

REGRESJA (jedna zmienna)

$\hat{Y} = b_0 + b_1 X$ prosta regresji zmiennej Y względem X .

przybliżone wartości parametrów strukturalnych

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} =$$

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

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

Uwaga:

- a) $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y}$
- b) $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$
- c) $\text{cov}(X, Y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}$ (kowariancja)
- d) $r = \frac{\text{cov}(X, Y)}{s_X s_Y}$ (współczynnik korelacji)

Wariancja resztowa.

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

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

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

Standardowe błędy oszacowania współczynników prostej regresji.

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

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

Własność: $\frac{s^2(b_0)}{s^2(b_1)} = \frac{\sum x_i^2}{n}$

Współczynnik determinacji

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

(określa jaką część całkowitej zmienności cechy Y wyjaśnia model regresji liniowej)

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = \frac{b_0 \sum y_i + b_1 \sum x_i y_i - n \bar{y}^2}{\sum y_i^2 - n(\bar{y})^2} =$$

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

Wnioskowanie dla współczynników regresji

Niech $y = \beta_0 + \beta_1 x$ będzie prostą regresji, a $y = b_0 + b_1 x$ prostą regresji wyznaczoną na podstawie próby.

Przedziały ufności dla $\beta_i, i = 0, 1$;

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 $\beta_i, i = 0, 1$;

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

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 $\beta_i^0 = 0$ tzn. rozpatrujemy hipotezę $H_0(\beta_i = 0)$

W modelach regresji pożądane jest odrzucenie hipotezy $H_0(\beta_i = 0)$, w przeciwnym przypadku mielibyśmy do czynienia z sytuacją gdy zmienna X nie ma wpływu na zmienną Y .

Badanie losowości reszt – test serii

Resztom przypisujemy symbol a lub b :

a – gdy $e_i > 0$

b – gdy $e_i < 0$

(elementów $e_i = 0$ nie rozpatrujemy).

Serie to podciągi złożone z jednakowych symboli.

Rozpatrujemy hipotezy

H_0 (reszty modelu mają charakter losowy),

H_1 (reszty modelu nie mają charakteru losowego),

Stosujemy statystykę:

$$U_n = \text{liczba serii}$$

Zbiór krytyczny:

$$K = (0; k >$$

gdzie k odczytujemy z tablicy rozkładu serii dla poziomu istotności α i liczb n_1 oraz n_2 ,
gdzie n_1 – liczba symboli a , n_2 – liczba symboli b ,

Tablica rozkładu serii
 Tablica dla $\alpha = 0,05$: (tablica jest symetryczna)

$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4			2																
5		2	2	3															
6		2	3	3	3														
7		2	3	3	4	4													
8	2	2	3	3	4	4	5												
9	2	2	3	4	4	5	5	6											
10	2	3	3	4	5	5	6	6	6										
11	2	3	3	4	5	5	6	6	7	7									
12	2	3	4	4	5	6	6	7	7	8	8								
13	2	3	4	4	5	6	6	7	8	8	9	9							
14	2	3	4	5	5	6	7	7	8	8	9	9	10						
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11					
16	2	3	4	5	6	6	7	8	8	9	10	10	11	11	11				
17	2	3	4	5	6	7	7	8	9	9	10	10	11	11	12	12			
18	2	3	4	5	6	7	8	8	9	10	10	11	11	12	12	13	13		
19	2	3	4	5	6	7	8	8	9	10	10	11	12	12	13	13	14	14	
20	2	3	4	5	6	7	8	9	9	10	11	11	12	12	13	13	14	14	15

Decyzje:

Jeśli $u_n \in K$ to H_0 odrzucamy,

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

Badanie symetrii składnika losowego

Niech

n – liczba obserwacji,

m – liczba reszt dodatnich.

Wysuwamy dwie hipotezy:

$$H_0\left(\frac{m}{n} = \frac{1}{2}\right) \qquad H_1\left(\frac{m}{n} \neq \frac{1}{2}\right)$$

Stosujemy statystykę

$$U_n = \frac{\frac{m}{n} - \frac{1}{2}}{\sqrt{\frac{\frac{m}{n}\left(1 - \frac{m}{n}\right)}{n-1}}}$$

Rozpatrujemy zbiór krytyczny:

$$K = (-\infty; -k) \cup (k; +\infty)$$

gdzie k odczytujemy dla poziomu istotności α z tablicy rozkładu Studenta:

$$P(|T_{n-1}| > k) = \alpha.$$

Decyzje:

Jeśli $u_n \in K$ to H_0 odrzucamy,

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

Badanie normalności rozkładu reszt. Test normalności (test Shapiro-Wilka)

Wysuwamy dwie hipotezy:

H_0 – reszty mają rozkład normalny,

H_1 – reszty nie mają rozkładu normalnego.

Reszty porządkujemy niemalejąco: $e_{(1)}, e_{(2)}, \dots, e_{(n)}$

Stosujemy statystykę

$$U_n = \frac{\left[\sum_{i=1}^{\lfloor n/2 \rfloor} a_{n,i} (e_{(n-i+1)} - e_{(i)}) \right]^2}{\sum_{i=1}^n (e_i - \bar{e})^2}$$

gdzie $\lfloor n/2 \rfloor$ jest częścią całkowitą liczby $n/2$,

$\bar{e} = 0$ dla modeli liniowych.

$a_{n,i}$ – współczynniki Shapiro-Wilka odczytane z tablicy:

$n \backslash i$	1	2	3	4	5	6	7	8	9	10
4	0,6872	0,1677	—	—	—	—	—	—	—	—
6	0,6431	0,2806	0,0875	—	—	—	—	—	—	—
8	0,6052	0,3164	0,1743	0,0561	—	—	—	—	—	—
10	0,5739	0,3291	0,2141	0,1224	0,0399	—	—	—	—	—
12	0,5475	0,3325	0,2347	0,1586	0,0922	0,0303	—	—	—	—
14	0,5251	0,3318	0,2460	0,1802	0,1240	0,0727	0,0240	—	—	—
15	0,5150	0,3306	0,2495	0,1878	0,1353	0,0880	0,0433	0	—	—
16	0,5056	0,3290	0,2521	0,1939	0,1447	0,1005	0,0593	0,0196	—	—
18	0,4886	0,3253	0,2553	0,2027	0,1587	0,1197	0,0837	0,0496	0,0163	—
20	0,4734	0,3211	0,2565	0,2085	0,1686	0,1334	0,1013	0,0711	0,0422	0,0140

Rozpatrujemy zbiór krytyczny: $K = \{k < 0; k > \}$

gdzie k odczytujemy dla poziomu istotności α i danego n z tablicy testu Shapiro-Wilka:

(tablica testu Shapiro-Wilka dla $\alpha = 0,05$)

n	4	6	8	10	12	14	15	16	18	20
k	0,767	0,788	0,818	0,842	0,859	0,874	0,881	0,887	0,897	0,905

Decyzje:

Jeśli $u_n \in K$ to H_0 odrzucamy.

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

Badanie jednorodności wariancji składnika losowego

Jednorodność wariancji składnika losowego jest jednym z założeń klasycznej metody najmniejszych kwadratów.

Niespełnienie tego założenia obniża efektywność estymatorów parametrów strukturalnych (nie wpływa na zgodność i nieobciążoność).

Zastosujemy **test Goldfelda-Quandta**.

W teście tym dzielimy próbę na dwie równoliczne podpróby o liczebnościach

$n_1 = n_2$ (gdy liczba obserwacji jest nieparzysta – środkowa lub środkowe obserwacje nie biorą udziału w dalszych obliczeniach). Na podstawie tych podprób szacujemy parametry strukturalne modelu i obliczamy wariancje

resztowe $S_{e_1}^2$, $S_{e_2}^2$. Próby numerujemy tak aby $S_{e_2}^2 \geq S_{e_1}^2$.

Wysuwamy dwie hipotezy:

$$H_0(\sigma_1^2 = \sigma_2^2) \quad H_1(\sigma_2^2 > \sigma_1^2)$$

Stosujemy statystykę

$$U_n = \frac{S_{e2}^2}{S_{e1}^2}$$

Rozpatrujemy zbiór krytyczny:

$$K = < k; +\infty)$$

gdzie k odczytujemy dla poziomu istotności α z tablicy rozkładu F-Snedecora dla $(n_2 - (k + 1), n_1 - (k + 1))$ stopni swobody.

Decyzje:

Jeśli $u_n \in K$ to H_0 odrzucamy ,

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

Badanie autokorelacji reszt (test Durбина-Watsona)

Rozpatrujemy hipotezę: H_0 (reszty nie są skorelowane) tzn $H_0(\rho = 0)$

Obliczamy wartość statystyki

$$U_n = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

Uwaga

$$u_n \in < 0; 4 >$$

Dla reszt nieskorelowanych $u_n \approx 2$

Z tablicy rozkładu D-W odczytuje się dla ustalonego α, n dwie liczby k_L i k_U .

Tablica rozkładu D-W dla $\alpha = 0,05$:

n	k_L	k_U
6	0,610	1,400
7	0,700	1,356
8	0,730	1,332
9	0,824	1,320
10	0,879	1,320
11	0,927	1,324
12	0,971	1,331
13	1,010	1,340
14	1,045	1,350
15	1,077	1,361

Jeśli $u_n < 2$ to rozpatrujemy hipotezę alternatywną:

H_1 (reszty są skorelowane dodatnio) tzn $H_1(\rho > 0)$.

Przyjmuje się następującą regułę decyzyjną:

Jeśli $u_n < k_L$ to H_0 odrzucamy.

Jeśli $u_n > k_U$ to nie ma podstaw do odrzucenia H_0 .

Jeśli $k_L \leq u_n \leq k_U$ to nie podejmujemy decyzji.

Jeśli $U_n > 2$ to rozpatrujemy hipotezę alternatywną:

H_1 (reszty są skorelowane ujemnie) tzn $H_1(\rho < 0)$.

Przyjmuje się następującą regułę decyzyjną:

Jeśli $u_n > 4 - k_L$ to H_0 odrzucamy.

Jeśli $u_n < 4 - k_U$ to nie ma podstaw do odrzucenia H_0 .

Jeśli $4 - k_U \leq u_n \leq 4 - k_L$ to nie podejmujemy decyzji.

Prognoza.

Prognoza punktowa.

Niech x_τ – przewidywana wartość cechy X w okresie prognozy.

Prognoza punktowa y_τ^* to przewidywana wartość cechy Y odpowiadająca wartości x_τ cechy X .

$$y_\tau^* = b_0 + b_1 x_\tau$$

Standardowy błąd prognozy

$$s_\tau = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_\tau - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = s_e \sqrt{1 + \frac{\sum_{i=1}^n x_i^2 + n x_\tau^2 - 2 x_\tau \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}}$$

Zatem należy traktować wartość prognozy jako $y_\tau^* \pm s_\tau$

Jakość prognozy punktowej możemy ocenić względnym błędem prognozy punktowej

$$\delta_{punkt} = \frac{s_\tau}{y_\tau^*} \cdot 100\%$$

Prognoza przedziałowa.

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

$$\left\langle y_\tau^* - u_\alpha s_\tau; y_\tau^* + u_\alpha s_\tau \right\rangle$$

gdzie u_α odczytujemy z tablicy rozkładu Studenta:

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

Jakość prognozy przedziałowej możemy ocenić względnym błędem prognozy przedziałowej

$$\delta_{prz} = \frac{u_\alpha s_\tau}{y_\tau^*} \cdot 100\%$$