

# Statystyka matematyczna - wykład siódmy<sup>1</sup>

## Klasyczny model regresji liniowej.

kierunek: informatyka i ekonometria I<sup>o</sup>

dr Jarosław Kotowicz

Instytut Informatyki, Uniwersytet w Białymstoku

---

<sup>1</sup>©J.Kotowicz

# Spis treści

- 1 **Klasyczny model regresji liniowej**
  - Sformułowanie modelu
  - Estymacja parametrów klasycznego modelu regresji liniowej
  - Analiza wariancji w modelu regresji
  - Macierzowe ujęcie modelu regresji liniowej
- 2 Klasyczny model regresji liniowej z wieloma zmiennymi niezależnymi

# Motywacja

Ostateczny cel analizy regresji to narzędzie predykcji, czyli przewidywanie jakie wartości przyjmie zmienna zależna przy ustalonych wartościach zmiennej (zmiennych) uznanych za niezależną (niezależne).

Stosowana jest konstrukcja tzw. **modeli regresji**, które wyjaśnia w sposób analityczny kształtowanie się wartości jednej zmiennej losowej pod wpływem innej lub innych zmiennych losowych.

Spośród wielu możliwych postaci modelu regresji podstawowe znacznie ma tzw. **model klasycznej regresji liniowej**.

# Klasyczny model regresji liniowej – przypadek dwuwymiarowy

## Model

Dla każdej ustalonej wartości jednej zmiennej losowej (np.  $X$  - zmienna niezależna) druga zmienna losowa ( $Y$  - zmienna zależna) ma warunkowy rozkład z wartością oczekiwaną

$$\mathbb{E}(Y|X = x) = \beta_1 x + \beta_0, \quad (1)$$

gdzie funkcja regresji  $l$ -go rodzaju zmiennej  $Y$  względem zmiennej  $X$  jest liniowa ( $\beta_1$  – współczynnik regresji liniowej) oraz stałą wariancję

$$\mathbb{D}^2(Y|X = x) = \sigma^2 \quad (2)$$

niezależna od  $x$ .

# Uwaga

- 1 Zmienna  $Y$  traktujemy jako zmienną zależną, a zmienną  $X$  jako niezależną.
- 2 Współczynnik regresji liniowej  $\beta_1$  jest wielkością o jaką zmienia się warunkowa wartość oczekiwana zmiennej zależnej  $Y$ , gdy  $x$  wzrasta o jednostkę.
- 3 Istotą klasycznego podejścia do zagadnienia regresji jest traktowanie wartości zmiennej niezależnej, jako wartości z góry ustalonych, czyli nielosowych.

# Klasyczny model normalnej regresji liniowej

Oprócz klasycznego modelu regresji liniowej będziemy rozważać jeszcze jeden model.

## Definicja

*Jeżeli oprócz warunków (1) i (2) będziemy zakładać, że rozkłady warunkowe zmiennej  $Y$  są normalne, tzn.  $Y$  dla  $X = x$  ma rozkład  $\mathcal{N}(\beta_1 x + \beta_0, \sigma)$ , to będziemy mówili wtedy o klasycznym modelu normalnej regresji liniowej.*

# Alternatywne sformułowanie klasycznego modelu regresji liniowej

Niech ciąg par  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  będzie  $n$ -elementową próbą losową z populacji dwuwymiarowej, stanowiącą podstawę estymacji parametrów badanej zależności (wartości zmiennej  $X$  są w próbie ustalone).

Kształtowanie się wartości  $Y_i$  w próbie można wyjaśnić następująco

$$Y_i = \mathbb{E}(Y|X = x_i) + \varepsilon_i = \beta_1 x_i + \beta_0 + \varepsilon_i, \quad (3)$$

gdzie  $i \in \overline{1, n}$  i  $\varepsilon_i$  są zmiennymi losowymi takimi, że

$$\forall_{i \in \overline{1, n}} \mathbb{E}(\varepsilon_i) = 0 \quad (4)$$

$$\forall_{i \in \overline{1, n}} \mathbb{D}^2(\varepsilon_i) = \sigma^2, \quad (5)$$

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ dla dowolnych } i \neq j. \quad (6)$$

Jest to alternatywne sformułowanie klasycznego modelu regresji liniowej  $Y$  względem  $X$ .

# Alternatywne sformułowanie klasycznego modelu normalnej regresji liniowej

Jeżeli warunki (3), (4), (5), (6) uzupełnimy o założenie, że  $\varepsilon_i$ , dla  $i \in \overline{1, n}$ , mają rozkład  $\mathcal{N}(0, \sigma)$ , to otrzymujemy klasyczny model normalnej regresji liniowej zmiennej  $Y$  względem zmiennej  $X$ .



# Równoważność warunków modeli

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}(\beta_1 x_i + \beta_0 + \varepsilon_i) = \beta_1 x_i + \beta_0 + \mathbb{E}(\varepsilon_i) \\ &= \beta_1 x_i + \beta_0,\end{aligned}$$

gdzie w pierwszej równości skorzystaliśmy z (3), w drugiej z liniowości wartości oczekiwanej i faktu, że wartości zmiennej niezależnej są nielosowe (deterministyczne), a w trzeciej z (4).

Podobnie wykorzystując (5) otrzymujemy

$$\mathbb{D}^2(Y_i) = E[Y_i - \mathbb{E}(Y_i)]^2 = E[Y_i - \beta_1 x_i + \beta_0]^2 = \mathbb{E}(\varepsilon_i^2) = \sigma^2.$$

# Estymacja parametrów $\beta_1$ i $\beta_0$ . I

Założmy, że w populacji dwuwymiarowej  $(X, Y)$  pobieramy  $n$ -elementową próbę  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ . Wyniki konkretnej próby  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  można przedstawić w układzie współrzędnych otrzymując w ten sposób wykres rozrzutu punktów empirycznych. Szukamy wykresy prostej "*najlepiej dopasowanej*" do otrzymanych punktów, stosując metodę najmniejszych kwadratów (**MNK**). Będziemy minimalizować funkcję

$$S \equiv S(\beta_1, \beta_0) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [Y_i - (\beta_1 x_i + \beta_0)]^2. \quad (7)$$

Licząc pochodne cząstkowe i przyrównując je do zera (warunek konieczny istnienia ekstremum) otrzymujemy

## Estymacja parametrów $\beta_1$ i $\beta_0$ . II

$$\begin{cases} \frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (Y_i - \beta_1 x_i - \beta_0) = 0 \\ \frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_1 x_i - \beta_0) = 0 \end{cases}$$

Zastępujemy parametry  $\beta_1$  i  $\beta_0$  ich estymatorami  $\widehat{\beta}_1$  i  $\widehat{\beta}_0$  otrzymując

$$\begin{cases} \sum_{i=1}^n x_i Y_i = \widehat{\beta}_1 \sum_{i=1}^n x_i^2 + \widehat{\beta}_0 \sum_{i=1}^n x_i \\ \sum_{i=1}^n Y_i = \widehat{\beta}_1 \sum_{i=1}^n x_i + n \widehat{\beta}_0 \end{cases}$$

Pierwsze równanie przekształcamy, a z drugiego równania wyznaczamy parametr  $\widehat{\beta}_0$  mamy

Estymacja parametrów  $\beta_1$  i  $\beta_0$ . III

$$\left\{ \begin{array}{l} \sum_{i=1}^n x_i Y_i = \widehat{\beta}_1 \sum_{i=1}^n x_i^2 + \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n Y_i \right) - \widehat{\beta}_1 \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \\ \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \end{array} \right. .$$

Z pierwszego równania wyznaczamy parametr  $\widehat{\beta}_1$  otrzymując

$$\left\{ \begin{array}{l} \widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n Y_i \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2} \\ \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \end{array} \right. .$$

Ostatecznie otrzymujemy

Estymacja parametrów  $\beta_1$  i  $\beta_0$ . IV

$$\left\{ \begin{array}{l} \widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \end{array} \right. .$$

Dzieląc przez  $n$  licznik i mianownik w pierwszym z równań możemy zapisać inaczej rozwiązania

$$\left\{ \begin{array}{l} \widehat{\beta}_1 = \frac{\text{cov}(x, Y)}{s_x^2} \\ \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \end{array} \right. .$$

Otrzymane wzory przedstawiają estymatory parametrów  $\beta_1$  i  $\beta_0$  metodą **MNK**.

# Twierdzenie Gaussa-Markowa

Własności estymatorów parametrów  $\beta_1$  i  $\beta_0$  przedstawia twierdzenie Gaussa-Markowa.

## Twierdzenie

*W klasycznym modelu regresji liniowej najefektywniejszym nieobciążonym estymatorami współczynników regresji są estymatory uzyskane metodą najmniejszych kwadratów.*

# Odchylenia standardowe estymatorów $\widehat{\beta}_1$ i $\widehat{\beta}_0$

Miarą wielkości błędu losowego przy estymacji parametru przy pomocy estymatora jest odchylenie standardowe estymatora nazywane również **standardowym błędem oceny** parametru.

W naszym przypadku mamy

$$\mathbb{D}(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad (8)$$

$$\mathbb{D}(\widehat{\beta}_0) = \sqrt{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}. \quad (9)$$

# Uwagi

- 1 Obie wielkości zależą od  $\sigma^2$ .
- 2 Obie wielkości zależą od liczebności próby  $n$ .
- 3 Obie wielkości zależą od rozproszenia w próbie obserwowanych wartości zmiennej niezależnej  $\sum_{i=1}^n (x_i - \bar{x})^2$
- 4 Mogą być oszacowane dopiero po oszacowaniu  $\sigma^2$ .



## Teoretyczne wartości zmiennej $Y$ i reszty modelu

Liniowa funkcja regresji po oszacowaniu parametrów na podstawie próby wyraża się wzorem

$$\widehat{Y}_i = \widehat{\beta}_1 x_i + \widehat{\beta}_0. \quad (10)$$

### Definicja

Wartości  $\widehat{Y}_i$  nazywamy **teoretycznymi wartościami zmiennej  $Y$** .

### Definicja

Zmienne losowe  $e_i$ , dla  $i \in \overline{1, n}$ , zadane warunkiem

$$e_i \stackrel{\text{def}}{=} Y_i - \widehat{Y}_i$$

nazywamy **resztami modelu**.

# Własności reszt modelu

Biorąc równanie

$$\sum_{i=1}^n Y_i = \hat{\beta}_1 \sum_{i=1}^n x_i + n\hat{\beta}_0,$$

widzimy, że

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i. \quad (11)$$

Stąd suma reszt model spełnia równanie

$$\sum_{i=1}^n e_i = 0. \quad (12)$$

# Uwaga

Rozważając równanie określające estymator  $\widehat{\beta}_0$  otrzymujemy następujący fakt

## Fakt

*Wykres funkcji regresji z próby przechodzi przez punkty  $(\bar{x}, \bar{Y})$ .*

# Estymacja $\sigma^2$

Podstawą estymacji wariancji składników losowych  $\sigma^2$  są reszty

$$e_i = Y_i - \widehat{Y}_i.$$

Obliczając

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^n e_i^2\right) &= \mathbb{E}\left(\sum_{i=1}^n Y_i^2 - 2\sum_{i=1}^n Y_i \widehat{Y}_i + \sum_{i=1}^n \widehat{Y}_i^2\right) = \dots \\ &= \sigma^2(n-2).\end{aligned}$$

Tak więc nieobciążonym estymatorem parametru  $\sigma^2$  jest wariancja reszt

$$S_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}. \quad (13)$$

# Estymacja $\mathbb{D}(\widehat{\beta}_1), \mathbb{D}(\widehat{\beta}_0)$

Natomiast odchylenie standardowe reszt

$$S_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$$

można wykorzystać do estymacji standardowych błędów ocen parametrów  $\beta_1$  i  $\beta_0$ , czyli  $\mathbb{D}(\widehat{\beta}_1)$  i  $\mathbb{D}(\widehat{\beta}_0)$ .

Otrzymujemy wtedy

$$S_{\widehat{\beta}_1} = \sqrt{\frac{S_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (14)$$

$$S_{\widehat{\beta}_0} = \sqrt{\frac{S_e^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}. \quad (15)$$

# Uwagi

- 1 Dokładność estymacji parametrów  $\beta_1$  i  $\beta_0$  jest tym większa,
  - im mniejsza jest wariancja reszt,
  - im większa jest próba,
  - im większy zakres zmienności zmiennej niezależnej  $X$ .

# Dokładność dopasowania prostej MNK. I

Odchylenie obserwowane wartości  $Y_i$  od średniej  $\bar{Y}$  może być przedstawione, jako suma dwóch składników, z których pierwszy jest wyjaśniany regresją liniową  $Y$  względem  $X$  i reszt modelu ( $e_i$ ) tzw. losowej części odchylenia nie wyjaśnianej regresją.

Zapisujemy to

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i). \quad (16)$$

Podnosząc obie strony równości do kwadratu, a następnie sumując po  $i$  otrzymujemy równanie

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (17)$$

Udowodnimy, że środkowy składnik sumy równa się zero.

## Dokładność dopasowania prostej MNK. II

Skorzystamy w tym celu z warunków

$$\begin{cases} \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \\ \widehat{Y}_i = \widehat{\beta}_1 x_i + \widehat{\beta}_0 \end{cases}.$$

Stąd

$$\widehat{Y}_i - \bar{Y} = \widehat{\beta}_1 (x_i - \bar{x}) \quad \text{oraz} \quad \widehat{Y}_i = \bar{Y} + \widehat{\beta}_1 (x_i - \bar{x}).$$

Mamy wtedy

$$\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})(Y_i - \widehat{Y}_i) = \widehat{\beta}_1 \left[ \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) - \widehat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2 \right].$$

Wstawiając wartość estymatora  $\widehat{\beta}_1$  otrzymujemy żadaną tezę.



## Dokładność dopasowania prostej MNK. III

Stąd ostatecznie otrzymujemy równanie

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (18)$$

# Współczynnik deterministyczny

Miarą dokładności dopasowania prostej jest współczynnik deterministyczny, który definiujemy jedną z równości

$$r^2 \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \equiv 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}. \quad (19)$$

Współczynnik ten ma następujące własności

- $r^2 \in [0, 1]$ ,
- $r^2 = 1$  wtedy, gdy między zmiennymi  $X$  i  $Y$  zachodzi zależność liniowa (wszystkie punkty empiryczne leżą na prostej),
- $r^2 = 0$ , gdy  $\widehat{\beta}_1 = 0$ , czyli  $\widehat{Y}_i = \widehat{\beta}_0 = \bar{Y}$  (znajomość wartości zmiennej  $X$  nie dostarcza żadnych informacji na temat wartości zmiennej zależnej  $Y$ ).

# Wnioskowanie o klasycznym modelu normalnej regresji liniowej

Założmy, że warunkowe rozkłady zmiennej zależnej są normalne (składniki losowe modelu  $\varepsilon_i$  mają rozkład  $\mathcal{N}(0, \sigma)$ ).

Parametry  $\widehat{\beta}_1$  i  $\widehat{\beta}_0$  mają rozkłady  $\mathcal{N}(\beta_1, \mathbb{D}(\widehat{\beta}_1))$  i  $\mathcal{N}(\beta_0, \mathbb{D}(\widehat{\beta}_0))$ .

Konstruujemy statystyki dla nich

$$\begin{cases} t = \frac{\widehat{\beta}_1 - \beta_1}{s^{\widehat{\beta}_1}} \\ t = \frac{\widehat{\beta}_0 - \beta_0}{s^{\widehat{\beta}_0}} \end{cases} . \quad (20)$$

Są one rozkładami  $t$ -Studenta o  $n - 2$  stopniach swobody.

Dla współczynnika ufności  $1 - \alpha$  odpowiadające im przedział ufności wynoszą

$$\begin{aligned} &] \widehat{\beta}_1 - t_{\alpha, n-2} S_{\widehat{\beta}_1}, \widehat{\beta}_1 + t_{\alpha, n-2} S_{\widehat{\beta}_1} [ , \\ &] \widehat{\beta}_0 - t_{\alpha, n-2} S_{\widehat{\beta}_0}, \widehat{\beta}_0 + t_{\alpha, n-2} S_{\widehat{\beta}_0} [ . \end{aligned}$$

# Test do weryfikacji hipotezy o parametrze $\beta_1$

$$H_0 : \beta_1 = \beta_1^0$$

$$H_1 : \beta_1 \neq \beta_1^0.$$

Przy założeniu prawdziwości hipotezy zerowej statystyka ma postać

$$t = \frac{\widehat{\beta}_1 - \beta_1^0}{s_{\widehat{\beta}_1}},$$

zaś obszar krytyczny dla poziomu istotności  $\alpha$  opisany jest równaniem

$$P(\{|t| \geq t_{\alpha, n-2}\}) = \alpha.$$

# Test do weryfikacji hipotezy o parametrze $\beta_0$

$$H_0 : \beta_0 = \beta_0^0$$

$$H_1 : \beta_0 \neq \beta_0^0.$$

Przy założeniu prawdziwości hipotezy zerowej statystka ma postać

$$t = \frac{\widehat{\beta}_0 - \beta_0^0}{s_{\widehat{\beta}_0}},$$

zaś obszar krytyczny dla poziomu istotności  $\alpha$  opisany jest równaniem

$$P(\{|t| \geq t_{\alpha, n-2}\}) = \alpha.$$

# Uwagi

- 1 Najczęściej stosowaną wersją testu istotności dla  $\beta_1$  jest  $\beta_1^0 = 0$ .
- 2 Najczęściej hipotezę dotyczącą wyrazu wolnego ( $\beta_0$ ) pomijamy.

# Analiza wariancji w modelu regresji

Podstawą analizy wariancji jest równanie

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (21)$$

Otrzymujemy z niego tzw. **tablicę analizy wariancji**.

# Tablica analizy wariancji

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średni kwadrat	Statystyka $F$
Regresja	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}$	$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{S_e^2}$
Reszta	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 2$	$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$	
Całkowita	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$		



# Testowanie istotności parametrów modelu.

Hipoteza testowana to

$$H_0 : \beta_1 = 0,$$

$$H_1 : \beta_1 \neq 0.$$

Statystyka z jaką mamy do czynienia, to statystyka  $F$ -Snedecora

$$\frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-2}}$$

z liczbą stopni swobody licznika 1 i mianownika  $n - 2$ .

Obszar krytyczny przy poziomie istotności  $\alpha$  zadaje równość

$$P(\{F_{1,n-2} \geq F_{\alpha;1,n-2}\}) = \alpha.$$

Można udowodnić, że  $F_{1,n-2} = t_{n-2}^2$ .

# Sformułowanie modelu. I

Klasyczny model regresji liniowej może być zapisany w następującej postaci macierzowej

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \quad (22)$$

## Sformułowanie modelu. II

W skróconym zapisie macierzowym mamy

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (23)$$

gdzie

$\mathbf{Y}$  jest wektorem obserwacji zmiennej losowej  $Y$  o wymiarach  $n \times 1$ ;

$\mathbf{X}$  jest macierzą obserwacji dla zmiennej niezależnej  $X$  o wymiarach  $n \times 2$ ;

$\boldsymbol{\beta}$  jest wektorem współczynników o wymiarach  $2 \times 1$ ;

$\boldsymbol{\varepsilon}$  jest wektorem składników losowych o wymiarach  $n \times 1$ .

Założenia klasycznego modelu regresji liniowej mają postać

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta} \quad (24)$$

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{I}, \quad (25)$$

gdzie zero w pierwszym równaniu jest wektorem zerowym o wymiarze  $n \times 1$ , zaś  $\mathbf{I}$  jest macierzą jednostkową stopnia  $n$ , a  $\cdot^T$  jest transponowaniem macierzy.

# Macierz kowariancji składników losowych

## Uwaga

Macierz  $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$  nazywamy macierzą kowariancji składników losowych.

Zauważmy, że dla dowolnych  $i, j \in \overline{1, n}$  mamy

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)_{ij} = \mathbb{E}(\varepsilon_i\varepsilon_j) = \text{cov}(\varepsilon_i, \varepsilon_j).$$

## Warunek nielosowości zmiennej niezależnej

Ponieważ mamy założone, że wartości zmiennej niezależnej są nielosowe (deterministyczne), więc należy ten warunek ująć w ujęciu macierzowy modelu regresji liniowej.

$\mathbf{X}$  jest macierzą o wymiarach  $n \times 2$  o ustalonych elementach. (26)

Aby ustalić wartość współczynników występujących w regresji liniowej musimy założyć, że rząd macierzy  $\mathbf{X}$  jest równy 2, co odpowiada założeniu, że w próbie są co najmniej dwie obserwacje dokonane dla różnych wartości  $x$ .

# Wyznaczanie parametrów $\beta$ . I

W ujęciu macierzowym wyrażenie podlegające minimalizacji metodą najmniejszych kwadratów jest postaci

$$S = \varepsilon^T \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta). \quad (27)$$

Różniczkując względem wektora  $\beta$  otrzymujemy

$$\frac{\partial S}{\partial \beta} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X}\beta. \quad (28)$$

Korzystając z warunku koniecznego istnienia ekstremum otrzymujemy równanie

$$\mathbf{X}^T \mathbf{X}\hat{\beta} = \mathbf{X}^T \mathbf{Y}, \quad (29)$$

## Wyznaczanie parametrów $\beta$ . II

które można zapisać w jawnej postaci macierzowej

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i Y_i \\ \sum_{i=1}^n Y_i \end{bmatrix}. \quad (30)$$

Wyznaczając z równania (29) wektor  $\widehat{\beta}$  otrzymujemy

$$\widehat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}, \quad (31)$$

gdzie macierz  $(\mathbf{X}^T \mathbf{X})^{-1}$  jest postaci

Wyznaczanie parametrów  $\beta$ . III

$$\begin{bmatrix} \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\sum_{i=1}^n x_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\sum_{i=1}^n x_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix} \cdot \quad (32)$$



## Wartości teoretyczne modelu i wektor reszt.

Na podstawie wyznaczonej z próby wektora  $\hat{\beta}$  wyznaczamy wektor  $\hat{\mathbf{Y}}$  teoretycznych wartości zmiennej losowej  $Y$  i wektor reszt  $\mathbf{e}$

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} \\ \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}}.\end{aligned}$$

Ponieważ sumę kwadratów reszt można przedstawić wzorem

$$\sum_{i=1}^n e_i^2 = \mathbf{e}^T \mathbf{e},$$

więc nieobciążony estymator wariacji jest postaci

$$S_e^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - 2}.$$

# Macierz kowariancji wektora losowego $\hat{\beta}$ . I

Macierz kowariancji wektora losowego  $\hat{\beta}$  definiujemy

$$V(\hat{\beta}) = \mathbb{E}((\hat{\beta} - \beta)^T (\hat{\beta} - \beta)) \equiv \begin{bmatrix} \mathbb{D}^2(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \mathbb{D}^2(\hat{\beta}_2) \end{bmatrix}.$$

## Stwierdzenie

W klasycznym modelu regresji liniowej macierz  $V(\hat{\beta})$  jest postaci  $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$ .

# Macierz kowariancji wektora losowego $\hat{\beta}$ . II

Na podstawie tego mamy

$$V(\hat{\beta}) = \begin{bmatrix} \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\sigma^2 \sum_{i=1}^n x_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\sigma^2 \sum_{i=1}^n x_i}{n \sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}.$$

Nieobciążonym estymatorem macierzy  $V(\hat{\beta})$  jest macierz

$$\hat{V}(\hat{\beta}) = s_e^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

# Spis treści

- 1 Klasyczny model regresji liniowej
  - Sformułowanie modelu
  - Estymacja parametrów klasycznego modelu regresji liniowej
  - Analiza wariancji w modelu regresji
  - Macierzowe ujęcie modelu regresji liniowej
- 2 Klasyczny model regresji liniowej z wieloma zmiennymi niezależnymi

## Sformułowanie zagadnienia

Rozważamy zmienną  $(k + 1)$ -wymiarową  $(Y, X_1, \dots, X_k)$ , gdzie  $X_1, \dots, X_k$  są zmiennymi niezależnymi, a  $Y$  zmienną zależną.

Do opisu stosujemy klasyczny model regresji liniowej, o ile dla każdego układu wartości  $x_1, \dots, x_k$  warunkowe rozkłady zmiennej  $Y$  mają średnie

$$\mathbb{E}(Y|x_1, \dots, x_k) = \beta_1 x_1 + \dots + \beta_k x_k + \beta_{k+1}$$

oraz wariancję

$$\mathbb{D}^2(Y|x_1, \dots, x_k) = \sigma^2.$$

Jeżeli dodatkowo warunkowe rozkłady zmiennej  $Y$  mająby rozkład normalny, to mówilibyśmy o normalnej regresji liniowej.

Próbę losową stanowiącą podstawę sformułowania i oszacowania modelu określa  $n$  łącznych obserwacji postaci

$$(Y_i, x_{i1}, \dots, x_{ik}), \quad i \in \overline{1, n}.$$

# Model

Będziemy więc rozpatrywać model zadany warunkami

$$Y_i = \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \beta_{k+1} + \varepsilon_i, \quad (33)$$

$$\mathbb{E}(\varepsilon_i) = 0 \quad (34)$$

$$\mathbb{D}^2(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \sigma^2, \quad (35)$$

$$\text{cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j) = 0 \text{ dla dowolnych } i \neq j, \quad (36)$$

gdzie  $i, j \in \overline{1, n}$ .

# Postać macierzowa. I

Klasyczny model regresji liniowej może być zapisany w następującej postaci macierzowej

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} & 1 \\ x_{21} & \dots & x_{2k} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \quad (37)$$

## Postać macierzowa. II

W skróconym zapisie macierzowym mamy

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (38)$$

gdzie

$\mathbf{Y}$  jest wektorem obserwacji zmiennej losowej  $Y$  o wymiarach  $n \times 1$ ;

$\mathbf{X}$  jest macierzą obserwacji dla zmiennej niezależnej  $X$  o wymiarach  $n \times (k + 1)$ ;

$\boldsymbol{\beta}$  jest wektorem współczynników o wymiarach  $(k + 1) \times 1$ ;

$\boldsymbol{\varepsilon}$  jest wektorem składników losowych o wymiarach  $n \times 1$ .

Założenia klasycznego modelu regresji liniowej mają postać

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta} \quad (39)$$

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 I, \quad (40)$$

gdzie zero w pierwszym równaniu jest wektorem zerowym o wymiarze  $n \times 1$ , zaś  $I$  jest macierzą jednostkową stopnia  $n$ .



## Warunek nielosowości zmiennej niezależnej.

Ponieważ mamy założone, że wartości zmiennych niezależnych są nielosowe (deterministyczne), więc należy ten warunek ująć w ujęciu macierzowy modelu regresji liniowej.

$\mathbf{X}$  jest macierzą o wymiarach  $n \times (k + 1)$  o ustalonych elementach. (41)

Aby ustalić wartość współczynników występujących w regresji liniowej musimy założyć, że rząd macierzy  $\mathbf{X}$  jest równy  $k + 1$ , co odpowiada założeniu, że w próbie są co najmniej  $k + 1$  obserwacje dokonane dla różnych wartości  $x$ .

# Parametry modelu. I

Podobnie jak w przypadku dwóch zmiennych wyrażenie podlegające minimalizacji metodą najmniejszych kwadratów jest postaci

$$S = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \quad (42)$$

Otrzymujemy

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}, \quad (43)$$

które można zapisać w jawnej postaci macierzowej

$$\begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \dots & \sum x_{i1}x_{ik} & \sum x_{i1} \\ \sum x_{i2}x_{i1} & \sum x_{i2}^2 & \dots & \sum x_{i2}x_{ik} & \sum x_{i2} \\ \vdots & \vdots & & \vdots & \vdots \\ \sum x_{i1} & \sum x_{i2} & \dots & \sum x_{ik} & n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{bmatrix} = \begin{bmatrix} \sum x_{i1} Y_i \\ \sum x_{i2} Y_i \\ \vdots \\ \sum Y_i \end{bmatrix}.$$

## Parametry modelu. II

Wyznaczając z ostatniego równania wektor  $\hat{\beta}$  otrzymujemy

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}, \quad (44)$$

Na podstawie wyznaczonej z próby wektora  $\hat{\beta}$  wyznaczamy wektor  $\hat{\mathbf{Y}}$  teoretycznych wartości zmiennej losowej  $Y$  i wektor reszt  $\mathbf{e}$

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X} \hat{\beta} \\ \mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}}. \end{aligned}$$

Nieobciążony estymator wariancji jest postaci

$$S_e^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - k - 1}.$$

## Parametry modelu. III

Macierz kowariancji wektora losowego  $\hat{\beta}$  definiujemy

$$V(\hat{\beta}) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1},$$

a jej estymator to

$$V(\hat{\beta}) = S_e^2(\mathbf{X}^T \mathbf{X})^{-1}.$$

## Współczynnik korelacji wielorakiej

Podobnie jak w przypadku dwóch zmiennych mamy współczynnik determinacji

$$r^2 \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \equiv 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}. \quad (45)$$

Natomiast dodatni pierwiastek z współczynnika determinacji nazywany jest współczynnikiem korelacji wielorakiej.

Współczynnik determinacji ma następujące własności

- $r^2 \in [0, 1]$ ,
- $r^2 = 1$  wtedy, gdy wszystkie punkty leżą w hiperpłaszczyźnie,
- $r^2 = 0$  – znajomość wartości zmiennych  $X_1, \dots, X_k$  nie dostarczają żadnych informacji na temat wartości zmiennej zależnej  $Y$ .

# Uwagi. I

Założenia i ich weryfikacja (testowanie).

- 1 Zależność między zmienną zależną, a zmiennymi niezależnymi jest liniowa.
- 2 Liczba obserwacji jest większa bądź równa liczbie parametrów wyprowadzonych z analizy regresji (współczynniki dla zmiennych niezależnych (predyktorów), wyraz wolny).
- 3 Zmienne niezależne nie są ze sobą silnie skorelowane . Brak współliniowości zmiennych niezależnych (sposób weryfikacji: test VIF).
- 4 Nie występuje autokorelacja reszt, składnika losowego (sposób weryfikacji: test Durбина-Watsona).
- 5 Brak znaczących obserwacji odstających (sposób weryfikacji: inspekcja wykresów punktowych, statystyka opisowa, odległość Cooka).
- 6 Reszty mają rozkład zbliżony do rozkładu normalnego (sposób weryfikacji: test Shapiro-Wilka, Kołmogorowa-Smirnowa, Lillieforsa).

## Uwagi. II

- 7 Wariancja reszt, składnika losowego jest taka sama dla wszystkich obserwacji homoskedastyczność (sposób weryfikacji dla dwóch prób: test Fishera  $F^2$ , sposób weryfikacji dla wielu prób: testy Barletta<sup>3</sup>, Flingera-Killeena, Levene'a<sup>4</sup>, Browna-Forsythe'a, Hartley'a<sup>5</sup>),

Jeśli wiele z założeń jest niespełnionych nie korzystaj z przedstawionych metod weryfikacji.

- Bardziej adekwatny skorygowany współczynnik determinacji (także stosowalny gdy nie ma wyrazu wolnego).

---

<sup>2</sup>Założenia: normalność.

<sup>3</sup>Założenia: normalność, równa liczebność grup.

<sup>4</sup>Założenia: niezależność prób.

<sup>5</sup>Założenia: normalność, równa liczebność grup.

# Metody doboru zmiennych do modelu.

- Zmienne wybiera się na podstawie wiedzy dziedzinowej.
- Wymagania dotyczące własności zmiennych niezależnych:
  - 1 są silnie skorelowanych ze zmienną, którą objaśniają,
  - 2 są nieskorelowane lub co najwyżej słabo skorelowane ze sobą,
  - 3 charakteryzują się dużą zmiennością.



# Regresja nieliniowa i transformacje do modelu liniowego. I

- Między zmienną objaśnianą a zmiennymi objaśniającymi mogą zachodzić związki nieliniowe.
- W wielu przypadkach można dokonać transformacji do modelu liniowego poprzez odpowiednie przekształcenia zmiennych.
- Model  $Y = f(X, b)$  jest liniowy względem parametrów, jeśli można go przedstawić jako liniową funkcję jednoznacznych przekształceń  $X$ , przy czym współczynniki tych przekształceń muszą być znane.

# Typowe modele nieliniowe i ich transformacje do modelu liniowego. I

Poza modelem regresji liniowej występują także modele regresji nieliniowej. Są nimi między innymi

- model wielomianowy (wielomian stopnia  $k$ )

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \varepsilon.$$

Podstawiając  $V_j = X^j$  dla  $j = 1, 2, \dots, k$ , model sprowadza się do modelu liniowego.

# Typowe modele nieliniowe i ich transformacje do modelu liniowego. II

- model potęgowy

$$Y = \beta_0 X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot \dots \cdot X_k^{\beta_k} e^{\varepsilon}.$$

Logarytmując otrzymujemy

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \dots + \beta_k \ln X_k + \varepsilon.$$

Podstawiając  $V_j = \ln X_j$  dla  $j = 1, 2, \dots, k$  i  $Z = \ln Y$ , model sprowadza się do modelu liniowego.

# Typowe modele nieliniowe i ich transformacje do modelu liniowego. III

- model wykładniczy

$$Y = \beta_0 \cdot \beta_1^{X_1} \cdot \beta_2^{X_2} \cdot \dots \cdot \beta_k^{X_k} \cdot e^\varepsilon.$$

Logarytmując otrzymujemy

$$\ln Y = \ln \beta_0 + X_1 \ln \beta_1 + X_2 \ln \beta_2 + \dots + X_k \ln \beta_k + \varepsilon.$$

Podstawiając  $Z = \ln Y$  i  $\tilde{\beta}_j = \ln \beta_j$  dla  $j = 1, 2, \dots, k$ , model sprowadza się do modelu liniowego.

# Regresja krokowa

- 1 Postępująca (forward).
  - Zakłada kolejne dołączanie do listy zmiennych objaśniających tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną.
- 2 Wsteczna (backward).
  - Usuwamy ze zbioru zmiennych, te które mają najmniejszy wpływ na zmienną zależną.
  - Stosując  $r^2$  lub testy istotności współczynników modelu ( $F$ ).