

# Elementarna statystyka Wnioskowanie o regresji (Inference for regression)

Alexander Bendikov

Uniwersytet Wrocławski

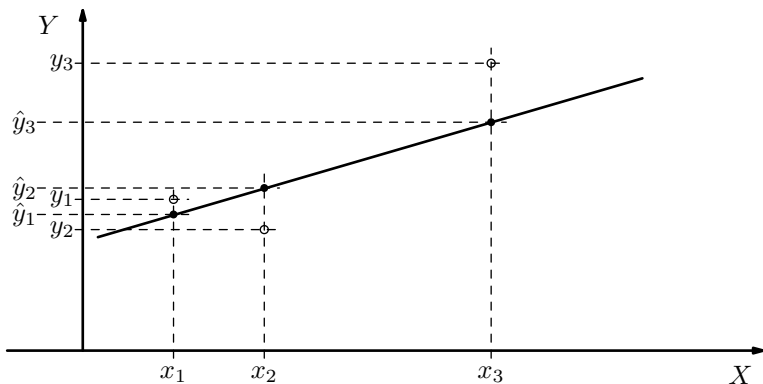
2 czerwca 2016

# Wnioskowanie o regresji

## Dane:

$Y : y_1 y_2 \dots y_n \longrightarrow \bar{y}, s_y,$

$X : x_1 x_2 \dots x_n \longrightarrow \bar{x}, s_x.$



## Równanie regresji:

$$\hat{y} = a + b x,$$

gdzie:

- $b = r \cdot s_y / s_x$ ;
- $a = \bar{y} - b \bar{x}$ ;
- $r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$ .

## Równanie regresji:

$$\hat{y} = a + b x,$$

gdzie:

- $b = r \cdot s_y / s_x$ ;
- $a = \bar{y} - b \bar{x}$ ;
- $r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$ .

**Wniosek:**  $a$ ,  $b$  i  $\hat{y}$  są zmiennymi losowymi. Tak więc linia regresji najmniejszych kwadratów jest *losowa*

## Równanie regresji:

$$\hat{y} = a + b x,$$

gdzie:

- $b = r \cdot s_y / s_x$ ;
- $a = \bar{y} - b \bar{x}$ ;
- $r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$ .

**Wniosek:**  $a$ ,  $b$  i  $\hat{y}$  są zmiennymi losowymi. Tak więc linia regresji najmniejszych kwadratów jest *losowa*

**Główne założenia:**

1) Dla dowolnego  $X$ ,  $Y \in N(\mu, \sigma)$ ;  $\mu = \mu_y$ ,

## Równanie regresji:

$$\hat{y} = a + b x,$$

gdzie:

- $b = r \cdot s_y / s_x$ ;
- $a = \bar{y} - b \bar{x}$ ;
- $r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$ .

**Wniosek:**  $a$ ,  $b$  i  $\hat{y}$  są zmiennymi losowymi. Tak więc linia regresji najmniejszych kwadratów jest *losowa*

## Główne założenia:

- 1) Dla dowolnego  $X$ ,  $Y \in N(\mu, \sigma)$ ;  $\mu = \mu_y$ ,
- 2) Średnia odpowiedź  $\mu_y$  jest związana liniowo z  $x$ :

$$\mu_y = \alpha + \beta x \quad (= \mu_{\hat{y}}),$$

gdzie  $\alpha$ ,  $\beta$  są nieznanymi parametrami.

## Równanie regresji:

$$\hat{y} = a + b x,$$

gdzie:

- $b = r \cdot s_y / s_x$ ;
- $a = \bar{y} - b \bar{x}$ ;
- $r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$ .

**Wniosek:**  $a$ ,  $b$  i  $\hat{y}$  są zmiennymi losowymi. Tak więc linia regresji najmniejszych kwadratów jest *losowa*

### Główne założenia:

- 1) Dla dowolnego  $X, Y \in N(\mu, \sigma)$ ;  $\mu = \mu_y$ ,
- 2) Średnia odpowiedź  $\mu_y$  jest związana liniowo z  $x$ :

$$\mu_y = \alpha + \beta x \quad (= \mu_{\hat{y}}),$$

gdzie  $\alpha, \beta$  są nieznanymi parametrami.

- 3) Odchylenie standardowe  $\sigma$  zmiennej  $Y$  jest takie samo dla dowolnych wartości  $x$ .

## Estymatory punktowe:

$$\hat{Y} = a + b X;$$

$$E \hat{Y} = E a + E b X;$$

$$\mu_{\hat{Y}} = \alpha + \beta X;$$



## Estymatory punktowe:

$$\hat{Y} = a + b X;$$

$$E \hat{Y} = E a + E b X;$$

$$\mu_{\hat{Y}} = \alpha + \beta X;$$

- 1)  $\alpha = E a$ ;  $a$  jest estymatorem  $\alpha$ ;
- 2)  $\beta = E b$ ;  $b$  jest estymatorem  $\beta$ ;
- 3)  $\sigma$  jest szacowana przez błąd standardowy  $s$  około linii regresji,

## Estymatory punktowe:

$$\hat{Y} = a + b X;$$

$$E \hat{Y} = E a + E b X;$$

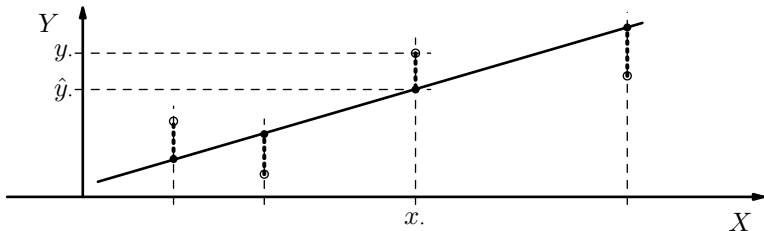
$$\mu_{\hat{Y}} = \alpha + \beta X;$$

1)  $\alpha = E a$ ;  $a$  jest estymatorem  $\alpha$ ;

2)  $\beta = E b$ ;  $b$  jest estymatorem  $\beta$ ;

3)  $\sigma$  jest szacowana przez błąd standardowy  $s$  około linii regresji,

$$s = \sqrt{\frac{1}{n-2} \sum (\hat{y}_i - y_i)^2}.$$



## Przedział ufności dla $\beta$ (dla $\alpha$ podobnie):

1)  $T$ -statystyka:

$$T = \frac{b - \beta}{SE_b},$$

gdzie

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{s_x \sqrt{n-1}}.$$

## Przedział ufności dla $\beta$ (dla $\alpha$ podobnie):

1)  $T$ -statystyka:

$$T = \frac{b - \beta}{SE_b},$$

gdzie

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{s_x \sqrt{n-1}}.$$

Zmienna  $T$  ma rozkład Studenta  $t(n-2)$  z  $df = n-2$ .

## Przedział ufności dla $\beta$ (dla $\alpha$ podobnie):

### 1) $T$ -statystyka:

$$T = \frac{b - \beta}{SE_b},$$

gdzie

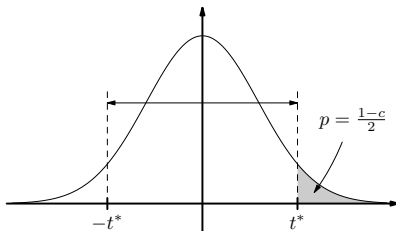
$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{s_x \sqrt{n-1}}.$$

Zmienna  $T$  ma rozkład Studenta  $t(n-2)$  z  $df = n-2$ .

### 2) Przedział ufności dla $\beta$ na poziomie $C = 0,95$ :

$$b \pm t^* SE_b,$$

gdzie, jak zwykle,  $t^*$  jest górną  $(1-c)/2$  krytyczną wartością  $t(n-2)$ .



**Przykład:** Autokomis ma na swoim parkingu 62 Mazdy, i badamy związek pomiędzy ich ceną (w dolarach) i wiekiem (w latach). Przyjmijmy, że prawdziwa linia regresji ma postać

$$\mu_{\text{cena}} = \alpha + \beta \cdot \text{wiek}.$$

**Przykład:** Autokomis ma na swoim parkingu 62 Mazdy, i badamy związek pomiędzy ich ceną (w dolarach) i wiekiem (w latach). Przyjmijmy, że prawdziwa linia regresji ma postać

$$\mu_{\text{cena}} = \alpha + \beta \cdot \text{wiek}.$$

Ten model został dopasowany do danych, używając metody najmniejszych kwadratów.

**Przykład:** Autokomis ma na swoim parkingu 62 Mazdy, i badamy związek pomiędzy ich ceną (w dolarach) i wiekiem (w latach). Przyjmijmy, że prawdziwa linia regresji ma postać

$$\mu_{\text{cena}} = \alpha + \beta \cdot \text{wiek}.$$

Ten model został dopasowany do danych, używając metody najmniejszych kwadratów. Program Minitab wygenerował następujące wyniki:

$$\overline{\text{wiek}} = 7,63$$

$$\text{odch. stand. wieku} = 4,30$$

$$\overline{\text{cena}} = 10283,1$$

$$\text{och. stand. ceny} = 7922,3$$

$$r^2 = 0,677$$

$$s = 4540,97$$

Otrzymujemy następujące współczynniki:

| Zmienna | Wartość | Odch. stand. |
|---------|---------|--------------|
| $a$     | 21833,6 | 1180,8       |
| $b$     | -1514,0 | 135,1        |



**Przykład:** Autokomis ma na swoim parkingu 62 Mazdy, i badamy związek pomiędzy ich ceną (w dolarach) i wiekiem (w latach). Przyjmijmy, że prawdziwa linia regresji ma postać

$$\mu_{\text{cena}} = \alpha + \beta \cdot \text{wiek}.$$

Ten model został dopasowany do danych, używając metody najmniejszych kwadratów. Program Minitab wygenerował następujące wyniki:

$$\overline{\text{wiek}} = 7,63$$

$$\text{odch. stand. wieku} = 4,30$$

$$\overline{\text{cena}} = 10283,1$$

$$\text{och. stand. ceny} = 7922,3$$

$$r^2 = 0,677$$

$$s = 4540,97$$

Otrzymujemy następujące współczynniki:

| Zmienna | Wartość | Odch. stand. |
|---------|---------|--------------|
| $a$     | 21833,6 | 1180,8       |
| $b$     | -1514,0 | 135,1        |

Pytanie: wyjaśnij, w tym kontekście, jakie jest znaczenie zmiennej  $\beta$ , reprezentującej nachylenie linii regresji i wyznacz przedział ufności dla  $\beta$  na poziomie  $C = 0,9$ .

## Rozwiązanie:

1. Nachylenie prawdziwej linii regresji  $\beta$  jest wartością, średnio, o którą wartość samochodu zmienia się kiedy samochód staje się o 1 rok starszy. W tym przykładzie nasze oszacowanie  $\beta$  wynosi  $b = -1514$ , a więc cena samochodu spada o \$1514 co roku.
2. Odchylenie standardowe w naszym oszacowaniu  $\beta$  wynosi  $SE_b = 135,1$ . W takim razie, przedział ufności dla  $\beta$  na poziomie  $C = 0,9$  ma postać

$$b \pm t^* SE_b,$$

gdzie  $t^*$  jest wyznaczone np z tablic dla  $df = 62 - 2 = 60$ , i wynosi  $t^* = 1,671$ , a więc otrzymujemy

$$-1514 \pm (1,671) \cdot (135,1),$$

czyli

$$-1740 \leq \beta \leq -1288.$$

## Przykład:

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

## Przykład:

|   |    |    |    |    |    |
|---|----|----|----|----|----|
| X | 38 | 56 | 59 | 64 | 74 |
| Y | 41 | 63 | 70 | 72 | 84 |

1) Linia regresji najmniejszych kwadratów ma równanie:

$$\hat{Y} = -3,6596 + 1,1969 X.$$

## Przykład:

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

1) Linia regresji najmniejszych kwadratów ma równanie:

$$\hat{Y} = -3,6596 + 1,1969 X.$$

2) Błąd standardowy

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = 0,0751.$$

## Przykład:

|   |    |    |    |    |    |
|---|----|----|----|----|----|
| X | 38 | 56 | 59 | 64 | 74 |
| Y | 41 | 63 | 70 | 72 | 84 |

1) Linia regresji najmniejszych kwadratów ma równanie:

$$\hat{Y} = -3,6596 + 1,1969 X.$$

2) Błąd standardowy

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = 0,0751.$$

3)  $df = 5 - 2 = 3$ , 2,5% krytyczna wartość  $t^*$ :

$$t^* = t^*(3) = 3,1820$$

4) 95% przedział ufności dla  $\beta$ :

$$\begin{aligned} & 1,1969 \pm 3,1820 \cdot 0,0751 = \\ & = 1,1969 \pm 0,2590 \\ & 0,9379 < \beta < 1,4559. \end{aligned}$$

4) 95% przedział ufności dla  $\beta$ :

$$\begin{aligned} & 1,1969 \pm 3,1820 \cdot 0,0751 = \\ & = 1,1969 \pm 0,2590 \\ & 0,9379 < \beta < 1,4559. \end{aligned}$$

## Testowanie hipotezy braku zależności liniowej

$H_0$  : nie ma liniowej zależności pomiędzy  $X$  i  $Y$ .



4) 95% przedział ufności dla  $\beta$ :

$$\begin{aligned} & 1,1969 \pm 3,1820 \cdot 0,0751 = \\ & = 1,1969 \pm 0,2590 \\ & 0,9379 < \beta < 1,4559. \end{aligned}$$

## Testowanie hipotezy braku zależności liniowej

$H_0$  : nie ma liniowej zależności pomiędzy  $X$  i  $Y$ .

To znaczy, że funkcja liniowa zmiennej  $X$  nie ma wartości dla przewidywania  $Y \Rightarrow$  korelacja jest 0.

4) 95% przedział ufności dla  $\beta$ :

$$\begin{aligned} & 1,1969 \pm 3,1820 \cdot 0,0751 = \\ & = 1,1969 \pm 0,2590 \\ & 0,9379 < \beta < 1,4559. \end{aligned}$$

## Testowanie hipotezy braku zależności liniowej

$H_0$  : nie ma liniowej zależności pomiędzy  $X$  i  $Y$ .

To znaczy, że funkcja liniowa zmiennej  $X$  nie ma wartości dla przewidywania  $Y \Rightarrow$  korelacja jest 0.

Skoro  $r \approx 0$  (zgodnie z  $H_0$ ),  $b = r \cdot s_Y / s_X \approx 0$  i  $\beta = E b = 0$ . W końcu nasza hipoteza ma postać:

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0 \quad (> 0, < 0).$$

4) 95% przedział ufności dla  $\beta$ :

$$\begin{aligned} & 1,1969 \pm 3,1820 \cdot 0,0751 = \\ & = 1,1969 \pm 0,2590 \\ & 0,9379 < \beta < 1,4559. \end{aligned}$$

## Testowanie hipotezy braku zależności liniowej

$H_0$  : nie ma liniowej zależności pomiędzy  $X$  i  $Y$ .

To znaczy, że funkcja liniowa zmiennej  $X$  nie ma wartości dla przewidywania  $Y \Rightarrow$  korelacja jest 0.

Skoro  $r \approx 0$  (zgodnie z  $H_0$ ),  $b = r \cdot s_Y / s_X \approx 0$  i  $\beta = E b = 0$ . W końcu nasza hipoteza ma postać:

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0 \quad (> 0, < 0).$$

Używamy  $T$ -statystyki do znalezienia  $t$ -wartości:

$$t = \frac{b}{SE_b} = \frac{b}{s / \sqrt{\sum (x_i - \bar{x})^2}}.$$

Hipotezę  $H_0$  następnie przyjmujemy bądź odrzucamy na podstawie jej relacji do wartości krytycznej  $t^*$  (wyznaczonej dla konkretnego poziomu istotności).

**Przykład:**

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

- $\hat{Y} = -3,6596 + 1,1969 X$
- $SE_b = 0,0751$
- $\mu_Y = \alpha + \beta X$ .

Hipotezę  $H_0$  następnie przyjmujemy bądź odrzucamy na podstawie jej relacji do wartości krytycznej  $t^*$  (wyznaczonej dla konkretnego poziomu istotności).

**Przykład:**

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

- $\hat{Y} = -3,6596 + 1,1969 X$
- $SE_b = 0,0751$
- $\mu_Y = \alpha + \beta X$ .

1. Ustalamy  $H_0 : \beta = 0$ ,  $H_a : \beta > 0$ ,  $\alpha = 0,05$

Hipotezę  $H_0$  następnie przyjmujemy bądź odrzucamy na podstawie jej relacji do wartości krytycznej  $t^*$  (wyznaczonej dla konkretnego poziomu istotności).

### Przykład:

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

- $\hat{Y} = -3,6596 + 1,1969 X$
- $SE_b = 0,0751$
- $\mu_Y = \alpha + \beta X$ .

1. Ustalamy  $H_0 : \beta = 0$ ,  $H_a : \beta > 0$ ,  $\alpha = 0,05$

2.  $t$ -wartość:

$$t = \frac{b}{SE_b} = \frac{1,1969}{0,0751} = 15,9374.$$

Hipotezę  $H_0$  następnie przyjmujemy bądź odrzucamy na podstawie jej relacji do wartości krytycznej  $t^*$  (wyznaczonej dla konkretnego poziomu istotności).

### Przykład:

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

- $\hat{Y} = -3,6596 + 1,1969 X$
- $SE_b = 0,0751$
- $\mu_Y = \alpha + \beta X$ .

1. Ustalamy  $H_0 : \beta = 0, H_a : \beta > 0, \alpha = 0,05$

2.  $t$ -wartość:

$$t = \frac{b}{SE_b} = \frac{1,1969}{0,0751} = 15,9374.$$

3.  $p$ -wartość:

$$df = 3, \quad t > t^* = 12,92, \quad p_* > p = 0,0005$$

Hipotezę  $H_0$  następnie przyjmujemy bądź odrzucamy na podstawie jej relacji do wartości krytycznej  $t^*$  (wyznaczonej dla konkretnego poziomu istotności).

**Przykład:**

|     |    |    |    |    |    |
|-----|----|----|----|----|----|
| $X$ | 38 | 56 | 59 | 64 | 74 |
| $Y$ | 41 | 63 | 70 | 72 | 84 |

- $\hat{Y} = -3,6596 + 1,1969 X$
- $SE_b = 0,0751$
- $\mu_Y = \alpha + \beta X$ .

1. Ustalamy  $H_0 : \beta = 0$ ,  $H_a : \beta > 0$ ,  $\alpha = 0,05$

2.  $t$ -wartość:

$$t = \frac{b}{SE_b} = \frac{1,1969}{0,0751} = 15,9374.$$

3.  $p$ -wartość:

$$df = 3, \quad t > t^* = 12,92, \quad p_* > p = 0,0005$$

4. Wniosek: dane stanowią mocny statystyczny dowód na to, że  $\beta > 0$ .



## Przedział ufności dla odpowiedzi regresji $\mu_y$ dla danego

$$x = x^*$$

Aby oszacować średnią odpowiedź  $\mu_y$  dla danego  $x = x^*$  w modelu regresji liniowej używamy przedziału wokół  $\hat{y} = a + b x$  dla  $x = x^*$ :

$$\hat{y} \pm t^* SE_\mu,$$

gdzie błąd standardowy  $SE_\mu$  jest dany przez:

$$SE_\mu = s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{s^2}{n} + (x^* - \bar{x})^2 SE_b^2}.$$

Zwykle do wyznaczenia przedziału ufności używamy programów:

Minitab

Fit

Stdev Fit

95% C.I.

4,8692

0,0855

(4,6858, 5,0526)



$\hat{y}|_{x=x^*}$



$SE_\mu$



$\hat{y} \pm t^* SE_\mu$

**Uwaga:** Dla danego  $x = x^*$  mamy  $y = \alpha + \beta x^* + \epsilon^*$ ,  $\epsilon^* \in N(0, \sigma^2)$ .

Wyjaśnienie: Używamy  $T$ -statystyki

$$T = \frac{\hat{y} - \mu_y}{SE_\mu},$$

gdzie

- $\hat{y} = a + b x^*$ ,
- $SE_\mu^2 = \text{Var}(\hat{y})$ .

$T$ -statystyka ma rozkład  $t(n - 2)$ , oraz

$$SE_\mu^2 \approx s^2 \left( \frac{1}{n} + \frac{(\bar{x} - x^*)^2}{\sum (x_i - \bar{x})^2} \right).$$

## Twierdzenie

$$SE_{\mu}^2 = \sigma^2 \left( \frac{1}{n} + \frac{(\bar{x} - x^*)^2}{\sum (x_i - \bar{x})^2} \right).$$

## Dowód.

Mamy  $y_i = \alpha + \beta x_i + \epsilon_i$ ,  $\epsilon_i$  są niezależne, jednakowo rozłożone  $N(0, \sigma^2)$ .

Mamy

$$\hat{y} = a + b x^* = \bar{y} - b \bar{x} + b x^* = \bar{y} - b(\bar{x} - x^*),$$

oraz, co więcej,

$$b = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \quad \text{oraz} \quad \bar{y} = \alpha + \beta \bar{x} + \bar{\epsilon}$$

są niezależne. W takim razie

$$\begin{aligned} \text{Var}(\hat{y}) &= \text{Var}(a + b x^*) = \text{Var}(\bar{y}) + \text{Var } b \cdot (\bar{x} - x^*)^2 \\ &= \frac{\sigma^2}{n} + (\bar{x} - x^*)^2 \sigma_b^2 = \frac{\sigma^2}{n} + (\bar{x} - x^*)^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2}. \end{aligned}$$