Zestaw 4

Statystyczna Analiza Danych

Emilia Wiśnios

11 kwietnia 2021

1. **Regresja prosta**

   Drużyna informatyków testuje nowy program rozciągania pleców w przerwach od programowania zespołowego, by ograniczyć liczbę dolegliwości. Poniższa tabelka pokazuje codzienną liczbę minut rozciągania się przez uczestników i powiązaną z nimi liczbę dolegliwości przez cały semestr.

Czas rozciągania w minutach	0	30	10	15	5	25	35	40
Dolegliwości	4	1	2	2	3	1	0	1

   (a) Znajdź prostą regresji, pokazującą zależność ilości dolegliwości od czasu rozciągania.

   (b) O ile spada ilość dolegliwości z każdą minutą rozciągania?

   (c) Ile minut rozciągania jest potrzebne, by zawodnik uniknął dolegliwości?

   Uwaga: We wszystkich zadaniach poniżej rozważany jest model liniowy:

$$Y = X\beta + \varepsilon,$$

   gdzie $Y \in \mathbb{R}^n$ jest zmienną objaśnianą, $X \in \mathbb{R}^{n\times p}$ jest macierzą planu, $\beta \in \mathbb{R}^p$ p wektorem nieznanych współczynników oraz $\varepsilon \in \mathbb{R}^n$ wektorem nieskorelowanych błędów, czyli $\mathbb{E}\varepsilon = 0, Var\varepsilon = \sigma^2 Id$.

   Rozwiązanie:

$$y = \beta_0 + \beta_1 X + \varepsilon$$
$$\hat{\beta_1} = \frac{\overline{XY} - \overline{XY}}{\overline{X^2} - \overline{X}^2}$$
$$\hat{\beta_0} = \overline{Y} - \beta_1 \overline{X}$$

   X - czas rozciągania, Y - dolegliwości

$$\overline{X} = 20, \quad \overline{Y} = \frac{14}{8}, \quad \overline{XY} = 20, \quad \overline{X^2} = 587,5$$
$$\hat{\beta_1} = \frac{20 \cdot 20 \cdot \frac{14}{8}}{587,5 - 20 \cdot 20} = -0.08$$
$$\hat{\beta_0} = \frac{14}{8} + 10 \cdot 0.08 = 3.35$$
$$y = 3.35 - 0.08x$$

   b) Z każdą minutą spada o 0.08.

   c) $y = 3.35 - 0.08x = 0, \quad x = 3.35 \cdot \frac{100}{8} = 41.875$

2. Pokaż, że jeżeli $n \geq p$ oraz $rank(X) = n$ to estymator β metodą najmniejszych kwadratów

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||^2,$$

jest postaci

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Dodatkowo, jeżeli $\varepsilon \sim N(0, \sigma^2 Id)$ to pokaż, że $\hat{\beta}$ jest estymatorem największej wiarygodności oraz znajdź estymator największej wiarygodności σ^2.

Rozwiązanie:

$$\min_{\beta_0,\beta_1} RSS(\beta) = \min_{\beta_0,\beta_1} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$$

Chcemy znaleźć minimum lokalne (funkcja jest wypukła, więc jest takie jedno).

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Rozwiązujemy układ równań

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

$$\sum X_i Y_i - \beta_0 \sum X_i - \beta_1 \sum X_i^2 = 0$$

$$\sum X_i Y_i - \overline{Y} \sum X_i + \beta_1 \sum X_i \overline{X} - \beta_1 \sum X_i^2 = 0$$

$$\overline{XY} - \overline{YX} + \beta_1 \overline{X}^2 - \beta_1 \overline{X^2} = 0$$

$$\beta_1 = \frac{\overline{XY} - \overline{YX}}{\overline{X^2}\overline{X}^2} = \frac{\hat{Cov}(X,Y)}{\hat{Var}(X)}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

Teraz w wersji macierzowej

$$RSS(\beta) = \sum_i (Y_i - X_i^T \beta)^2 = ||Y - X\beta||^2$$

$$\nabla RSS(\beta) = \left[\frac{\partial RSS}{\partial \beta_i}\right]_{i=1,\ldots,p}$$

$$f(\beta) = Y - \beta X$$

$$f(x) = g(x)^T h(x), \quad g, h : \mathbb{R}^p \to \mathbb{R}^n$$

$$\nabla f(x) = \nabla g^T h + \nabla h^T g$$

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta)$$

$$\nabla RSS(\beta) = -2X^T (Y - X\beta) = 0$$

$$X^T X \beta = X^T Y$$

Zatem macierz X musi być odwracalna, czyli kiedy

$$rank(X^T X) = p \;\Rightarrow\; rank(X) = p$$

oraz

$$n \geq p$$

$$\beta = (X^T X)^{-1} X^T Y$$

Na koniec pokażemy że $\hat{\beta}$ jest estymatorem największej wiarygodności

$$\varepsilon = [\varepsilon_1, ..., \varepsilon_n]$$

$$\frac{1}{\sqrt{2\pi}} \sigma^n \exp\left(-\frac{1}{2\sigma^2} \sum \varepsilon_i^2\right)$$

$$Y = X\beta + \varepsilon$$

$$f_Y(Y) = \left(\frac{1}{\sqrt{2\pi}}\sigma\right)^n \left(-\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta)\right)$$

$$\arg\max_\beta f_Y(Y) = \arg\max_\beta (Y - X\beta)^T (Y - X\beta)$$

$$\arg\max_\beta f_Y(Y)$$

$$\hat{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{n}$$

3. Oblicz wartość oczekiwaną i wariancję $\hat{\beta}$. Jaki ma rozkład $\hat{\beta}$, jeśli $\varepsilon \sim (0, \sigma^2 Id)$?

Rozwiązanie:

$$Y = X\beta + \varepsilon, \quad X \text{ jest deteriministyczne}$$

$$Var\varepsilon = \sigma^2 I$$

$$E\epsilon = 0$$

$$E\hat{\beta} = E((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T E(X\beta + \varepsilon) = (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T 0 = \beta$$

Zatem jest estymatorem nieobciążonym.

X zmienna losowa $X \in R$, a stała

$$Var(aX) = a^2 VarX$$

X zmienna losowa $X \in R$, A macierz $p \times d$

$$Var(AX) = A(VarX)A^T$$

Pokażemy, że tak jest

$$Var(X - EX) = VarX = E(X - EX)(X - EX)^T$$

$$EX = 0$$

$$Var(X) = E(XX^T)$$

$$E(AX) = AE(X) = 0$$

$$Var(AX) = E(AX)(AX)^T = AE(XX^T)A^T = AVar(X)A^T$$

Skorzystamy z tej zależności i obliczymy wariancję β

$$Var\hat{\beta} = Var(X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T VarY \left[(X^T X)^{-1} X^T\right]^T =$$

$$= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

$$Y = X\beta + \varepsilon$$

Teraz zajmiemy się rozkładem β

$$\varepsilon_i \sim iid \quad N(0, \sigma^2)$$
$$\varepsilon \sim N(0, \sigma^2 I)$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

4. Pokaż, że predykcja $\hat{Y} = X\hat{\beta}$ jest rzutem prostopadłym Y na przestrzeń liniową rozpiętą przez kolumny macierzy X.

   Rozwiązanie:

$$\hat{\sigma^2} = \frac{||Y - \hat{Y}||^2}{n}$$
$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY, \quad H = X(X^T X)^{-1} X^T$$
$$H^T = X(X^T X)^{-1} X^T = H$$
$$H^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = H$$

   Wynika z tego, że H jest rzutem ortogonalnym

$$HX = X(X^T X)^{-1} X^T X = X$$

   Zatem kolumny macierzy X rozpinają przestrzeń na którą rzutujemy. Czyli $\hat{Y} = X\hat{\beta}$ jest rzutem Y na X.

5. Niech $n > p$ oraz $\varepsilon \sim N(0, \sigma^2 Id_n)$. Rozpatrzmy rozkład QR macierzy X, czyli

$$X = QR,$$

   gdzie $Q \in \mathbb{R}^{n \times p}$ ortogonalna $R \in \mathbb{R}^{p \times p}$ górnotrójkątna. Korzystając z rozkładu QR pokaż, że

$$||Y - X\hat{\beta}||^2 = ||\overline{Q}^T Y||^2$$

   gdzie $\overline{Q}$ to dopełnienie ortogonalne Q do bazy ortogonalnej w $\mathbb{R}^n$. Stąd wynika, że $||Y - X\hat{\beta}||^2/\sigma^2$ ma rozkład $\chi^2(n-p)$ oraz jest niezależny od $\hat{\beta}$. Podaj nieobciążony estymator σ^2.

   Rozwiązanie:

6. Niech $n > p$ oraz $\varepsilon \sim N(0, \sigma^2 Id_n)$. Niech x_*, będzie nową obserwacją oraz $y_* = x_*^T \beta + \varepsilon_*$, gdzie $\varepsilon_* \sim N(0, \sigma^2)$ niezależny od ε. Skonstruuj przedział ufności dla predykcji $\hat{y}_* = x_*^T \hat{\beta}$.

   Rozwiązanie: Strony 137-138.

7. **Algorytm MNK**

   Rozpatrujemy model $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ oszacowany na próbie n-elementowej.

   - Pokazać przy użyciu wzoru $(X^T X)^{-1} X^T y$, że w tym modelu estymator MNK parametru β_2 ma postać:
$$b_2 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i^2)}$$
   - Pokazać, że w tym modelu estymator MNK parametru β_1 spełnia $\overline{y} = b_1 + b_2 \overline{x}$.
   - Ile wynosi liczebność próby, jeżeli $b_1 = 25, b_2 = -0.5, \sum_{i=1}^n y_i = 100, \sum_{i=1}^n x_i = 200$.

- Znaleźć wektor reszt, wektor wartości dopasowanych, estymator σ^2, estymator macierzy wariancj-kowariancji oraz R^2.

Rozwiązanie:

8. Znaleźć estymator MNK wektora dla modelu $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, w którym $n = 4, x_1^T = [1, 1, 2, -4], x_2^T = [-3, -3, 5, 1], y^T = [1, 2, 3, 1]$. Znaleźć wektor reszt, wektor wartości dopasowanych, estymator σ^2, estymator macierzy wariancji-kowariancji oraz R^2.

Rozwizanie:

$$X = \begin{bmatrix} 1 & 1 & -3 \\ 1 & 1 & -3 \\ 1 & 1 & 5 \\ 1 & -4 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \end{bmatrix}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = X^T \beta$$

$$X^T X = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 22 & 0 \\ 0 & 0 & 44 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 7 \\ 5 \\ 7 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{22} & 0 \\ 0 & 0 & \frac{1}{44} \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \frac{7}{4} \\ \frac{5}{22} \\ \frac{7}{44} \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} 1 & 1 & -3 \\ 1 & 1 & -3 \\ 1 & 2 & 5 \\ 1 & -4 & 1 \end{bmatrix} \begin{bmatrix} \frac{77}{44} \\ \frac{10}{44} \\ \frac{7}{44} \end{bmatrix} = \begin{bmatrix} 1\frac{1}{2} \\ 1\frac{1}{2} \\ 3 \\ 1 \end{bmatrix}$$

$$Y - \hat{Y} = \left[-\frac{1}{2}, -\frac{1}{2}, 0, 0 \right]^T$$

$$\hat{\sigma^2} = \frac{\frac{1}{4} + \frac{1}{4}}{4} = \frac{1}{8}$$

$$\hat{\sigma^2}(X^T X)^{-1} = \frac{1}{8} \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{22} & 0 \\ 0 & 0 & \frac{1}{44} \end{bmatrix}$$

9. Dany jest następujący model:

$$y_i = \beta_1 + \beta_2 d_i + \varepsilon_i, \text{ dla } i = 1, ..., N$$

$$d_i = \begin{cases} 1 & \text{dla } i \leq q \\ 0 & \text{gdy } i > q \end{cases}$$

$$Var(\varepsilon_i) = \sigma^2 Id$$

- Podać estymatory MNK dla parametrów β_1, β_2 dla $N = 60, q = 40, \sum_{i=1}^{N} y_i = 50, \sum_{i=1}^{q} y_i = 30$.
- Udowodnić, że estymatory te są nieobciążone.
- Podać postać macierzy wariancji-kowariancji dla estymatorów β_1, β_2, jeżeli spełnione są założenia KMRL.

Rozwiązanie:

$$d = \begin{bmatrix} 1 \\ 1 \\ 1 \\ ... \\ 0 \end{bmatrix}, \quad y_i = \beta_0 + \beta_1 d_i$$

$$\beta_1 = \frac{\overline{XY} - \overline{XY}}{\overline{X^2} - \overline{X}^2}$$

$$\overline{X} = \frac{q}{n}, \quad \overline{XY} = \frac{\sum_i^q y_i}{n} = \overline{Y_q}\frac{q}{n}, \quad \overline{X^2} = \frac{q}{n}$$

$$\beta_1 = \frac{\frac{1}{60}30 - \frac{4}{6}\frac{50}{60}}{\frac{4}{6} - \frac{16}{36}} = \frac{\frac{18}{36} - \frac{20}{36}}{\frac{24}{36} - \frac{16}{36}} = -\frac{1}{4}$$

$$\beta_0 = \frac{5}{6} - \left(-\frac{1}{4}\right)\frac{4}{6} = \frac{20}{24} + \frac{4}{24} = 1$$

Zatem

$$y = 1 - \frac{1}{4}d$$

Udowodniliśmy wcześniej że te estymatory są nieobciążone.

$$X^T X = \begin{bmatrix} n & q \\ q & q \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{nq - q^2}\begin{bmatrix} q & -q \\ -q & n \end{bmatrix}$$

Wystarczy policzyć σ^2 oraz $\hat{y}$