zmienna objaśniana             Dlug                      Dlug
                    Liczba poziomów odpowiedzi     2
                    Model                          logit binarny
                    Technika optymalizacji         Ocena Fishera

                              Wczytano obserwacji                 196
                              Użyto obserwacji                    196


                                      Profil odpowiedzi

                                 Wartość                   Całkowita
                            uporządkowana         Dlug     liczebność

                                        1            1             57
                                        2            0            139

                         Modelowane prawdopodobieństwo wynosi Dlug=1.

                                       Status zbieżności

                         Kryterium zbieżności (GCONV=1E-8) spełnione.


                                   Statystyki dopasowania

                                            Tylko      Wyraz wolny
                                            wyraz                i
                              Kryterium     wolny     współzmienne

                              AIC         238.329          222.872
                              SC          241.607          242.541
                              -2 log L    236.329          210.872

                        Testowanie globalnej hipotezy zerowej: BETA=0

                                                   St.
                    Test                Chi-kwadrat      sw.     Pr. > chi-kw.

                    Iloraz wiarygod.        25.4570        5            0.0001
                    Ocena                   26.1643        5            <.0001
                    Wald                    22.2486        5            0.0005
```

Dlug_logit z interakcją CxPc

2

21:37 Monday, February 10, 2014

Procedura LOGISTIC

Analiza ocen maksymalnej wiarygodności

Parametr	St. sw.	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.
Intercept	1	-1.8785	0.4687	16.0624	<.0001
Czas	1	0.0216	0.0126	2.9708	0.0848
Pc	1	**0.9635**	0.5902	2.6647	0.1026
L1	1	-0.2787	0.4087	0.4648	0.4954
L2	1	-0.2341	0.4568	0.2626	0.6083
CPc	1	**0.0103**	0.0176	0.3447	0.5571

Oceny ilorazu szans

Efekt	Ocena punktowa	Przedział ufności Walda 95%	
Czas	1.022	0.997	1.047
Pc	2.621	0.824	8.334
L1	0.757	0.340	1.686
L2	0.791	0.323	1.937
CPc	1.010	0.976	1.046

Przedział ufności Walda dla parametrów

Parametr	Ocena	Przedział ufności 95%	
Intercept	-1.8785	-2.7972	-0.9599
Czas	0.0216	-0.00297	0.0462
Pc	0.9635	-0.1933	2.1203
L1	-0.2787	-1.0797	0.5224
L2	-0.2341	-1.1294	0.6612
CPc	0.0103	-0.0242	0.0448

Przedział ufności Walda dla ilorazów szans

Efekt	Jednostka	Ocena	Przedział ufności 95%	
Czas	1.0000	1.022	0.997	1.047
Pc	1.0000	2.621	0.824	8.334
L1	1.0000	0.757	0.340	1.686
L2	1.0000	0.791	0.323	1.937
CPc	1.0000	1.010	0.976	1.046

Dlug_logit z interakcją CxPc

3

21:37 Monday, February 10, 2014

Procedura LOGISTIC

Macierz kowariancji szacunkowych

Parametr	Intercept	Czas	Pc	L1	L2	CPc
Intercept	0.219696	-0.00447	-0.16565	-0.07349	-0.08414	0.004709
Czas	-0.00447	0.000158	0.004258	0.000145	0.000506	-0.00016
Pc	-0.16565	0.004258	**0.348362**	-0.01915	-0.01591	-0.00832
L1	-0.07349	0.000145	-0.01915	0.167054	0.093327	-0.00075
L2	-0.08414	0.000506	-0.01591	0.093327	0.208678	-0.00075
CPc	0.004709	-0.00016	**-0.00832**	-0.00075	-0.00075	**0.00031**

Korzystając z powyższego raportu możemy wyznaczyć iloraz szans niespłacenia długu przez kobiety w stosunku do mężczyzn (zatem czynnikiem głównym jest P_c (Płeć)), podczas gdy pod kontrolą są zmienne

Czas oraz „Lokalizacja” (a ich wartości dla jednostek A i B są takie same). Skorzystajmy z ogólnej postaci dopasowanego ilorazu szans, zestawiającego szanse osoby A i B:

$$OR_{X_A vs. X_B} = \exp\left(\sum_{j=1}^{k} \beta_j (x_{jA} - x_{jB}) \right) \quad . \qquad (4\text{-}1\text{-}3.23')$$

Analizując model (5-2.41) z interakcją musimy wziąć pod uwagę człon iloczynowy $CP_c = Czas \times P_c$.

Ponieważ zmienna P_c przyjmuje wartość 1 lub 0, zatem zmienna CP_c przyjmuje wartość $CP_{cA} = Czas_A \times 1$, gdy osoba A jest kobietą ($P_c = 1$) oraz $CP_{cB} = Czas_B \times 0 = 0$, gdy osoba B jest mężczyzną ($P_c = 0$). Stąd wykładnik w (4-1-3.23’) ma postać:

$$\begin{aligned} &\sum_{j=1}^{k=5} \beta_j (x_{jA} - x_{jB}) = \\ &= \beta_1 (Czas_A - Czas_B) + \beta_2 (P_{cA} - P_{cB}) + \beta_3 (L_{1A} - L_{1B}) + \beta_4 (L_{2A} - L_{2B}) + \beta_5 (CP_{cA} - CP_{cB}), \end{aligned} \qquad (5\text{-}2.42)$$

która dla ustalonej wartości czynnika „*Czas*”, $Czas_A = Czas_B = Czas$, i „Lokalizacji”, $L_{1A} = L_{1B} = L_1$, $L_{2A} = L_{2B} = L_2$, jest następująca:

$$\begin{aligned} &\sum_{j=1}^{k=5} \beta_j (x_{jA} - x_{jB}) = \\ &= \beta_1 (Czas - Czas) + \beta_2 (1-0) + \beta_3 (L_1 - L_1) + \beta_4 (L_2 - L_2) + \beta_5 (Czas - 0) \\ &= \beta_2 + \beta_5 \times Czas \quad . \end{aligned} \qquad (5\text{-}2.43)$$

Ponieważ z Raportu 4 odczytujemy następujące wartości oszacowań parametrów β_2 i β_5:

$$\hat{\beta}_2 = 0{,}9635 \ \text{ oraz } \ \hat{\beta}_5 = 0{,}0103 \ , \qquad (5\text{-}2.44)$$

zatem oszacowanie interesującego nas ilorazu szans w pobranej próbce wynosi:

$$\hat{O}R_{(Pc=1\, vs.\, Pc=0\,|\,Czas, Lok)} = e^{\hat{\beta}_2 + \hat{\beta}_5 \times Czas} = e^{0{,}9635 \,+\, 0{,}0103 \times Czas} \ . \qquad (5\text{-}2.45)$$

Wynik ten oznacza, że w przypadku interakcji czynników „Płeć” i „*Czas*”, iloraz szans zależy od konkretnej wartości czasu zatrudnienia. Zatem czynnik czasu zatrudnienia (włączając w to czas na emeryturze) jest wpływem modyfikującym związek pomiędzy płcią a szansą niespłacenia długu (kredytu). Otrzymany wynik jest zilustrowany w poniższej Tabeli 5-1-1.2.

Wyznaczmy 95%-owe przedziały ufności dla $OR_{(Pc=1\, vs.\, Pc=0\,|\,Czas, Lok)} = e^{\beta_2 + \beta_5 \times Czas}$:

$$\exp\left(\hat{L} \pm u_{1-\alpha/2}\ \hat{\sigma}(\hat{L})\right) \equiv \exp\left((\hat{\beta}_2 + \hat{\beta}_5 \times Czas) \pm u_{1-\alpha/2}\ \hat{\sigma}(\hat{\beta}_2 + \hat{\beta}_5 \times Czas)\right) . \qquad (5\text{-}2.46)$$

Ponieważ dla zmiennych losowych X i Y oraz stałych a i b, zachodzi (pokazać):

$$\hat{\sigma}^2 (aX + bY) = a^2 \hat{\sigma}^2 (X) + 2\, a\, b \cdot \hat{\text{cov}}(X, Y) + b^2\ \hat{\sigma}^2 (Y), \qquad (5\text{-}2.47)$$

zatem odchylenie standardowe $\hat{\sigma}(\hat{L}) = \sqrt{\hat{\sigma}^2(\hat{L})}$ estymatora $\hat{L} \equiv \hat{\beta}_2 + \hat{\beta}_5 \times Czas$ wyznaczamy korzystając ze związku:

$$\hat{\sigma}^2(\hat{L}) = \hat{\sigma}^2(\hat{\beta}_2 + Czas\ \hat{\beta}_5) = \hat{\sigma}^2(\hat{\beta}_2) + 2\ Czas \cdot \hat{\text{cov}}(\hat{\beta}_2, \hat{\beta}_5) + Czas^2\ \hat{\sigma}^2(\hat{\beta}_5) \qquad (5\text{-}2.48)$$

Oszacowania dla $\sigma^2(\hat{\beta}_2)$, $\text{cov}(\hat{\beta}_2, \hat{\beta}_5)$ oraz $\sigma^2(\hat{\beta}_5)$ podane w powyższym raporcie SAS'a wynoszą $\hat{\sigma}^2(\hat{\beta}_2) = 0{,}34836$, $\hat{\text{cov}}(\hat{\beta}_2, \hat{\beta}_5) = -0.00832$, $\hat{\sigma}^2(\hat{\beta}_5) = 0{,}00031$, skąd:

$$\hat{\sigma}^2(\hat{L}) = 0{,}34836 - 2 \cdot 0{,}00832 \cdot Czas + 0{,}00031 \cdot Czas^2 \ . \qquad (5\text{-}2.49)$$

Aby aktywować ich wyznaczenie, w opcjach Analyst'a analizy logistycznej SAS'a otwieramy okno wyboru modelu i po kliknięciu opcji „Statistics”, w zakładce „Statistics” zaznaczamy możliwość wyboru „Covariance matrix of estimates” (Macierz kowariancji dla estymatorów (2-2-8.45)) dla estymatorów parametrów β_j.

Numeryczne wartości powyższego ilorazu szans wraz z 95%-owymi przedziałami ufności, zostały podane w poniższej Tabeli.

Tabela 5-1-1.2. Wpływ czasu zatrudnienia (włączając w to czas na emeryturze) na oszacowaną wartość dopasowanego ilorazu szans niespłacenia długu kobiet względem mężczyzn w przypadku interakcji czynników „Płeć” i „*Czas*”.

Zmiana czasu zatrudnienia *Czas*	Oszacowanie punktowe, $\hat{OR}_{(Pc=1\ vs.\ Pc=0\,|\,Czas,Lok)} = e^{\hat{\beta}_2 + \hat{\beta}_5 \times Czas}$	95%-owe przedziały ufności $\exp\left(\hat{L} \pm u_{1-\alpha/2}\ \hat{\sigma}(\hat{L})\right)$ dla $OR_{(Pc=1\ vs.\ Pc=0\,|\,Czas,Lok)}$
10	2,91	(1,18, 7,18)
20	3,22	(1,55, 6,70)
30	3,57	(1,77, 7,20)
40	3,96	(1,73, 9,06)
50	4,39	(1,52, 12,63)
60	4,86	(1,28, 18,53)
70	5,39	(1,04, 27,86)
80	5,97	(0,84, 42,46)
85	6,29	(0,75, 52,58)

Aż do wartości $Czas < 73$ lat zatrudnienia (łącznie z okresem emerytalnym), wszystkie otrzymane 95%-owe przedziały ufności nie obejmują jedynki hipotezy zerowej dla dopasowanego ilorazu szans $OR_{(Pc=1\ vs.\ Pc=0\,|\,Czas,Lok)}$. Zatem dla tych lat zatrudnienia, na poziomie istotności α=0,05, dopasowany iloraz szans różni się istotnie statystycznie od 1. Niestety dla wartości $Czas > 40$ lat, dokładność oszacowania ilorazu szans wraz ze wzrostem czasu zatrudnienia mocno spada.

Poniżej pokażemy, że rozszerzenie modelu (5-1-1.20) bez kowarianta „Lokalizacja” oraz bez interakcji czynników „Płeć” i „*Czas*” do omawianego, wyższego modelu () z interakcją, jest nieistotne statystycznie z punktu widzenia dopasowania modelu do danych empirycznych. Zatem dodawanie nowych czynników do modelu (5-1-1.20) mogłoby być jednak nieuzasadnione, również z powodu pewnego wzrostu niestabilności

estymatorów parametrów w modelu rozszerzonym (Wniosek ten płynie z ogólnych rozważań nad konsekwencjami nierówności Rao-Cramera [5]). Jednakże wstrzymamy się na razie z podjęciem ostatecznej decyzji co do wyboru modelu z (5-1-1.20) bez interakcji. Powodem jest zaobserwowany znaczny wzrost dopasowanego ilorazu szans niespłacenia długu kobiet względem mężczyzn w przypadku interakcji czynników „Płeć" i „*Czas*" na skutek modyfikującego wpływu czasu zatrudnienia.

Rozważmy zatem istotność statystyczną rozszerzenia poprzednio wyselekcjonowanego modelu niższego z $r = 2$ i bez interakcji:

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_2 \cdot P_c \qquad (5\text{-}1\text{-}1.20')$$

do modelu wyższego (5-2.41) z $\beta \equiv \beta_{(k=5)} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ z $k = 5$.

Testowana hipoteza ma postać (5-1-1.12). Można ją ująć następująco:

$$\mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0, \qquad (5\text{-}2.50)$$

a hipoteza alternatywna to:

$$\mathrm{H}_1 : \beta_3 \neq 0 \vee \beta_4 \neq 0 \vee \beta_5 \neq 0 \quad . \qquad (5\text{-}2.51)$$

Powyższa hipoteza zerowa H_0 oznacza statystyczną nieistotność zarówno kowarianta „Lokalizacja" jak i interakcji czynników „Płeć" i „*Czas*". Jest ona równocześnie hipotezą o nie występowaniu braku dopasowania do danych empirycznych w modelu zredukowanym (5-1-1.20) w porównaniu z modelem (5-2.41).

Statystyka testowa (5-1-1.13) służąca do weryfikacji hipotezy (5-2.50) ma postać:

$$L_{2/5} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r=2)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}_{(k=5)}\right)}\right] = -2\ln P(\tilde{Y} \mid \hat{\beta}_{(r=2)}) - \left(-2\ln P(\tilde{Y} \mid \hat{\beta}_{(k=5)})\right), \qquad (5\text{-}2.52)$$

Statystyka $L_{2/5}$ ma przy prawdziwości hipotezy zerowej (5-2.50) asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody *l.st.sw.*=*k*–*r*=5–2=3. Dla modelu (5-1-1.20), otrzymaliśmy w próbce $-2\ln P\left(\tilde{Y}|\hat{\beta}_{(r=2)}\right) = 211{,}639$, (5-1-1.22), a z Raportu 4 dla modelu (5-2.41) z interakcją otrzymaliśmy $-2\ln P\left(\tilde{Y}|\hat{\beta}_{(k=5)}\right) = 210{,}872$. Obserwowana w próbce wartość statystyki ilorazu wiarygodności wynosi więc:

$$L_{2/5} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r=2)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}_{(k=5)}\right)}\right] = 211{,}639 - 210{,}872 = 0{,}767, \qquad (5\text{-}2.53)$$

a odpowiadający jej empiryczny poziom istotności jest równy:

$$p = P(LR_{2/5} \sim \chi_3^2 \geq 0{,}767\,) \approx 0{,}857\,, \qquad (5\text{-}2.54)$$

Zatem, na każdym poziomie istotności $\alpha < p = 0{,}857$ nie ma podstaw do odrzucenia hipotezy H_0 (5-2.50) o nie występowaniu braku dopasowania do danych empirycznych modelu zredukowanego (5-1-1.20) w porównaniu z modelem (5-2.41), w którym występował zarówno czynnik „Lokalizacja" jak i człon interakcji CP_c czynników „Płeć" i „*Czas*". Jak powyżej wspomniano, z tego punktu widzenia (i tym razem) można by

się zdecydować na wybór modelu zredukowanego (5-1-1.20) jako prawie tak samo dobrze dopasowującego się do danych empirycznych jak model wyższy (5-2.41).

Porównajmy model zredukowany (5-1-1.20) z modelem rozszerzonym o człon iloczynowy $CP_c = Czas \times P_c$:

$$\text{logit}[pr(Y=1)] = \ln odds(Y=1) = \beta_0 + \beta_1 \cdot Czas + \beta_2 \cdot P_c + \beta_5 \cdot CP_c \,. \qquad (5\text{-}2.55)$$

Pominięto w nim zmienne L_1 i L_2 wskazujące lokalizację. Dlatego analiza przebiega jak dla modelu (5-2.41) tyle, że z konsekwentnie należy pominąć we wzorach z obszaru (5-2.41)-(5-2.54) zmienne L_1 i L_2.

Raport 5 dla modelu (5-2.55). Odpowiedni raport SAS'a ma postać:

```
Dlug_logit bez Lokalizacji, z interakcją CPc                        1
                                                   23:46 Tuesday, February 11, 2014

                              Procedura LOGISTIC

                                 Informacje

                Zbiór                          JACEK.DLUG_KIERUNKOWE
                Zmienna objaśniana             Dlug                      Dlug
                Liczba poziomów odpowiedzi     2
                Model                          logit binarny
                Technika optymalizacji         Ocena Fishera

                     Wczytano obserwacji                 196
                     Użyto obserwacji                    196

                                 Profil odpowiedzi

                           Wartość                   Całkowita
                     uporządkowana         Dlug     liczebność

                                 1            1             57
                                 2            0            139

                Modelowane prawdopodobieństwo wynosi Dlug=1.

                                Status zbieżności

                Kryterium zbieżności (GCONV=1E-8) spełnione.

                            Statystyki dopasowania

                                      Tylko      Wyraz wolny
                                      wyraz               i
                     Kryterium        wolny     współzmienne

                     AIC            238.329          219.376
                     SC             241.607          232.489
                     -2 log L       236.329          211.376


               Testowanie globalnej hipotezy zerowej: BETA=0

                                                St.
                Test                 Chi-kwadrat      sw.     Pr. > chi-kw.

                Iloraz wiarygod.         24.9530        3            <.0001
                Ocena                    25.6877        3            <.0001
                Wald                     21.8992        3            <.0001
                         Dlug_logit bez Lokalizacji, z interakcją CPc
2
                                                                23:46 Tuesday, February 11,
2014
```

Procedura LOGISTIC

Analiza ocen maksymalnej wiarygodności

Parametr	St. sw.	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.
Intercept	1	-2.0296	0.4209	23.2472	<.0001
Czas	1	0.0222	0.0125	3.1172	0.0775
Plec	1	0.9326	0.5898	2.5002	0.1138
CPc	1	0.00894	0.0175	0.2609	0.6095

Oceny ilorazu szans

Efekt	Ocena punktowa	Przedział ufności Walda 95%	
Czas	1.022	0.998	1.048
Pc	2.541	0.800	8.074
CPc	1.009	0.975	1.044

Przedział ufności Walda dla parametrów

Parametr	Ocena	Przedział ufności 95%	
Intercept	-2.0296	-2.8546	-1.2046
Czas	0.0222	-0.00244	0.0467
Pc	0.9326	-0.2234	2.0887
CPc	0.00894	-0.0254	0.0433

Przedział ufności Walda dla ilorazów szans

Efekt	Jednostka	Ocena	Przedział ufności 95%	
Czas	1.0000	1.022	0.998	1.048
Pc	1.0000	2.541	0.800	8.074
CPc	1.0000	1.009	0.975	1.044

Dlug_logit bez Lokalizacji, z interakcją CPc

3

23:46 Tuesday, February 11, 2014

Procedura LOGISTIC

Macierz kowariancji szacunkowych

Parametr	Intercept	Czas	Pc	CPc
Intercept	0.177195	-0.00434	-0.17719	0.004339
Czas	-0.00434	0.000157	0.004339	-0.00016
Pc	-0.17719	0.004339	**0.347893**	**-0.00847**
CPc	0.004339	-0.00016	-0.00847	**0.000307**

Testowana hipoteza ma postać:

$$H_0 : \beta_5 = 0, \tag{5-2.56}$$

a hipoteza alternatywna:

$$H_1 : \beta_5 \neq 0 \quad . \tag{5-2.57}$$

Powyższa hipoteza zerowa oznacza statystyczną nieistotność interakcji czynników „Płeć” i „*Czas*”. Statystyka testowa (5-1-1.13) służąca do weryfikacji hipotezy (5-2.56) ma postać:

$$L_{2/3} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r=2)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}_{(k=3)}\right)}\right] = -2\ln P(\tilde{Y} \mid \hat{\beta}_{(r=2)}) - \left(-2\ln P(\tilde{Y} \mid \hat{\beta}_{(k=3)})\right), \qquad (5\text{-}2.58)$$

Statystyka $L_{2/3}$ ma przy prawdziwości hipotezy zerowej (5-2.56) asymptotycznie rozkład chi-kwadrat z liczbą stopni swobody *l.st.sw.*=*k*–*r*=3–2=1. Dla modelu (5-1-1.20), otrzymaliśmy w próbce $-2\ln P\left(\tilde{Y}|\hat{\beta}_{(r=2)}\right) = 211{,}639$, (5-1-1.22), a z Raportu 5 dla modelu (5-2.55) otrzymaliśmy $-2\ln P\left(\tilde{Y}|\hat{\beta}_{(k=3)}\right)$ $= 211{,}376$. Obserwowana w próbce wartość statystyki ilorazu wiarygodności wynosi więc:

$$L_{2/5} = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r=2)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}_{(k=3)}\right)}\right] = 211{,}639 - 211{,}376 = 0{,}263, \qquad (5\text{-}2.59)$$

a odpowiadający jej empiryczny poziom istotności jest równy:

$$p = P(LR_{2/3} \sim \chi_1^2 \geq 0{,}263) \approx 0{,}608, \qquad (5\text{-}2.60)$$

Zatem, na żadnym poziomie istotności $\alpha < p = 0{,}608$ nie ma podstaw do odrzucenia hipotezy H_0 (5-2.56) o niewystępowaniu braku dopasowania do danych empirycznych modelu zredukowanego (5-1-1.20) w porównaniu z modelem (5-2.55), w którym występował człon interakcji czynników „Płeć" i „*Czas*". Z tego punktu widzenia należałoby zdecydować się na wybór modelu zredukowanego (5-1-1.20), bez interakcji.

Z Raportu 5 odczytujemy $\hat{\sigma}^2(\hat{\beta}_2) = 0{,}34789$, $\hat{\text{cov}}(\hat{\beta}_2, \hat{\beta}_5) = -0{,}00847$, $\hat{\sigma}^2(\hat{\beta}_5) = 0{,}00031$, skąd:

$$\hat{\sigma}^2(\hat{L}) = 0{,}34789 - 2 \cdot 0{,}00847 \cdot Czas + 0{,}00031 \cdot Czas^2 \qquad (5\text{-}2.61)$$

Numeryczne wartości powyższego ilorazu szans wraz z 95%-owymi przedziałami ufności, zostały podane w poniższej Tabeli.

Tabela 5-1-1.3. Wpływ czasu zatrudnienia (włączając w to czas na emeryturze) na oszacowaną wartość dopasowanego ilorazu szans niespłacenia długu kobiet względem mężczyzn w przypadku interakcji czynników „Płeć" i „*Czas*".

Zmiana czasu zatrudnienia *Czas*	Oszacowanie punktowe, $\hat{OR}_{(Pc=1\, vs.\, Pc=0 \mid Czas)} = e^{\hat{\beta}_2 + \hat{\beta}_5 \times Czas}$	95%-owe przedziały ufności $\exp\left(\hat{L} \pm u_{1-\alpha/2}\, \hat{\sigma}(\hat{L})\right)$ dla $OR_{(Pc=1\, vs.\, Pc=0 \mid Czas)}$
10	2,78	(1,13, 6,81)
20	3,04	(1,49, 6,19)
30	3,32	(1,70, 6,48)
40	3,63	(1,65, 7,99)
50	3,97	(1,44, 10,97)
60	4,34	(1,19, 15,87)
70	4,75	(0,96, 23,53)
80	5,20	(0,76, 35,36)
85	5,43	(0,68, 43,48)

Dla wartości $Czas > 40$ lat, dokładność oszacowania ilorazu szans wraz ze wzrostem czasu zatrudnienia spada, aczkolwiek nieco wolnie niż dla modelu (). Aż do wartości $Czas < 69$ lat zatrudnienia (łącznie z

okresem emerytalnym), wszystkie otrzymane 95%-owe przedziały ufności nie obejmują jedynki hipotezy zerowej dla dopasowanego ilorazu szans $OR_{(Pc=1\,vs.\,Pc=0\,|\,Czas)}$. Zatem dla tych lat zatrudnienia, na poziomie istotności α=0,05, dopasowany iloraz szans różni się istotnie statystycznie od 1.

Zwróćmy uwagę, że dla modelu (5-1-1.20) otrzymaliśmy oszacowanie dopasowanego, ogólnego ilorazu szans równe $\hat{O}R_{(Pc=1\,vs.\,Pc=0\,|\,Czas)} = \frac{odds(grupa\ \text{kobiet})}{odds(grupa\ \text{mezczyzn})} = 3{,}26$, (5-1-1.26). Jego wartość nie jest modyfikowana przez czynnik czasu zatrudnienia. Fakt ten w sposób zasadniczy odróżnia własności modelu zredukowanego (5-1-1.20) od modelu (5-2.55) z interakcją czynników „Płeć" i „*Czas*". Chociaż dopasowanie modelu z interakcją nie jest istotnie statystycznie lepsze niż modelu bez interakcji ($p \approx 0{,}608$), to wyniki zawarte w powyższej Tabeli sugerują, że ze względu na *znaczącą różnicę* wartości ilorazu szans dla różnych wariantów czasu zatrudnienia, pominięcie modyfikującego wpływu czasu zatrudnienia redukuje użyteczność modelu bez interakcji do okresu 20-35 lat zatrudnienia. (Nie wykonuje się testu statystycznego dotyczącego „znaczącej różnicy", a na temat wielkości „znaczącej różnicy" powinni wypowiedzieć się eksperci z branży).

Rozdział 5-3. Dane dla przykładu z Rozdziału 5 „Spłata długu”.

Opis wszystkich zmiennych znajduje sie na początku Rozdziału 5. Kolumna dla zmiennej „Lokalizacja” (*Lok*) jest dana jedynie dla celów łatwej lokalizacji dzielnicy miasta i w prezentowanej analizie SAS'a nie bierze udziału.

Nr	*Dlug*	*Czas*	*Pc*	*L1*	*L2*	*CPc*	*Lok*
1	0	33	0	1	0	0	1
2	0	35	0	1	0	0	1
3	0	6	0	1	0	0	1
4	0	60	0	1	0	0	1
5	1	18	0	0	0	0	3
6	0	26	0	0	0	0	3
7	0	6	0	0	0	0	3
8	1	31	0	0	1	0	2
9	1	26	0	0	1	0	2
10	0	37	0	0	1	0	2
11	0	23	0	1	0	0	1
12	0	23	0	1	0	0	1
13	0	27	0	1	0	0	1
14	1	9	0	1	0	0	1
15	1	37	1	1	0	37	1
16	1	22	1	1	0	22	1
17	1	67	1	1	0	67	1
18	0	8	1	1	0	8	1
19	1	6	1	1	0	6	1
20	1	15	1	1	0	15	1
21	1	21	1	0	1	21	2
22	1	32	1	0	1	32	2
23	1	16	1	1	0	16	1
24	0	11	1	0	1	11	2
25	0	14	1	0	0	14	3
26	0	9	1	0	1	9	2
27	0	18	1	0	1	18	2
28	0	2	0	0	0	0	3
29	0	61	0	0	0	0	3
30	0	20	0	0	0	0	3
31	0	16	0	0	0	0	3
32	0	9	0	0	1	0	2
33	0	35	0	0	1	0	2
34	0	4	0	1	0	0	1
35	0	44	1	0	0	44	3
36	1	11	1	0	0	11	3
37	0	3	1	0	1	3	2
38	0	6	1	0	0	6	3
39	1	17	1	0	1	17	2
40	0	1	1	0	0	1	3
41	1	53	1	0	1	53	2
42	1	13	1	1	0	13	1
43	0	24	1	1	0	24	1

44	1	70	1	1	0	70	1
45	1	16	1	0	0	16	3
46	0	12	1	0	1	12	2
47	1	20	1	0	0	20	3
48	0	65	1	0	0	65	3
49	1	40	1	0	1	40	2
50	1	38	1	0	1	38	2
51	1	68	1	0	1	68	2
52	1	74	1	1	0	74	1
53	1	14	1	1	0	14	1
54	1	27	1	1	0	27	1
55	0	31	1	1	0	31	1
56	0	18	1	1	0	18	1
57	0	39	1	1	0	39	1
58	0	50	1	1	0	50	1
59	0	31	1	1	0	31	1
60	0	61	1	1	0	61	1
61	0	18	0	0	0	0	3
62	0	5	0	0	0	0	3
63	0	2	0	0	0	0	3
64	0	16	0	0	0	0	3
65	1	59	0	0	0	0	3
66	0	22	0	0	0	0	3
67	0	24	0	1	0	0	1
68	0	30	0	1	0	0	1
69	0	46	0	1	0	0	1
70	0	28	0	1	0	0	1
71	0	27	0	1	0	0	1
72	1	27	0	1	0	0	1
73	0	28	0	1	0	0	1
74	1	52	0	1	0	0	1
75	0	11	0	0	0	0	3
76	0	6	0	0	1	0	2
77	0	46	0	0	0	0	3
78	1	20	0	0	1	0	2
79	0	3	0	1	0	0	1
80	0	18	0	0	1	0	2
81	0	25	0	0	1	0	2
82	0	6	0	0	0	0	3
83	1	65	0	0	0	0	3
84	0	51	0	0	0	0	3
85	0	39	0	0	1	0	2
86	0	8	0	1	0	0	1
87	0	8	0	0	1	0	2
88	0	14	0	0	0	0	3
89	0	6	0	0	0	0	3
90	0	6	0	0	0	0	3
91	0	7	0	0	0	0	3
92	0	4	0	0	0	0	3
93	0	8	0	0	0	0	3
94	0	9	0	0	1	0	2
95	1	32	0	0	0	0	3
96	0	19	0	0	0	0	3
97	0	11	0	0	0	0	3

98	0	35	0	0	0	0	3
99	0	16	0	1	0	0	1
100	0	1	0	1	0	0	1
101	0	6	0	1	0	0	1
102	0	27	0	1	0	0	1
103	0	25	0	1	0	0	1
104	0	18	0	1	0	0	1
105	0	37	0	0	0	0	3
106	1	33	0	0	0	0	3
107	0	27	0	0	1	0	2
108	0	2	0	1	0	0	1
109	0	8	0	0	1	0	2
110	0	5	0	1	0	0	1
111	0	1	0	1	0	0	1
112	0	32	0	1	0	0	1
113	1	25	0	1	0	0	1
114	0	15	1	1	0	15	1
115	0	15	1	1	0	15	1
116	0	26	1	1	0	26	1
117	1	42	1	1	0	42	1
118	0	7	1	1	0	7	1
119	0	2	1	1	0	2	1
120	1	65	1	1	0	65	1
121	0	33	1	0	1	33	2
122	1	8	1	0	1	8	2
123	0	30	1	0	1	30	2
124	0	5	1	0	0	5	3
125	0	15	1	0	0	15	3
126	1	60	1	0	0	60	3
127	1	13	1	0	0	13	3
128	0	70	0	0	0	0	3
129	0	5	0	0	0	0	3
130	0	3	0	0	0	0	3
131	0	50	0	0	1	0	2
132	0	6	0	0	1	0	2
133	0	12	0	0	1	0	2
134	1	39	1	0	0	39	3
135	0	15	1	0	1	15	2
136	1	35	1	0	1	35	2
137	0	2	1	0	1	2	2
138	0	17	1	0	0	17	3
139	1	43	1	0	0	43	3
140	0	30	1	0	1	30	2
141	0	11	1	1	0	11	1
142	1	39	1	1	0	39	1
143	0	32	1	1	0	32	1
144	0	17	1	1	0	17	1
145	0	3	1	0	0	3	3
146	0	7	1	0	0	7	3
147	0	2	1	0	1	2	2
148	1	64	1	0	1	64	2
149	1	13	1	1	0	13	1
150	1	15	1	0	1	15	2
151	0	48	1	0	1	48	2

152	0	23	1	1	0	23	1
153	1	48	1	1	0	48	1
154	0	25	1	1	0	25	1
155	0	12	1	1	0	12	1
156	1	46	1	1	0	46	1
157	0	79	1	1	0	79	1
158	0	56	1	1	0	56	1
159	0	8	1	1	0	8	1
160	1	29	0	0	0	0	3
161	1	35	0	0	0	0	3
162	1	11	0	0	0	0	3
163	0	69	0	0	0	0	3
164	1	21	0	0	0	0	3
165	0	13	0	0	0	0	3
166	0	21	0	1	0	0	1
167	1	32	0	1	0	0	1
168	1	24	0	1	0	0	1
169	0	24	0	1	0	0	1
170	0	73	0	1	0	0	1
171	0	42	0	1	0	0	1
172	1	34	0	1	0	0	1
173	0	30	0	0	1	0	2
174	0	7	0	0	1	0	2
175	1	29	0	0	0	0	3
176	1	22	0	0	0	0	3
177	0	38	0	0	1	0	2
178	0	13	0	0	1	0	2
179	0	12	0	0	1	0	2
180	0	42	0	0	0	0	3
181	1	17	0	0	0	0	3
182	0	21	0	0	0	0	3
183	0	34	0	1	0	0	1
184	0	1	0	0	0	0	3
185	0	14	0	0	1	0	2
186	0	16	0	0	1	0	2
187	0	9	0	0	0	0	3
188	0	53	0	0	0	0	3
189	0	27	0	0	0	0	3
190	0	15	0	0	0	0	3
191	0	9	0	0	0	0	3
192	0	4	0	0	1	0	2
193	0	10	0	0	0	0	3
194	0	31	0	0	0	0	3
195	0	85	0	0	0	0	3
196	0	24	0	0	1	0	2

A. Rozdział 6. Podsumowanie regresji logistycznej.

Powyższe rozważania dotyczyły zastosowania MNW w analizie regresji logistycznej, która umożliwia wyciąganiu wniosków z danych empirycznych w przypadku, gdy zmienna objaśniana ma charakter dychotomiczny, dając oszacowanie siły i kierunku zależności pomiędzy odpowiedzią, a niemal każdym typem czynnika. Estymatory MNW posiadają *asymptotycznie* własność zgodności i efektywności oraz zbieżność do rozkładu normalnego. Ponadto, jeśli istnieje estymator dostateczny parametru (tzn. niosący taką samą informację na temat parametru co cała próba), to jest to estymator MNW. Zastosowano estymację MNW do szacowania współczynników regresji logistycznej, wykorzystującą bezwarunkową (pełną) funkcję wiarygodności.

Podano założenia modelowe regresji logistycznej, począwszy od podania intuicyjnego wprowadzenia funkcji logistycznej do opisu prawdopodobieństwa zajścia sukcesu. Prawdopodobieństwo to jest równe wartości oczekiwanej zmiennej dychotomicznej. Podstawową wielkością, dla której zapisuje się równanie regresji jest szansa (odds) zajścia zdarzenia. Logit'owa postać dla szansy jest modelowana jako liniowa kombinacja czynników, z których niektóre mogą być czynnikami głównymi, a niektóre pobocznymi, wziętymi pod kontrolę w celu otrzymania dopasowanego oszacowania podstawowego parametru analizy, jakim jest iloraz szans. W praktyce logistycznej, przy dokonywaniu porównania dwóch grup (lub jednostek), wariant czynnika zasadniczego zmienia się pomiędzy porównywanymi grupami, podczas gdy kontrolowane czynniki poboczne (kowarianty) mają wartości ustalone. Ze względu na postać modelu logistycznego, okazuje się, że przeprowadzenie weryfikacji hipotezy o nieistotności różnicy ilorazu szans od wartości jeden (będącej zerową wartością hipotezy zerowej) pociąga za sobą testowanie hipotez o nieistotności współczynników kierunkowych stojących przy zmiennych uważanych w analizie jako zasadnicze. Przy założeniu dużej próby, wykonano testy tych hipotez wykorzystując statystykę Wald'a oraz statystykę chi-kwadrat Wald'a. Natomiast testy nieistotności statystycznej występowania braku dopasowania do danych empirycznych modeli niższych w hierarchii w stosunku do modeli wyższych wykonano stosując statystykę ilorazu wiarygodności, która dla dużej wielkości próby ma w przybliżeniu rozkład chi-kwadrat z liczbą stopni swobody będącą różnicą liczby parametrów porównywanych modeli. W końcu podkreślono fakt, że ze względu na różnorodność zastosowanych kryteriów doboru modelu oraz cel prowadzonej analizy, wybór odpowiedniego modelu jest efektem kompromisu pomiędzy wyborem modelu o najprostszej możliwej strukturze (z punktu widzenia niewystępowania braku dopasowania), a nie utraceniem ważnych dla branży informacji niesionych przez model, związanych np. z pojęciem znaczącej różnicy eksperckiej.

Chociaż regresja logistyczna znalazła głównie zastosowanie w badaniach medycznych, co wynika z charakteru dychotomicznego zmiennej objaśnianej, to problem dychotomicznej odpowiedzi, może pojawić się również w zjawiskach ekonomicznych. W Rozdziale został podany przykład takiej statystycznej analizy, wykorzystującej możliwości procedury SAS'a LOGISTIC, wywoływanej z poziomu pakietu Analyst. Odwołując się do otrzymanych raportów SAS'a, zaprezentowano przykład modelowania zmiany szansy zajścia sukcesu, którym było niespłacenie kredytu (długu), na skutek wpływu czynnika głównego, którym

była płeć osoby, przy kontrolowanym wpływie innych czynników, to znaczy czasu zatrudnienia i dzielnicy zamieszkania osoby w wybranym do analizy mieście.

Wadą bezwarunkowej estymacji MNW jest to, że estymatory MNW są na ogół obciążone. Ponieważ jednak asymptotycznie estymatory MNW są nieobciążone, zatem mogłoby sie wydawać, że lekarstwem na zmniejszenie obciążenia jest zwiększenie liczebności próby. Jednak w przypadku konieczności występowania jednoczesnych porównań, nie tylko pomiędzy jedną parą grup, lecz pomiędzy wieloma parami skojarzonych grup (które mogą być nawet jednoelementowe), dla każdego takiego porównania do modelu logistycznego dochodzi jeden parametr, co oznacza, że liczba parametrów zwiększa się wraz z liczebnością próby, tworząc stale (co do liczebności) znaczną jej część. Gdy liczba jednocześnie kontrastowanych grup jest porównywalna z liczebnością próby, wtedy przy posługiwaniu się pełną funkcją wiarygodności powstaje problem nieusuwalnego obciążenia estymatorów MNW. Stąd w pewnych przypadkach należało by zastosować tzw. warunkową estymację metodą największej wiarygodności, która wykorzystuje warunkowe funkcje wiarygodności [1], [38]. Co prawda warunkowa funkcja wiarygodności nie niesie na ogół tak dużej informacji o parametrze jak pełna funkcja wiarygodności, prowadząc do straty informacji dotyczącej parametru, jednak jej zastosowanie do konstrukcji estymatorów redukuje ich obciążenie. Problem ten sam w sobie mógłby być przedmiotem dalszej obszernej analizy [38], [1].

A. Rozdział 7. Uzupełnienia.

Rozdział 7-1. Uzupełnienie 1. Błąd statystyczny i statystyka Wald'a.

Wiarygodnościowe przedziały ufności wspomniane w Rozdziałach 4 do 6 są uzupełnieniem analizy MNW, pozwalającym na szybkie określenie niepewności oszacowania skalarnego parametru θ. Posługiwanie się nimi jest prostsze niż samą funkcją wiarygodności. Istotnym pojęciem przy ich konstrukcji jest obserwowana informacja Fishera $\mathbf{iF}(\hat{\theta})$ wprowadzona w Rozdziale 2-2-8 jako czynnik po prawej stronie wyrażenia (1-1.28):

$$\ln \frac{P(y \mid \theta)}{P(y \mid \hat{\theta})} \approx -\frac{1}{2}\mathbf{iF}(\hat{\theta})(\hat{\theta}-\theta)^2. \tag{7-1.1}$$

Wyrażenie to jest w przybliżeniu słuszne dla *rozkładów regularnych*, dla których funkcjonuje przybliżenie kwadratowe dla log ilorazu funkcji wiarygodności. Więcej na temat $\mathbf{iF}(\hat{\theta})$ można znaleźć w [5], [38].

W przypadku gdy pierwotna zmienna losowa Y ma rozkład normalny $N(\theta,\sigma^2)$, wtedy związek (1-1.28) staje się dokładny, a z (1-1.19) widać, że $\mathbf{iF}(\hat{\theta}) = \frac{N}{\sigma^2}$. Ponadto, dla rozkładu normalnego można rozkładem χ_1^2 dokonać dokładnego wyskalowania ilorazu funkcji wiarygodności, a zatem i wartości parametru obcięcia $c = e^{-\frac{1}{2}\chi^2_{1,(1-\alpha)}}$, zgodnie z (1-1.25).

Błąd standardowy: Obserwowana informacja Fishera definiuje tzw. *błąd standardowy* estymatora $\hat{\theta}$ jako równy:

$$\hat{\sigma}_{\hat{\theta}} \equiv se(\hat{\theta}) = \left(\mathbf{iF}(\hat{\theta})\right)^{-1/2}, \tag{7-1.2}$$

który w przypadku rozkładu normalnego wynosi:

$$\hat{\sigma}_{\hat{\theta}} = \sigma / \sqrt{N}\,. \tag{7-1.3}$$

Jeśli przyjąć, że iloraz wiarygodności jest równy wartości granicznej $P(y \mid \theta)/P(y \mid \hat{\theta}) = c$, wtedy dla przedziału wiarygodności (1-1.27), $\{\theta, P(\tilde{Y} \mid \theta)/P(\tilde{Y} \mid \hat{\theta}) > c\}$ otrzymujemy z (1-1.28) graniczną wartość parametru $\theta \approx \hat{\theta} \pm \sqrt{-2\ln c}\,\left(\mathbf{iF}(\hat{\theta})\right)^{-1/2}$. Stąd przybliżona postać[25] przedziału wiarygodności jest następująca

[25] Dla rozkładu normalnego jest to postać dokładna.

$\left(\hat{\theta}-\sqrt{-2\ln c}\ \hat{\sigma}_{\hat{\theta}}\ ,\quad \hat{\theta}+\sqrt{-2\ln c}\ \hat{\sigma}_{\hat{\theta}}\right)$, co w przypadku rozkładu normalnego daje dokładną postać $(1-\alpha)\cdot 100\%$-wego przedziału wiarygodności (CI):

$$\left(\hat{\theta}-\sqrt{\chi^2_{1,(1-\alpha)}}\ \hat{\sigma}_{\hat{\theta}}\ ,\quad \hat{\theta}+\sqrt{\chi^2_{1,(1-\alpha)}}\ \hat{\sigma}_{\hat{\theta}}\right) \tag{7-1.4}$$

Na przykład, gdy $(1-\alpha)=0{,}95$, wtedy 95%-wy dokładny przedział wiarygodności wynosi:

$$\hat{\theta}\pm 1{,}96\,\hat{\sigma}_{\hat{\theta}} \tag{7-1.5}$$

W przypadku reguralnym przedział (7-1.4) określa przybliżoną postać $(1-\alpha)\cdot 100\%$-wego przedziału wiarygodności.

Test Wald'a dla hipotezy zerowej o skalarnym parametrze θ: Zweryfikujmy hipotezę zerową $H_0:\theta=\theta_0$ wobec hipotezy alternatywnej $H_1:\theta\neq\theta_0$. W celu przeprowadzenia testu statystycznego wprowadźmy tzw. statystykę Wald'a.

Statystyka Wald'a ma postać:

$$U=\frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}}\quad , \tag{7-1.6}$$

gdzie wartość *u* zmiennej *U* jest wyznaczona na podstawie obserwacji i dla estymatora $\hat{\theta}$ MNW parametru θ. Z porównania (7-1.1) oraz (7-1.6), widać, że duża wartość $|u|$ statystyki $|U|$ jest związana z małą wiarygodnością modelu dla $H_0:\theta=\theta_0$.

Przykład: Niech na podstawie obserwacji wartość $|u|=3$. Zakładając regularność modelu, otrzymujemy z (7-1.1) oraz (7-1.6) i przy wartości granicznej $\frac{P(y\mid\theta)}{P(y\mid\hat{\theta})}=c$ ilorazu wiarygodności, związek pomiędzy parametrem c a wartością u:

$$c=\exp(-\frac{u^2}{2})\ . \tag{7-1.7}$$

Dla rozważanego przykładu z (7-1.7) otrzymujemy $c=\exp(-\frac{u^2}{2})\Big|_{|u|=3}=\exp(-4{,}5)=0{,}011$. Zatem zgodnie z (1-1.30) wartość empirycznego poziomu istotności wynosi $p=P(\chi^2_1>-2\ln c)\ =P(\chi^2_1>9)=0{,}0027$. Oznacza to, że na każdym poziomie istotności $\alpha\geq p=0{,}0027$, np. $\alpha=0{,}01$, odrzucamy hipotezę zerową $H_0:\theta=\theta_0$ na rzecz hipotezy alternatywnej.

W Rozdziałach 2-2-11-3 i 2-2-11-4 zastosowano statystykę Wald'a do estymacji i weryfikacji hipotezy zerowej dotyczącej parametru β Modelu 1, (2-2-10.47).

Rozdział 7-2. Uzupełnienie 2. Zasada niezmienniczości ilorazu funkcji wiarygodności.

Z powyższych rozważań wynika, że funkcja wiarygodności reprezentuje niepewność dla ustalonego parametru. Nie jest ona jednak gęstością rozkładu prawdopodobieństwa dla tego parametru. Pojęcie takie byłoby całkowicie obce statystyce klasycznej (nie włączając procesów stochastycznych). Inaczej ma się sprawa w tzw. statystyce Bayesowskiej. Aby zrozumieć różnicę pomiędzy podejściem klasycznym i Bayesowskim [41], [38] rozważmy transformację parametru.

Przykład transformacji parametru: Rozważmy eksperyment, w którym dokonujemy jednokrotnego pomiaru zmiennej o rozkładzie dwumianowym (1-1.11). Funkcja wiarygodności ma więc postać $P(\theta)=\binom{m}{x}\theta^x(1-\theta)^{m-x}$. Niech parametr $m=12$ a w pomiarze otrzymano $x=9$. Testujemy model, dla którego $\theta=\theta_1=3/4$ wobec modelu z $\theta=\theta_2=3/10$. Stosunek wiarygodności wynosi:

$$\frac{P(\theta_1=3/4)}{P(\theta_2=3/10)}=\frac{\binom{m}{x}\theta_1^9(1-\theta_1)^3}{\binom{m}{x}\theta_2^9(1-\theta_2)^3}=173.774 \tag{7-2.1}$$

Dokonajmy hiperbolicznego wzajemnie jednoznacznego przekształcenia parametru:

$$\psi=1/\theta. \tag{7-2.2}$$

Funkcja wiarygodności po transformacji parametru ma postać $\tilde{P}(\psi)=\binom{m}{x}(1/\psi)^x(1-1/\psi)^{m-x}$. Wartości parametru ψ odpowiadające wartościom θ_1 i θ_2 wynoszą odpowiednio $\psi_1=4/3$ oraz $\psi_2=10/3$. Łatwo sprawdzić, że transformacja (7-2.2) nie zmienia stosunku wiarygodności, tzn.:

$$\frac{\tilde{P}(\psi_1=4/3)}{\tilde{P}(\psi_2=10/3)}=\frac{P(\theta_1=3/4)}{P(\theta_1=3/10)}=173.774. \tag{7-2.3}$$

Niezmienniczość stosunku wiarygodności: Zatem widać, że stosunek wiarygodności jest niezmienniczy ze względu na wzajemnie jednoznaczą transformację parametru. Gdyby transformacja parametru była np. transformacją "logit" $\psi=\ln(\theta/(1-\theta))$ lub paraboliczną $\psi=\theta^2$, to sytuacja także nie uległaby zmianie. Również w ogólnym przypadku transformacji parametru własność *niezmienniczości stosunku wiarygodności* pozostaje słuszna. Oznacza to, że informacja zawarta w próbce jest niezmiennicza ze względu na wybór parametryzacji, tzn. powinniśmy być w takiej samej sytuacji niewiedzy niezależnie od tego jak zamodelujemy zjawisko, o ile różnica w modelowaniu sprowadza się jedynie do transformacji parametru. W omawianym przykładzie powinniśmy równie dobrze móc stosować parametr θ, jak $1/\theta$, θ^2, czy $\ln(\theta/(1-\theta))$.

Uwaga o transformacji parametru w statystyce Bayesowskiej: Natomiast sytuacja ma się zupełnie inaczej w przypadku Bayesowskiego podejścia do funkcji wiarygodności [41], [38], w którym funkcja wiarygodności uwzględnia (Bayesowski) rozkład prawdopodobieństwa $f(\theta | x)$ parametru θ. Oznacza to, że Jakobian transformacji $\theta \to \psi$ parametru, modyfikując rozkład parametru, zmienia również funkcję wiarygodności. Zmiana ta zależy od wartości parametru, różnie zmieniając licznik i mianownik w (7-2.1), co niszczy *intuicyjną* własność niezmienniczości ilorazu wiarygodności ze względu na transformację parametru [38].

Do zagadnienia użyteczności własności niezmienniczości ilorazu funkcji wiarygodności przy konstrukcji przedziału wiarygodności, powrócimy w dalszej części.

B. Rozdział 8. Kryterium AIC Akaike'a wyboru rzędu parametrów *p* i *q* w modelu ARIMA szeregów czasowych.

W [24], [42] przedstawiono procedurę modelowania szeregów czasowych za pomocą liniowych modeli ARMA(*p, q*) co niesie ze sobą korzyści ze względu na prostą budowę modelu, oraz ciekawą liniową predykcję. Sprawia to, że modelowanie to jest szeroko wykorzystywane w różnych dziedzinach np. ekonomii, medycynie i fizyce. Omówienie podstawowych właściwości procesów ARMA, wraz z identyfikacją „prostych" modeli AR(*p*) (proces z autoregresją), MA(*q*) (proces średniej ruchomej), można znaleźć w [42]. Wskazówkę, co do wartości parametrów *p* i *q* procesu można otrzymać z np. analizy zachowań dwóch funkcji: funkcji autokorelacji procesu ACF (autocorrelation function) i funkcji autokorelacji cząstkowej PACF (Partial Autocorrelation Function), których estymację z próby omówiono w [42], [24]. Poniżej podano jedynie wprowadzenie do kryterium AICC [24], które dzięki swoim własnościom naprowadza na wybór odpowiedniego modelu ARMA(*p, q*).

Niech { X_t }, (t $=0,\pm1,\pm2\ldots$) jest procesem spełniającym liniowe równanie różnicowe ze stałymi współczynnikami, oraz niech zmienne losowe X_t są niezależne i mają takie same rozkłady ze średnią zero i wariancją σ^2 tzn. { X_t } ~ iid(0, σ^2). Proces { X_t, $t=0,\pm1,\pm2\ldots$} jest procesem ARMA(*p, q*), jeśli { X_t } jest stacjonarny, oraz jeśli dla każdego *t*, proces { X_t } spełnia równanie różnicowe:

$$X_t-\phi_1X_{t-1}-\ldots-\phi_pX_{t-p}=Z_t+\theta_1Z_{t-1}+\ldots+\theta_qZ_{t-q} \tag{8.1}$$

gdzie { Z_t } ~WN(0, σ^2), co oznacza, że proces { Z_t } jest białym szumem ze średnią 0 i wariancją σ^2. W [24] omówiono najważniejszą część budowania modeli ARMA, czyli estymację parametrów (ϕ,θ).

Problemem jest znalezienie najbardziej satysfakcjonującego modelu ARMA(*p, q*) reprezentującego { X_t }. Jeśli *p* i *q* są znane z góry to należy skorzystać z techniki estymacji parametrów opisanych np. w [24], [42]. W przypadku procesów AR(*p*) może to być metoda Yule'a-Walker'a, a w przypadku procesów MA(*q*), tzw. algorytm innowacyjny.

Jednakże zazwyczaj nie mamy do czynienia z taką sytuacją, więc podstawową sprawą jest identyfikacja odpowiednich wartości *p* i *q*. Na pierwszy rzut oka mogłoby się wydawać, że im większe wartości *p* i *q*, tym lepiej będzie model dopasowany. Na przykład, jeśli dopasowaliśmy całą sekwencję procesów AR(*p*), *p* = 1,2..., wtedy wartość estymatora $\hat{\sigma}^2$ MNW parametru σ^2, będącego wariancją białego szumu w modelu ARMA, spada zasadniczo w sposób monotoniczny wraz ze wzrostem *p*. Jednakże musimy uważać na niebezpieczeństwo przefitowania, czyli na zbyt dokładne dopasowanie modelu do poszczególnych obserwacji. Skrajny przypadek przefitowania miałby miejsce, jeśli np. wielomian stopnia *n*-tego dopasujemy do *n*+1 obserwacji wygenerowanych z modelu $Y_t=a+bt+Z_t$, gdzie *a* oraz *b* są parametrami. *W takiej sytuacji*

dopasowanie będzie idealne dla obserwacji otrzymanych w próbie, ale obok skomplikowanej postaci modelu, użycie go do predykcji przyszłych wartości może być obarczone dużymi błędami.

Rozwinięto szereg kryteriów, w szczególności kryterium AIC Akaike'a i kryterium FPE_p (final prediction error) [24], które próbują zapobiegać przefitowaniu przez skuteczne wyznaczanie nakładu związanego z wyprowadzeniem każdego dodatkowego parametru.
FPE_p jest oszacowaniem średnio – kwadratowego błędu predykcji w jednym kroku dla realizacji procesu, która jest niezależna od realizacji właśnie obserwowanej. Jeśli dopasujemy proces AR(p) o stale rosnącym rzędzie p do obserwowanych danych, wtedy oszacowanie MNW $\hat{\sigma}^2$ dla wariancji białego szumu będzie zazwyczaj spadać wraz ze wzrostem p. Jednakże błędy oszacowań w poszerzanym ciągle zbiorze dopasowanych parametrów, spowodują w końcu wzrost FPE_p. Zgodnie z kryterium FPE_p [24]:

$$\mathrm{FPE}_p = \hat{\sigma}^2 \frac{n+p}{n-p} \ , \tag{8.1}$$

wybieramy jako rząd dopasowanego modelu AR(p) tą wartość p, dla której FPE_p jest *najmniejsze*. Sytuacja jest analogiczna do tej, która jest powodem zastąpienia w zwykłych modelach regresji współczynnika determinacji R^2 współczynnikiem R^2_{adj}, który (co wiązało się ze spadkiem średniej wariancji wewnątrzgrupowej *MSE*) w przeciwieństwie do R^2 nie zawsze wzrasta przy dodaniu do modelu nowej zmiennej (Rozdział 8, Część I). Działanie to jest więc wymuszone uniknięciem wzrostu średniej wariancji wewnątrzgrupowej *MSE*, spowodowanego wolniejszym tempem spadku wraz ze wzrostem liczby parametrów p sumy kwadratów błędu *SSE* niż tempo spadku wartości liczby stopni swobody $n - p$ dla reszt modelu.

W estymacji szeregów czasowych bardzo istotną rolę odgrywa metoda największej wiarygodności oraz standardowa metoda najmniejszych kwadratów (która jest jednym z przypadków MNW dla modelu gaussowskiego). Metody te okazują się uniwersalne nie tylko w estymacji szeregów czasowych, ale również w dowolnej estymacji parametrów funkcji. Wiarygodnościowe kryterium AIC Akaike'a [43] wyboru modelu zostało skonstruowane tak, aby być w przybliżeniu nieobciążonym estymatorem (zdefiniowanego poniżej) wskaźnika Kullback'a – Leibler'a dopasowanego modelu względem prawdziwego modelu.
Poniżej zostanie podana zmodyfikowana, nieobciążona wersja AIC, nazywana kryterium AICC, wprowadzona przez Hurvich'a i Tsai'a [44].

Jeśli $\vec{X}$ jest n wymiarowym wektorem losowym, którego gęstość prawdopodobieństwa należy do rodziny $\{f(\cdot\,;\psi), \psi \in \Psi\}$, wtedy różnica Kullback'a – Leibler'a [39], [5] między $f(\cdot\,;\psi)$ i $f(\cdot\,;\theta)$ jest zdefiniowana jako:

$$d(\psi \mid \theta) = \Delta(\psi \mid \theta) - \Delta(\theta \mid \theta) \ , \tag{8.3}$$

gdzie:

$$\Delta(\psi \mid \theta) = E_\theta(-2\ln f(\vec{X};\psi)) = \int_{R^n} -2\ln(f(\vec{x};\psi))\, f(\vec{x};\theta)\, d\vec{x} \tag{8.4}$$

jest *wskaźnikiem* Kullback'a – Leibler'a dla $f(\cdot\,;\psi)$ w stosunku do $f(\cdot\,;\theta)$ (w ogólności $\Delta(\psi \mid \theta) \neq \Delta(\theta \mid \psi)$).

Po zastosowaniu nierówności Jensen'a otrzymujemy [24]:

$$d(\psi \mid \theta) = \int_{R^n} -2\ln\left(\frac{f(\vec{x};\psi)}{f(\vec{x};\theta)}\right) f(\vec{x};\theta)\, d\vec{x} \;\geq -2\ln\int_{R^n} \frac{f(\vec{x};\psi)}{f(\vec{x};\theta)}\, f(\vec{x};\theta)\, d\vec{x} = -2\ln\int_{R^n} f(\vec{x};\psi)\, d\vec{x} = 0 \quad , \tag{8.5}$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy $f(\vec{x};\theta) = f(\vec{x};\psi)$ prawie zawsze z miarą [$f(\cdot\,;\theta)$].

Wprowadźmy oznaczenie $\vec{\beta} \equiv (\vec{\phi}, \vec{\theta})$, gdzie $\vec{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)$ i $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$. Mając obserwacje $X_1, \ldots, X_n$ z procesu ARMA z nieznanymi parametrami $\theta = (\vec{\beta}, \sigma^2)$, prawdziwy model może zostać zidentyfikowany wtedy, jeśli byłoby możliwe wyznaczenie różnicy Kullback'a – Leibler'a pomiędzy *prawdziwym* modelem, a wszystkimi kandydującymi modelami. Ponieważ model prawdziwy nie jest znany, więc nie jest to możliwe. Dlatego oszacowujemy różnice Kullback'a – Leibler'a i wybieramy model, dla którego oszacowana różnica (bądź wskaźnik) jest najmniejsza. W tym celu załóżmy, że prawdziwy model oraz wszystkie modele alternatywne do niego, są gaussowskie. Wtedy dla każdego zadanego $\theta = (\vec{\beta}, \sigma^2)$, funkcja $f(\cdot\,;\theta)$ jest gęstością prawdopodobieństwa dla $(Y_1, \ldots, Y_n)^T$, gdzie $\{Y_t\}$ jest gaussowskim procesem ARMA(*p*, *q*) z wektorem współczynników $\vec{\beta}$, oraz wariancją białego szumu σ^2. Zależność parametrów θ od *p* i *q* ma miejsce poprzez wymiar wektora współczynników autoregresji i ruchomej średniej w $\vec{\beta}$ [24].

Przypuśćmy, że nasze obserwacje $X_1, \ldots, X_n$ pochodzą z gaussowskiego procesu ARMA z parametrowym wektorem $\theta = (\vec{\beta}, \sigma^2)$, oraz załóżmy na chwilę, że prawdziwy rząd modelu wynosi (*p*, *q*). Niech $\hat{\theta} = (\hat{\vec{\beta}}, \hat{\sigma}^2)$ będzie estymatorem MNW dla θ wyznaczonym z $X_1, \ldots, X_n$, i niech $Y_1, \ldots, Y_n$ będzie niezależną realizacją *prawdziwego* procesu (z parametrem θ).

Wtedy [24]:

$$-2\ln L_Y(\hat{\vec{\beta}}, \hat{\sigma}^2) = -2\ln L_X(\hat{\vec{\beta}}, \hat{\sigma}^2) + \hat{\sigma}^{-2} S_Y(\hat{\vec{\beta}}) - n \tag{8.6}$$

co daje:

$$E_\theta(\Delta(\hat{\theta} \mid \theta)) = E_{\vec{\beta},\sigma^2}(-2\ln L_Y(\hat{\vec{\beta}}, \hat{\sigma}^2)) = E_{\vec{\beta},\sigma^2}(-2\ln L_X(\hat{\vec{\beta}}, \hat{\sigma}^2)) + E_{\vec{\beta},\sigma^2}\left(\frac{S_Y(\hat{\vec{\beta}})}{\hat{\sigma}^2}\right) - n \quad , \tag{8.7}$$

gdzie $S_Y(\hat{\vec{\beta}})$ jest ważoną sumą *kwadratów* odchyleń (reszt) dla realizacji procesu $Y_1,\ldots,Y_n$ przy parametrach oszacowanych dla procesu $X_1,\ldots,X_n$ [24]. Dla dużych *n* możemy pominąć wyrazy wyższych rzędów rozwinięcia $S_Y(\hat{\vec{\beta}})$ wokół prawdziwej wartości parametru $\vec{\beta}$, otrzymując [24]:

$$S_Y(\hat{\vec{\beta}}) \approx S_Y(\vec{\beta}) + (\hat{\vec{\beta}} - \vec{\beta})\frac{\partial S_Y(\vec{\beta})}{\partial \vec{\beta}} + \frac{1}{2}(\hat{\vec{\beta}} - \vec{\beta})^T \left[\frac{\partial^2 S_Y(\vec{\beta})}{\partial \beta_i \partial \beta_j}\right]_{i,j=1}^{n} (\hat{\vec{\beta}} - \vec{\beta}) \approx$$

$$\approx S_Y(\vec{\beta}) + (\hat{\vec{\beta}} - \vec{\beta}) 2\sum_{t=1}^{n} \frac{\partial Z_t(\vec{\beta})}{\partial \vec{\beta}} Z_t(\vec{\beta}) + \frac{1}{2}(\hat{\vec{\beta}} - \vec{\beta})^T \left[\frac{\partial^2 S_Y(\vec{\beta})}{\partial \beta_i \partial \beta_j}\right]_{i,j=1}^{n} (\hat{\vec{\beta}} - \vec{\beta}), \tag{8.8}$$

gdzie skorzystano z faktu, że $S_Y(\vec{\beta}) \approx \sum_{t=1}^{n} Z_t^2(\vec{\beta})$, [24]. Niech $V(\vec{\beta})$ jest macierzą kowariancji dla $(\hat{\vec{\beta}} - \vec{\beta})$.

Ponieważ $\frac{\partial Z_t}{\partial \vec{\beta}}(\vec{\beta}) Z_t(\vec{\beta})$ ma asymptotycznie (dla dużych *n*) średnią równą zero i jest niezależne od $(\hat{\vec{\beta}} - \vec{\beta})$ oraz ze względu na zbieżność[26] $n^{-1}\left[\frac{\partial^2 S_Y(\vec{\beta})}{\partial \beta_i \partial \beta_j}\right]_{i,j=1}^{n} \xrightarrow{P} 2\sigma^2 V^{-1}(\vec{\beta})$, skąd zastąpmy $\left[\frac{\partial^2 S_Y(\vec{\beta})}{\partial \beta_i \partial \beta_j}\right]_{i,j=1}^{n}$ przez $2n\sigma^2 V^{-1}(\beta)$, oraz ze względu na asymptotyczną normalność (*AN*) estymatora $\hat{\vec{\beta}} \sim AN\left(\vec{\beta}, n^{-1}V(\beta)\right)$, skąd $n^{1/2}(\hat{\vec{\beta}} - \vec{\beta}) \sim AN(0, V(\beta))$ [24], więc z (8.8) dla wartości oczekiwanej, otrzymujemy [24][27]:

$$\begin{aligned} E_{\vec{\beta},\sigma^2}\left[S_Y(\hat{\vec{\beta}})\right] &\approx E_{\vec{\beta},\sigma^2}\left[S_Y(\vec{\beta})\right] + \sigma^2 E_{\vec{\beta},\sigma^2}\left[(\hat{\vec{\beta}} - \vec{\beta})^T V^{-1}(\vec{\beta})(\hat{\vec{\beta}} - \vec{\beta})\right] \approx \sigma^2 n + \sigma^2(p+q) \\ &= \sigma^2(n+p+q) \quad . \end{aligned} \tag{8.9}$$

Statystyka $n\hat{\sigma}^2 = S_X(\hat{\vec{\beta}})$ ma asymptotycznie rozkład $\sigma^2\chi^2(n-p-q)$ oraz jest asymptotycznie niezależna od $\hat{\vec{\beta}}$. Wraz z niezależnością $\{X_1,\ldots,X_n\}$ oraz $\{Y_1,\ldots,Y_n\}$ implikuje to asymptotyczną niezależność $\hat{\sigma}^2$ od $S_Y(\hat{\vec{\beta}})$. W konsekwencji otrzymujemy [24]:

$$E_{\vec{\beta},\sigma^2}\left(\frac{S_Y(\hat{\vec{\beta}})}{\hat{\sigma}^2}\right) \cong \sigma^2(n+p+q)\left(E_{\vec{\beta},\sigma^2}(\hat{\sigma}^{-2})\right) \cong \sigma^2(n+p+q)\left(\sigma^2\,\frac{n-p-q-2}{n}\right)^{-1}$$

skąd

---

[26] $X_n \xrightarrow{P} X$ oznacza zbieżność w prawdopodobieństwie ciągu zmiennych losowy X_n do zmiennej X.

[27] Korzystając z faktu, że $E(\vec{U}^T\Sigma^{-1}\vec{U}) = Tr(\Sigma\Sigma^{-1}) = k$, dla każdego *k* – wymiarowego wektora losowego $\vec{U}$ ze średnią zero i z nieosobliwą macierzą kowariancji Σ.

$$E_{\vec{\beta},\sigma^2}\left(\frac{S_Y(\hat{\vec{\beta}})}{\hat{\sigma}^2}\right)-n \cong \sigma^2(n+p+q)\left(\sigma^2\,\frac{n-p-q-2}{n}\right)^{-1}-n=\frac{2(p+q+1)n}{n-p-q-2}\ . \tag{8.10}$$

Z (8.10) oraz (8.7) widać więc, że wielkość:

$$-2\ln L_X(\hat{\vec{\beta}},\hat{\sigma}^2)+2(p+q+1)n/(n-p-q-2) \tag{8.11}$$

jest w przybliżeniu nieobciążonym estymatorem wartości oczekiwanej wskaźnika Kullback'a – Leibler'a $E_\theta\left(\Delta(\hat{\theta}\,|\,\theta)\right)$, (8.7).

Ponieważ powyższe obliczenia (oraz estymatory MNW $\hat{\vec{\beta}}$ oraz $\hat{\sigma}^2$) są oparte na założeniu, iż prawdziwy model jest rzędu (*p*, *q*), dlatego wybieramy wartości *p* i *q* dla naszego dopasowanego modelu jako te, które minimalizują AICC($\hat{\vec{\beta}}$), gdzie:

$$\mathrm{AICC}(\hat{\vec{\beta}}):=-2\ln L_X(\hat{\vec{\beta}},S_X(\hat{\vec{\beta}})/n)+2(p+q+1)n/(n-p-q-2)\,. \tag{8.12}$$

W końcu, jeśli już znajdziemy model, który minimalizuje wartość AICC, to należy sprawdzić dobroć dopasowania. Zasadniczo oznacza to sprawdzenie czy reszty to biały szum gaussowski [42].

Statystyka AIC, zdefiniowana jako:

$$\mathrm{AIC}(\hat{\vec{\beta}}):=-2\ln L_X(\hat{\vec{\beta}},S_X(\hat{\vec{\beta}})/n)+2(p+q+1) \tag{8.13}$$

może być użyta w ten sam sposób.

Oba kryteria AICC($\vec{\beta},\sigma^2$) oraz AIC($\vec{\beta},\sigma^2$) mogą zostać zdefiniowane dla dowolnego σ^2 poprzez zastąpienie w powyższych definicjach $S_X(\vec{\beta})/n$ przez σ^2. Jednakże korzystniej jest używać AICC($\vec{\beta}$) oraz AIC($\vec{\beta}$) tak jak zostały zdefiniowane powyżej. Wynika to stąd, że obie wartości kryteriów AICC i AIC są minimalizowane dla dowolnych $\vec{\beta}$ przez podstawienie $\sigma^2=S_X(\vec{\beta})/n$ [24].

Rozdział 8-1. Zakończenie.

Po niezbędnym przekształceniu danych i otrzymaniu szeregu stacjonarnego [24], jedynym problemem staje się znalezienie satysfakcjonującego modelu ARMA(*p*, *q*) i zidentyfikowanie *p* oraz *q*. Przydatnych wskazówek w tym wyborze mogą dostarczyć funkcja autokorelacji i częściowej autokorelacji z próby oraz wstępne estymatory $\hat{\phi}_m$ i $\hat{\theta}_m$ parametrów ϕ i θ [24].

Jednakże najważniejszym kryterium w wyborze modelu jest AICC, zmodyfikowana wersja kryterium Akaike'a AIC, zgodnie z którym obliczamy estymatory największej wiarygodności parametrów ϕ i θ oraz σ^2 dla odpowiednie dobranego *p* i *q*, i wybieramy dopasowany model z najmniejszym AICC. Jeśli dopasowany model jest satysfakcjonujący, reszty powinny przypominać biały szum gaussowski (test

sprawdzający tę własność opisany jest w [24], [42]) i powinny być przeprowadzone dla minimalnego modelu AICC (tzn. takiego, dla którego kryterium AICC ma najmniejszą wartość) tak żeby mieć pewność, że reszty są zgodne z ich oczekiwanym zachowaniem w tym modelu. Jeśli reszty w minimalnym modelu AICC nie są zgodne z białym szumem gaussowskim, wtedy powinien zostać sprawdzony inny model (w którym kryterium AICC jest bliskie minimum); i tak aż do momentu znalezienia takiego modelu, który przechodzi testy dobroci dopasowania. W niektórych przypadkach mała różnica w wartości AICC (umownie mniej niż 2) między dwoma zadowalającymi modelami może zostać zignorowana po to, aby wybrać model prostszy.

Na koniec, okazało się, że aby dokonać analizy porównawczej dwóch szeregów czasowych zamodelowanych jako ARMA(p,q), koniecznym jest rozważanie własności całej przestrzeni tych szeregów wraz z ich strukturą geometryczną. Zainteresowanego czytelnika odsyłamy do pracy [39] oraz [45]. Poza tym, chociaż procesy ARMA można badać narzędziami (statycznej) analizy regresji, to należą one do obszaru zainteresowań dynamiki procesów czasowych, tzn. analiza procesów ARMA leży w dziale dynamiki stochastycznej z czasem dyskretnym, a nawet ciągłym. Np. proces AR(1) z czasem ciągłym można interpretować jako stacjonarne rozwiązanie stochastycznego równania różniczkowego Itô [46], [47]. Tym samym doszliśmy jakby do sedna właściwego sformułowania analizy regresji. Poza nią leży dynamika procesów przyczynowo-skutkowych dla obiektów ze strukturą wewnętrzną.

Cześć III. Zagadnienia do opracowania i zadania do rozwiązania.

Rozdział 1. Zagadnienia do opracowania.

1. Omówić podstawowe cztery własności estymatorów (nieobciążoność, zgodność, efektywność i dostateczność).

2. Niech będzie dany rozkład m-wymiarowej zmiennej losowej (X_1 , X_2 , ..., X_m).

a) Podać definicję i interpretację współczynników korelacji rzędu zerowego i wyższych (współczynniki korelacji cząstkowej i półcząstkowej). Podać związek cząstkowych współczynników korelacji rzędu pierwszego dla układu trzech zmiennych X,Y,Z (gdzie zmienna Z jest pod kontrolą) ze współczynnikiem rzędu zerowego dla reszt modeli X od Z oraz Y od Z.

b) Podać definicję i interpretację współczynnika korelacji wielorakiej.

c) W pewnym kraju przeprowadzono badania nad korelacją plonów pszenicy X a opadami Y i temperaturą Z w okresie zasiewów. Wyniki były następujące: r_{XY} = -0,66, r_{XZ} = 0,36, r_{YZ} = - 0,55

Wyznaczyć $r_{XY|Z}$ i $r_{XZ|Y}$ i zinterpretować wyniki. Wyznaczyć współczynnik korelacji wielorakiej zależności plonów X od opadów Y i temperatury Z oraz zinterpretować wynik.

3. a) Omówić zagadnienie transformacji wielomianów zwyczajnych do wielomianów ortogonalnych. Jaki jest powód wprowadzania zmiennych ortogonalnych do analizy regresji.

b) Niech liczba wszystkich obserwacji w próbce wynosi *n*.

Rozważmy układ zmiennych X oraz X^2 i przyjmijmy, że zmienna X ma równo rozstawione warianty. Załóżmy, że liczba obserwacji jest taka sama w każdym wariancie zmiennej X i wynosi n/l dla każdego wariantu.

Niech liczba poziomów zmiennej X wynosi l=3 tak, że zmienna X ma postać:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

Pokazać, że z dokładnością do różnych od zera multiplikatywnych stałych oraz z dokładnością do stałych addytywnych, układ zmiennych ortogonalnych X_1^* i X_2^* ma postać:

$$X_1^* = \begin{pmatrix} x_{11}^* \\ x_{12}^* \\ x_{13}^* \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \quad , \qquad X_2^* = \begin{pmatrix} x_{21}^* \\ x_{22}^* \\ x_{23}^* \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \quad .$$

(Uwaga: Przyjąć średnie arytmetyczne zmiennych X_1^* i X_2^* jako wartości stałych addytywnych).

4. Podać twierdzenie o składowych głównych wraz z omówieniem wszystkich występujących w nim pojęć oraz omówić ich znaczenie w analizie regresji. Na czym polega zastosowanie indeksu warunkowego w analizie regresji.

5. Podać i wyprowadzić nierówność Bonferroni'ego oraz omówić płynący z niej wniosek dla szczegółowego poziomu istotności α_s w przypadku jednoczesnej weryfikacji g hipotez statystycznych.

6. Omówić typy reszt i ich własności w klasycznym modelu regresji oraz pojęcie dźwignięcia i jego własności.
Omówić pojęcie obserwacji wpływowej oraz istotę diagnostyki regresji opartej o odległość Cook'a D_i.

7. Podać macierzowe ujęcie klasycznego liniowego modelu regresji wraz z własnościami macierzy rzutowej (dźwignięcia).
a) Pokazać, że rozkład wektora $\mathbf{Y}$ odpowiedzi układu na n-wymiarowy wektor teoretycznych średnich warunkowych $\hat{\mathbf{Y}}$ oraz wektor reszt $\mathbf{U}$, jest rozkładem ortogonalnym.
b) Pokazać, że macierz wariancji-kowariancji $\sigma^2_{\hat{\vec{\beta}}} \equiv \sigma^2(\hat{\vec{\beta}})$ estymatorów parametrów strukturalnych $\vec{\beta}$ klasycznego modelu regresji jest równa $\sigma^2_{\hat{\vec{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2_E$ i uzasadnić, że ich oszacowana z próby macierz wariancji-kowariancji wynosi $\hat{\sigma}^2_{\hat{\vec{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}MSE$, gdzie $\mathbf{X}$ jest macierzą planowania dla zmiennych objaśniających.
c) Pokazać, że wektor $\hat{\vec{\beta}}$ estymatorów MNK parametrów strukturalnych $\vec{\beta}$ modelu regresji jest nieobciążony, tzn.: $E(\hat{\vec{\beta}}) = \vec{\beta}$.
d) Niech wymiar próby wynosi n. Korzystając z nierówności Laguerre'a - Samuelson'a pokazać, że dźwignięcie spełnia warunek: $\frac{1}{n} \le h_i \le 1$.

8. Omówić model regresji dla jednoczynnikowej ANOVA. Pokazać, że *w przypadku, gdy* parametr przesunięcia $\hat{\mu}$ modelu regresji w próbie jest równy średniej ogólnej $\overline{Y}_{\bullet\bullet}$ w próbie, wtedy sposób kodowania dla układu zmiennych wskazujących jest następujący:

$$X_i = \begin{cases} 1 & dla \quad populacji \ i-tej \\ -1 & dla \quad populacji \ k-tej \\ 0 & w \quad pozostalych \quad przypadkach \end{cases} \qquad i = 1, 2, \ldots, k-1.$$

9. Omówić wszystkie hipotezy zerowe rozważane w trakcie przeprowadzania analizy jednoczynnikowej ANOVA wraz z stosowanymi statystykami testowymi.

Pokazać, że w przypadku prawdziwości hipotezy zerowej o jednorodności wariancji w populacjach, statystyka *MSE* występująca w mianowniku statystyki *F* jest nieobciążonym estymatorem wariancji składnika losowego σ_E^2.

Podać pojęcie kontrastu i wyjaśnić potrzebę porównań szczegółowych oraz omówić metodę Scheffe'ego.

10. Omówić zastosowanie metody największej wiarygodności w analizie doboru modelu regresji dla rozkładu Poissona.

Określić model podstawowy, a następnie dla modelu regresji Poissona:

$$\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta), \quad n = 1,2,...,N,$$

opisującego zmianę wartości oczekiwanej liczby zdarzeń Y_n wraz ze zmianą *wariantu* $x_n = (x_{1n}, x_{2n},..., x_{kn})$ czynników (gdzie $r(x_n, \beta)$ opisuje *tempo zdarzeń*), podać postać rozważanych hipotez zerowych i wyprowadzić odpowiadającą im postać dewiancji bądź statystyki ilorazu wiarygodności.

11. Wprowadzenie: Wiadomo, że dla modelu regresji logistycznej podstawową zmienną jest zmienna dychotomiczna *Y*, przyjmująca wartość 1 z prawdopodobieństwem θ oraz 0 z prawdopodobieństwem 1-θ. Rozkład prawdopodobieństwa zmiennej dychotomicznej jest więc następujący:

$$pr(Y;\theta) = \theta^Y (1-\theta)^{1-Y}, \quad Y = 0,1.$$

Niech $x_n = (x_{1n}, x_{2n},..., x_{kn})$ oznacza zbiór wartości *k* czynników X_j (*j* = 1,2,...,*k*) dla określonej *n* –tej jednostki ($n = 1,2,...,N$) w *N*-wymiarowej próbie (Y_1, Y_2, ..., Y_N). Model logistyczny zakłada następującą postać związku pomiędzy θ_n (czyli pomiędzy warunkowymi wartościami oczekiwanymi zmiennej dychotomicznej Y) a czynnikami X_j:

$$E(Y_n) = \theta_n = 1/(1+\exp(-\lambda_n^*)), \quad n = 1,2,...,N,$$

gdzie $\lambda_n^* \equiv \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn}$, a β_j, (*j*=0,1,2,...,*k*) są nieznanymi współczynnikami regresji logistycznej, które trzeba estymować.

Zadanie: Dla tak sformułowanego modelu regresji logistycznej, podaj postać funkcji wiarygodności oraz ogólną postać równań wiarygodności dla estymatorów $\hat{\beta}_j$ parametrów β_j.

Rozdział 2. Zadania do rozwiązania w SAS'ie.

1. Porównano *średnice pni* (Y) (w *cm*) trzech gatunków sosny (i = A, B, C) w czterech lokalizacjach (j = 1,2,3,4). Poniżej podano wyniki obserwacji.

Gatunek	Lokalizacja			
	1	**2**	**3**	**4**
A	23	25	21	14
	15	20	17	17
	26	21	16	19
	13	16	24	20
	21	18	27	24
B	28	30	19	17
	22	26	24	21
	25	26	19	18
	19	20	25	26
	26	28	29	23
C	18	15	23	18
	10	21	25	12
	12	22	19	23
	22	14	13	22
	13	12	22	19

A) **Analiza Regresji** (rachunki w SAS'ie)

1) Dodaj do zmiennych (*Gatunek*) i (*Lokalizacja)*, zmienne $(Gatunek)^2$, $(Lokalizacja)^2$, (*Gatunek* x *Lokalizacja*). Następnie wykorzystując odpowiedni program SAS'a i otrzymany raport, określ **metodą eliminacji wstecz**, najlepszy model regresji wiążący *średnicę pni* (Y) drzew z pięcioma powyższymi zmiennymi.
2) Zgodnie z podejściem, według którego pierwotnie określa się model podstawowy, w który wchodzą jedynie zmienne (*Gatunek*) i (*Lokalizacja*), określ na poziomie α =0,10 czy istnieje jakiś model drugiego rzędu, który powinien być dodany do modelu podstawowego.

B) **Dwuczynnikowa analiza ANOVA** (rachunki w SAS'ie).

1) Wypowiedz się w kwestii tego, czy czynniki w modelu należy uznać za ustalone czy losowe.
2) Utwórz tablicę średnich w pobranej próbce.
3) Z pomocą tablicy ANOVA, wykonaj właściwe testy F dla każdego z czterech możliwych schematów układów czynników (tzn. oba czynniki ustalone, oba czynniki losowe, pierwszy czynnik ustalony a drugi losowy, pierwszy czynnik losowy a drugi ustalony); sformułuj odpowiednie hipotezy zerowe.
4) Przeanalizuj dane w oparciu o każdy z czterech możliwych schematów klasyfikacji czynników. Porównaj wyniki.
5) Wykorzystując metodę Scheffe'go na poziomie istotności α =0,05 , wskaż dla obu wpływów głównych i wpływu interakcji pomiędzy nimi, średnie *średnic pni* w próbce różniące się istotnie statystycznie.
6) Wykorzystując metodę Scheffe'go, znajdź dla wpływów głównych obu czynników, 95% -owe przedziały ufności dla prawdziwych różnic pomiędzy średnimi *średnic pni* w populacjach.
7) Sprawdź czy można uznać wariancje w rozważanych populacjach (i, j) za takie same.

C) **Jednoczynnikowa analiza ANOVA** (rachunki w SAS'ie). Biorąc pod uwagę jedynie czynnik *Gatunek*), przeprowadź następującą analizę:

1) Wyznacz średnie w próbce i odchylenia standardowe, dla każdego gatunku.
2) Utwórz tablicę ANOVA.
3) Odpowiedz na pytanie: czy rozpatrywane trzy gatunki sosny różnią się znacząco pod względem średniej *średnicy pni*. Sformułuj właściwą hipotezę zerową i hipotezę alternatywną.
4) Wykorzystaj dla różnic pomiędzy parami średnich, metodę wielokrotnych porównań Scheffe'go do określenia istotności tych różnic, oraz wyznacz przedziały ufności dla odpowiednich kontrastów.

Literatura

[1] D. G. Kleinbaum, L. L. Kupper, K. E. Muller, A. Nizam, „Applied Regression Analysis and Other Multivariable Methods", Duxbury Press, (1998).
M. Czerwik, „Wykorzystanie programu SAS jako narzędzia do analizy współzależności zmiennych metodą regresji", Praca licencjacka, Uniwersytet Śląski, Instytut Fizyki, Jastrzębie Zdrój, (2004.)
Patrycja Kruczek, „Diagnostyka reszt w modelach regresji liniowej", Praca licencjacka, Uniwersytet Śląski, Instytut Fizyki, Rybnik, (2005).
A. Maryniok, „Wykorzystanie programu SAS jako narzędzia do analizy współzależności zmiennych metodą analizy wariancji", Praca licencjacka, Uniwersytet Śląski, Instytut Fizyki, Rybnik, (2004).
A. Rząsa, „Dwuczynnikowa analiza wariancji dla komórek z różną liczebnością oraz problem analizy regresji dla ANOVA w systemie SAS", Praca licencjacka, Uniwersytet Śląski, Instytut Fizyki, Rybnik, (2005).
D. Mroziakiewicz, „Analiza regresji Poissona z estymatorami metody największej wiarygodności z wykorzystaniem programu statystycznego SAS", Praca licencjacka, Inst. Fizyki, Uniwersytet Śląski, Rybnik, (2006).
M. Jaworski, Logistyczna analiza regresji z estymatorami metody największej wiarygodności z wykorzystaniem programu statystycznego SAS, Inst. Fizyki, Uniwersytet Śląski, Jastrzębie Zdrój, (2006).

[2] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewski, „Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach", Część II. „Statystyka matematyczna", Wydawnictwo Naukowe PWN, Warszawa, (1995).

[3] „SAS Products and Solutions", http://support.sas.com/software/index.html .

[4] M. Kurpas, „Wybrane zagadnienia ekonometrii z wykorzystaniem programu Statistica", skrypt dla studentów kierunku Ekonfizyka, Instytut Fizyki, Uniwersytet Śląski, (2014), http://ekonofizyka.pl/skrypty/Ekonometria/Ekonometria.pdf .

[5] J. Syska, „Metoda największej wiarygodności i informacja Fisher'a w fizyce i ekonofizyce", skrypt dla studentów kierunku Ekonfizyka, Instytut Fizyki, Uniwersytet Śląski, (2011), http://el.us.edu.pl/ekonofizyka/images/f/f2/Fisher.pdf.

[6] T.W. Anderson, „An introduction to multivariate statistical analysis",3rd edition, Wiley-Interscience, 2003.

[7] M. Beśka, Statystyka matematyczna, str.87-88, http://www.mif.pg.gda.pl/homepages/beska/.

[8] „The NLIN Procedure”, SAS/STAT 9.2 User’s Guide, http://support.sas.com/documentation/cdl/en/statugnlin/61811/PDF/default/statugnlin.pdf .

[9] M. Maliński, „Statystyka matematyczna wspomagana komputerowo”, Wydawnictwo Politechniki Śląskiej, Gliwice, (2000).

[10] „The Analyst Application”, http://www.math.wpi.edu/saspdf/analyst/pdfidx.htm .

[11] E. Frątczak, M. Pęczkowski, K. Sienkiewicz, K. Skaskiewicz, „Statystyka od podstaw z systemem SAS”, Szkoła Główna Handlowa, Warszawa, (2001).

[12] C. F. Ansley, R. Kohn, and T. S. Shively, „Computing p-Values for the Generalized Durbin-Watson and Other Invariant Test Statistics, *”Journal of Econometrics*, 54, 277–300, (1992).
D.A. Dickey, „Regression with Time Series Errors”, http://www2.sas.com/proceedings/sugi23/ Begtutor/p59.pdf.

[13] R.R. Hocking, „The analysis and selection of variables in linear regression”, „Biometrics 32” 1-49, 1976.

[14] C.L. Mallows, „Some Comments on Cp”, Technometrics, **15**, 661–675, (1973).

[15] A. Boisbunon, S. Canu, D. Fourdrinier, W. Strawderman and M.T. Wells, „AIC and Cp as estimators of loss for spherically symmetric distributions”, arXiv:1308.2766v1, (2013).

[16] D.C. Hoaglin, R.E. Welsch, „The Hat Matrix in Regression and ANOVA”, American Statistician **32**: 17-22, (1978).

[17] R.D. Cook, S. Weisberg, „Residuals and Influence in Regression”, New York: Chapman & Hall, (1982).

[18] R.L. Obenchain, Letter to Editor, Technometrics **19**: 348-49, (1997).

[19] B. Borkowski, H, Dudek, W. Szczesny „Ekonometria wybrane zagadnienia”, Wyda-wnictwo Naukowe PWN, (2003).

[20] The GLM Procedure, SAS/STAT(R) 9.2 User’s Guide, Second Edition, http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#glm_toc.htm , oraz Chapter 30 http://www.math.wpi.edu/saspdf/stat/chap30.pdf ze strony http://www.math.wpi.edu/saspdf/stat/ .

[21] J. Durbin, G.S. Watson, „Testing for Serial Correlation in Least Squares Regression II", Biometrika **38**, 159–178, (1951).
J. Durbin, G.S. Watson, „Testing for Serial Correlation in Least Squares Regression I", Biometrika **37**, 409–428, (1950).
J. Durbin, G.S. Watson, „Testing for Serial Correlation in Least Squares Regression III", Biometrika **58**, 1–19, (1971).

[22] T.S. Breusch, „Testing for Autocorrelation in Dynamic Linear Models", Australian Economic Papers, **17**, 334–355, (1979). L.G. Godfrey, „Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables", Econometrica, **46**, 1293–1302, (1978).

[23] Tablica wartości krytycznych d_l i d_u dla testu Durbina -Watsona ($\alpha = 0{,}05$),
www.statystyka.org/tablice_durbina_watsona.php ,
http://www.ekonometria.4me.pl/durbina.htm .

[24] P.J. Brockwell, R.A. Davis, „Time Series: Theory and methods", second edition, Springer-Verlag, (1991).
Z. Wierzbicka „Wykorzystanie programu statystycznego SAS jako narzędzia do analizy szeregów czasowych", Praca licencjacka, Uniwersytet Śląski, Rybnik, (2004).
S. Zając, „Modelowanie szeregów czasowych za pomocą procesów ARMA i ARIMA" (aplikacje w systemie SAS), Praca licencjacka, Uniwersytet Śląski, Rybnik, (2005).

[25] R.A. Bethel, D. Sheppard, B. Geffroy, E.Tam, J.A. Nadel, H.A. Boushey, „Effect of 0.25 ppm sulfur dioxide on airway resistance in freely breathing, heavily exercising, asthmatic subjects.", Am Rev Respir Dis. **131**:659-61, (1985).

[26] Tablice rozkładów (w tym rozkładu Kołmogorowa-Smirnowa),
http://www.math.uni.wroc.pl/~zpalma/Tablicestatystyczne.pdf ,
http://karasiewicz.az.pl/wp/statystyka/stattables/ ,
http://www.parlinski.pl/stat/tablice/tabl8.html .

[27] „Empirical distribution", A.V. Prokhorov (originator), Encyclopedia of Mathematics,
http://www.encyclopediaofmath.org/index.php?title=Empirical_distribution&oldid=11280 .

[28] V. Glivenko, „Sulla determinazione empirica della legge di probabilita", Giorn. Ist. Ital. Attuari **4**, 92-99, (1933); F.P. Cantelli, „Sulla determinazione empirica delle leggi di probabilita", Giorn. Ist. Ital. Attuari **4**, 221-424, (1933).

[29] „Kolmogorov test. Encyclopedia of Mathematics”, http://www.encyclopediaofmath.org/index.php?title=Kolmogorov_test&oldid=26541 **.**

[30] O. Korosteleva, „Nonparametric Methods in Statistics with SAS Applications”, Chapman & Hall/CRC Texts in Statistical Science, (Chapman and Hall/CRC, 2013) (195 Pages).
„How to Analyze Data with Low Quality or Small Samples, Nonparametric Statistics”, https://www.statsoft.com/Textbook/Nonparametric-Statistics .
„Nonparametric Analysis”, http://support.sas.com/rnd/app/stat/procedures/NonparametricAnalysis.html .
„NONPARAMETRIC METHODS”, http://www.env.gov.bc.ca/epd/remediation/guidance/technical/pdf/12/gd05_all.pdf .

[31] A.N. Kolmogorov, „Sulla determinizione empirica di una legge di distribuzione”, Giorn. Ist. Ital. Attuari **4,** 83–91, (1933).

[32] L.N. Bol'shev, „Asymptotically Pearson transformations”, Theory Probab. Appl. **8**(2), 121–146, (1963), http://dx.doi.org/10.1137/1108012 .

[33] N.V. Smirnov, „On estimating the discrepancy between empirical distribiution curves for two independent samples”, [In Russian], Byull. Moskov. Gos. Univ. Ser. A **2** : 2, 3–14, (1938).

[34] SAS/STAT(R) 9.2 User's Guide, Second Edition, https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#npar1way_toc.htm .

[35] R.A. Fisher „Statistical Methods for Research Workers”, Oliver and Boyd (1925); Twelfth ed., Oliver and Boyd, (1954).
A. Donner, J.J. Koval, „The Estimation of Intraclass Correlation in the Analysis of Family Data”, Biometrics, (International Biometric Society) **36** (1): 19–25, (1980).

[36] R. Nowak, „Statystyka dla fizyków”, Wydawnictwo Naukowe PWN, Warszawa, (2002).

[37] J. Jakubowski, R. Sztencel, „Wstep do teorii prawdopodobienstwa”, wydanie 2, Script, Warszawa, (2001).

[38] Y. Pawitan, „In all likelihood, Statistical Modeling and inference using likelihood”, Oxford, (2001).

[39] S. Amari, H. Nagaoka, „Methods of information geometry, translations of Mathematical monographs”, Vol.191, Oxford University Press, (2000).

[40] D.G. Kleinbium, “Logistic Regression- A Self-Learning Text”, New York: Springer-Verlag, (1994).

[41] M. Biesiada, „Statystyka w ujęciu Bayesowskim”, Skrypt dla studentów ekonofizyki, Uniwersytet Śląski, Instytut Fizyki, (2011), http://el.us.edu.pl/ekonofizyka/index.php/Statystyka_w_uj%C4%99ciu_Bayesowskim.

[42] Ł. Machura, „Analiza Szeregów Czasowych”, skrypt dla studentów kierunku Ekonfizyka, Instytut Fizyki, Uniwersytet Śląski, (2011), http://el.us.edu.pl/ekonofizyka/index.php/Analiza_Szereg%C3%B3w_Czasowych.

[43] H. Akaike, „Fitting autoregressive models for prediction”, Annals of the institute of statistical mathematics Tokyo, **21**, 243-247, (1969).

[44] C.M. Hurvich, C.L. Tsai, „Regression and time series model selection in small samples”, Biometrika **76**, 297-307, (1989).

[45] „Applications of Differential Geometry to Econometrics”, P. Marriott and M. Salmon (Editors), Cambridge University Press, (2011).

[46] J. Łuczka i Ł. Machura, „Dynamika stochastyczna”, skrypt dla studentów kierunku Ekonfizyka, Instytut Fizyki, Uniwersytet Śląski, (2014), http://ekonofizyka.pl/skrypty/StochDyn .

[47] P.J. Brockwell, R.A. Davis, „Introduction to Time Series and Forecasting”, second edition, Springer-Verlag, (2002).

[48] I.N. Bronsztejn, K.A. Siemiendiajew, G. Musiol, H. Mühlig, „Nowoczesne kompendium matematyki”, Warszawa, PWN, (2004).

[49] J. Łuczka, „Procesy i zjawiska losowe”, skrypt dla studentów kierunku Ekonfizyka, Instytut Fizyki, Uniwersytet Śląski, (2011), http://el.us.edu.pl/ekonofizyka/index.php/Procesy_i_Zjawiska_Losowe .

Część IV. Dodatek. Uzupełnienia teoretyczne. Strony 1 - 17 rękopisu.

Wektory losowe i ich rozkłady prawdopodobieństwa.

Niech Z będzie dowolnym zbiorem i niech $\mathscr{F}$ będzie σ - *algebrą podzbiorów zbioru* Z. Oznaczmy przez $\overline{\mathbf{R}}$ zbiór liczb rzeczywistych $\mathbf{R}$ uzupełniony o elementy $\{-\infty\}$ i $\{+\infty\}$ (tzn. prosta przedłużona). Symbol $\emptyset$ oznacza zbiór pusty.

Definicja miary. Funkcję $\mu : \mathscr{F} \to \overline{\mathbf{R}}_+ \equiv \mathbf{R} \cup \{+\infty\}$ zdefiniowaną na σ - algebrze $\mathscr{F}$ nazywamy *miarą*, jeśli [Bron]: a) $\mu(A) \geq 0$ (dla każdego $A \in \mathscr{F}$), b) $\mu(\emptyset) = 0$, c) jeśli $A_1, A_2, ..., A_n \in \mathscr{F}$ i $A_k \cap A_l = \emptyset$ $(k \neq l)$ implikuje $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.

Przykład. Gdy zbiór Z jest przestrzenią zdarzeń Ω, wtedy przykładem miary jest *funkcja prawdopodobieństwa*[1] $P : \mathscr{F} \to \langle 0, +\infty) \cup \{+\infty\}$.

Zbiory należące do $\mathscr{F}$ określamy mianem *mierzalnych* lub $\mathscr{F}$ - *mierzalnych*.

Określenie σ - algebry zbiorów borelowskich. Niech Z będzie *przestrzenią metryczną* [Bron], a $\mathscr{B}(Z)$ najmniejszą σ - algebrą podzbiorów zbioru Z, do której należą wszystkie podzbiory otwarte w Z. $\mathscr{B}(Z)$ istnieje jako część wspólna wszystkich σ - algebr, do których należą wszystkie zbiory otwarte.

$\mathscr{B}(Z)$ nazywamy σ - *algebrą zbiorów borelowskich w* Z, a każdy element algebry $\mathscr{B}(Z)$ nazywamy *zbiorem borelowskim.*

Przykład. Niech $Z = \mathbf{R}^n$ $(n \geq 1)$. Z twierdzenia Lebesgue'a-Caratheodory wynika, że na σ - algebrze $\mathscr{B}(\mathbf{R}^n)$ można skonstruować miarę, zgodną na wszystkich kostkach $\mathbf{R}^n$ z ich objętością. Objętość *n*-wymiarowej kostki Q w przestrzeni $\mathbf{R}^n$, $(Q = \{x \in \mathbf{R}^n ; a_k \leq x_k \leq b_k \ (k = 1, 2, ..., n)\})$, wynosi:

$$\prod_{k=1}^{n} (b_k - a_k).$$

Definicja funkcji mierzalnej. Funkcję $\mathscr{g} : Z \to \overline{\mathbf{R}}$ określamy jako *mierzalną względem* $\mathscr{F}$, jeśli dla dowolnego $\alpha \in \mathbf{R}$ zbiór $\mathscr{g}^{-1}(\langle -\infty, \alpha)) = \{z : z \in Z, \mathscr{g}(z) < \alpha\}$ należy do $\mathscr{F}$.

---

[1] **Uwaga**. W świetle *drugiego aksjomatu* aksjomatycznej definicji prawdopodobieństwa Kołmogorowa, $P(\Omega) = 1$, i w związku z zachodzeniem dla zdarzeń $A, B \in \Omega$ implikacji: $A \subset B \Rightarrow P(A) \leq P(B)$, *pierwszy aksjomat* $P : \mathscr{F} \to \langle 0, 1 \rangle$ można osłabić do $P : \mathscr{F} \to \langle 0, +\infty) \cup \{+\infty\}$. Istotnie: korzystając z tych faktów otrzymujemy: $0 \leq P(A) \leq P(\Omega) = 1$.

Przypomnijmy, że *trzeci aksjomat* Kołmogorowa brzmi: Jeśli zdarzenia *A* i *B* są zdarzeniami wykluczającymi się nawzajem (tzn. $A \cap B = \emptyset$), to prawdopodobieństwo zajścia zdarzenia *A* lub *B* jest równe $P(A \cup B) = P(A) + P(B)$. Z trzeciego aksjomatu oraz osłabionego aksjomatu pierwszego wynika zastosowana powyżej implikacja $A \subset B \Rightarrow P(A) \leq P(B)$.

Przestrzeń mierzalna $(\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n))$. Niech $(\Omega, \mathscr{F}, P)$ oznacza przestrzeń probabilistyczną [49], a $(\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n))$ przestrzeń mierzalną w *n*-wymiarowej przestrzeni euklidesowej $\mathbf{R}^n$, gdzie $\mathscr{B}(\mathbf{R}^n)$ jest σ- algebrą podzbiorów borelowskich w przestrzeni $\mathbf{R}^n$ $(n \geq 1)$.

Definicja wektora losowego. Wektorem losowym *n*-wymiarowym nazywamy funkcję $\vec{X} : \Omega \to \mathbf{R}^n$, która jest mierzalna względem σ- algebry $\mathscr{F}$ ($\mathscr{F}$ - mierzalna), tzn. taką, że dla każdego zbioru $\mathrm{B} \in \mathscr{B}(\mathbf{R}^n)$ zachodzi:

$$\vec{X}^{-1}(\mathrm{B}) \in \mathscr{F}.$$

Mierzalność odwzorowania $\vec{X}$ można symbolicznie oznaczyć jako:

$$\vec{X} : (\Omega, \mathscr{F}) \to (\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n))$$

Definicja zmiennej losowej. Jednowymiarowy wektor losowy nazywamy zmienną losową.

Zarówno wartości zmiennej losowej (np. X) jak i wektora losowego (np. $\vec{X}$), oznaczać będziemy małymi literami, tzn.:

$$x = X(\omega), \quad \vec{x} = \vec{X}(\omega), \qquad \text{gdzie} \quad \omega \in \Omega.$$

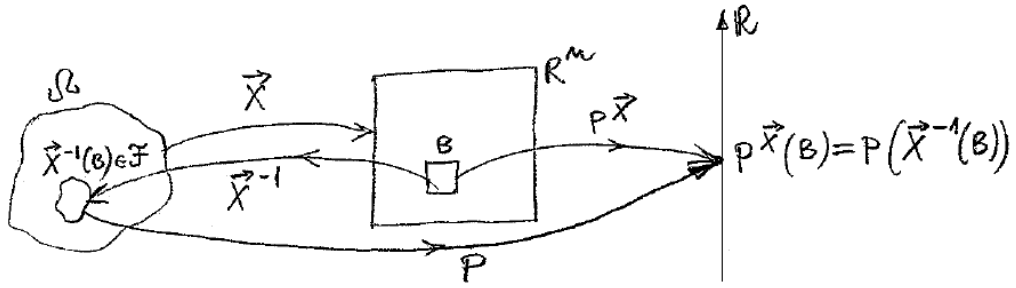
Wektory losowe zapisujemy w postaci kolumn (czyli $n \times 1$-wymiarowych macierzy), tzn. $\vec{X} = (X_1, X_2, ..., X_n)^T$, gdzie T oznacza transpozycję. Wartość wektora losowego w punkcie $\omega \in \Omega$ zapisujemy więc w zmiennych losowych X_i, $i = 1, 2, ..., n$, będących współrzędnymi wektora losowego następująco: $\vec{x} = \vec{X}(\omega) = (X_1(\omega), X_2(\omega), ..., X_n(\omega))^T = (x_1, x_2, ..., x_n)^T$. Wektor $(x_1, x_2, ..., x_n)^T$ jest więc konkretną realizacją wektora losowego $\vec{X}$.

Definicja rozkładu prawdopodobieństwa wektora losowego $\vec{X}$. Rozkładem prawdopodobieństwa wektora losowego $\vec{X}$ nazywamy miarę $P^{\vec{X}}$ na przestrzeni mierzalnej $(\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n))$, określoną nastepująco:

$$P^{\vec{X}}(\mathrm{B}) = P(\vec{X}^{-1}(\mathrm{B})) \qquad \text{dla każdego} \quad \mathrm{B} \in \mathscr{B}(\mathbf{R}^n).$$

Widać więc, że wektor losowy $\vec{X}$ indukuje przestrzeń probabilistyczną $\left(\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n), P^{\vec{X}}\right)$ w przestrzeni euklidesowej $\mathbf{R}^n$.

Poniższy rysunek podsumowuje rozważania obecnego rozdziału.



Rysunek 1.1. Konstrukcja rozkładu prawdopodobieństwa wektora losowego (Opis w tekście).

Przykład. (Niezbędne informacje dotyczące wartości oczekiwanej zmiennej losowej). W przypadku, gdy zmienna losowa X jest typu dyskretnego, przyjmuje ona wartość (realizację) x_i, $i = 1, 2, \ldots$, ze zbioru skończonego lub przeliczalnego, z prawdopodobieństwem $p_i = P(X = x_i)$. W przypadku, gdy zmienna losowa X jest typu ciągłego wtedy, o ile istnieje funkcja $f(x)$ zwana gęstością prawdopodobieństwa, określa się prawdopodobieństwo $P(x_a \le X \le x_b) = \int_{x_a}^{x_b} f(x)\, dx$ zdarzenia oznaczającego, że zmienna losowa X przyjmie wartość z ustalonego, skończonego przedziału $\langle x_a, x_b \rangle$. Prawdopodobieństwo zajścia zdarzenia $X \le \mathbf{x}$, $-\infty < \mathbf{x} < +\infty$, określa dystrybuantę $F(\mathbf{x})$ zmiennej losowej X, $F(\mathbf{x}) = P(X \le \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(x)\, dx$. W ten sposób określony jest rozkład prawdopodobieństwa zmiennej losowej X.

Następnie, niech $h(X)$ jest jednoznaczną funkcją zmiennej losowej X. Wartość oczekiwana zmiennej losowej $h(X)$ jest określona następująco:

$E(h(X)) = \sum_i h(x_i)\, p_i$, dla zmiennej losowej X typu dyskretnego, która przyjmuje realizacje x_i $i = 1, 2, \ldots$,

$E(h(X)) = \int_{-\infty}^{+\infty} h(x) f(x)\, dx$, dla zmiennej losowej X typu ciągłego, która przyjmuje realizacje $-\infty < x < +\infty$,

o ile szereg $\sum_i |h(x_i)|\, p_i$ (w przypadku przeliczalnego zbioru wartości x_i) lub całka $\int_{-\infty}^{+\infty} |h(x)|\, f(x)\, dx$ są zbieżne.

Jeśli $h(X) = X^k$ (k = 1,2,…), to powyższe (ogólne) wartości oczekiwane wyznaczają momenty zwykłe rzędu k zmiennej losowej X. Np. moment zwykły pierwszego rzędu (dla k = 1) jest wartością oczekiwaną $E(X)$ zmiennej losowej X. Jeśli $h(X) = (X - E(X))^k$ (k = 1,2,…), to powyższe wartości oczekiwane wyznaczają momenty centralne rzędu k zmiennej losowej X. Np. moment centralny drugiego rzędu (k = 2) jest wariancją $\sigma^2(X) \equiv E\left((X - E(X))^2\right)$ zmiennej losowej X, której pierwiastek $\sigma(X) = \sqrt{\sigma^2(X)}$ jest nazywany odchyleniem standardowym zmiennej losowej X. Wartości oczekiwane $E(h(X))$ różnych typów funkcji $h(X)$ są *parametrami* charakteryzującymi rozkład zmiennej losowej X. Szczegółowe rozważania dotyczące zarówno parametrów opisujących rozkład zmiennej losowej jak i wektora losowego można znaleźć np. w [2], [6], [36], [48], [49]. W pozycjach tych można również znaleźć przykłady różnych typów rozkładów. Znajdują się tam także informacje dotyczące wartości oczekiwanych i wariancji warunkowych wyznaczanych dla rozkładów warunkowych zmiennych losowych.

Na koniec, jako ilustrację podajmy kilka przykładów wartości oczekiwanych pojawiających się w obecnym skrypcie. Niech zmienne losowe X oraz Y mają łączny rozkład gęstości prawdopodobieństwa $f(x,y)$. Wtedy:

a) ich kowariancja jest określona następująco:

$$\mathrm{cov}(X,Y) \equiv E\big((X-E(X))(Y-E(Y))\big) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(X))(y-E(Y))f(x,y)\,dx\,dy$$

b) warunkowa wartość oczekiwana oraz warunkowa wariancja zmiennej losowej Y pod warunkiem, że zmienna losowa X przyjęła wartość x, są równe kolejno:

$$\mu_{Y|x} \equiv E(Y \mid X=x) = \int_{-\infty}^{+\infty} y f(y \mid x)\,dy$$

oraz

$$\sigma^2_{Y|x} \equiv E\Big((Y-\mu_{Y|x})^2 \mid X=x\Big) = \int_{-\infty}^{+\infty}(y-\mu_{Y|x})^2 f(y \mid x)\,dy\,,$$

gdzie $f(y \mid x)$ jest warunkowym rozkładem gęstości prawdopodobieństwa zmiennej losowej Y (pod warunkiem, że zmienna losowa X przyjęła wartość x), określonym następująco:

$$f(y \mid x) = \frac{f(x,y)}{f(x)}\,,$$

oraz $f(x)$ jest brzegowym rozkładem gęstości prawdopodobieństwa zmiennej losowej X, określonym następująco:

$$f(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy\,.$$

c) wartości oczekiwane zmiennych losowych X i Y, wyznaczone z ich rozkładów brzegowych lub z rozkładu łącznego, są określone następująco:

$$\mu_X \equiv E(X) = \int_{-\infty}^{+\infty} x f(x)\,dx = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\, f(x,y)\,dx\,dy$$

$$\mu_Y \equiv E(Y) = \int_{-\infty}^{+\infty} y f(y)\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\, f(x,y)\,dx\,dy\,,$$

gdzie $f(y)$ jest brzegowym rozkładem gęstości prawdopodobieństwa zmiennej losowej Y, określonym następująco:

$$f(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx\,.$$

Wspomnijmy jeszcze tylko, że w przeciwieństwie do (ogólnych) wartości oczekiwanych omówionych tuż powyżej, warunkowe momenty są w ogólności zmiennymi losowymi. Uogólnienie powyższych pojęć na przypadek więcej wymiarowy jest natychmiastowe. Proste jest również zapisanie powyższych wielkości dla przypadku wielowymiarowego rozkładu dyskretnego [2].

Wyznaczenie funkcji regresji $E(Y|x)$ dla dwuwymiarowego rozkładu normalnego zmiennych (X,Y):

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} \exp\left\{-\frac{1}{2(1-\varrho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\varrho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right\}$$

gdzie: $\mu_X \equiv E(X)$, $\mu_Y \equiv E(Y)$, $\varrho \equiv \varrho_{XY}$ $\qquad (x,y) \in \mathbb{R}^2$

$\sigma_X \equiv \sigma(X)$, $\sigma_Y \equiv \sigma(Y)$

Gęstość warunkowa rozkładu: $f(y|x) = \dfrac{f(x,y)}{f(x)}$

Rozkład brzegowy $f(x) = \displaystyle\int_{-\infty}^{+\infty} f(x,y)\,dy =$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2(1-\varrho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\varrho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right\} dy =$$

$$= \left\{\begin{array}{l} t = \dfrac{y-\mu_Y}{\sigma_Y} \\ y = \mu_Y + \sigma_Y t \\ dy = \sigma_Y\,dt \end{array}\right\} = \frac{e^{-\frac{1}{2(1-\varrho^2)}\frac{(x-\mu_X)^2}{\sigma_X^2}}}{2\pi\sigma_X\cancel{\sigma_Y}\sqrt{1-\varrho^2}}\cancel{\sigma_Y}\int_{-\infty}^{+\infty} e^{-\frac{1}{2(1-\varrho^2)}\left(-2\varrho\frac{x-\mu_X}{\sigma_X}t + t^2\right)} dt =$$

$$= \frac{e^{-\frac{1}{2(1-\varrho^2)}\frac{(x-\mu_X)^2}{\sigma_X^2}}}{2\pi\sigma_X\sqrt{1-\varrho^2}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2(1-\varrho^2)}\left(\varrho^2\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\varrho\frac{(x-\mu_X)}{\sigma_X}t + t^2 - \varrho^2\frac{(x-\mu_X)^2}{\sigma_X^2}\right)} dt$$

$$= \frac{e^{-\frac{1}{2(1-\varrho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \varrho^2\frac{(x-\mu_X)^2}{\sigma_X^2}\right]}}{2\pi\sigma_X\sqrt{1-\varrho^2}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2(1-\varrho^2)}\left(t - \frac{\varrho(x-\mu_X)}{\sigma_X}\right)^2} dt =$$

$$= \left\{\frac{t - \frac{\varrho(x-\mu_X)}{\sigma_X}}{\sqrt{1-\varrho^2}} = u, \quad t = \sqrt{1-\varrho^2}\,u + \varrho\frac{(x-\mu_X)}{\sigma_X}, \quad dt = \sqrt{1-\varrho^2}\,du\right\} =$$

$$= \frac{e^{-\frac{1}{\cancel{2(1-\varrho^2)}}\left[\cancel{(1-\varrho^2)}\cdot\frac{(x-\mu_X)^2}{\sigma_X^2}\right]}\cancel{\sqrt{1-\varrho^2}}}{2\pi\sigma_X\cancel{\sqrt{1-\varrho^2}}} \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} du =$$

$$= \left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} du = 1 \quad \left(\text{z norm. standardowego rozkł. normalnego}\right)\right\} = \frac{e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}\cdot\sqrt{2\pi}}{2\pi\sigma_X} = \frac{1}{\sqrt{2\pi}\,\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}$$

Zatem rozkład brzegowy zmiennej X ma postać:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma_X} e^{-(x-\mu_X)^2/(2\sigma_X^2)}.$$

Wyznaczmy rozkład warunkowy: $f(y|x) = \dfrac{f(x,y)}{f(x)} =$

$$= \frac{\dfrac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\dfrac{1}{2(1-\rho^2)}\left(\dfrac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\dfrac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \dfrac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right\}}{\dfrac{1}{\sqrt{2\pi}\,\sigma_X}\exp\left(-\dfrac{(x-\mu_X)^2}{2\sigma_X^2}\right)} =$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right) + \frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} =$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_Y^2}\left(\frac{\sigma_Y^2}{\sigma_X^2}(x-\mu_X)^2 - 2\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)(y-\mu_Y) + (y-\mu_Y)^2 - \frac{(1-\rho^2)\sigma_Y^2}{\sigma_X^2}(x-\mu_X)^2\right)\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_Y^2}\left((y-\mu_Y)^2 - 2\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)(y-\mu_Y) - \rho^2\frac{\sigma_Y^2}{\sigma_X^2}(x-\mu_X)^2\right)\right\} =$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_Y^2(1-\rho^2)}}\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left((y-\mu_Y) - \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)^2\right\} =$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_Y^2(1-\rho^2)}}\exp\left\{-\left(y-\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)\right)^2 \Big/ \left(2\sigma_Y^2(1-\rho^2)\right)\right\} =$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_{Y|X}} e^{-(y-\mu_{Y|X})^2/(2\sigma_{Y|X}^2)}$$

gdzie warunkowa wartość oczekiwana wynosi:

$$\boxed{E(Y|x) = \mu_{Y|X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}$$

natomiast wariancja warunkowa:

$$\boxed{\sigma_{Y|X}^2 = \sigma_Y^2(1-\rho^2)}$$

Widać, że $E(Y|x)$ jest liniową funkcją x, więc funkcja regresji $E(Y|x)$ jest funkcją liniową, oraz $\sigma_{Y|X}^2$ jest taka sama dla wszystkich wartości x.

Wniosek: W analizie regresji zmiennej Y względem X w przypadku, gdy rozkład dwuwymiarowy (X,Y) jest normalny, funkcja regresji $E(Y|x)$ jest liniowa, a przed przystąpieniem do analizy należy przeprowadzić test jednorodności wariancji.

Składowe główne (zasadnicze).

(1). Składowe zasadnicze w populacji.

Rozważmy wektor losowy $\vec{X}$ o p składowych. Jego macierz kowariancji to:

$$\Sigma \equiv \mathrm{cov}(\vec{X}) = E(\vec{X}\vec{X}^T) = \left(E(X_i - \bar{X}_i)(X_j - \bar{X}_j)\right) = (\sigma_{ij}), \quad i,j = 1,2,\dots,p. \tag{1}$$

Pomijając rozważać będziemy wariancje i kowariancje zmiennych losowych, więc załóżmy, że $\bar{X}_j = 0$, $j = 1,2,\dots,p$.

Dla poniższych rozważań nie jest też istotny typ rozkładu zmiennej losowej $\vec{X}$. (Niemniej w przypadku gdy $\vec{X}$ ma rozkład normalny, składowe zasadnicze stają się szczególnie użyteczne).

Poniższe podejście [6] pozwala na uwzględnienie przypadków gdy macierz $\Sigma = (\sigma_{ij})$ jest dodatnio półokreślona oraz Σ może mieć pierwiastki wielokrotne.

Rozważmy pewną kombinację czynników $\vec{X}$ (inną niż ta, która występuje w $E(Y|\vec{X})$):

$$\vec{\beta}^T \vec{X} \tag{2}$$

gdzie

$$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \tag{3}$$

Niech wektor $\vec{\beta}$ będzie unormowany, tzn.:

$$\vec{\beta}^T \vec{\beta} = 1 \tag{4}$$

(A)

Wariancja zmiennej losowej $\vec{\beta}^T\vec{X}$ wynosi:

$$E\left\{(\vec{\beta}^T\vec{X})^2\right\} = E\left\{(\vec{\beta}^T X)(\vec{\beta}^T X)^T\right\} = E\left\{\vec{\beta}^T X X^T \vec{\beta}\right\} =$$
$$= \vec{\beta}^T E\left\{X X^T\right\}\vec{\beta} = \vec{\beta}^T \Sigma \vec{\beta} \tag{5}$$

Obecnie unormowaną kombinację liniową $\vec{\beta}^T\vec{X}$, która posiada maksymalną wariancję. W tym celu znajdziemy wektor $\vec{\beta}$ spełniający warunek $\vec{\beta}^T\vec{\beta} = 1$, (4), który jednocześnie maksymalizuje wariancję (5). Zatem szukane rozwiązanie odpowiada znalezieniu maksimum warunkowego. Szukamy więc maksimum funkcji:

$$\phi = \vec{\beta}^T \Sigma \vec{\beta} - \lambda\underbrace{(\vec{\beta}^T\vec{\beta} - 1)}_{\text{z war. (4)}} = \sum_{ij} \beta_i \sigma_{ij} \beta_j - \lambda\left(\sum_i \beta_i^2 - 1\right), \tag{6}$$

gdzie λ jest mnożnikiem Lagrange'a.

Wektor pochodnych cząstkowych ϕ ma j-tą składową równą:

$$\frac{\partial \phi}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}\left(\sum_{l,s} \beta_l \sigma_{ls} \beta_s - \lambda\left(\sum_l \beta_l^2 - 1\right)\right) =$$

$$= \sum_{l,s} \frac{\partial \beta_l}{\partial \beta_j} \sigma_{ls} \beta_s + \sum_{l,s} \beta_l \sigma_{ls} \frac{\partial \beta_s}{\partial \beta_j} - \lambda\left(\sum_l 2\beta_l \frac{\partial \beta_l}{\partial \beta_j}\right) =$$

$$= \sum_{l,s} \delta_{lj} \sigma_{ls} \beta_s + \sum_{l,s} \beta_l \sigma_{ls} \delta_{sj} - 2\lambda \sum_l \beta_l \delta_{lj} =$$

$$= \sum_s \sigma_{js} \beta_s + \sum_l \beta_l \sigma_{lj} - 2\lambda \beta_j = \{ \sigma_{lj} = \sigma_{jl} \} =$$

$$= 2\sum_l \sigma_{jl} \beta_l - 2\lambda \beta_j \qquad (7)$$

Zatem:

$$\frac{\partial \phi}{\partial \vec{\beta}} \equiv \left(\frac{\partial \phi}{\partial \beta_j}\right) = 2\Sigma\vec{\beta} - 2\lambda\vec{\beta} = 2(\Sigma - \lambda I)\vec{\beta} \qquad (8)$$

Ponieważ $\vec{\beta}^T \Sigma \vec{\beta}$ oraz $\vec{\beta}^T\vec{\beta}$ mają pochodne w całym obszarze, w którym spełniony jest warunek $\vec{\beta}^T\vec{\beta} = 1$, (4), zatem wektor $\vec{\beta}$ maksymalizujący $\vec{\beta}^T \Sigma \vec{\beta}$, (5), musi spełniać układ równań:

$$\frac{\partial \phi}{\partial \vec{\beta}} = 2(\Sigma - \lambda I)\vec{\beta} = 0 \qquad (9)$$

Aby rozwiązanie równania (9) było nietrywialne ($\vec{\beta} \neq 0$), macierz $\Sigma - \lambda I$ musi być osobliwa, co oznacza, że musi zachodzić warunek:

$$\det(\Sigma - \lambda I) = 0 \qquad (10)$$

Funkcja $\det(\Sigma - \lambda I)$ jest wielomianem λ rzędu p. Zatem (10) ma p pierwiastków. Oznaczmy i uporządkujmy je następująco:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \qquad (11)$$

Uwaga: Macierz kowariancji dla zmiennych losowych rzeczywistych jest rzeczywista i symetryczna:

$$\Sigma = \Sigma^T \qquad (12)$$

W przypadku gdy zmienne losowe $\vec{X}$ są zespolone wtedy macierz kowariancji

$$\Sigma = E\{(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^+\} \equiv (\sigma_{ij}) \qquad (13)$$

jest hermitowska:

$$\Sigma^+ = \Sigma \qquad (14)$$

gdzie "+" oznacza sprzężenie hermitowskie: $\Sigma^+ = (\Sigma^*)^T$. (15)

Przypadek macierzy rzeczywistej, symetrycznej jest szczególnym przykładem macierzy hermitowskiej.

Wartości własne λ macierzy Σ (która jest hermitowska) są rzeczywiste.
W ogólnym przypadku zmiennej losowej $\vec{X}$ zespolonej warunek unormowania (4) ma postać:

$$\vec{\beta}^{+}\beta = 1 \qquad \text{(koniec dowodu)} \tag{16}$$

Mnożąc (9) lewostronnie przez β^T otrzymujemy równanie:

$$\vec{\beta}^T(\Sigma - \lambda I)\vec{\beta} = 0 \tag{17}$$

skąd

$$\vec{\beta}^T \Sigma \vec{\beta} = \lambda \vec{\beta}^T\vec{\beta} = \left\{ \underset{\text{warunek (4)}}{\vec{\beta}^T\vec{\beta} = 1} \right\} = \lambda$$

Zatem pokazaliśmy, że o ile $\vec{\beta}$ spełnia warunek unormowania (4), wtedy wariancja (5) wynosi:

$$E\{(\vec{\beta}^T\vec{X})^2\} = \vec{\beta}^T \vec{X}\vec{\beta} = \lambda \tag{18}$$

Ponieważ wariancja ~~to~~ była maksymalizowana, zatem jest ona pierwszą wartością λ_1 stojącą w ciągu (11).

Jeśli przez $\vec{\beta}^{(1)}$ oznaczymy unormowane rozwiązanie równania $(\Sigma - \lambda_1 I)\vec{\beta} = \vec{0}$, (9), dla $\lambda = \lambda_1$, wtedy unormowana ($\vec{\beta}^{(1)T}\vec{\beta}^{(1)} = 1$) liniowa kombinacja

$$U_1 = \vec{\beta}^{(1)T}\vec{X} \tag{19}$$

ma maksymalną wariancję.

Dowód: Jeśli rząd (rząd) macierzy $\Sigma - \lambda_1 I$ wynosi $p-1$, wtedy istnieje tylko jedno rozwiązanie równania $(\Sigma - \lambda_1 I)\vec{\beta} = \vec{0}$ z warunkiem $\vec{\beta}^T\vec{\beta} = 1$.

(B)

Znajdźmy unormowaną kombinację $\vec{\beta}^T\vec{X}$, która ma maksymalną wariancję z pośród wszystkich liniowych kombinacji nieskorelowanych z U_1, (19).
Brak korelacji oznacza (przypomnijmy, że założyliśmy $\vec{\bar{X}} = \vec{0}$):

$$0 = E\{(\vec{\beta}^T\vec{X})U_1^T\} = E\{(\vec{\beta}^T\vec{X})(\vec{\beta}^{(1)T}\vec{X})^T\} = \tag{20}$$

$$= E\{(\vec{\beta}^T\vec{X})(\vec{X}^T\vec{\beta}^{(1)})\} = \vec{\beta}^T E\{\vec{X}\vec{X}^T\}\vec{\beta}^{(1)} = \vec{\beta}^T\Sigma\vec{\beta}^{(1)} \tag{21}$$

$$= \lambda_1 \vec{\beta}^T\vec{\beta}^{(1)} \tag{22}$$

gdzie przy przejściu od (21) do (22) skorzystaliśmy z (9), tzn.:

$$\Sigma\vec{\beta}^{(1)} = \lambda_1\vec{\beta}^{(1)} \tag{23}$$

Stąd widać, że $\vec{\beta}^T\vec{X}$ jest ortogonalne do U_1 zarówno w znaczeniu ~~statystycznym~~ ($E\{(\vec{\beta}^T\vec{X})U_1\} = 0$) jak i w znaczeniu geometrycznym, ~~to~~ znany w znaczeniu ortogonalności wektorów $\vec{\beta}$ oraz $\vec{\beta}^{(1)}$:

$$\vec{\beta}^T\vec{\beta}^{(1)} = 0 \qquad \text{dla } \lambda_1 \neq 0 \tag{24}$$

przy czym $\lambda_1 \neq 0$ gdy $\Sigma \neq 0$. (25)

Z kolei maksymalizujemy funkcję (w celu znalezienia kolejnego $\vec{\beta}$):

$$\phi_2 = \underbrace{\vec{\beta}^T \Sigma \vec{\beta}}_{\substack{\text{wariancja} \\ E\{(\vec{\beta}^T X)^2\}}} - \lambda\underbrace{(\vec{\beta}^T\vec{\beta} - 1)}_{\substack{\text{normalizacja} \\ \vec{\beta}^T\vec{\beta}=1}} - 2\nu_1 \underbrace{\vec{\beta}^T \Sigma \vec{\beta}^{(1)}}_{\substack{\text{z warunku (21)} \\ \vec{\beta}^T\Sigma\vec{\beta}^{(1)}=0}} \tag{26}$$

(27)

gdzie λ oraz ν_1 są mnożnikami Lagrange'a.

Wektor pochodnych cząstkowych ma postać (pokazać, porównaj (7)–(8))

$$\frac{\partial \phi_2}{\partial \vec{\beta}} = 2\Sigma\vec{\beta} - 2\lambda\vec{\beta} - 2\nu_1 \Sigma \vec{\beta}^{(1)} \tag{28}$$

(29)

Z żądania maksimum warunkowego $\frac{\partial \phi_2}{\partial \vec{\beta}} = \vec{0}$, otrzymujemy:

$$\Sigma\vec{\beta} - \lambda\vec{\beta} - \nu_1 \Sigma \vec{\beta}^{(1)} = \vec{0} \tag{30}$$

Mnożąc (28) lewostronnie przez $\vec{\beta}^{(1)T}$, otrzymujemy:

$$\underbrace{\vec{\beta}^{(1)T}\Sigma\vec{\beta}}_{\substack{= 0 \text{ z (21) oraz} \\ (12),\ \Sigma^T=\Sigma}} - \underbrace{\lambda\vec{\beta}^{(1)T}\vec{\beta}}_{\substack{= 0 \text{ z (22)}}} - \underbrace{\nu_1\vec{\beta}^{(1)T}\Sigma\vec{\beta}^{(1)}}_{\substack{= \nu_1\vec{\beta}^{(1)T}\lambda_1\vec{\beta}^{(1)} \\ \text{z (23)}}} = 0 \tag{31}$$

(32)

Zatem z (31), otrzymujemy:

$$\nu_1\lambda_1\underbrace{\vec{\beta}^{(1)T}\vec{\beta}^{(1)}}_{\substack{= 1 \text{ z unormowania}}} = 0 , \tag{33}$$

skąd:

$$\nu_1\lambda_1 = 0 \tag{33}$$

Ze względu na $\lambda_1 \neq 0$, (25), z (33) otrzymujemy:

$$\nu_1 = 0 , \tag{34}$$

więc z (30) widać, że:

$$(\Sigma - \lambda I)\vec{\beta} = 0 \tag{35}$$

tak jak w (9), przy czym λ spełnia (10).

Niech $\lambda_{(2)}$ będzie największą z $\lambda_1, \lambda_2, \dots, \lambda_p$, dla której istnieje $\vec{\beta}$ spełniające trzy warunki:

z (35) $\Rightarrow$

$$(\Sigma - \lambda_{(2)} I)\vec{\beta} = 0 \tag{36}$$

z normalizacji $\Rightarrow$

$$\vec{\beta}^T\vec{\beta} = 1 \tag{37}$$

z ortogonalizacji $\vec{\beta}^T\vec{X}$ do U_1

$$0 = E(\vec{\beta}^T X U_1) = E(\vec{\beta}^T X X^T \vec{\beta}^{(1)}) = \vec{\beta}^T\Sigma\vec{\beta}^{(1)} = \lambda_1\vec{\beta}^T\vec{\beta}^{(1)} \tag{38}$$

To nowe $\vec{\beta}$ oznaczamy $\vec{\beta}^{(2)}$, a odpowiadającą mu liniową kombinację, to:

$$U_2 = \vec{\beta}^{(2)T}\vec{X} \tag{39}$$

Zdefiniujmy $\lambda_{(1)}$ jako λ_1: $\lambda_{(1)} = \lambda_1$ (40)

W ~~ostatecznosci~~ pokażemy, że:

$$\lambda_{(2)} = \lambda_2 . \tag{41}$$

Procedura ~~ta~~ jest kontynuowana. W $(r+1)$ kroku szukamy wektora $\vec{\beta}$ ~~takiego~~, że $\vec{\beta}^T \vec{X}$ ma maksymalną wariancję i jest nieskorelowane z unormowanymi liniowymi kombinacjami $U_1, U_2, \dots, U_r$, tzn:

$$0 = E\left(\vec{\beta}^T \vec{X}\, U_i\right) = E\left(\vec{\beta}^T \vec{X} \underbrace{\vec{X}^T \vec{\beta}^{(i)}}_{U_i}\right) = \vec{\beta}^T \Sigma \vec{\beta}^{(i)} = \tag{42}$$

$$= \lambda_{(i)} \vec{\beta}^T \vec{\beta}^{(i)}, \quad i = 1, 2, \dots, r \tag{43}$$

Tak jak poprzednio maksymalizujemy funkcję:

$$\phi_{r+1} = \vec{\beta}^T \Sigma \vec{\beta} - \lambda\left(\vec{\beta}^T \vec{\beta} - 1\right) - 2 \sum_{i=1}^{r} \nu_i \vec{\beta}^T \Sigma \vec{\beta}^{(i)} \tag{44}$$

gdzie λ oraz $\nu_1, \dots, \nu_r$ są mnożnikami Lagrange'a.

Wektor pochodnych cząstkowych wynosi:

$$\frac{\partial \phi_{r+1}}{\partial \vec{\beta}} = \underbrace{2\Sigma\vec{\beta} - 2\lambda\vec{\beta}} - 2 \sum_{i=1} \nu_i \Sigma \vec{\beta}^{(i)} \tag{45}$$

i z warunku maksymalizacji oraz po pomnożeniu przez $\vec{\beta}^{(j)T}$, mamy:

$$\vec{\beta}^{(j)T} \Sigma \vec{\beta} - \lambda \vec{\beta}^{(j)T} \vec{\beta} - \nu_j \vec{\beta}^{(j)T} \Sigma \vec{\beta}^{(j)} = 0 \tag{46}$$

gdzie w ostatnim składniku skorzystano z:

$$\vec{\beta}^{(j)T} \sum_{i=1}^{r} \nu_i \Sigma \vec{\beta}^{(i)} = \nu_j \vec{\beta}^{(j)T} \Sigma \vec{\beta}^{(j)} + \sum_{\substack{i=1 \\ i \neq j}}^{r} \nu_i \underbrace{\vec{\beta}^{(j)T} \Sigma \vec{\beta}^{(i)}}_{=0} \tag{47}$$

Jeśli $\lambda_{(j)} \neq 0$, wtedy równanie (46) daje:

$$\nu_j \lambda_{(j)} = 0 \tag{48}$$

skąd:

$$\nu_j = 0$$

Jeśli $\lambda_{(j)} = 0$ to $\Sigma \vec{\beta}^{(j)} = \lambda_{(j)} \vec{\beta}^{(j)} = 0$ (49)

i w (45) znika j-ty składnik w $\sum_{i=1} \nu_i \Sigma \vec{\beta}^{(i)}$. Zatem zarówno dla $\lambda_{(j)} \neq 0$ jak i $\lambda_{(j)} = 0$, otrzymujemy:

$$\left(\Sigma - \lambda I\right)\vec{\beta} = 0 \tag{50}$$

dla $\lambda = \lambda_{(r+1)}$ oraz $\vec{\beta} = \vec{\beta}^{(r+1)}$.

Niech teraz $\lambda_{(r+1)}$ jest wartością maksymalną pośród $\lambda_1, \lambda_2, \dots, \lambda_p$ takiego, że $\vec{\beta}$ spełnia warunki:

$$\begin{cases} \left(\Sigma - \lambda_{(r+1)} I\right)\vec{\beta} = \vec{0} \\ \vec{\beta}^T \vec{\beta} = 1 \\ 0 = E\left(\vec{\beta}^T X U_i^T\right) = E\left(\vec{\beta}^T X X^T \vec{\beta}^{(i)}\right) = \vec{\beta}^T \Sigma \vec{\beta}^{(i)} = \\ \qquad = \lambda_{(i)} \vec{\beta}^T \vec{\beta}^{(i)} \end{cases} \tag{51}$$

Wektor $\vec{\beta}$ spełniający powyższe warunki nazywamy $\vec{\beta}^{(r+1)}$, a odpowiadającą mu liniową kombinacją, to: $U_{r+1} = \vec{\beta}^{(r+1)T} \vec{X}$.

Uwaga! Gdyby $\lambda_{(r+1)} = 0$ oraz dla niektórych $j \neq r+1$, $\lambda_j = 0$, wtedy $\vec{\beta}^{(j)T} \Sigma \vec{\beta}^{(r+1)} = 0$ nie pociąga za sobą $\vec{\beta}^{(j)T} \vec{\beta}^{(r+1)} = 0$.
Wtedy $\vec{\beta}^{(r+1)}$ należy zastąpić przez kombinację (np. sumę) $\vec{\beta}^{(r+1)}$ oraz $\vec{\beta}^{(j)}$, dla których $\lambda_j = 0$, tak aby wektor $\vec{\beta}^{(r+1)}$ ortogonalny do $\vec{\beta}^{(j)}$, $j = 1, 2, \ldots, r$.
Procedura ta jest kontynuowana do czasu, aż w kroku $(r+1)$-szym nie można już znaleźć $\vec{\beta}$ spełniającego warunku $\vec{\beta}^T \vec{\beta} = 1$.
Ponieważ można pokazać, że dla ~~ostatniego~~ r nie może zachodzić warunek $r < p$, zatem ostatnie r spełnia warunek [6]:

$$r_{(ostatnie)} = p. \tag{52}$$

Nie może być też tak, aby $r > p$, gdyż wektory $\vec{\beta}^{(1)}, \vec{\beta}^{(2)}, \ldots, \vec{\beta}^{(r)}$ muszą być niezależne.

Niech $\mathbb{\beta} = (\vec{\beta}^{(1)}, \dots, \vec{\beta}^{(p)})$ (53)

oraz

$$\Lambda = \begin{pmatrix} \lambda_{(1)} & 0 & \cdots & 0 \\ 0 & \lambda_{(2)} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{(p)} \end{pmatrix} \tag{54}$$

Równania $\underset{p\times p}{\Sigma}\, \underset{p\times 1}{\vec{\beta}^{(r)}} = \lambda_{(r)} \underset{p\times 1}{\vec{\beta}^{(r)}}$ można łącznie zapisać, postępując:

$$\Sigma\, \mathbb{\beta} = \mathbb{\beta}\, \Lambda \tag{55}$$

natomiast równania $\vec{\beta}^{(r)T} \vec{\beta}^{(r)} = 1$, $\vec{\beta}^{(r)T} \vec{\beta}^{(s)} = 0$ dla $r \neq s$ (56)

można zapisać następująco:

$$\mathbb{\beta}^T \mathbb{\beta} = I \tag{57}$$

Z równań (55) oraz (57) otrzymujemy:

$$\mathbb{\beta}^T \Sigma\, \mathbb{\beta} = \mathbb{\beta}^T \mathbb{\beta}\, \Lambda = I \Lambda = \Lambda$$

czyli

$$\mathbb{\beta}^T \Sigma\, \mathbb{\beta} = \Lambda \tag{58}$$

Rozważmy wyznacznik $\det(\Sigma - \lambda I)$:

$$\det(\Sigma - \lambda I) = \det\left(I(\Sigma - \lambda I)\right) = \det\left[\overbrace{(\mathbb{\beta}^T \mathbb{\beta})}^{I}(\Sigma - \lambda I)\right] =$$

$$= \det(\mathbb{\beta}^T) \det(\Sigma - \lambda I) \det(\mathbb{\beta}) =$$

$$= \det\left[\mathbb{\beta}^T (\Sigma - \lambda I)\, \mathbb{\beta}\right] =$$

$$= \det\left[\mathbb{\beta}^T \Sigma\, \mathbb{\beta} - \lambda\, \mathbb{\beta}^T \mathbb{\beta}\right] = \det\left[\Lambda - \lambda I\right] =$$

$$= \prod_{j=1}^{p} (\lambda_{(j)} - \lambda) \tag{59}$$

Widzimy więc, że pierwiastki wyznacznika $\det(\Sigma - \lambda I)$, czyli λ_j, są ze względu na warunek $\det(\Sigma - \lambda I) = 0$, (10), równe $\lambda_{(j)}$, tzn.:

$$\lambda_{(1)} = \lambda_1, \quad \lambda_{(2)} = \lambda_2, \dots, \quad \lambda_{(p)} = \lambda_p \tag{60}$$

W ten sposób udowodniono [6] twierdzenie:

Twierdzenie (o składowych głównych (zasadniczych))

Niech p-wymiarowy wektor losowy $\vec{X}$ ma wartość oczekiwaną $E(\vec{X})=\vec{0}$ (61)
oraz macierz kowariancji $E(\vec{X}\vec{X}^T)=\Sigma$. (62)
Istnieje wtedy ortogonalna liniowa transformacja:

$$\vec{U} = \mathbb{P}^T \vec{X} \tag{63}$$

taka, że macierz kowariancji dla $\vec{U}$ jest równa:

$$E(\vec{U}\vec{U}^T) = \Lambda \tag{64}$$

gdzie:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \tag{65}$$

przy czym:

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0 \tag{66}$$

są pierwiastkami równania $\det(\Sigma - \lambda I)=0$, (10). (67)
Kolumna r-ta macierzy $\mathbb{P}$, czyli $\vec{p}^{(r)}$ spełnia równanie (9):

$$(\Sigma - \lambda_r I)\, p^{(r)} = \vec{0} \tag{68}$$

Składową r-tą wektora $\vec{U}$ jest $U_r = \vec{p}^{(r)T}\vec{X}$. (69)
Liniowa kombinacja U_r ma maksymalną wariancję pośród wszystkich liniowych kombinacji nieskorelowanych ze zmiennymi $U_1, U_2, \ldots, U_{r-1}$.

Wektor $\vec{U}$ jest tak zwanym wektorem składowych zasadniczych (głównych) wektora losowego $\vec{X}$.

Znalezienie składowych zasadniczych polega na znalezieniu takich liniowych kombinacji zmiennych wektora losowego $\vec{X}$, które mają maksymalną wariancję i są z sobą nieskorelowane. W praktycznych zastosowaniach zdarza się, że liczba zmiennych branych pod uwagę jest za duża. Ponieważ istotna sprawa w tej analizie jest rozmiar wartości czynników, dlatego metoda składowych zasadniczych pozwala na odrzucenie tych liniowych kombinacji zmiennych wektora losowego $\vec{X}$, które mają małą wariancję i na poddaniu analizie kombinacji z dużą wariancją.

Przykład Badany jest wpływ różnych cech określających standard mieszkania, na jego cenę. Z cech tych, metodą składowych zasadniczych tworzy liniowe ich kombinacje, które różnicują jakoś jednostki badanej zbiorowości mieszkań. Te z kombinacji, które mają największy rozrzut wartości przy zmianie mieszkania są interesujące. Natomiast kombinacje, które zmieniają się nieznacznie od mieszkania do mieszkania można mało o zmienności pomiędzy mieszkaniami i można je usunąć z analizy.

(2). Estymatory MNW składowych zasadniczych oraz ich wariancji [6].

Ostatecznie podlegają wektory $\vec{\beta}^{(1)}, \vec{\beta}^{(2)}, \dots, \vec{\beta}^{(p)}$ oraz skalary $\lambda_1, \lambda_2, \dots, \lambda_p$.

Zastosujemy algebrę poprzedniego rozdziału do estymatora macierzy kowariancji

Twierdzenie. Niech $\vec{x}_1, \dots, \vec{x}_n$ są próbką n obserwacji $(n > p)$ otrzymanymi z rozkładu $N(\vec{\mu}, \Sigma)$, gdzie Σ is macierzą z p różnymi pierwiastkami równania charakterystycznego (10).

Wtedy zbiór estymatorów parametrów $\lambda_1, \dots, \lambda_p$ oraz wektorów $\vec{\beta}_1^{(1)}, \dots, \vec{\beta}^{(p)}$ zdefiniowanych w Twierdzeniu (o składowych zasadniczych) trzyma pierwiastki:

$$\hat{\lambda}_1 > \hat{\lambda}_2 > \dots > \hat{\lambda}_p \qquad (1)^*$$

równanie charakterystyczne:

$$\det(\hat{\Sigma} - \hat{\lambda} I) = 0, \qquad (2)^*$$

a zbiór odpowiednich wektorów $\hat{\vec{\beta}}^{(1)}, \dots, \hat{\vec{\beta}}^{(p)}$ spełnia równania własne:

$$(\hat{\Sigma} - \hat{\lambda}_j I)\hat{\vec{\beta}}^{(j)} = 0 \qquad (3)^*$$

oraz warunek normalizacji:

$$\hat{\vec{\beta}}^{(j)T}\hat{\vec{\beta}}^{(j)} = 1, \qquad (4)^*$$

gdzie $\hat{\Sigma}$ jest estymatorem MNW macierzy Σ równym:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i - \bar{\vec{x}})(\vec{x}_i - \bar{\vec{x}})^T \qquad (5)^*$$

Uwaga: Powyższa funkcja wiarygodności zależy jedynie od $\hat{\Sigma}$, zatem maksimum funkcji wiarygodności jest osiągnięte przez podstawienie obserwowanego zbioru rozwiązań równań $(3)^*$ i $(4)^*$. W naszej bazie składowych zasadniczych mamy:

$$\hat{\Sigma} = \hat{\beta}\hat{\Lambda}\hat{\beta}^T = \sum_j \hat{\lambda}_j \hat{\vec{\beta}}^{(j)}\hat{\vec{\beta}}^{(j)T}$$

Uwaga: Funkcja wiarygodności n obserwacji zmiennej $\vec{X}$ ma postać:

$$L = \prod_{i=1}^{n} f_i(\vec{x}_i \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}pn}(\det\Sigma)^{\frac{1}{2}n}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\vec{x}_i - \vec{\mu})^T\Sigma^{-1}(\vec{x}_i - \vec{\mu})\right]$$

Uwaga: Można zdefiniować tzw. indeks warunkowy

$$CI_j = \sqrt{\frac{\hat{\lambda}_1}{\hat{\lambda}_j}} \qquad j = 1, 2, \dots, p$$

Przyjmuje się, że $CI_j > 10$ oznacza występowanie kowariancji czynników, które można pominąć w analizie regresji Y od $\vec{X}$. Gdy dla pewnego j, $\lambda_j = 0$, wtedy w układzie zmiennych początkowych $\vec{X}$ występuje idealna współliniowość i któraś ze zmiennych $X_1, \dots, X_p$ należy pominąć.

Obszar ufności [2].

Niech $I(X_1, \dots, X_n)$ będzie dwuwymiarowym zbiorem zależnym od próby (tzn. obszar I jest zmienną losową i zmienia się od próbki do próbki).
Obszar I jest tak dobrany, aby prawdopodobieństwo, że pokryje on parę parametrów (θ_1, θ_2) (tzn. punkt o współrzędnych θ_1, θ_2), było równe $(1-\alpha)$, tzn:

$$\boxed{P\left((\theta_1, \theta_2) \in I\right) = 1-\alpha}$$

Każdy zbiór I spełniający powyższy warunek nazywamy $(1-\alpha)\cdot 100\%$ -owym obszarem ufności dla pary parametrów (θ_1, θ_2).

Przykład. Obszar ufności dla pary (μ, σ^2).
Rozpatrzmy przypadek, gdy wyznaczamy łączny dwuwymiarowy zbiór ufności dla wartości oczekiwanej $\mu \equiv E(X)$ i wariancji $\sigma^2 \equiv \sigma^2(X)$, dla przypadku gdy badana cecha X ma rozkład $N(\mu, \sigma^2)$ o nieznanych μ i σ.

Można pokazać, że statystyki [2]:

$$G_1 = \frac{(\bar{X}-\mu)^2}{\frac{\sigma^2}{n}} \quad \text{oraz} \quad G_2 = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

są niezależne i mają rozkład χ^2 z liczbą stopni swobody równą 1 dla G_1 oraz $(n-1)$ dla G_2 (pokazać).

Statystyki G_1 i G_2 wykorzystujemy do budowy obszaru ufności, tzn. wyznaczamy takie wartości a, b, c, aby [2]:

$$P(0 \leq G_1 < a,\ b < G_2 < c) = P(0 \leq G_1 < a)\cdot P(b < G_2 < c) = 1-\alpha$$

Wartości a, b, c można wybrać na wiele sposobów. Zazwyczaj wybieramy je tak, aby w zgodzie z powyższą zależnością, zachodziły związki:

$$P(0 \leq G_1 < a) = \sqrt{1-\alpha} \underset{\alpha\text{-małe}}{\approx} 1 - \frac{1}{2}\alpha - \frac{1}{8}\alpha^2 \qquad (P_{G_1})$$

oraz

$$P(b < G_2 < c) = \sqrt{1-\alpha} \approx 1 - \frac{1}{2}\alpha - \frac{1}{8}\alpha^2 \qquad (P^a_{G_2})$$

i dodatkowo:

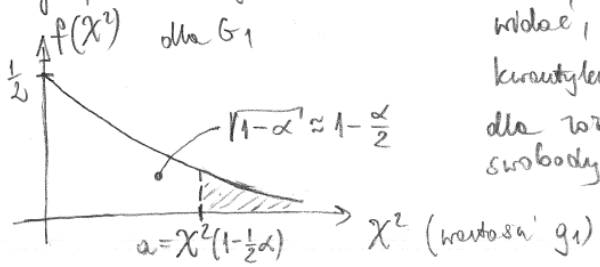
$$P(G_2 < b) = P(G_2 > c) \qquad (P^b_{G_2})$$

Uwaga. Z $(P^a_{G_2})$ wynika, że prawdziwy szczegółowy poziom istotności dla testu dotyczącego G^2 wynosi nie $\frac{\alpha}{2}$ ale około $\frac{\alpha}{2} + \frac{\alpha^2}{8}$ (Porównaj Rozdział 9, Część I, (9.7)).

(Poniżej pomijamy człon kwadratowy, $\alpha^2/8$).

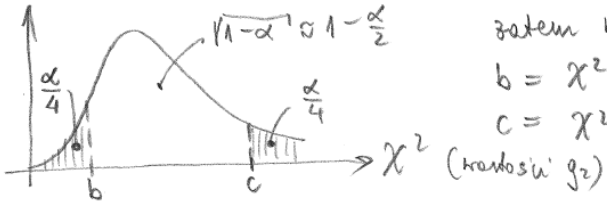
Niech statystyka G_1 przyjmuje wartości g_1, a statystyka G_2 wartości g_2.

Korzystając z wykresu funkcji rozkładu dla G_1:



widać, że $a = \chi^2(1-\frac{1}{2}\alpha)$, jest kwantylem rzędu $(1-\frac{1}{2}\alpha)$ wyznaczonym dla rozkładu χ^2 z jednym stopniem swobody, tzn. $a = \chi^2(1-\frac{\alpha}{2}, 1)$.

i podobnie dla G_2:



zatem widać, że
$b = \chi^2(\frac{\alpha}{4}, n-1)$ oraz
$c = \chi^2(1-\frac{\alpha}{4}, n-1)$

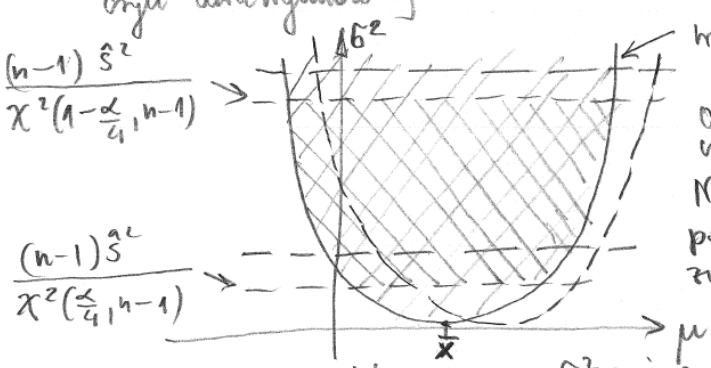
Rozwiązując nierówności pod oznaczeniem prawdopodobieństw w (P_{G_1}) i $(P_{G_2}^{a,b})$ (str 290), tzn.:

$$\begin{cases} 0 \leq g_1 = \dfrac{(\bar{x}-\mu)^2}{\frac{\sigma^2}{n}} < \chi^2(1-\frac{\alpha}{2}, 1) \\ \chi^2(\frac{\alpha}{4}, n-1) < \dfrac{(n-1)\hat{s}^2}{\sigma^2} = g_2 < \chi^2(1-\frac{\alpha}{4}, n-1) \end{cases}$$

względem parametrów μ i σ^2, otrzymujemy [2]:

$$\begin{cases} \sigma^2 > \dfrac{n(\bar{x}-\mu)^2}{\chi^2(1-\frac{\alpha}{2}, 1)} \\ \dfrac{(n-1)\hat{s}^2}{\chi^2(1-\frac{\alpha}{4}, n-1)} < \sigma^2 < \dfrac{(n-1)\hat{s}^2}{\chi^2(\frac{\alpha}{4}, n-1)} \end{cases}$$

czyli dwuwymiarowy obszar ufności dla parametrów μ i σ^2.



wykres krzywej granicznej $\sigma^2 = \dfrac{n(\bar{x}-\mu)^2}{\chi^2(1-\frac{\alpha}{2}, 1)}$

obszar zacieniowany ▨ jest wyznaczonym dla próbki obszarem ufności. Na poziomie ufności $(1-\alpha)$ pokrywa on parę parametrów (μ, σ^2) dla zmiennej losowej $X: N(\mu, \sigma^2)$ [2].

Uwaga: Wartości parametrów μ oraz σ^2 <u>nie są znane</u>. Na rysunku [2] zaznaczono również linie graniczne obszaru otrzymanego dla innej próbki.