

# Część IV

## Regresja



# Rozdział 9

## Modele regresji

### 9.1 Wstęp

Modele regresji zajmują szczególne miejsce w statystyce. Mają niebywałą ilość różnorodnych zastosowań. Używa się ich powszechnie w chemii, biologii, ekonomii, doświadczałnictwie rolniczym i właściwie w każdej z nauk empirycznych. Z konieczności ograniczymy się do paru najprostszych modeli i nasza dyskusja będzie bardzo pobieżna. Regresja opisuje, mówiąc najogólniej, statystyczną zależność tak zwanej zmiennej „objaśnianej” od zmiennych „objaśniających”.

Przypuśćmy, że interesuje nas związek pomiędzy dwiema zmiennymi, które oznaczymy przez  $x$  i  $Y$ . Mierzmy lub obserwujemy wielokrotnie wartości tych zmiennych. Dane mają postać par  $(x_i, Y_i)$  i możemy je zapisać w takiej tabelce:

przypadki \ zmienne	niezależna (objaśniająca) $x$	zależna (objaśniana) $Y$
1	$x_1$	$Y_1$
2	$x_2$	$Y_2$
$\vdots$	$\vdots$	$\vdots$
$n$	$x_n$	$Y_n$

Na przykład, możemy badać zależność pomiędzy parami zmiennych  $(x, Y)$  takiego typu:

$x$	$Y$
wielkość produkcji	– zużycie energii
wiek dziecka	– wzrost
stężenie katalizatora	– wydajność procesu
dawka nawozu	– plony
...	– ...

„Przypadki” odpowiadają pomiarom lub obserwacjom zmiennej  $Y$  dla różnych wartości zmiennej  $x$ . Poszczególne pomiary mogą dotyczyć różnych obiektów lub tego samego, ewoluującego procesu.

Przypuszczamy, że zmienna  $Y$  jest „w zasadzie” funkcją  $x$ , ale „zaburzoną losowymi błędami”. Nasz model zależności będzie taki:

$$Y = \phi(x) + \varepsilon,$$

gdzie  $\varepsilon$  jest *błędem losowym*. Funkcję  $y = \phi(x)$  nazywamy *funkcją regresji*. Dla poszczególnych „przypadków”, czyli uzyskanych doświadczalnie punktów  $(x_i, Y_i)$  mamy

$$Y_i = \phi(x_i) + \varepsilon_i, \quad (i = 1, \dots, n).$$

Punkty doświadczalne  $(x_i, Y_i)$  nie leżą dokładnie na krzywej regresji, ale znajdują się „w pobliżu” wykresu funkcji  $y = \phi(x)$ . Zakładamy, że wielkości  $x_i$  są znane i nielosowe. Odpowiada to sytuacji, gdy zmienna  $x$  jest „pod kontrolą eksperymentatora” i jest mierzona bezbłędnie. Wartości zmiennej  $Y$  są losowymi obserwacjami (ze względu na wpływ losowego składnika  $\varepsilon$ ). Funkcja regresji  $\phi$  jest nieznaną i będziemy ją estymować na podstawie danych.

Oznaczenie zmiennej niezależnej *małą* literą  $x$ , a zmiennej zależnej *dużą* literą  $Y$  ma nam stale przypominać, gdzie „tkwi losowość”. Czytelnik powinien wiedzieć, że w literaturze ta konwencja nie jest powszechnie przyjęta. Istnieją również inne modele regresji, w których zmienna objaśniająca też jest losowa, ale nie będziemy ich rozważać.

## Metoda najmniejszych kwadratów

Sprecyzowanie modelu regresji wymaga przyjęcia konkretnych założeń o funkcji  $\phi$  oraz o błędach losowych  $\varepsilon_i$ . Założymy, że funkcja regresji ma znaną postać, natomiast zależy od nieznanego parametru  $\beta$ . Napiszemy zatem  $\phi(x) = \phi(\beta, x)$ . Zwróćmy uwagę, że wartość  $\beta$  dla poszczególnych „przypadków”  $i = 1, \dots, n$  jest taka sama (zależność opisuje jedna funkcja, tylko błędy losowe są różne). W ten sposób powstają *parametryczne* modele regresji.

Przyjmujemy klasyczne założenie, że błędy są niezależne i mają jednakowy rozkład normalny. Podsumujmy i uzupełnijmy opis modelu:

$$(9.1.1) \quad Y_i = \phi(\beta, x_i) + \varepsilon_i, \quad (i = 1, \dots, n)$$

gdzie

$i$  - numer „przypadku”,

$x_i$  - wartość zmiennej „objaśniającej” (znana i nielosowa),

$\varepsilon_i$  - błąd losowy (nieobserwowana zmienna losowa),

$Y_i$  - obserwowana zmienna losowa „objaśniana”,

$\beta$  - nieznan parametr (nielosowy).

**9.1.2 Założenie.** *Spełniona jest zależność (9.1.1). Błędy  $\varepsilon_1, \dots, \varepsilon_n$  są niezależnymi zmiennymi losowymi o jednakowym rozkładzie normalnym  $N(0, \sigma^2)$ .*

Schemat opisany powyżej można łatwo uogólnić uwzględniając wpływ *wielu* zmiennych objaśniających na zmienną objaśnianą. Na przykład, wydajność procesu chemicznego może zależeć od stężenia katalizatora i od ciśnienia. Na wysokość plonów może mieć wpływ intensywność nawożenia, poziom opadów i jeszcze inne czynniki (zmienne). Nie musimy zakładać, że  $x_i$  są skalarami; mogą to być wektory. Również parametr  $\beta$  może być wektorem. Pozostaniemy natomiast przy założeniu, że wartości zmiennej objaśnianej  $Y_i$  są skalarne.

Łączna gęstość prawdopodobieństwa obserwacji  $Y_1, \dots, Y_n$  jest następująca:

$$f_{\beta, \sigma}(y_1, \dots, y_n) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \phi(\beta, x_i))^2 \right].$$

W ten sposób określona jest rodzina rozkładów prawdopodobieństwa na przestrzeni próbkowej  $\Omega = \mathbb{R}^n$ ; przestrzenią parametrów jest  $\Theta = \mathbb{R}^p \times ]0, \infty[$ , gdzie  $p$  jest wymiarem parametru  $\beta$ . Ten opis modelu mieści się w ogólnym schemacie wprowadzonym w Rozdziale 2.

Ze wzoru na postać gęstości natychmiast wynika prosty wniosek.

**9.1.3 Stwierdzenie.** *Jeśli spełnione jest Założenie 9.1.2, to estymator największej wiarygodności parametru  $\beta$  jest rozwiązaniem zadania minimalizacji*

$$\text{RSS}(\beta) = \sum_{i=1}^n (Y_i - \phi(\beta, x_i))^2 \rightarrow \min_{\beta}.$$

Skrót „RSS” pochodzi od angielskiego zwrotu „*Residual Sum of Squares*”. Będziemy nazywać  $\text{RSS} = \min_{\beta} \text{RSS}(\beta)$  *resztową sumą kwadratów*. Estymator wprowadzony w Stwierdzeniu 9.1.3 nazywamy *estymatorem najmniejszych kwadratów* i w skrócie napiszemy  $\hat{\beta} = \text{LSE}(\beta)$ . Niezależnie od Założenia 9.1.2, LSE ma bardzo przekonującą interpretację. Dopasowujemy krzywą do punktów doświadczalnych w ten sposób, żeby suma kwadratów „odchyłek” punktów od krzywej była minimalna. Przy tym „odchyłki” mierzymy *wzdłuż osi Y*. Metoda najmniejszych kwadratów sprowadza się do metody największej wiarygodności przy założeniu o normalnym rozkładzie błędów, ale ma samodzielny sens i może być stosowana bez tego założenia.

## 9.2 Model liniowy

Ograniczymy się do najprostszej, liniowej postaci funkcji regresji. Mimo, że założenie o liniowości wydaje się bardzo ograniczające, różnorodność i zakres zastosowań modeli liniowych są zaskakująco duże.

### Prosta regresja liniowa

Rozpatrzmy na początek model z jedną (skalarną) zmienną objaśniającą. Model liniowy z wyrazem wolnym ma postać

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (i = 1, \dots, n).$$

Wykresem funkcji regresji jest linia prosta  $y = \beta_0 + \beta_1 x$ . Wzory przybierają prostą i przejrzystą formę. Estymatory najmniejszych kwadratów parametrów  $\beta_0$  i  $\beta_1$  są następujące:

$$(9.2.1) \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

gdzie

$$\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{Y} = \frac{1}{n} \sum Y_i.$$

Istotnie,

$$\text{RSS}(\beta) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Rozwiązujemy układ równań:

$$\frac{1}{2} \frac{\partial \text{RSS}(\beta)}{\partial \beta_0} = \sum_{i=1}^n (\beta_0 + \beta_1 x_i - Y_i) = 0,$$

$$\frac{1}{2} \frac{\partial \text{RSS}(\beta)}{\partial \beta_1} = \sum_{i=1}^n (\beta_0 + \beta_1 x_i - Y_i) x_i = 0.$$

Rachunki są elementarne i łatwe (Zadanie 9.2).

Niech  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ , gdzie  $\hat{\beta}_0$  i  $\hat{\beta}_1$  są LSE danymi wzorem (9.2.1). Punkty  $(x_i, \hat{Y}_i)$  leżą na *dopasowanej* (wyestymowanej) prostej regresji. Mówimy, że  $\hat{Y}_i$  są *przewidywanymi* wartościami zmiennej objaśnianej. Różnice  $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$  pomiędzy wartościami obserwowanymi i przewidywanymi nazywamy *resztami* albo *residuami*.

**9.2.2 Przykład** (Ilość produktu i stężenie katalizatora). Badamy zależność ilości produktu w pewnej reakcji chemicznej (zmienna  $Y$ ) od stężenia katalizatora (zmienna  $x$ ). Przeprowadzono doświadczenie 15 razy, wybierając różne stężenia katalizatora i otrzymano takie wyniki:

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$x_i$	3	5	7	8	10	11	12	12	13	15	15	16	18	19	20
$Y_i$	14	40	40	30	50	34	40	70	52	50	70	64	88	72	90

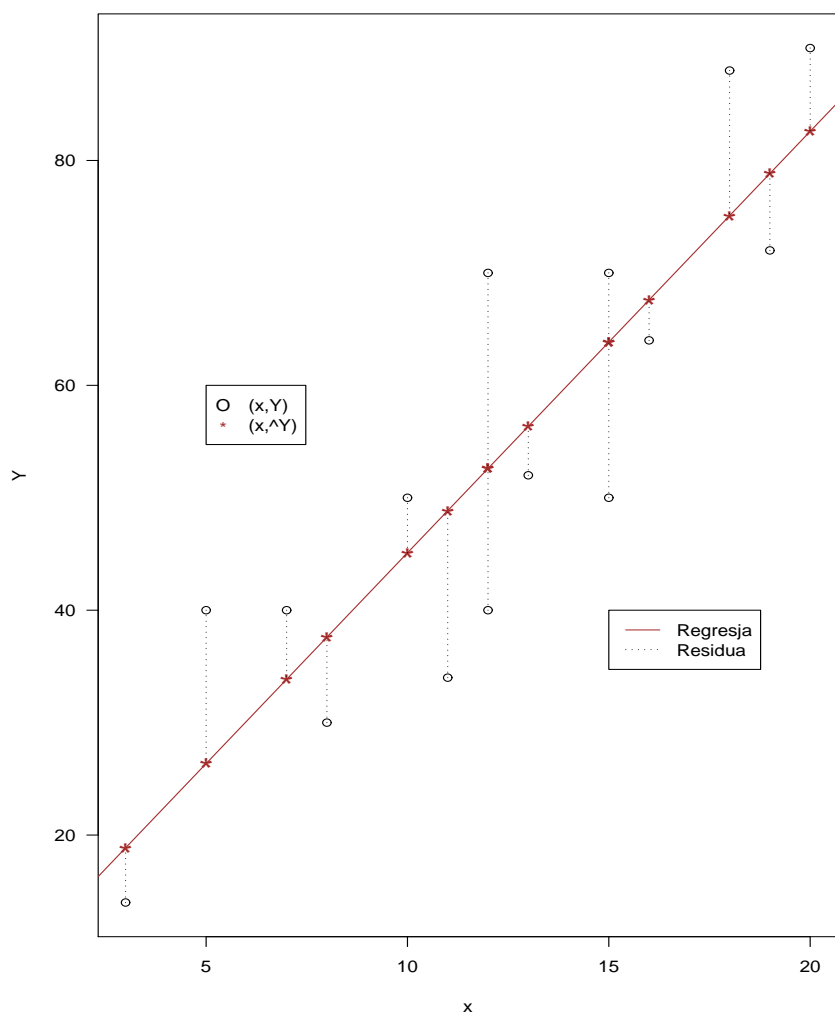
Zakładamy, że ilość produktu zależy w sposób liniowy od stężenia katalizatora (w interesującym nas zakresie wartości obu zmiennych). Odchylenia od dokładnej zależności liniowej traktujemy jako „błędy losowe”. Mówiąc porządniej, decydujemy się na opis zależności  $Y$  od  $x$  przy pomocy modelu prostej regresji liniowej.

Estymowane wartości współczynników są, dla naszych danych, równe  $\hat{\beta}_0 = 7.61$  i  $\hat{\beta}_1 = 3.75$ . Przyjmujemy więc, że funkcja

$$\hat{Y} = 7.61 + 3.75x$$

opisuje w przybliżeniu interesującą nas zależność. Obliczyliśmy to przy pomocy programiku napisanego w języku R, który wygląda tak:

Punkty doświadczalne wraz z dopasowaną prostą regresji pokazuje następujący rysunek.



Rysunek 9.1: Dane i regresja liniowa w Przykładzie 9.2.2.



## Regresja liniowa wieloraka

Rozpatrzmy teraz model z *wieloma* zmiennymi objaśniającymi. Ich liczbę oznaczmy przez  $r$ . Zmienna objasniata jest jedna, skalarna, tak jak poprzednio. Wskaźnik  $i = 1, \dots, n$  będzie, tak jak dotąd, numerował kolejne „przypadki” lub „obiekty”. Zmienne opisujące  $i$ -ty obiekt oznaczmy przez  $x_{i1}, \dots, x_{ir}$  i  $Y_i$ . Model regresji liniowej z wyrazem wolnym przybiera postać

$$Y_i = \beta_0 + \sum_{j=1}^r \beta_j x_{ij} + \varepsilon_i, \quad (i = 1, \dots, n).$$

Prosty chwyt pozwala włączyć wyraz wolny do funkcji liniowej. Przyjmijmy umownie, że  $x_{i0} = 1$ . Zmienne objaśniające dla  $i$ -tego obiektu ustawimy w wektor wierszowy, dołączając jedynekę:  $x_i^\top = (1, x_{i1}, \dots, x_{ir})$ . Można teraz zapisać bardziej zwięźle model w postaci wektorowej:

$$Y_i = \sum_{j=0}^r \beta_j x_{ij} + \varepsilon_i = x_i^\top \beta, \quad (i = 1, \dots, n),$$

gdzie  $\beta = (\beta_0, \beta_1, \dots, \beta_r)^\top$ . W postaci macierzowej to można przepisać tak:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nr} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Będziemy konsekwentnie stosowali notację wektorowo-macierzową. Wektory i macierze w powyższym wzorze oznaczmy pojedynczymi literami  $Y$ ,  $\mathbf{X}$ ,  $\beta$  i  $\varepsilon$ . Przyjmijmy, dla jednoznaczności oznaczeń, że symbol  $p$  oznaczać będzie wymiar wektora  $\beta$ . Dla regresji liniowej z  $r$  zmiennymi objaśniającymi i wyrazem wolnym mamy zatem

$$p = r + 1.$$

Model liniowy przybiera zwięźłą postać:

$$\begin{array}{ccccccc} Y & = & \mathbf{X} & \cdot & \beta & + & \varepsilon. \\ n \times 1 & & n \times p & & p \times 1 & & n \times 1 \end{array}$$

Pod spodem napisaliśmy *wymiary* poszczególnych obiektów. Znana i nielosowa macierz  $\mathbf{X}$  jest zwana *macierzą planu*,  $\beta$  jest wektorem nieznanych parametrów,  $Y$  jest wektorem obserwacji,  $\varepsilon$  jest losowym wektorem „błędów”.

*Uwaga.* Zauważmy, że do macierzy  $\mathbf{X}$  dołączyliśmy „zerową” kolumnę złożoną z samych jedynek. W większości zastosowań jest to naturalne (ta operacja jest wykonywana w R „domyślnie”). Czasami trzeba rozważyć model regresji bez wyrazu wolnego. Należy wtedy pamiętać, że  $p = r$ , a nie  $p = r + 1$ . Przyjmijmy umowę, że liczba kolumn macierzy  $\mathbf{X}$  i wymiar wektora  $\beta$  będą zawsze równe  $p$ . W ogólnych, teoretycznych rozważaniach, będziemy pisać  $\beta = (\beta_1, \dots, \beta_p)^\top$ , bo wygodniej numerować współrzędne wektora od 1, nie od 0. Wzory dla regresji z wyrazem wolnym wymagają oczywistej modyfikacji.

W dalszym ciągu ograniczymy się do rozważania następującej sytuacji.

**9.2.3 Założenie.** *Mamy  $p < n$  i macierz  $\mathbf{X}$  jest pełnego rzędu, to znaczy  $\text{rz}(\mathbf{X}) = p$ .*

Sens powyższego założenia jest jasny. Wydaje się, że do wyestymowania  $p$  nieznanych parametrów, potrzeba więcej niż  $p$  obserwacji <sup>1</sup>.

Ważna część teorii wymaga wprowadzonego w Założeniu 9.1.2 warunku:  $\varepsilon_1, \dots, \varepsilon_n$  są niezależnymi zmiennymi losowymi o jednakowym rozkładzie  $N(0, \sigma^2)$ . Zreasumujmy nasze rozważania w następującej postaci.

**9.2.4 Założenie.** *Model jest opisany równaniem  $Y = \mathbf{X}\beta + \varepsilon$ , gdzie  $\varepsilon \sim N(0, \sigma^2\mathbf{I})$ .*

Część teorii nie wymaga założenia o normalności. Wystarczy, że zmienne losowe  $\varepsilon_1, \dots, \varepsilon_n$  spełniają warunki  $\mathbb{E}\varepsilon_i = 0$  i  $\text{Var}\varepsilon_i = \sigma^2$  dla  $i = 1, \dots, n$  oraz  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  dla  $i \neq j$ . Sformułujmy to w postaci następującego, słabszego założenia (w tym rozdziale będziemy trochę mniej pedantyczni niż poprzednio i odstąpimy od jawnego zaznaczania zależności  $\mathbb{P}$  i  $\mathbb{E}$  od nieznanego rozkładu).

**9.2.5 Założenie.** *Model jest opisany równaniem  $Y = \mathbf{X}\beta + \varepsilon$ , gdzie  $\mathbb{E}\varepsilon = 0$  i  $\text{VAR}\varepsilon \sim \sigma^2\mathbf{I}$ .*

Poniższy przykład pokazuje, że założenie o liniowości funkcji regresji jest mniej ograniczające, niż się wydaje.

**9.2.6 Przykład** (Regresja wielomianowa). Rozpatrzmy model z pojedynczą zmienną objaśniającą, w którym funkcja regresji jest wielomianem  $r$ -tego stopnia:

$$Y_i = \beta_0 + \sum_{j=1}^r \beta_j x_i^j + \varepsilon_i, \quad (i = 1, \dots, n).$$

To jest *model liniowy*, w którym  $i$ -ty wiersz macierzy planu jest równy

$$x_i^\top = (1, x_i, \dots, x_i^j, \dots, x_i^r) \quad (i = 1, \dots, n).$$

<sup>1</sup>W ostatnich latach coraz więcej uwagi poświęca się w statystyce modelom, w których  $p > n$ . Ale to już inna historia, wykraczająca poza zakres naszych rozważań.

## Estymacja w modelu liniowym

Pracujemy w ogólnym modelu liniowym  $Y = \mathbf{X}\beta + \varepsilon$ . Przy Założeniach 9.2.5 i 9.2.3 można napisać jawne, macierzowe wzory na estymator najmniejszych kwadratów, LSE( $\beta$ ). Rozwiązujemy zadanie minimalizacji

$$\text{RSS}(\beta) = \sum_{i=1}^n (Y_i - x_i^\top \beta)^2 = (\mathbf{X}\beta - Y)^\top (\mathbf{X}\beta - Y) = \min_{\beta}.$$

Obliczając gradient lewej strony względem  $\beta$  dostajemy  $\mathbf{X}^\top (\mathbf{X}\beta - Y) = 0$ , czyli

$$\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top Y.$$

Jest to tak zwany układ równań *normalnych* w postaci macierzowej. Założenie 9.2.3 gwarantuje, że macierz  $\mathbf{X}^\top \mathbf{X}$  jest odwracalna i mamy prosty wzór:

$$\text{LSE}(\beta) = \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y.$$

Ponieważ  $\mathbb{E}Y = \mathbf{X}\beta$ , więc  $\mathbb{E}\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}Y = \beta$ . LSE( $\beta$ ) jest estymatorem *nieobciążonym*. Policzymy macierz kowariancji LSE. Mamy

$$\text{VAR}(\hat{\beta}) = (\text{Cov}(\beta_j, \beta_k); j, k = 1, \dots, p) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Istotnie,

$$\begin{aligned} \text{VAR}(\hat{\beta}) &= \mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}(Y - \mathbf{X}\beta)(Y - \mathbf{X}\beta)^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\sigma^2 \mathbf{I}) \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

W Rozdziale 3 wprowadziliśmy pojęcie estymatora nieobciążonego o minimalnej wariancji. Obecnie mamy do czynienia z *wektorowym* parametrem  $\beta = (\beta_1, \dots, \beta_p)$ . Zajmiemy się estymacją *liniowej funkcji* tego parametru, to znaczy wyrażenia postaci

$$g^\top \beta = \sum_{j=1}^p g_j \beta_j,$$

sposób przechodzimy do *jednowymiarowego* zagadnienia estymacji i możemy odwołać się do znanych pojęć. Z definicji,

$$\text{LSE}(g^\top \beta) = \hat{g} = g^\top \hat{\beta}$$

jest *estymatorem najmniejszych kwadratów* funkcji  $g^\top \beta$  (po prostu, wstawiamy LSE( $\beta$ ) w miejsce nieznanego  $\beta$ ). Okazuje się, że LSE mają *najmniejszą wariancję spośród estymatorów liniowych i nieobciążonych*. Mówi się, że  $\hat{\beta}$  jest najlepszym liniowym nieobciążonym estymatorem  $\beta$ , w skrócie BLUE (*Best Linear Unbiased Estimator*). Taka jest treść klasycznego twierdzenia, które teraz sformułujemy dokładniej.

**9.2.7 TWIERDZENIE** (Gausa – Markowa). *Przyjmijmy Założenie 9.2.5. Rozważmy dowolny nieobciążony i liniowy estymator funkcji  $g^\top \beta$ , to znaczy estymator postaci  $\tilde{g} = c^\top Y$  taki, że  $\mathbb{E}\tilde{g} = g^\top \beta$ . Jeżeli  $\hat{g} = \text{LSE}(g^\top \beta)$ , to*

$$\text{Var}\hat{g} \leq \text{Var}\tilde{g}.$$

*Dowód.* Ponieważ  $\mathbb{E}\tilde{g} = c^\top \mathbf{X}\beta = g^\top \beta$  dla każdego  $\beta$ , więc  $c^\top \mathbf{X} = g^\top$ . Oczywiście,  $\text{Var}\tilde{g} = \sigma^2 c^\top c$ . Dowód będzie zakończony gdy pokażemy, że

$$0 \leq c^\top c - g^\top (\mathbf{X}^\top \mathbf{X})^{-1} g.$$

Mozemy tę nierówność przepisać w postaci

$$0 \leq c^\top c - c^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top c = c^\top (\mathbf{I} - \mathbf{H})c,$$

gdzie  $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ . Wystarczy teraz zauważyć, że macierz  $\mathbf{I} - \mathbf{H}$  jest *nieujemnie określona*. Jest tak, bo jest ona *symetryczna i idempotentna*.  $\square$

Jeśli przyjmiemy silniejsze Założenie 9.2.4 zamiast 9.2.5 to można pokazać, że LSE jest nie tylko BLUE (najlepszy wśród liniowych estymatorów nieobciążonych) ale także BUE (najlepszy wśród *wszystkich* estymatorów nieobciążonych). Przyjmiemy ten fakt bez dowodu.

## Geometria ENK

W dalszym ciągu rozważamy ogólny model  $Y = \mathbf{X}\beta + \varepsilon$ . Będziemy w istotny sposób korzystać z Założeń 9.2.4 i 9.2.3. Współrzędne wektorów  $p$ -wymiarowych numerujemy od 1 do  $p$ . Zauważmy, że  $\hat{Y}_i = x_i^\top \hat{\beta}$  jest współrzędną  $Y$ -ową punktu odpowiadającego wektorowi  $x_i$  i leżącemu na wykresie *dopasowanej* (estymowanej metodą najmniejszych Kwadratów) funkcji regresji. Odpowiednią resztą jest  $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ . Wektorowo napiszemy  $\hat{Y} = \mathbf{X}\hat{\beta} = (\hat{Y}_1, \dots, \hat{Y}_n)^\top$  i  $\hat{\varepsilon} = Y - \hat{Y}$ . Mamy

$$\hat{Y} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y = \mathbf{H}Y,$$

gdzie  $\mathbf{H}$  jest macierzą rzutu ortogonalnego (w przestrzeni  $\mathbb{R}^n$ ) na  $p$ -wymiarową podprzestrzeń liniową  $\mathcal{R}(\mathbf{X})$  generowaną przez kolumny macierzy  $\mathbf{X}$  (czyli obraz przekształcenia liniowego o macierzy  $\mathbf{X}$ ). Wystarczy sprawdzić, że  $\mathbf{H}$  jest macierzą *symetryczną* ( $\mathbf{H}^\top = \mathbf{H}$ ) i *idempotentną* ( $\mathbf{H}^2 = \mathbf{H}\mathbf{H} = \mathbf{H}$ ). Rzut na dopełnienie ortogonalne  $\mathcal{R}(\mathbf{X})^\perp$  ma macierz  $\mathbf{I} - \mathbf{H}$ .

Geometryczna interpretacja metody najmniejszych kwadratów staje się przejrzysta, jeśli przejdziemy do takiego ortogonalnego układu współrzędnych, którego pierwsze  $p$  wektorów jest bazą podprzestrzeni  $\mathcal{R}(\mathbf{X})$  a następne  $n - p$  wektorów jest bazą  $\mathcal{R}(\mathbf{X})^\perp$ . Taki układ można napisać w jawnej postaci stosując procedurę ortogonalizacji Hilberta-Schmidta do bazy (nieortogonalnej) przestrzeni  $\mathbb{R}^n$ , złożonej z  $p$  kolumn macierzy  $\mathbf{X}$  oraz  $n - p$  innych wektorów. Pamiętajmy, że macierz  $\mathbf{X}$  o wymiarach  $n \times p$  jest pełnego rzędu  $p$ .

Potrzebny fakt sformułujemy w następującej postaci <sup>2</sup>.

<sup>2</sup>Dla naszych celów istotne będą tylko współrzędne kolumn  $\mathbf{X}$  w nowej bazie.

**9.2.8 Stwierdzenie.** *Istnieje macierz ortogonalna  $\mathbf{Q}$  o wymiarze  $n \times n$  oraz macierz górna trójkątna  $\mathbf{R}$  o wymiarze  $p \times p$  takie, że*

$$\mathbf{X} = \mathbf{Q} \begin{pmatrix} \mathbf{R} \\ \cdots \\ \mathbf{O} \end{pmatrix}.$$

W tym wzorze  $\mathbf{O}$  jest zerową macierzą o wymiarze  $(n - p) \times p$ .

Kolumny macierzy  $\mathbf{Q}$  tworzą ortonormalną bazę  $\mathcal{R}(\mathbf{X})$ . Element  $r_{jk}$  macierzy  $\mathbf{R}$  jest iloczynem skalarnym  $k$ -tej kolumny  $\mathbf{X}$  i  $j$ -tej kolumny  $\mathbf{Q}$ . Tak więc  $\mathbf{R}$  zawiera współrzędne kolumn  $\mathbf{X}$  w nowej bazie. Fakt, że macierz  $\mathbf{R}$  jest trójkątna,  $r_{jk} = 0$  dla  $k < j$  oznacza, że początkowe  $k$  kolumn  $\mathbf{Q}$  jest bazą w przestrzeni rozpiętej przez  $k$  pierwszych kolumn  $\mathbf{X}$ .

Współrzędne wektora  $Y$  w nowej bazie oznaczmy  $\underline{Y} = \mathbf{Q}^\top Y$ . Rzut na przestrzeń  $\mathcal{R}(\mathbf{X})$  ma w nowej bazie macierz współrzędnych  $\underline{\mathbf{H}} = \mathbf{Q}^\top \mathbf{H} \mathbf{Q}$ :

$$\underline{Y} \mapsto Y = \mathbf{Q} \underline{Y} \mapsto \hat{Y} = \mathbf{H} Y = \mathbf{H} \mathbf{Q} \underline{Y} \mapsto \hat{\underline{Y}} = \mathbf{Q}^\top \mathbf{H} \mathbf{Q} \underline{Y}.$$

Zauważmy, że

$$\begin{aligned} \underline{\mathbf{H}} &= \mathbf{Q}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Q} = \begin{pmatrix} \mathbf{R} \\ \cdots \\ \mathbf{O} \end{pmatrix} (\mathbf{R}^\top \mathbf{R})^{-1} \begin{pmatrix} \mathbf{R}^\top : \mathbf{O} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & : & \mathbf{O} \\ \cdots & & \cdots \\ \mathbf{O} & : & \mathbf{O} \end{pmatrix}, \end{aligned}$$

gdzie  $\mathbf{I}$  jest macierzą jednostkową wymiaru  $p \times p$ . To znaczy, że w nowym układzie współrzędnych, rzutowanie wektora  $Y$  na podprzestrzeń  $\mathcal{R}(\mathbf{X})$  polega na zastąpieniu  $n - p$  ostatnich współrzędnych zerami:

$$\text{jeśli } \underline{Y} = (\underline{Y}_1, \dots, \underline{Y}_p, \dots, \underline{Y}_n)^\top \text{ to } \hat{\underline{Y}} = (\underline{Y}_1, \dots, \underline{Y}_p, 0, \dots, 0)^\top.$$

Pamiętajmy przy tym, że  $Y = \mathbf{Q} \underline{Y}$  i  $\hat{Y} = \mathbf{Q} \hat{\underline{Y}}$ . Wzór  $Y = \mathbf{X} \beta + \varepsilon$  w nowym układzie współrzędnych przybiera postać

$$\underline{Y} = \begin{pmatrix} \mathbf{R} \\ \cdots \\ \mathbf{O} \end{pmatrix} \beta + \underline{\varepsilon},$$

gdzie  $\underline{\varepsilon} = \mathbf{Q} \varepsilon$ . Następujące proste spostrzeżenie odgrywa w dalszych rozważaniach zasadniczą rolę. Przy Założeniu 9.2.4, wektor losowy  $\underline{\varepsilon}$  ma łączny rozkład normalny  $N(0, \sigma^2 \mathbf{I})$ .

Geometryczne rozważania prowadzą do bardzo prostego dowodu następującej ogólnej wersji Twierdzenia Fishera.

**9.2.9 TWIERDZENIE** (Fishera). *Jeśli spełnione jest Założenie 9.2.4 i  $\hat{\beta} = \text{LSE}(\beta)$  to  $\hat{\beta}$  jest zmienną losową niezależną od  $Y - \hat{Y}$ . Ponadto mamy  $\|Y - \hat{Y}\|^2 \sim \chi^2(n - p)$  i  $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$ .*

*Dowód.* Ponieważ  $\mathbf{Q}$  jest macierzą ortogonalną, więc jest izometrią, stąd

$$\begin{aligned} \|Y - \hat{Y}\|^2 &= \|\mathbf{Q}^\top Y - \mathbf{Q}^\top \hat{Y}\|^2 = \|\mathbf{Q}^\top (\mathbf{I} - \mathbf{H}) \mathbf{Q} \underline{Y}\|^2 = \|(\mathbf{I} - \mathbf{H}) \underline{Y}\|^2 \\ &= \varepsilon_{p+1}^2 + \dots + \varepsilon_n^2 \sim \chi^2(n - p). \end{aligned}$$

Z kolei  $\hat{Y} = \mathbf{Q} \hat{Y} = \mathbf{Q} \mathbf{H} \underline{Y} = \mathbf{Q}(\varepsilon_1, \dots, \varepsilon_p, 0, \dots, 0)^\top$ . Stąd widać, że  $\hat{Y}$  jest zmienną niezależną od  $Y - \hat{Y}$ . Oczywiście,  $\hat{\beta}$  jest funkcją  $\hat{Y}$ , a więc też jest zmienną niezależną od  $Y - \hat{Y}$ . Wreszcie,  $\hat{\beta}$  jest to liniową funkcją wektora  $Y$ , a więc ma rozkład normalny. Wiemy, że  $\mathbb{E}\hat{\beta} = \beta$  i  $\text{VAR}(\hat{\beta}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ , co kończy dowód.  $\square$

**9.2.10 Wniosek.** *Nieobciążonym estymatorem wariancji błędu,  $\sigma^2$ , jest*

$$S^2 = \frac{\|Y - \hat{Y}\|^2}{n - p} = \frac{\text{RSS}}{n - p}.$$

Estymatory najmniejszych kwadratów  $\hat{\beta}_j$  można „uzupełnić” konstrukcją przedziałów ufności.

**9.2.11 Wniosek.** *Przedział ufności dla parametru  $\beta_j$  jest określony wzorem*

$$\left[ \hat{\beta}_j - \text{Std}_j, \hat{\beta}_j + \text{Std}_j \right],$$

gdzie  $S = \sqrt{S^2}$ ,  $t = t_{1-\alpha/2}(n - p)$  jest kwantylem rozkładu  $t$ -Studenta z  $n - p$  stopniami swobody, zaś  $d_j = \sqrt{(\mathbf{X}^\top \mathbf{X})_{jj}^{-1}}$  (wskaźnik  $jj$  odpowiada  $j$ -temu elementowi na przekątnej macierzy).

Żeby ten wniosek uzasadnić, wystarczy zauważyć, że  $\text{Var}\beta_j = \sigma^2 d_j$ , a zatem

$$\frac{\hat{\beta}_j - \beta_j}{d_j S} \sim t(n - p),$$

na mocy twierdzenia Fishera.

## Predykcja

Po co właściwie dopasowujemy funkcję do punktów doświadczalnych? Rzecz jasna, jest przyjemnie mieć prosty, liniowy model opisujący zależność. Ostatecznym sprawdzianem wartości poznawczej modelu jest możliwość *przewidywania wyników doświadczeń*. W przypadku modelu regresji, chodzi o przewidywanie wartości zmiennej  $Y$  dla *danej wartości  $x$* . Tak jak dotąd, mamy dane punkty  $(x_i, Y_i)$  dla  $i = 1, \dots, n$ . Dla ustalenia uwagi umówmy się, że wracamy do modelu regresji liniowej z wyrazem wolnym i do oznaczenia  $\beta = (\beta_0, \beta_1, \dots, \beta_r)$  na wektor współczynników, gdzie  $p = r + 1$ . Rozważamy „nowy” wektor zmiennych objaśniających, który oznaczmy  $x_*^\top = (1, x_1^*, \dots, x_r^*)$  i uważamy za znany. Jeśli przeprowadzimy *nowe* doświadczenie, to pojawi się odpowiednia wartość  $Y_*$ . Naszym zadaniem jest *predykcja* nieznannej wartości  $Y_*$  *przed* dokonaniem tego dodatkowego doświadczenia. Nasz model przewiduje, że

$$Y_* = x_*^\top \beta + \varepsilon_*,$$

gdzie współczynniki  $\beta$  są te same, co we wzorze  $Y_i = x_i^\top \beta + \varepsilon_i$ , zaś  $\varepsilon_* \sim N(0, \sigma^2)$  jest błędem losowym niezależnym od poprzednich błędów  $\varepsilon_i$ . Musimy zmierzyć się z dwiema trudnościami. Po pierwsze, nie znamy współczynników  $\beta$ . Po drugie, musimy się liczyć z nowym, losowym odchyleniem  $\varepsilon_*$  od prostej regresji. Niemniej, nasuwa się dość oczywiste rozwiązanie. Za *przewidywany* wynik doświadczenia możemy przyjąć

$$\hat{Y}_* = x_*^\top \hat{\beta},$$

gdzie  $\hat{\beta}$  jest estymatorem obliczonym na podstawie *poprzednich* punktów doświadczalnych  $(x_i, Y_i)$ . Spróbujemy teraz oszacować dokładność predykcji. Mamy  $\mathbb{E}\hat{Y}_* = \mathbb{E}Y_* = x_*^\top \beta$  i możemy powiedzieć, że *predyktor*  $\hat{Y}_*$  jest nieobciążony<sup>3</sup>. Obliczmy jego wariancję:

$$\text{Var}\hat{Y}_* = \sigma^2 x_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_*.$$

Łatwo stąd wywnioskować ważny wzór na błąd średniokwadratowy predykcji:

$$\mathbb{E}(\hat{Y}_* - Y_*)^2 = \sigma^2 [1 + x_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_*].$$

„Dodatkowa jedynka” w tym wzorze pochodzi stąd, że musimy uwzględnić wpływ błędu  $\varepsilon_*$ , czyli losowe odchylenie punktu  $Y_*$  od funkcji regresji.

Bardzo podobnie jak we Wniosku 9.2.11 konstruuje się przedziały ufności dla wartości funkcji regresji i predykcji.

**9.2.12 Wniosek.** *Przedział ufności dla wartości funkcji regresji w punkcie  $x_*$ , czyli dla  $\beta^\top x_*$  jest określony wzorem*

$$\left[ x_*^\top \hat{\beta} - \text{Std}_*, x_*^\top \hat{\beta} + \text{Std}_* \right],$$

gdzie  $d_* = \sqrt{x_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_*}$  i  $t = t_{1-\alpha/2}(n-p)$ .

<sup>3</sup>Zauważmy, że przewidywana wielkość  $Y_*$  jest zmienną losową, a zatem nieobciążoność predyktora wymaga osobnej definicji.

Przejdźmy do przedziałów ufności dla predykcji. Ustalamy poziom ufności  $1 - \alpha$  i chcemy skonstruować takie statystyki  $\underline{Y}_*$  i  $\overline{Y}_*$ , żeby, dla dowolnych  $\beta$  i  $\sigma$ ,

$$\mathbb{P}(\underline{Y}_* \leq Y_* \leq \overline{Y}_*) = 1 - \alpha.$$

W powyższym wzorze występuje rozkład prawdopodobieństwa na przestrzeni próbkowej  $\mathbb{R}^{n+1}$ . Jest to łączny rozkład zmiennych losowych  $Y_1, \dots, Y_n$  oraz  $Y_*$ . Statystyki  $\underline{Y}_*$  i  $\overline{Y}_*$  są to funkcje *obserwacji*, czyli zmiennych losowych  $Y_i$  dla  $i = 1, \dots, n$ . Poza tym mogą zależeć od znanych liczb  $x_i$  oraz  $x_*$ , ale nie mogą zależeć od  $Y_*$ . Przedział  $[\underline{Y}_*, \overline{Y}_*]$  będziemy dla uproszczenia nazywać *przedziałem ufności dla predykcji*, ale nie jest to przedział ufności w rozumieniu Definicji 6.0.1. Wartość, którą staramy się przewidzieć,  $Y_*$ , nie jest funkcją nieznanego parametru, tylko *nieobserwowaną zmienną losową*.

**9.2.13 Wniosek.** *Przedział ufności dla predykcji  $Y_*$  jest określony wzorem*

$$\left[ x_*^\top \hat{\beta} - StD_*, x_*^\top \hat{\beta} + StD_* \right],$$

gdzie  $D_* = \sqrt{1 + x_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_*^\top}$  i  $t = t_{1-\alpha/2}(n-p)$ .

Uzasadnienie Wniosków 9.2.12 i 9.2.13 jest analogiczne jak Wniosku 9.2.11. Wystarczy powołać się na twierdzenie Fishera i wykorzystać wzory  $\text{Var} \hat{Y}_* = \sigma^2 d_* = \sigma^2 x_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_*$  i  $\mathbb{E}(\hat{Y}_* - Y_*)^2 = \sigma^2 D_* = \sigma^2 [1 + x_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_*]$ .

W szczególnym przypadku *prostej regresji liniowej z wyrazem wolnym* wzory na przedziały ufności mają wyjątkowo intuicyjną interpretację i warto je przytoczyć. Wprowadźmy wygodne oznaczenie  $SS_x = \sum (x_i - \bar{x})^2$ . Mamy

$$d_* = \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SS_x},$$

$$D_* = 1 + d_* = 1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SS_x}.$$

To można sprawdzić wykorzystując ogólne wzory (trzeba obliczyć macierz odwrotną  $(\mathbf{X}^\top \mathbf{X})^{-1}$  wymiaru  $2 \times 2$ ). Można też obliczyć bezpośrednio  $\text{Var} \hat{Y}_*$  i  $\text{Var}(Y_* - \hat{Y}_*)$  w rozważanym szczególnym przypadku (Zadanie 9.5). Tak czy inaczej, rachunki są łatwe. Zwróćmy uwagę, że liczba stopni swobody resztowej sumy kwadratów RSS jest równa  $n - 2$ , ze względu na obecność wyrazu wolnego. Zatem  $t = t_{1-\alpha/2}(n-2)$ .

Jeśli prawe strony wzorów na przedziały ufności,

$$\beta_0 + \beta_1 x_* = \hat{\beta}_0 + \hat{\beta}_1 x_* \pm t S d_*,$$

$$Y_* = \hat{\beta}_0 + \hat{\beta}_1 x_* \pm t \hat{S} D_*,$$

potraktujemy jako funkcje  $x_*$ , to otrzymamy krzywe wyznaczające „pasy ufności” odpowiednio, dla funkcji regresji i predykcji.



**9.2.14 Przykład** (Produkt i katalizator, kontynuacja). Wróćmy do Przykładu 9.2.14. Przypomnijmy, że na podstawie  $n = 15$  punktów doświadczalnych obliczyliśmy estymatory współczynników równe  $\hat{\beta}_0 = 7.61$  i  $\hat{\beta}_1 = 3.75$ . Dopasowana prosta regresji jest więc taka:

$$\hat{Y} = 7.61 + 3.75x.$$

Przypuśćmy teraz, że chcemy przewidzieć, jaką uzyskamy ilość produktu w nowym doświadczeniu, przy stężeniu katalizatora  $x_* = 10.5$ . Oczywiście, 24.39

$$\hat{Y}_* = 7.61 + 3.75 * 10.5 = 46.98.$$

Szerokość przedziału ufności, równa 24.39, została obliczona według wzoru  $t_{0.975}(13) * \sqrt{\text{RSS}/13 * \sqrt{1 + 1/15 + (10.5 - \bar{x})^2/\text{SS}_x}} = 24.39$ , gdzie  $\bar{x} = 12.27$ ,  $\text{SS}_x = 358.93$  i  $\text{RSS} = 1541.1$ . Na poziomie ufności 0.95 możemy twierdzić, że doświadczenie da wynik  $46.98 \pm 24.39$ , czyli  $Y_*$  zmieści się w przedziale

$$[22.58, 71.37]$$

Zatrzymajmy się jeszcze nad interpretacją przedziału ufności dla funkcji regresji,  $\beta_0 + \beta_1 * 10.5$ . Ten przedział w naszym przykładzie przybiera postać  $46.98 \pm 6.46$ , czyli

$$[40.52, 53.43].$$

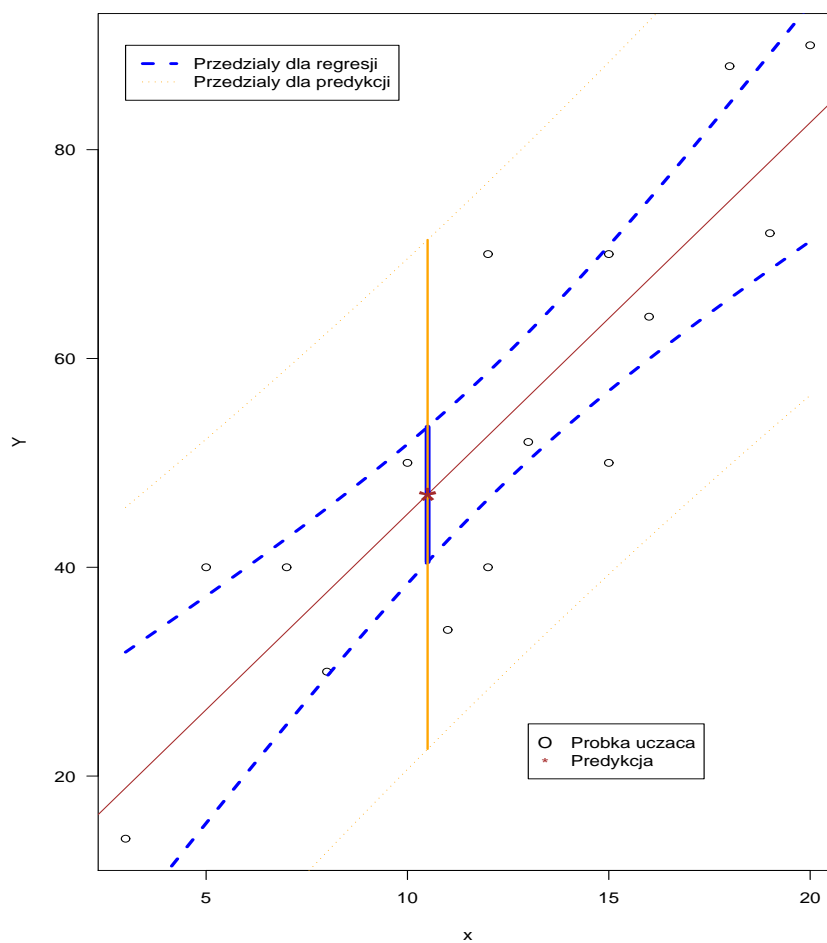
Powiedzmy, że zdecydujemy się uruchomić produkcję na większą skalę i powtarzać wielokrotnie reakcję przy tym samym „roboczym” stężeniu  $x_* = 10.5$ . Wtedy *średnia* ilość otrzymanego produktu będzie równa  $\beta_0 + 10.5\beta_1$ . Ponieważ parametry zależności  $\beta_0$  i  $\beta_1$  są nieznanne (wartości  $\hat{\beta}_0 = 7.61$  i  $\hat{\beta}_1 = 3.75$  są tylko *estymatorami* !) to średnią ilość produktu możemy oszacować tylko „z dokładnością  $\pm 6.46$ ”.

Przedziały ufności dla  $Y_*$ , dla wartości funkcji regresji w punkcie  $x_* = 10.5$  oraz pasy ufności widać na następującym rysunku:

## 9.3 Testowanie hipotez

Najprostsze i najważniejsze zagadnienie testowania hipotez w modelu liniowym zmierza do odpowiedzi na pytanie: „czy wszystkie zmienne objaśniające mają istotny wpływ na zmienną objaśnianą?”. Czy może pewien podzbiór zmiennych  $x$ , można pominąć? Formalnie, niech  $\beta = (\beta_1, \dots, \beta_q, \beta_{q+1}, \dots, \beta_p)^\top$  dla  $q < p$ . Weryfikujemy hipotezę zerową

$$H_0 : (\beta_{q+1}, \dots, \beta_p)^\top = (0, \dots, 0)^\top.$$



Rysunek 9.2: Przedziały ufności i „pasy ufności”.

przeciwko alternatywie

$$H_1 : (\beta_{q+1}, \dots, \beta_p)^\top \neq (0, \dots, 0)^\top.$$

Niech  $\mathbf{X}_0$  oznacza macierz planu  $\mathbf{X}$  z pominiętymi kolumnami  $q+1, \dots, p$ . Jest to więc macierz  $n \times q$ , która odpowiada modelowi regresji zbudowanemu przy założeniu prawdziwości  $H_0$ . Zauważmy, że geometryczne rozważania poprzedniego punktu przenoszą się bez zmian, jeśli zastąpimy  $\mathbf{X}$  przez  $\mathbf{X}_0$ . Co więcej, jeśli rozpatrzmy dekompozycję macierzy  $\mathbf{X}$  podaną w Stwierdzeniu 9.2.8 to automatycznie otrzymujemy dekompozycję  $\mathbf{X}_0$ . Wystarczy wybrać za  $\mathbf{R}_0$  podmacierz trójkątną o wymiarach  $q \times q$  stojącą „w lewym górnym rogu  $\mathbf{R}$ ”. Macierz  $\mathbf{Q}$  pozostaje ta sama, czyli możemy pracować w tym samym wygodnym ortogonalnym układzie współrzędnych. Niech  $\mathbf{H}_0$  oznacza rzut na  $\mathcal{R}(\mathbf{X}_0) \subset \mathcal{R}(\mathbf{X})$ . Mamy  $\mathbf{H}_0 = \mathbf{H}_0 \mathbf{H}$ : rzut rzutu jest rzutem. Najlepiej to widać w nowym układzie współrzędnych:

$$\underline{\mathbf{H}}_0 = \begin{pmatrix} \mathbf{I}_q & \vdots & \mathbf{O} & \vdots & \mathbf{O} \\ \dots & & \dots & & \dots \\ \mathbf{O} & \vdots & \mathbf{O}_{p-q} & \vdots & \mathbf{O} \\ \dots & & \dots & & \dots \\ \mathbf{O} & \vdots & \mathbf{O} & \vdots & \mathbf{O}_{n-p} \end{pmatrix}, \quad \underline{\mathbf{H}} = \begin{pmatrix} \mathbf{I}_q & \vdots & \mathbf{O} & \vdots & \mathbf{O} \\ \dots & & \dots & & \dots \\ \mathbf{O} & \vdots & \mathbf{I}_{p-q} & \vdots & \mathbf{O} \\ \dots & & \dots & & \dots \\ \mathbf{O} & \vdots & \mathbf{O} & \vdots & \mathbf{O}_{n-p} \end{pmatrix}.$$

W tym wzorze indeksy oznaczają wymiary kwadratowych bloków. Rzuty na  $\mathcal{R}(\mathbf{X})$  i  $\mathcal{R}(\mathbf{X}_0)$  opisują wzory

$$\begin{aligned} \underline{Y} &= (\underline{Y}_1, \dots, \underline{Y}_q, \underline{Y}_{q+1}, \dots, \underline{Y}_p, \underline{Y}_{p+1}, \dots, \underline{Y}_n)^\top, \\ \underline{\hat{Y}} &= (\underline{Y}_1, \dots, \underline{Y}_q, \underline{Y}_{q+1}, \dots, \underline{Y}_p, 0, \dots, 0)^\top, \\ \underline{\hat{Y}}_0 &= (\underline{Y}_1, \dots, \underline{Y}_q, 0, \dots, 0, 0, \dots, 0)^\top. \end{aligned}$$

Wektory

$$\hat{Y}_0, \quad \hat{Y} - \hat{Y}_0, \quad Y - \hat{Y}$$

są wzajemnie prostopadłe. Z twierdzenia Pitagorasa wynikają tożsamości

$$\|\hat{Y}\|^2 + \|Y - \hat{Y}\|^2 = \|Y\|^2.$$

oraz

$$\|\hat{Y} - \hat{Y}_0\|^2 + \|Y - \hat{Y}\|^2 = \|Y - \hat{Y}_0\|^2.$$

Wprowadźmy oznaczenia

$$\text{RSS} = \|Y - \hat{Y}\|^2, \quad \text{RSS}_0 = \|Y - \hat{Y}_0\|^2.$$

Wiemy, że  $\text{RSS} = \underline{\varepsilon}_{p+1}^2 + \dots + \underline{\varepsilon}_n^2$  (przy założeniu, że model jest poprawny). Jeśli że  $H_0$  jest prawdziwa (czyli model w istocie zawiera tylko  $q$  zmiennych  $x$ ) to mamy analogicznie  $\text{RSS}_0 = \underline{\varepsilon}_{q+1}^2 + \dots + \underline{\varepsilon}_p^2 + \dots + \underline{\varepsilon}_n^2$ . Stąd wyciągamy wniosek, pozwalający na skonstruowanie testu  $H_0$ :

### 9.3.1 Wniosek. Przy prawdziwości $H_0$ , statystyka

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(p - q)}{\text{RSS}/(n - p)}$$

ma rozkład  $F(p - q, n - p)$  (rozkład Fishera-Snedecora z  $p - q$  stopniami swobody w liczniku i  $p$  stopniami swobody w mianowniku).

Warto zauważyć, że ten test jest niczym innym jak testem ilorazu wiarygodności dla hipotez złożonych. W istocie, wiarygodność jest dana wzorem

$$\begin{aligned} \mathcal{L}(\beta, \sigma) &= \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum (Y_i - x_i^\top \beta)^2 \right) \\ &\propto \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \|Y - \mathbf{X}\beta\|^2 \right), \end{aligned}$$

więc

$$\ell(\beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \|Y - \mathbf{X}\beta\|^2 + \text{const.}$$

Ponieważ  $\text{MLE}(\sigma) = \hat{\sigma} = \text{RSS}/n$ , więc

$$\ell(\hat{\beta}, \hat{\sigma}) = -n \log \text{RSS} + \text{const.}$$

Analogicznie, dla „mniejszego modelu” otrzymujemy estymator z ograniczeniami  $\text{MLE}_0(\sigma) = \hat{\sigma}_0 = \text{RSS}_0/n$ . W rezultacie,

$$\ell(\hat{\beta}, \hat{\sigma}) - \ell(\hat{\beta}_0, \hat{\sigma}_0) = n \log \frac{\text{RSS}_0}{\text{RSS}}.$$

Statystyka  $F$  jest rosnącą funkcją obliczonej powyżej statystyki ilorazu wiarygodności.

*Uwaga* (Ogólne hipotezy liniowe). Skoncentrowaliśmy się na zagadnieniu testowania hipotezy  $H_0 : (\beta_{q+1}, \dots, \beta_p)^\top = (0, \dots, 0)^\top$  po pierwsze dlatego, że to jest ważne w zastosowaniach: chcemy wyeliminować „niepotrzebne zmienne” i uprościć model. Po drugie, macierz  $\mathbf{Q}$  w Stwierdzeniu 9.2.8 daje ortogonalny układ współrzędnych idealnie pasujący do tej postaci hipotez. Oczywiście, należy wpięrow ustawić zmienne objaśniające w odpowiedniej kolejności, zaczynając numerację od tych, które wydają się ważniejsze a kończąc na tych, które podejrzewamy o „bycie zbędnymi”. Co więcej, cała teoria bez zmian stosuje się do ogólnych hipotez liniowych postaci

$$H_0 : \mathbf{C}\beta = 0,$$

gdzie  $\mathbf{C}$  jest macierzą  $(p - q) \times p$  pełnego rzędu. Taka hipoteza stwierdza, że  $\beta$  należy do podprzestrzeni liniowej wymiaru  $q$ . Wektor  $\hat{Y}_0$  jest rzutem ortogonalnym na  $\{y : y = \mathbf{X}\beta, \mathbf{C}\beta = 0\}$  i definiujemy, tak jak poprzednio,  $\text{RSS}_0 = \|Y - \hat{Y}_0\|^2$ . Wniosek 9.3.1 pozostaje prawdziwy.

*Uwaga* (Test t i test F). Rozumowanie uzasadniające Wniosek 9.2.11 można wykorzystać do konstrukcji testu hipotezy  $H_0 : \beta_j = 0$ . Używamy statystyki testowej Studenta,

$$T = \frac{\hat{\beta}_j}{d_j S}, \quad d_j = \sqrt{(\mathbf{X}^\top \mathbf{X})_{jj}^{-1}}$$

i odrzucamy  $H_0$  jeśli  $|T| > t$ . Jeśli  $H_0$  jest prawdziwa, to  $T \sim t(n-p)$ , więc próg odrzuceń jest odpowiednim kwantylem rozkładu t-Studenta,  $t = t_{1-\alpha/2}(n-p)$  (zamiast ustalać próg, możemy obliczać  $P$ -wartości testu).

Z drugiej strony,  $H_0 : \beta_p = 0$  jest szczególnym przypadkiem hipotezy rozpatrywanej we Wniosku 9.3.1 i może być użyty test F. Jeśli  $H_0$  jest prawdziwa, to  $F \sim F(1, n-p)$ . Zamiast  $p$  możemy wziąć dowolne  $j$ , zmieniając porządek współczynników. Oba testy, t i F, są *równoważne*, bo  $T^2 = F$  (Zadanie 9.8).

Test t ma tę przewagę, że nadaje się do testowania  $H_0$  przeciw alternatywie jednostronnej, powiedzmy  $H_1 : \beta_j > 0$ , podczas gdy F jest dostosowany do alternatywy dwustronnej  $H_1 : \beta_j \neq 0$ .

## Analiza wariancji

Rozważmy zagadnienie porównywania kilku próbek. Chodzi o sprawdzenie, czy wszystkie pochodzą z tej samej populacji, czy też z populacji o różnych średnich. Najprostszy model zakłada, że mamy  $p$  niezależnych próbek z rozkładów normalnych:

$$\begin{aligned} \text{próbka 1:} & \quad Y_{11}, \dots, Y_{1n_1} \sim N(\mu_1, \sigma^2); \\ \dots & \quad \dots \quad \dots \\ \text{próbka } j: & \quad Y_{j1}, \dots, Y_{jn_j} \sim N(\mu_j, \sigma^2); \\ \dots & \quad \dots \quad \dots \\ \text{próbka } p: & \quad Y_{p1}, \dots, Y_{pn_p} \sim N(\mu_p, \sigma^2). \end{aligned}$$

Zakłada się przy tym, że w poszczególnych próbkach wariancja  $\sigma^2$  jest jednakowa, natomiast wartości średnie  $\mu_j$  mogą być różne. Jest to szczególny przypadek modelu liniowego. W rzeczy samej, napiszmy

$$Y_{ji} = \mu_1 + \alpha_j + \varepsilon_{ji},$$

gdzie  $\alpha_j = \mu_j - \mu_1$  dla  $j = 1, \dots, p$  oraz  $\varepsilon_{ji} = Y_{ji} - \mu_j$ . Oczywiście,  $\varepsilon_{ji} \sim N(0, \sigma^2)$  są niezależnymi zmiennymi losowymi. Wprowadźmy sztuczne, „nieme” zmienne objaśniające  $x_1, \dots, x_p$ . Przyjmiemy *umownie*, że dla obserwacji z  $j$ -tej próbki mamy  $x_1 = 1, x_j = 1$ , zaś wszystkie inne zmienne  $x$ -owe są zerami. Obrazuje to taka tabelka:

próbka \ zmienne	$x_1$	$x_2$	...	$x_p$
1	1	0	...	0
2	1	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p$	1	0	...	1

Niech  $Y$  oznacza „długi” wektor o  $n = \sum_{j=1}^p n_j$  współrzędnych, powstały przez ustawienie kolejnych próbek „jedna nad drugą”. Podobnie określamy wektor błędów  $\varepsilon$ . „Nieme zmienne” umieścimy w macierzy  $\mathbf{X}$ . Model możemy napisać w postaci macierzowej tak:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix},$$

a w skrócie

$$\begin{matrix} Y & = & \mathbf{X} & \cdot & \beta & + & \varepsilon \\ n \times 1 & & n \times p & & p \times 1 & & n \times 1 \end{matrix}$$

gdzie  $\beta = (\mu_1, \alpha_2, \dots, \alpha_p)^\top$ . Zauważmy, że w tym modelu  $\mu_1$  odgrywa rolę „wyrazu wolnego”. Można sobie wyobrazić, że średnią  $\mu_1$  traktujemy jako „poziom bazowy” zaś pozostałe parametry uznajemy za „odchylenia od poziomu bazowego”.

Hipoteza

$$H_0 : \alpha_2 = \dots = \alpha_p = 0$$

sprowadza się do stwierdzenia, że wszystkie próbki pochodzą z tego samego rozkładu. Alternatywa jest bardzo ogólna:  $H_1$  : nie jest prawdą, że  $\alpha_2 = \dots = \alpha_p = 0$  (czyli nie wszystkie średnie  $\mu_j$  są jednakowe).

Statystyka testowa  $F$  wyprowadzona w ogólnej sytuacji we Wniosku 9.3.1 przybiera dla modelu wielu próbek szczególnie prostą postać. Niech

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ji}$$

będzie średnią w  $j$ -tej grupie, zaś

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^{n_j} Y_{ji} = \frac{1}{n} \sum_{j=1}^p n_j \bar{Y}_j$$

oznacza średnią „globalną”, obliczoną z połączonych próbek. Wprowadźmy oznaczenia

$$\text{TSS} = \sum_{j=1}^p \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y})^2, \quad \text{BSS} = \sum_{j=1}^p n_j (\bar{Y}_j - \bar{Y})^2,$$

$$\text{WSS} = \sum_{j=1}^p \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2.$$

Te skróty są związane ze specyfiką modelu kilku próbek: TSS jest *całkowitą* sumą kwadratów (ang. „Total Sum of Squares”) BSS jest sumą kwadratów *pomiędzy* próbkami (ang. „Between”), zaś WSS jest sumą kwadratów *wewnątrz* próbek (ang. „Within”).

Rozpatrujemy tylko szczególny przypadek ogólnego modelu liniowego. Łatwo zauważyć związek naszych nowych oznaczeń z używanymi poprzednio. Mamy

$$\hat{Y}_0 = \underbrace{(\bar{Y}, \dots, \bar{Y})}_n,$$

$$\hat{Y} = \underbrace{(\bar{Y}_1, \dots, \bar{Y}_1)}_{n_1}, \dots, \underbrace{(\bar{Y}_p, \dots, \bar{Y}_p)}_{n_p}^\top.$$

Stąd

$$\text{TSS} = \|Y - \hat{Y}_0\|^2, \quad \text{BSS} = \|\hat{Y} - \hat{Y}_0\|^2, \quad \text{WSS} = \text{RSS} = \|Y - \hat{Y}\|^2$$

Otrzymujemy podstawową tożsamość analizy wariancji:

$$\text{TSS} = \text{BSS} + \text{WSS}.$$

Wiemy również, że  $\text{WSS} \sim \chi^2(n-p)$ . Przy założeniu prawdziwości  $H_0$  mamy  $\text{BSS} \sim \chi^2(p-1)$ . Statystyka testowa przyjmuje postać

$$F = \frac{\text{BSS}/(p-1)}{\text{WSS}/(n-p)}.$$

Hipotezę  $H_0$  odrzucamy, jeśli  $F > F_{1-\alpha}(p-1, n-p)$ . W praktyce, zamiast ustalać próg odrzuceń, podaje się  $P$ -wartość testu. Intuicyjny sens statystyki  $F$  wydaje się przy tym zrozumiały niezależnie od formalnej konstrukcji testu.

Związłym podsumowaniem obliczeń jest tak zwaną tabelką analizy wariancji (ANOVA) dla modelu  $p$  próbek:

Źródło zmienności	Sumy kwadratów	Stopnie swobody	Średnie kwadraty	Statystyka
między próbkami	BSS	$p - 1$	$BSS/(p - 1)$	$F$
wewnątrz próbek	WSS	$n - p$	$WSS/(n - p)$	
razem	TSS	$n - 1$	$TSS/(n - 1)$	

*Uwaga.* Opisany powyżej sposób „zakodowania” modelu kilku próbek nie jest najbardziej naturalny. Można by zdefiniować „nieme” zmienne objaśniające  $x_1, \dots, x_p$  inaczej, według takiej tabelki:

próbka \ zmienne	$x_1$	$x_2$	...	$x_p$
1	1	0	...	0
2	0	1	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$p$	0	0	...	1

Odpowiada to przyjęciu, że dla obserwacji z  $j$ -tej próbki mamy,  $x_j = 1$  i wszystkie inne zmienne  $x$ -owe są zerami. Wtedy w roli wektora współczynników mielibyśmy po prostu wektor średnich w próbkach:  $\beta = (\mu_1, \dots, \mu_p)$ . Interesującą nas hipotezę napisalibyśmy w postaci  $H_0 : \mu_1 = \dots = \mu_p$ . Nietrudno zauważyć, że oba sposoby kodowania są całkowicie równoważne.

**9.3.2 Przykład.** Rozważmy trzech klientów towarzystwa ubezpieczeniowego. Powiedzmy, że są to firmy wynajmujące samochody. Wyobraźmy sobie, że roczne sumy szkód są takie:

Lata:	1	2	3	4	5	6	7	średnie „indywidualne”
1 firma	10	15	16	14	8	17	11	13
2 firma	7	16	10	8	9			10
3 firma	9	18	14	15	11	17		14

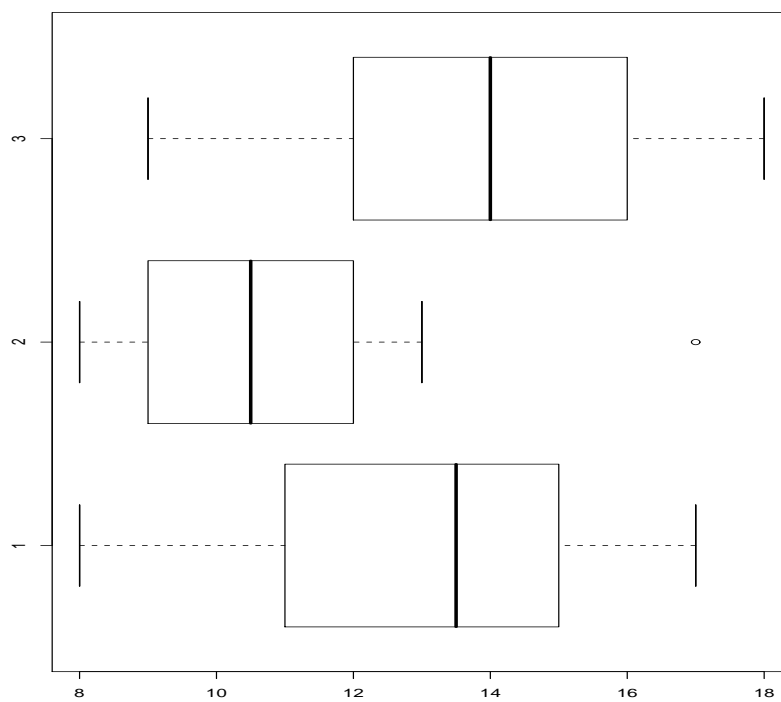
„Średnia globalna” jest równa  $(7/18) * 13 + (5/18) * 10 + (6/18) * 14 = 12.5$ .

Chcemy sprawdzić, czy wysokości szkód w trzech firmach są „istotnie różne”. Całkiem heurystycznie, możemy wizualnie porównać nasze trzy próbki przy pomocy wykresów pudełkowych. Rezultat widoczny jest na rysunku.

Wydaje się, że różnice pomiędzy rozkładami szkód (średnimi) dla trzech firm są dość wyraźne.

Formalny test prowadzi do przeciwnego wniosku. Testujemy hipotezę zerową: „wartości oczekiwane wszystkich trzech próbek są równe”. Oto tabelka analizy wariancji:





Rysunek 9.3: Wykresy pudełkowe dla 3 próbek w Przykładzie 9.3.2.

Źródło zmienności	Sumy kwadratów	Stopnie swobody	Średnie kwadraty	Statystyka
między próbkami	46.5	2	23.25	$F = 1.96$
wewnątrz próbek	178	15	11.87	
razem	224.5	17		

Test F na poziomie istotności  $\alpha = 0.05$  nie odrzuca hipotezy zerowej, bo  $F_{0.95}(2, 15) = 3.69 > 1.96$  (odpowiedni kwantyl został odczytany z tablicy rozkładu Snedecora). Możemy przypuszczać, że wysokość zgłaszanych w przyszłości szkód będzie podobna dla wszystkich trzech firm. Nasze dane nie dają dostatecznych podstaw, by zwątpić w to przypuszczenie.

Dodajmy jeszcze komentarz na temat założeń, które są wymagane przy stosowaniu testu F. Możemy dość spokojnie przyjąć, że sumaryczne (lub średnie) szkody w kolejnych latach są zmiennymi losowymi o rozkładzie zbliżonym do normalnego. To jest pierwsze z podstawowych założeń modelu. Gorzej jest z drugim założeniem: o równości wariancji poszczególnych próbek. Jest ono uzasadnione właściwie tylko wtedy, gdy liczba ubezpieczonych samochodów dla trzech firm (i dla kolejnych lat) jest w przybliżeniu równa.  $\diamond$

## Hipoteza o braku zależności

Wróćmy do liniowej regresji wielorakiej z wyrazem wolnym. Wektor współczynników zapisujemy w tej sytuacji jako  $(\beta_0, \beta_1, \dots, \beta_r)$ , gdzie  $r + 1 = p$ . Jeśli znikają wszystkie współczynniki funkcji regresji z wyjątkiem wyrazu wolnego, to wartości zmiennej  $Y$  nie są powiązane z wartościami zmiennych objaśniających  $x$ . Ważną kwestią jest więc weryfikacja hipotezy

$$H_0 : \beta_1 = \dots = \beta_r = 0 \text{ przeciw } H_1 : \beta_1 \neq 0 \text{ lub } \dots \text{ lub } \beta_r \neq 0.$$

W istocie, gdy nie ma podstaw do odrzucenia  $H_0$ , to model traci swoją użyteczność. Jest to tylko szczególny przypadek zagadnienia testowania hipotezy liniowej i możemy skorzystać z ogólnych wyników. Trzeba tylko uważnie liczyć stopnie swobody. Oprócz używanego już oznaczenia

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

wprowadźmy nowe nazwy sum kwadratów:

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

TSS nazywamy całkowitą sumą kwadratów, zaś ESS sumą kwadratów związana z regresją (ang. „Sum of Squares, Regression”). Zauważmy, że  $(\bar{Y}, \dots, \bar{Y})$  jest predykcją przy założeniu  $H_0$ , a  $(\hat{Y}_1, \dots, \hat{Y}_n)$  jest predykcją w „dużym modelu” z  $r + 1$  współczynnikami. Stąd natychmiast wynika tożsamość analizy wariancji,

$$TSS = ESS + RSS.$$

Tę równość interpretuje się w taki sposób:

$$\begin{aligned} \text{całkowita zmienność } Y &= \text{zmienność „wyjaśniona regresją”} \\ &+ \text{zmienność „resztowa”}. \end{aligned}$$

Pozostawimy Czytelnikowi wytłumaczenie intuicji stojących za tą sugestywną terminologią. *Współczynnikiem dopasowania* nazywamy

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Zgodnie z przytoczoną wyżej interpretacją,  $R^2$  jest „częścią zmienności, wyjaśnioną przez regresję”. Zazwyczaj współczynnik dopasowania wyraża się w procentach. Im większe  $R^2$ , tym lepiej (estymowana) prosta regresji „pasuje” do punktów doświadczalnych – stąd nazwa.

Z ogólnej teorii wynika, że  $\text{RSS} \sim \chi^2(n - r - 1)$ . Przy założeniu prawdziwości  $H_0$  mamy  $\text{ESS} \sim \chi^2(r)$ . Statystyka testu Snedecora jest następująca

$$F = \frac{\text{ESS}/r}{\text{RSS}/(n - r - 1)}.$$

Hipotezę  $H_0$  odrzucamy, jeśli  $F > F_{1-\alpha}(r, n - r - 1)$  lub, równoważnie, jeśli  $P$ -wartość testu jest poniżej  $\alpha$ . Tabelka „analizy wariancji” przybiera postać:

Źródło zmienności	Sumy kwadratów	Stopnie swobody	Średnie kwadraty	Statystyka
regresja	ESS	$r$	ESS/ $r$	$F$
błąd	RSS	$n - r - 1$	RSS/ $(n - r - 1)$	
razem	TSS	$n - 1$	TSS/ $(n - 1)$	

Wartości statystyki  $F$  interpretuje się jako wskaźnik „istotnej zależności” zmiennej  $Y$  od zmiennych  $x_1, \dots, x_r$ . Mówi się w żargonie statystycznym, że zależność „jest istotna na poziomie  $\alpha$ ”, jeśli test  $F$  na tym poziomie istotności odrzuca hipotezę o braku zależności.

## 9.4 Zadania

**Zadania dotyczące modelu z jedną zmienną objaśniającą (prosta regresja liniowa).**

**9.1.** W celu zbadania zależności pomiędzy liczbą urodzonych wiosną dzieci (cecha  $Y$ ) a liczbą bocianów (cecha  $x$ ) zebrano wyniki ze 100 miejscowości. Otrzymano następujące wyniki

$\sum_{i=1}^n x_i$	200
$\sum_{i=1}^n x_i^2$	1400
$\sum_{i=1}^n Y_i$	1000
$\sum_{i=1}^n x_i Y_i$	7000

Wyznaczyć estymatory współczynników prostej regresji liniowej z wyrazem wolnym traktując liczbę dzieci jako zmienna zależną (objaśnianą).

**9.2.** Wyprowadzić wzory (9.2.1) na  $\hat{\beta}_0$  i  $\hat{\beta}_1$ .

**9.3.** Wyprowadzić bezpośrednio wzory na  $\text{Var}\hat{\beta}_1$  i  $\text{Var}\hat{\beta}_0$ .

**9.4.** Pokazać, że zmienne losowe  $\bar{Y}$  i  $\hat{\beta}_1$  są niezależne.

**9.5.** Wyprowadzić bezpośrednio wzory na  $\text{Var}\hat{Y}^*$  i  $\text{Var}(Y^* - \hat{Y}^*)$ .

*Wskazówka:* Skorzystać z poprzednich zadań.

**9.6.** Udowodnić bezpośrednio (nie korzystając z geometrycznych rozważań w przestrzeni  $\mathbb{R}^n$ ) podstawową tożsamość analizy wariancji:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2.$$

**9.7.** Współczynnik korelacji<sup>4</sup>  $R$  określamy wzorem

$$R = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (Y_i - \bar{Y})^2}}.$$

Pokazać, że kwadrat współczynnika korelacji jest współczynnikiem dopasowania.

**9.8.** Udowodnić fakt sformułowany w Uwadze 9.3:  $T^2 = F$ .

Pokazać, że test F odrzuca  $H_0 : \beta_1 = 0$  (na poziomie istotności  $\alpha$ ) wtedy i tylko wtedy, gdy przedział ufności dla  $\beta_1$  (na poziomie  $1 - \alpha$ ) nie zawiera zera.

**9.9.** Wyprowadzić wzory na estymatory największej wiarygodności w modelu prostej regresji liniowej bez wyrazu wolnego,

$$Y_i = \beta x_i + \varepsilon_i, \quad (i = 1, \dots, n),$$

przyjmując Założenie 9.1.2.

<sup>4</sup>Związek  $R$  z pojęciem korelacji zmiennych losowych staje się jasny, gdy rozpatrujemy model z losową zmienną objaśniającą. W modelu z deterministycznym  $x$ , przyjmijmy po prostu, że „tak się mówi”.

Następujące zadania dotyczą analizy wariancji w modelu kilku próbek.

**9.10.** Uzasadnić elementarnie, bez odwoływania się do geometrii przestrzeni  $\mathbb{R}^n$ , następujące fakty.

1. Wyprowadzić tożsamość analizy wariancji:  $TSS = BSS + WSS$ .
2. Wykazać, że  $WSS/\sigma^2 \sim \chi^2(n - r)$ , niezależnie od tego, czy  $H_0$  jest prawdziwa, czy nie.
3. Wykazać, że  $BSS/\sigma^2 \sim \chi^2(r - 1)$ , przy założeniu, że  $H_0$  jest prawdziwa.

*Wskazówka:* Skorzystać ze Stwierdzenia 2.2.3.

**9.11.** Pokazać, że statystyka  $F$  testu analizy wariancji jest równoważna statystyce ilorazu wiarygodności dla modeli zagnieżdżonych (Punkt 8.3.1 w Podrozdziale 8.3).

*Wskazówka:* Bardzo podobne rozważania przeprowadziliśmy w Przykładzie 8.3.2.

**9.12.** Wykazać, że dwupróbkowy test Studenta 7.2.2 w wersji z dwustronną alternatywą jest równoważny testowi analizy wariancji dla  $k = 2$ . Dokładniej, statystyka  $F$  jest kwadratem statystyki  $t$  i odpowiednie kwantyle są związane relacją  $F_{1-\alpha}(1, n - 2) = t_{1-\alpha/2}^2(n - 2)$ .

Następujące zadania dotyczą innych kwestii związanych z modelem liniowym.

**9.13.** Niech obserwacje  $Y_1, \dots, Y_n$  będą niezależnymi zmiennymi losowymi takimi, że  $Y_i$  ma rozkład normalny  $N(\beta_0, \sigma^2)$  dla  $i = 1, \dots, n$  oraz  $N(\beta_0 + \beta_1, \sigma^2)$  dla  $i = n + 1, \dots, 2n$ . Zakładamy, że  $\beta_0, \beta_1, \sigma$  są nieznanymi parametrami.

1. Podać estymator największej wiarygodności parametrów  $\beta_0, \beta_1$ .
2. Podać estymator największej wiarygodności parametru  $\sigma^2$ .

*Uwaga:* Podać wzory nie zawierające działań na wektorach i macierzach (wraz z uzasadnieniem tych wzorów)

**9.14.** Zakładamy (błędnie), że badane przez nas zjawisko opisuje model regresji liniowej dany równaniem:

$$Y = \mathbf{X}_1 \cdot \beta_1 + \varepsilon,$$

gdzie  $Y$  to wektor obserwacji zmiennej objaśnianej,  $\mathbf{X}_1$  to macierz obserwacji zmiennych objaśniających, a  $\beta_1$  jest wektorem nieznanymi parametrami. Oznaczmy przez  $\hat{\beta}_1$  estymator  $\beta_1$  uzyskany metodą najmniejszych kwadratów w tym modelu.

W rzeczywistości, badane zjawisko opisuje prawdziwie model regresji liniowej z dodatkowymi zmiennymi objaśniającymi  $\mathbf{X}_2$ , dany równaniem:

$$Y = \mathbf{X}_1 \cdot \beta_1 + \mathbf{X}_2 \cdot \beta_2 + \varepsilon.$$

Wektor błędów losowych  $\varepsilon$  ma zerową wartość oczekiwaną i macierz wariancji-kowariancji równą  $\sigma^2 \mathbf{I}$ .

1. Obliczyć obciążenie estymatora  $\hat{\beta}_1$ , czyli wielkość  $\mathbb{E}\hat{\beta}_1 - \beta_1$ , biorąc pod uwagę prawdziwy mechanizm generowania zmiennej objaśnianej  $Y$ .
2. Obliczyć macierz wariancji-kowariancji estymatora  $\hat{\beta}_1$ .

