

STATYSTYKA MATEMATYCZNA

WYKŁAD 6

Wnioskowanie statystyczne dla korelacji i regresji.

Analiza korelacji

Założenie: zmienna losowa dwuwymiarowa (X, Y) ma rozkład normalny o współczynniku korelacji ρ .

X, Y – cechy badane równocześnie.

X_i	X_1	X_2		X_n
Y_i	Y_1	Y_2		Y_n

$$\begin{aligned}
 R &= \frac{\text{Cov}(X, Y)}{S_X \cdot S_Y} = \\
 &= \frac{\frac{1}{n} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n} \sum (X_i - \bar{X})^2} \cdot \sqrt{\frac{1}{n} \sum (Y_i - \bar{Y})^2}}
 \end{aligned}$$

Uwaga:

$$\text{a) } \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y}$$

$$\text{b) } \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

$$R = \frac{\frac{1}{n} \sum X_i Y_i - \bar{X} \bar{Y}}{\sqrt{\frac{1}{n} \sum X_i^2 - (\bar{X})^2} \cdot \sqrt{\frac{1}{n} \sum Y_i^2 - (\bar{Y})^2}}$$

Estymator R jest estymatorem zgodnym i asymptotycznie nieobciążonym współczynnika korelacji ρ .

Estymator ten ma asymptotycznie rozkład normalny

$$N\left(\rho, \frac{1-\rho^2}{\sqrt{n}}\right), \text{ (próba musi być bardzo liczna)}.$$

Uwaga.

$$U_n = \frac{1}{2} \ln \frac{1+R}{1-R} \approx N\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{\sqrt{n-3}}\right), n \geq 20).$$

Jeśli $\rho = 0$, to $U_n = \frac{R}{\sqrt{1-R^2}} \sqrt{n-2}$ ma rozkład T_{n-2} .

Zatem **przedział ufności dla współczynnika korelacji ρ** (próba bardzo liczna) ma postać:

$$\left\langle R - u_{\alpha} \frac{1 - R^2}{\sqrt{n}}; R + u_{\alpha} \frac{1 - R^2}{\sqrt{n}} \right\rangle$$

R – współczynnik korelacji z próby,

n – liczebność próby,

$1 - \alpha$ – poziom ufności,

u_{α} odczytujemy w tablicy z zależności $\Phi(u_{\alpha}) = 1 - \alpha/2$

Dla małych prób ($n \geq 20$) **przedział ufności dla współczynnika korelacji ρ** ma postać:

$$\left\langle \frac{e^{2A} - 1}{e^{2A} + 1} ; \frac{e^{2B} - 1}{e^{2B} + 1} \right\rangle ,$$

gdzie

$$A = \frac{1}{2} \ln \frac{1+R}{1-R} - \frac{u_\alpha}{\sqrt{n-3}}, \quad B = \frac{1}{2} \ln \frac{1+R}{1-R} + \frac{u_\alpha}{\sqrt{n-3}}$$

u_α odczytujemy w tablicy z zależności $\Phi(u_\alpha) = 1 - \alpha/2$

Weryfikacja hipotez dotyczących współczynnika korelacji.

Zakładamy, że cecha (X,Y) ma rozkład normalny, $n \geq 20$.

Wysuwamy dwie hipotezy:

$H_0(\rho = \rho_0)$, H_1 - jedną z trzech poniższych hipotez.

Rozpatrujemy statystykę i zbiór krytyczny wg tabeli:

H_1	Statystyka U	Zbiór krytyczny	Odczyt k
$\rho \neq \rho_0$	$\left(\frac{1}{2} \ln \frac{1+R}{1-R} - \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} \right) \sqrt{n-3}$	$K = (-\infty; -k > \cup < k; +\infty)$	$\Phi(k) = 1 - \alpha / 2$
$\rho > \rho_0$		$K = < k; +\infty)$	$\Phi(k) = 1 - \alpha$
$\rho < \rho_0$		$K = (-\infty; -k >$	$\Phi(k) = 1 - \alpha$

Decyzje:

Jeśli $U_n \in K$ to H_0 odrzucamy ,

Jeśli $U_n \notin K$ to nie ma podstaw do odrzucenia H_0 .

Przykład.

Badano zależność między wydatkami na reklamę a wysokością sprzedaży.

Dla próby 40 elementowej otrzymano $r = 0,812$.

Sprawdzimy na poziomie istotności 0,05 hipotezy:

$$H_0(\rho = 0,85) \text{ i } H_0(\rho \neq 0,85).$$

Rozwiązanie.

Wartość statystyki wynosi

$$u = -0,816.$$

Zbiór krytyczny ma postać

$$K = (-\infty; -1,96] \cup [1,96; +\infty)$$

Ponieważ $u \notin K$ to nie ma podstaw do odrzucenia H_0 .

B adanie istotności współczynnika korelacji

Wysuwamy dwie hipotezy:

$H_0(\rho = 0)$ (brak korelacji pomiędzy zmiennymi),

H_1 – jedną z trzech poniższych hipotez.

Rozpatrujemy statystykę i zbiór krytyczny wg tabeli:

H_1	Statystyka U	Zbiór krytyczny	Odczyt k
$\rho \neq 0$	$\frac{R}{\sqrt{1-R^2}} \sqrt{n-2}$	$K = (-\infty; -k > \cup < k; +\infty)$	$P(T_{n-2} > k) = \alpha$
$\rho > 0$		$K = [k; +\infty)$	$P(T_{n-2} > k) = \alpha$
$\rho < 0$		$K = (-\infty; -k]$	$P(T_{n-2} < -k) = \alpha$

Decyzje:

Jeśli $u_n \in K$ to H_0 odrzucamy,

Jeśli $u_n \notin K$ to nie ma podstaw do odrzucenia H_0 .

Wnioskowanie dla współczynników regresji

Niech $y = \beta_0 + \beta_1 x$ będzie prostą regresji,
a

$\hat{y} = b_0 + b_1 x$ prostą regresji wyznaczoną na podstawie próby.

Współczynniki b_0 , b_1 prostej regresji wyznaczamy stosując **metodę najmniejszych kwadratów (MNK)**.

Chcemy wyznaczyć minimum funkcji:

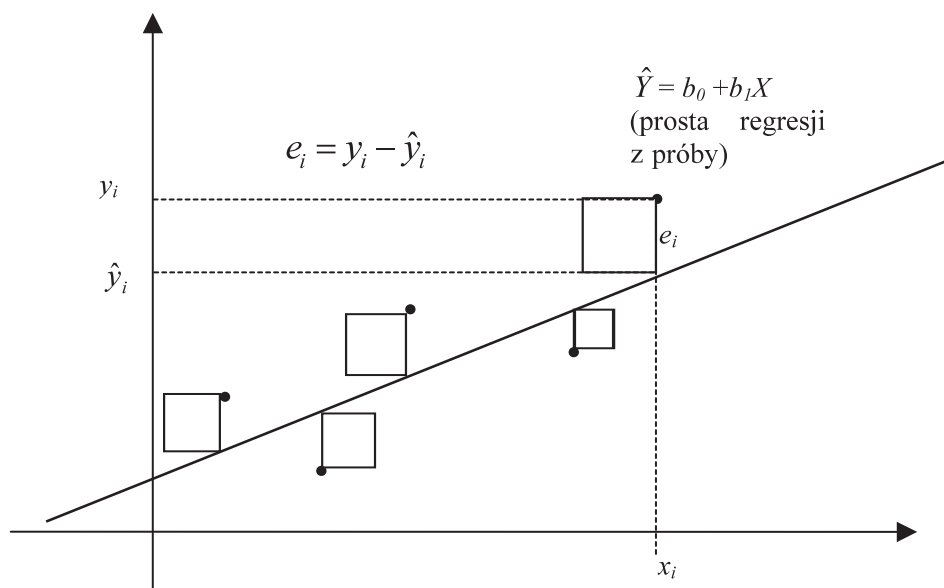
$$S(b_0, b_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 =$$

$$(*) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

gdzie $e_i = y_i - \hat{y}_i$ nazywamy **resztami** modelu regresji.

Uwaga. $\sum_{i=1}^n e_i = 0$

MNK:



Należy wyznaczyć prostą regresji tak aby suma pól kwadratów była minimalna.

**Obliczając pochodne cząstkowe funkcji (*)
i przyrównując je do zera otrzymujemy układ
równań normalnych:**

$$\begin{aligned}\frac{\partial S}{\partial b_0} &= 2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i)(-1) = \\ &= -2 \left(\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i - n b_0 \right) = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial S}{\partial b_1} &= 2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i)(-x_i) = \\ &= -2 \left(\sum_{i=1}^n y_i x_i - b_1 \sum_{i=1}^n x_i^2 - b_0 \sum_{i=1}^n x_i \right) = 0\end{aligned}$$

Rozwiązując otrzymany układ równań otrzymamy wzory na przybliżone wartości współczynników

$$\begin{aligned} b_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \\ &= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n (\bar{x})^2} = \frac{s_Y}{s_X} r = \frac{\text{cov}(X, Y)}{s_X^2} \end{aligned}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Wariancja resztowa:

Wariancja resztowa to średnia z pól kwadratów zbudowanych na resztach odzwierciedlająca stopień dopasowania prostej regresji do danych statystycznych.

Niech $e_i = y_i - \hat{y}_i$, gdzie $\hat{y}_i = b_0 + b_1 x_i$ wtedy

$$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} \text{ czyli}$$

$$s_e^2 = \frac{\sum_{i=1}^n y_i^2 - b_0 \sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i y_i}{n-2}$$

$$s_e^2 = \frac{ns_Y^2(1-r^2)}{n-2}$$

$s_e = \sqrt{s_e^2}$ oznacza średnie (standardowe) odchylenie od prostej regresji.

Współczynnik determinacji

$$\begin{aligned} R^2 &= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = \\ &= \frac{b_0 \sum y_i + b_1 \sum x_i y_i - n\bar{y}^2}{\sum y_i^2 - n(\bar{y})^2} = \\ &= \frac{b_1 (\sum x_i y_i - n\bar{x}\bar{y})}{\sum y_i^2 - n(\bar{y})^2} = \frac{\text{cov}^2(X, Y)}{S_X^2 S_Y^2} = r^2 \end{aligned}$$

Uwaga: $R^2 \in \langle 0, 1 \rangle$

Standardowe błędy oszacowania współczynników prostej regresji.

$$s(b_1) = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$s(b_0) = \frac{s_e \sqrt{\sum x_i^2}}{\sqrt{n \sum (x_i - \bar{x})^2}} = s(b_1) \cdot \sqrt{\frac{1}{n} \sum x_i^2} = s(b_1) \cdot \sqrt{s_X^2 + (\bar{x})^2}$$

Stosujemy niekiedy zapis

$$\hat{Y} = \underset{(\pm s(b_0))}{b_0} + \underset{(\pm s(b_1))}{b_1} X$$

lub

$$\hat{Y} = \underset{(\pm s(b_0))}{b_0} + \underset{(\pm s(b_1))}{b_1} X \quad (\pm s_e)$$

Przedziały ufności dla β_i , $i = 0, 1$;

dla poziomu ufności $1 - \alpha$ mamy:

$$\beta_i \in \langle b_i - u_\alpha S(b_i); b_i + u_\alpha S(b_i) \rangle$$

gdzie u_α odczytujemy z tablicy rozkładu Studenta:

$$P(|T_{n-2}| > u_\alpha) = \alpha.$$

$S(b_i)$ – standardowe błędy współczynników prostej regresji.

Weryfikacja hipotez dla β_i , $i = 0, 1$;

dla poziomu istotności α rozpatrujemy test dla poszczególnych parametrów β_i , $i = 0, 1$.

Wysuwamy dwie hipotezy:

$H_0(\beta_i = \beta_i^0)$, H_1 – jedną z trzech poniższych hipotez.

Rozpatrujemy statystykę i zbiór krytyczny wg tabeli:

H_1	Statystyka	Zbiór krytyczny	Odczyt k
$\beta_i \neq \beta_i^0$	$U_n = \frac{b_i - \beta_i^0}{S(b_i)}$	$K = (-\infty; -k > \cup < k; +\infty)$	$P(T_{n-2} > k) = \alpha$
$\beta_i > \beta_i^0$		$K = < k; +\infty)$	$P(T_{n-2} > k) = 2\alpha$
$\beta_i < \beta_i^0$		$K = (-\infty; -k >$	$P(T_{n-2} < -k) = 2\alpha$

Decyzje:

Jeśli $u_n \in K$ to H_0 odrzucamy ,

Jeśli $u_n \notin K$ to nie ma podstaw do odrzucenia H_0 .

Uwaga

Jeśli badamy istotność parametru β_i to przyjmujemy

$$\beta_i^0 = 0 \text{ tzn. } H_0(\beta_i = 0)$$

W modelach regresji pożądane jest odrzucenie hipotezy $H_0(\beta_1 = 0)$.

Przykład

Z populacji dla której (X, Y) ma rozkład normalny pobrano próbę 100 elementową i obliczono, że prosta regresji z próby ma postać:

$$Y = 1,35 - \frac{0,4}{(0,5)} X$$

Na poziomie istotności 0,05 sprawdź istotność współczynnika β_1 .

Rozpatrujemy hipotezy $H_0(\beta_1 = 0)$ $H_1(\beta_1 < 0)$

Wartość statystyki wynosi $u = -4.$,

Zbiór krytyczny ma postać $K = (-\infty; -1,66 > .$

Ponieważ obliczona wartość statystyki należy do zbioru krytycznego to odrzucamy hipotezę $H_0(\beta_1 = 0)$, zatem wynik testu nie podważa istotności współczynnika β_1 .

Prognoza punktowa to przewidywana wartość zmiennej Y odpowiadająca wartości x_τ zmiennej X .

$$y^*(\tau) = b_0 + b_1 x_\tau$$

Standardowy błąd prognozy

$$\begin{aligned} S_{\tau} &= S_e \sqrt{1 + \frac{1}{n} + \frac{(x_{\tau} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \\ &= S_e \sqrt{1 + \frac{\sum_{i=1}^n x_i^2 + nx_{\tau}^2 - 2x_{\tau} \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}} \end{aligned}$$

gdzie S_e – odchylenie resztowe

błąd względny prognozy punktowej:

$$\delta_{\tau} = \frac{S_{\tau}}{|y_{\tau}^*|} 100\%$$

Prognoza przedziałowa dla poz. ufności $1 - \alpha$

$$\left\langle y^*(\tau) - u_\alpha S_\tau; y^*(\tau) + u_\alpha S_\tau \right\rangle$$

u_α odczytujemy z tablicy rozkładu Studenta:

$$P(|T_{n-2}| > u_\alpha) = \alpha.$$

Badanie liniowości funkcji regresji - test serii.

Badamy populację ze względu na dwie cechy X i Y .

Losujemy próbę n - elementową (x_i, y_i) .

Na podstawie tej próby chcemy sprawdzić czy funkcja regresji jest liniowa.

Prostym testem do weryfikacji tej własności jest test serii.

Niech $y = b_0 + b_1x$ będzie prostą regresji wyznaczoną na podstawie próby metodą najmniejszych kwadratów.

Elementom próby przypisujemy symbol a lub b :

a - gdy $y_i > b_0 + b_1x_i$, (punkt leży nad prostą)

b - gdy $y_i < b_0 + b_1x_i$ (punkt leży pod prostą)

(elementów dla których zachodzi równość nie rozpatrujemy).

Serie to podciągi złożone z jednakowych symboli.

Rozpatrujemy hipotezy

H_0 (funkcja regresji jest liniowa),

H_1 (funkcja regresji nie jest liniowa),

Stosujemy statystykę:

U_n = liczba serii

Zbiór krytyczny:

$$K = (0; k>$$

gdzie k odczytujemy z tablicy dla poziomu istotności α i liczb n_1 oraz n_2 ,

gdzie n_1 - liczba symboli a , n_2 - liczba symboli b ,

Decyzje:

Jeśli $U_n \in K$ to H_0 odrzucamy,

Jeśli $U_n \notin K$ to nie ma podstaw do odrzucenia H_0 .

Przykład.

Badano zależność między wynikami testów z dwóch przedmiotów przeprowadzonymi w pewnej uczelni. Dla próby 12 wylosowanych studentów otrzymano następujące liczby punktów:

X	16	20	22	24	33	47	55	70	77	82	90	94
Y	25	34	60	83	92	104	110	124	133	150	145	170

Prosta regresji z próby ma postać $Y = 2x + 5$.

Sprawdzimy na poziomie istotności 0,05 hipotezę że zależność między tymi cechami jest liniowa.

Rozwiązanie.

Otrzymamy następujący ciąg symboli:

b, b, a, a, a, a, b, b, b, b, b, b

Liczba serii wynosi $u = 3$

Z tablic rozkładu serii odczytujemy

$$K = (0; 3>$$

Ponieważ $u \in K$ to odrzucamy hipotezę H_0 , zatem możemy sądzić, że zależność między wynikami testów nie jest liniowa.