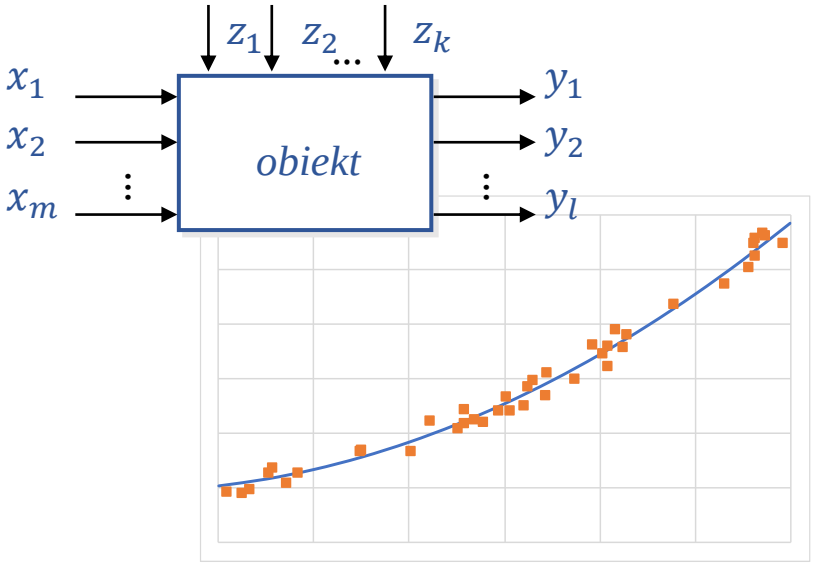
Planowanie doświadczeń

Analiza regresji



Materiały

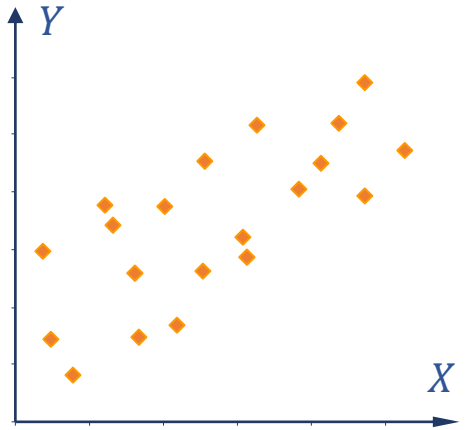
http://pracownicy.uz.zgora.pl/ipajak/

Statystyka – dwuwymiarowe zmienne losowe

Dwuwymiarową zmienną losową jest nazywany układ dwóch zmiennych losowych (X, Y) jeśli znane są prawdopodobieństwa wystąpienia zdarzenia (x, y) lub prawdopodobieństwa, że (x, y) przyjmie wartość z określonego dwuwymiarowego przedziału (tzn. z określonego prostokąta).

Graficzna prezentacja danych

Obserwacje pochodzące z obserwacji rozkładu dwuwymiarowego przedstawiane są na tzw. *wykresach rozrzutu*. Wykres rozrzutu to wykres punktowy, poszczególne punkty wykresu odpowiadają wartościom uzyskanym z kolejnych obserwacji zmiennych X i Y.



Miary zależności

Dla dwuwymiarowych zmiennych losowych, oprócz miar wprowadzonych dla zmiennych jednowymiarowych, definiowane są miary związane z istnieniem (lub nieistnieniem) zależności pomiędzy zmiennymi tworzącymi ten rozkład: *kowariancja* i *współczynnik korelacji*.

Dla zmiennych losowych X i Y

$$E(X+Y) = E(X) + E(Y)$$

Dodatkowo, jeśli zmienne są niezależne spełniona jest zależność:

$$E(XY) = E(X)E(Y)$$

Różnica $E(XY) - E(X)E(Y)$ jest miarą zależności zmiennych jest to tzw. *kowariancja*

$$\sigma_{XY} = E(XY) - E(X)E(Y)$$

jest ona również definiowana w równoważny sposób jako

$$\sigma_{XY} = E\big(X - E(X)\big)\,E\big(Y - E(Y)\big)$$

Własności

jeżeli zmienne X i Y są niezależne to $\sigma_{XY} = 0$,

jeżeli wartości X większe od średniej pojawiają się najczęściej z wartościami Y większymi od średniej to $\sigma_{XY} > 0$,

jeżeli wartości X większe od średniej pojawiają się najczęściej z wartościami Y mniejszymi od średniej to $\sigma_{XY} < 0$.

Współczynnik korelacji liniowej Paersona to standaryzowana kowariancja, miara została zdefiniowana w celu uniezależnienia *kowariancji* od wariancji zmiennych X i Y

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Własności

- przyjmuje wartości

$$-1 \leq \rho_{XY} \leq 1$$

- opisuje siłę związku zmiennych

Wartości $\lvert\rho_{XY}\rvert$		Siła związku (korelacja)
od	do	
0	0,2	brak
0,2	0,4	słaba
0,4	0,7	średnia
0,7	0,9	silna
0,9	1	bardzo silna

σ_X, σ_Y to odchylenia standardowe zmiennych X i Y

Kowariancja i *korelacja* mogą być szacowane na podstawie próby

kowariancja σ_{XY} jest szacowana jako:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left((x_i - \overline{x})(y_i - \overline{y})\right) \quad \text{lub} \quad s_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left((x_i - \overline{x})(y_i - \overline{y})\right)$$

korelację ρ_{XY} przybliża się z wzoru:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_X, s_Y to oszacowane z próby odchylenia standardowe σ_X i σ_Y,
do szacowania σ_X wykorzystywane są zależności:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{lub} \quad s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Statystyka – wielowymiarowe zmienne losowe

Rozkład prawdopodobieństwa *wielowymiarowej zmiennej losowej* $(X_1, X_2, \cdots, X_k)$ jest opisany w postaci łącznego rozkładu zmiennych losowych X_1, X_2, $\cdots$, X_k.

Zależności pomiędzy zmiennymi tworzącymi ten rozkład opisuje się *macierzami kowariancji i korelacji*, które są uogólnionymi na k wymiarów kowariancją i korelacją:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix}$$

$$\rho = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\sigma_2} & \cdots & \frac{\sigma_{1k}}{\sigma_1\sigma_k} \\ \frac{\sigma_{21}}{\sigma_2\sigma_1} & 1 & \cdots & \frac{\sigma_{2k}}{\sigma_2\sigma_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{k1}}{\sigma_k\sigma_1} & \frac{\sigma_{k2}}{\sigma_k\sigma_2} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{bmatrix}$$

σ_i^2 to wariancja zmiennej X_i

σ_{ij} to kowariancja zmiennych X_i i X_j, macierz kowariancji jest symetryczna ponieważ: $\sigma_{ij} = \sigma_{ji}$

Statystyka – wielowymiarowy rozkład normalny

Wielowymiarowy rozkład normalny jest jednym z najważniejszych rozkładów wielowymiarowych zmiennych losowych, jest on uogólnieniem rozkładu normalnego $\mathcal{N}(\mu, \sigma)$ o funkcji gęstości:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Wielowymiarowy rozkład normalny k–wymiarowej zmiennej losowej $(X_1, X_2, \cdots, X_k)$ jest zapisywany jako $\mathcal{N}(\mu, \Sigma)$ a funkcja gęstości tego rozkładu dana jest wzorem:

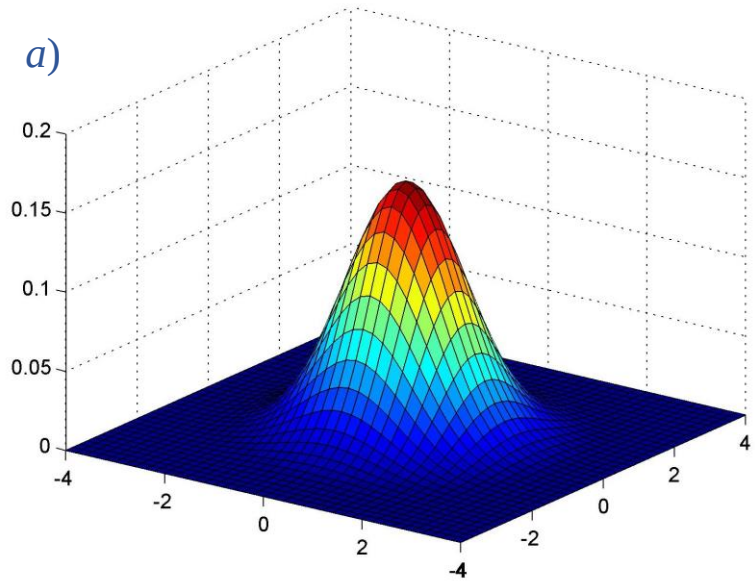
$$f(\mathrm{x}) = \frac{1}{\sqrt{(2\pi)^k}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\mathrm{x}-\mu)^T \Sigma^{-1}(\mathrm{x}-\mu)}$$

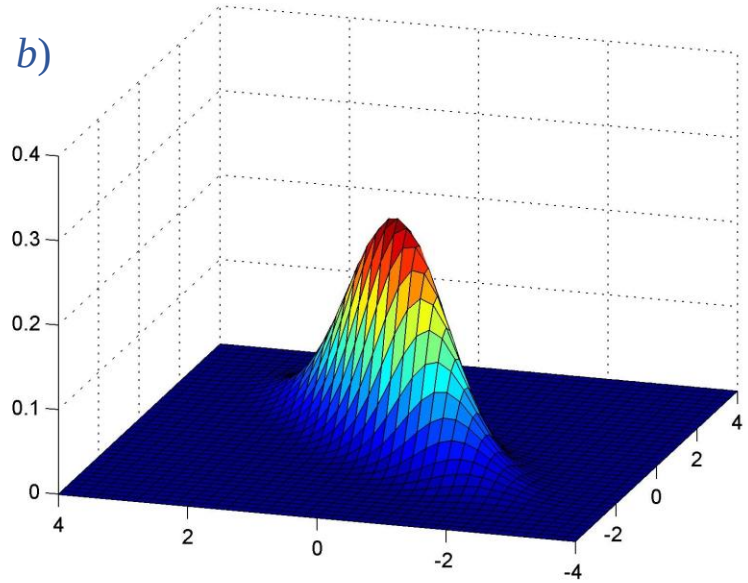
μ to wektor średnich $\mu = [\mu_1, \mu_2, \cdots \mu_k]^T$, Σ to $(k \times k)$ wymiarowa macierz kowariancji, parametry μ i Σ mogą być szacowane z próby, oznaczane są jako $\bar{\mathrm{x}}$ i S:

$$\bar{\mathrm{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i\,, \qquad \mathrm{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathrm{x}_i - \bar{\mathrm{x}})(\mathrm{x}_i - \bar{\mathrm{x}})^T$$

gdzie: n – rozmiar próby, tzn. ilość obserwacji, $\mathrm{x}_i = [x_{i1}, \ldots, x_{ik}]$ to i-ta obserwacja

Statystyka – dwuwymiarowy rozkład normalny





Dwuwymiarowy rozkład normalny o średnich $\mu = [0 \quad 0]^T$ *i*

macierzy kowariancji: a) $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ b) $\Sigma = \begin{bmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}$

Statystyka – dwuwymiarowe zmienne losowe

W tabeli zebrane zostały wartości cech X i Y wylosowane z dwuwymiarowego rozkładu $N(\mu, \Sigma)$ o parametrach:

$$\mu = \begin{bmatrix} 100 \\ 50 \end{bmatrix} \quad \text{i} \quad \Sigma = \begin{bmatrix} 2 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}.$$

Wyznaczyć macierze kowariancji i korelacji, narysować wykres rozrzutu.

$$\overline{x} = \frac{1982{,}8}{20} \approx 99{,}14 \qquad \overline{y} = \frac{993{,}66}{20} \approx 49{,}68$$

$$\sum_{i=1}^{20} (x_i - \overline{x})^2 \approx 40{,}63 \qquad s_x^2 = \frac{40{,}63}{19} \approx 2{,}14$$

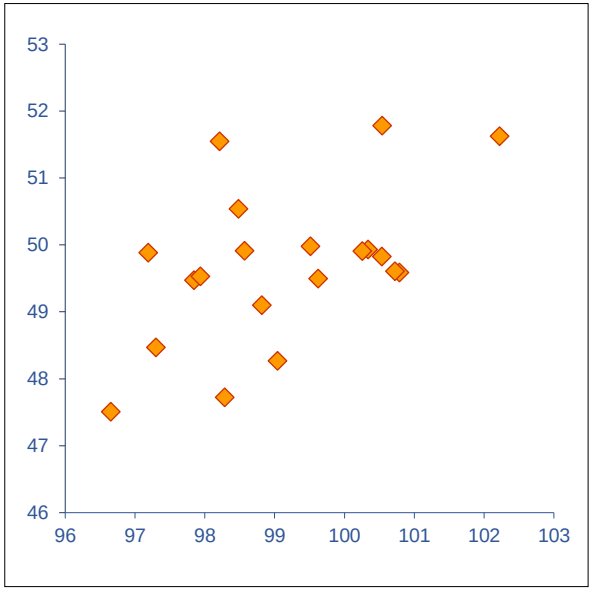
$$\sum_{i=1}^{20} (y_i - \overline{y})^2 \approx 25{,}20 \qquad s_y^2 = \frac{25{,}20}{19} \approx 1{,}33$$

$$\sum_{i=1}^{20} (x_i - \overline{x})(y_i - \overline{y}) = 16{,}69$$

$$s_{XY} = \frac{16{,}69}{19} = 0{,}88,$$

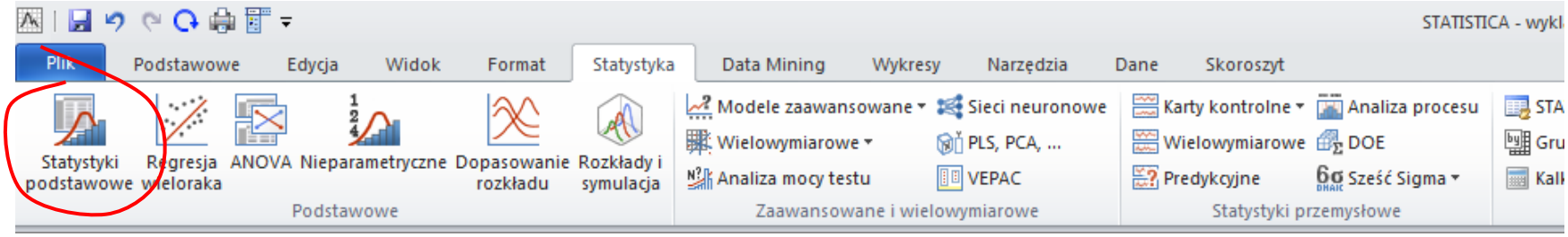
$$r_{XY} = \frac{0{,}47}{0{,}91 \cdot 0{,}66} \approx 0{,}79$$

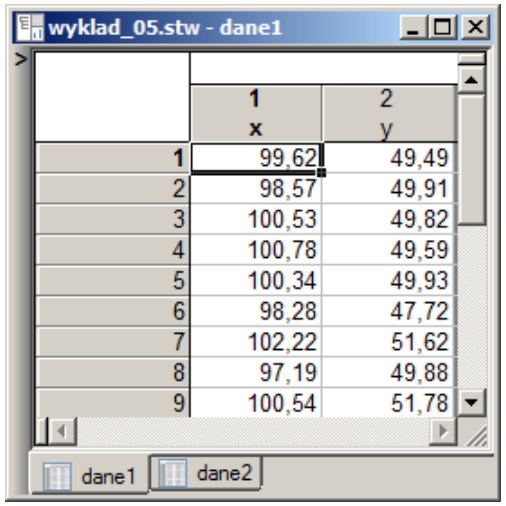
$$S = \begin{bmatrix} 2{,}14 & 0{,}88 \\ 0{,}88 & 1{,}33 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0{,}79 \\ 0{,}79 & 1 \end{bmatrix}$$

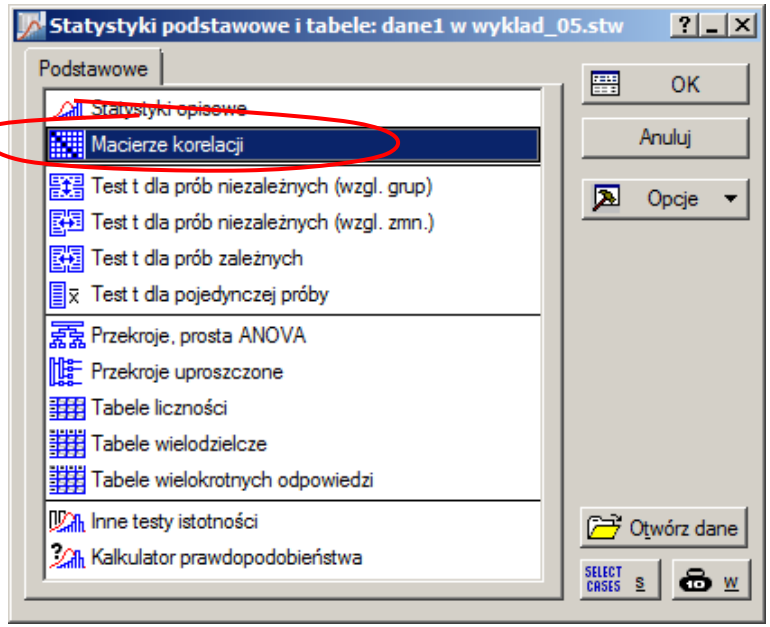


Lp.	x_i	y_i
1	99,62	49,49
2	98,57	49,91
3	100,53	49,82
4	100,78	49,59
5	100,34	49,93
6	98,28	47,72
7	102,22	51,62
8	97,19	49,88
9	100,54	51,78
10	98,81	49,10
11	97,85	49,47
12	97,94	49,53
13	100,72	49,60
14	98,48	50,54
15	99,51	49,98
16	97,30	48,47
17	99,04	48,27
18	100,26	49,91
19	98,21	51,55
20	96,65	47,51
Σ	**1982,83**	**993,66**

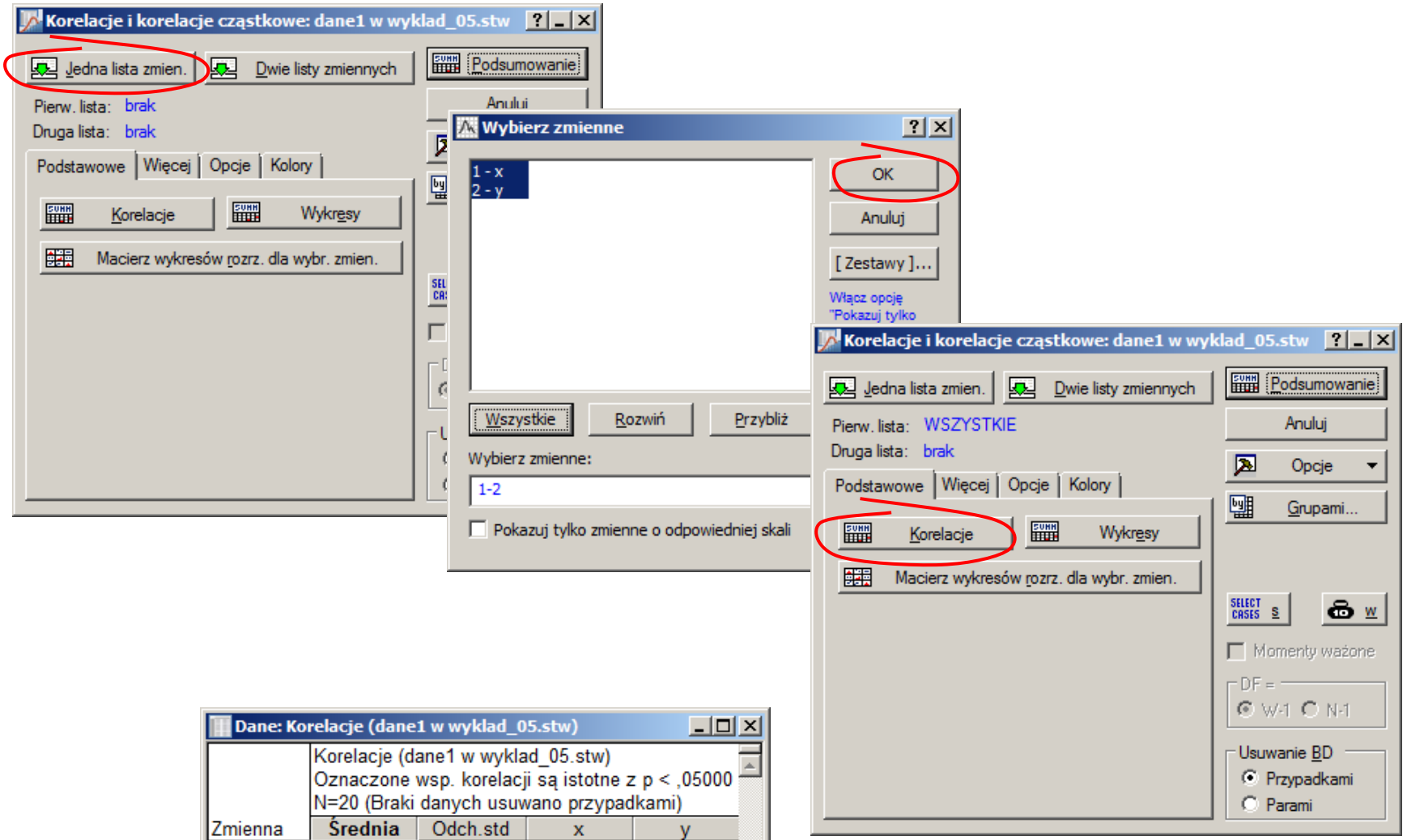
STATISTICA – macierz korelacji







STATISTICA – macierz korelacji



Dane: Korelacje (dane1 w wyklad_05.stw)

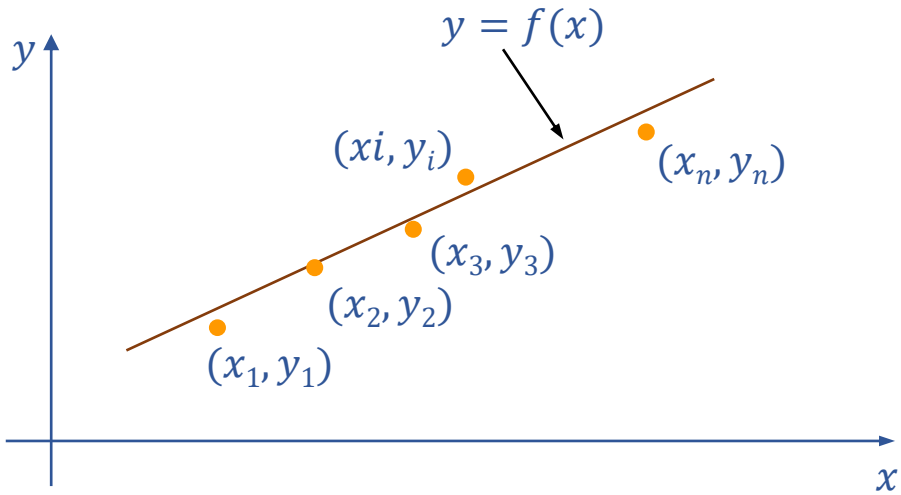
Korelacje (dane1 w wyklad_05.stw)
Oznaczone wsp. korelacji są istotne z p < ,05000
N=20 (Braki danych usuwano przypadkami)

Zmienna	Średnia	Odch.std	x	y
x	99,14200	1,461836	1,000000	0,521745
y	49,68350	1,151293	0,521745	1,000000

Aproksymacja funkcji obiektu

Aproksymacja – to inaczej przybliżanie, polega na zastępowaniu jednych wielkości innymi, bliskimi im w pewnym określonym sensie.

Typowym zastosowaniem aproksymacji jest poszukiwanie zależności funkcyjnej pomiędzy pochodzącymi z pomiarów: wartościami wyjściowymi a wejściowymi badanego obiektu. Ze względu na to, że uzyskane dane mogą być obarczone błędami pomiarowymi – wymaganie dokładnego przyjmowania przez poszukiwaną funkcję określonych wartości nie ma sensu – gdyż nie odpowiada właściwej zależności funkcyjnej.

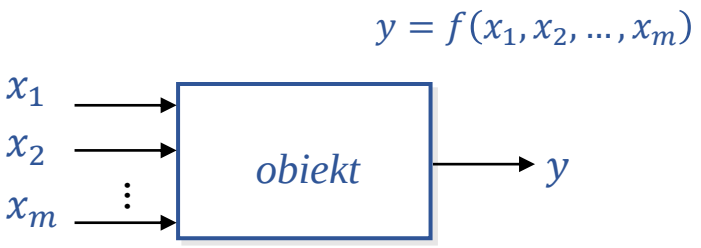


Aproksymacja funkcją liniową zależności funkcyjnej obiektu o jednym wejściu i jednym wyjściu.

Analiza regresji

Analiza regresji pozwala na ilościowe określenie związków pomiędzy kilkoma zmiennymi niezależnymi a zmienną zależną – metoda nazywana jest analizą regresji wielorakiej (wielowymiarowej lub wielokrotnej), w przypadku dwóch zmiennych (jedna zmienna niezależna i jedna zależna) nazywana jest analizą regresji*.

Analiza regresji pozwala na wyznaczenie parametrów teoretycznego modelu obiektu na podstawie obserwacji jego wejść i wyjść. Zależność wyjścia od wejść jest, ze względu na występujące zawsze zakłócenia losowe, zależnością stochastyczną a nie funkcyjną.



Schemat obiektu badań (x_1, x_2, ..., x_m – wielkości wejściowe, y – wielkość wyjściowa)

*Termin regresja wprowadził Sir Francis Galton w pracy *„Naturalna dziedziczność”* (1889). Opisał on zjawisko regresji do średniej:

- dzieci rodziców uzdolnionych są przeciętnie mniej uzdolnione,
- rozmiary nasion groszku w kolejnych pokoleniach wracają do średniego rozmiaru, itp.

Sir Francis Galton – regresja do średniej

Dziedziczność wzrostu – 'dziedziczna regresja do przeciętności'

id_f	h_f	h_m	h_{ch}	g_{ch}
1	78,5	67	73,2	M
1	78,5	67	69,2	F
1	78,5	67	69	F
1	78,5	67	69	F
2	75,5	66,5	65,5	F
2	75,5	66,5	65,5	F
⋮	⋮	⋮	⋮	⋮
205	68,5	65	72	M

id_f	h_p	h_t
1	75,43	73,2
1	75,43	74,736
1	75,43	74,52
1	75,43	74,52
2	73,66	70,74
2	73,66	70,74
⋮	⋮	⋮
205	69,35	72

Zbiór danych

informacje o wzroście 928 dorosłych dzieci z 205 rodzin, gdzie:

id_f – identyfikator rodziny

g_{ch} – płeć dziecka

h_f, h_m, h_{ch} – wzrost ojca, matki i dziecka (w calach)

mężczyźni są średnio o 8% wyżsi od kobiet:

$$\overline{h}_f = 1{,}08\,\overline{h}_m, \quad \overline{h}_{ch(M)} = 1{,}08\,\overline{h}_{ch(F)}$$

h_p – średni wzrost rodzica (mid-parentage height),

$$h_p = \tfrac{1}{2}(h_f + 1{,}08\,h_m)$$

h_t – ujednolicony wzrost dziecka

(dla kobiet $h_t = 1{,}08\,h_{ch}$, dla mężczyzn $h_t = h_{ch}$)

Sir Francis Galton – regresja do średniej

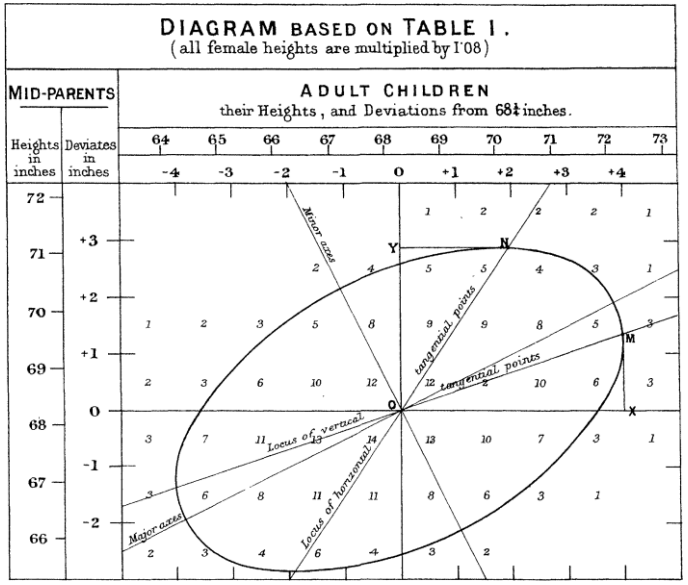
TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

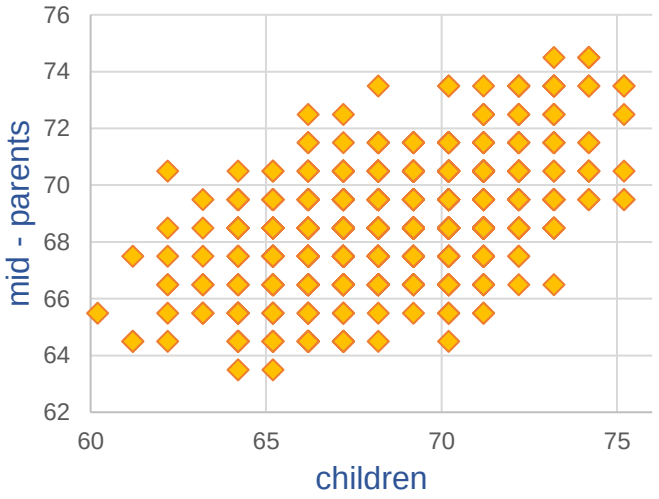
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..

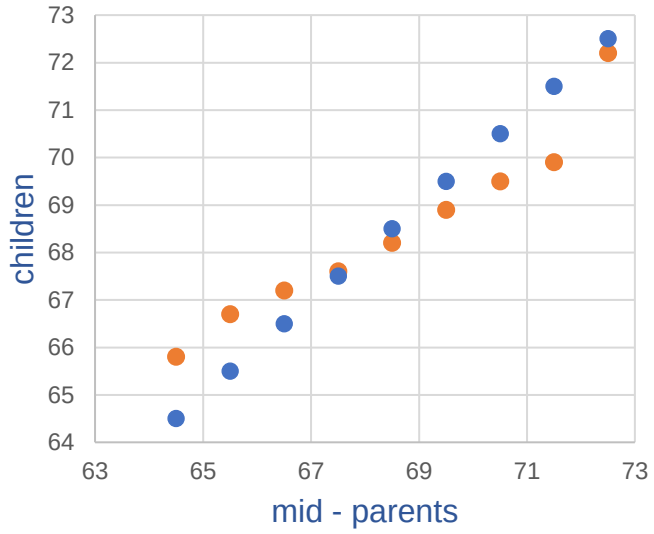
NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

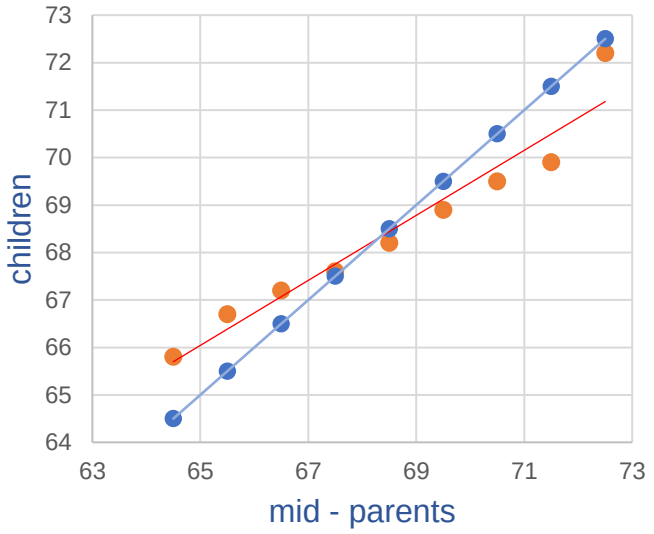


h_p	h_t
72,5	72,2
71,5	69,9
70,5	69,5
69,5	68,9
68,5	68,2
67,5	67,6
66,5	67,2
65,5	66,7
64,5	65,8



Sir Francis Galton – regresja do średniej





- $h_t = h_p$
-

h_p	h_t
72,5	72,2
71,5	69,9
70,5	69,5
69,5	68,9
68,5	68,2
67,5	67,6
66,5	67,2
65,5	66,7
64,5	65,8

Regresja (cofanie) w kierunku przeciętności

$h_t = 21{,}5 + 0{,}685\, h_p$

Odchylenie od przeciętnego w przypadku wzrostu dzieci wynosi 2/3 odchylenia wzrostu rodziców ...

Wysocy rodzice mają dzieci niższe od siebie, niscy rodzice mają dzieci wyższe od siebie.

$\frac{2}{3}$

The Deviates of the Children are to those of their Mid-Parents as 2 to 3 ...

When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they. When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.

Celem analizy regresji jest wyznaczenie parametrów modelu matematycznego obiektu:

$$y = f(x_1, x_2, \dots, x_m, b_0, b_1, \dots, b_p) + e$$

gdzie:

f – funkcja regresji

y – wartość wyjścia obiektu

$x_1, x_2, \dots, x_m$ – wejścia obiektu

$b_0, b_1, \dots, b_p$ – nieznane parametry modelu

e – błąd losowy zawierający wszystkie pozostałe składowe zmienności zmiennej wyjściowej (poza wpływem zmiennych wejściowych), zakłada się, że: $E[e] = 0$ i $D^2[e] = \sigma^2$

Regresja liniowa to jeden z najbardziej popularnych modeli regresji zakładający, że związek pomiędzy zmienną zależną a zmiennymi niezależnymi ma charakter liniowy. W przypadku regresji liniowej, funkcja regresji jest funkcją liniową:

$$y = b_0 + b_1 x_1 + \cdots + b_m x_m + e$$

a w przypadku **uogólnionej regresji liniowej** zakłada się, że funkcja regresji jest liniową kombinacją **funkcji bazowych**:

$$y = b_0 \phi_0(\boldsymbol{x}) + b_1 \phi_1(\boldsymbol{x}) + \cdots + b_p \phi_p(\boldsymbol{x}) + e$$

gdzie:

$\boldsymbol{x} = [x_1, x_2, \ \ldots, x_m]^T$ – wektor wejść obiektu

$\phi_i(\boldsymbol{x})$ – z góry zadane funkcje bazowe

Zakładając, że funkcja regresji jest liniową kombinacją funkcji bazowych

$$y = b_0\phi_0(\boldsymbol{x}) + b_1\phi_1(\boldsymbol{x}) + \cdots + b_p\phi_p(\boldsymbol{x}) + e$$

zadanie analizy polega na wyznaczeniu nieznanych wartości parametrów b_i w oparciu o wyniki przeprowadzanych doświadczeń.

Wykorzystując dane pomiarowe z n obserwacji:

$$\begin{matrix} x_{11}, & x_{12}, & \dots & x_{1m}, & y_1 \\ x_{21}, & x_{22}, & \dots & x_{2m}, & y_2 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ x_{n1}, & x_{n2}, & \dots & x_{nm} & y_n \end{matrix}$$

można wyznaczyć tzw. macierz wejść:

$$\mathbf{X} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_p(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_p(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_n) & \phi_1(\boldsymbol{x}_n) & \cdots & \phi_p(\boldsymbol{x}_n) \end{bmatrix}$$

gdzie: $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \dots, x_{im}]^T$ to wektor zawierający wartości wejść obiektu i -tej obserwacji; $i = 1, 2, \dots, n$.

Metoda najmniejszych kwadratów

Zapisując przybliżone wyniki wyjść oraz szacowane wartości parametrów poszukiwanej funkcji w postaci wektorów:

$$\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n]^T \qquad \boldsymbol{b} = [b_0, b_1, \dots, b_p]^T$$

rezultat przybliżania funkcji obiektu funkcją aproksymującą dla wykonanych obserwacji można zapisać w postaci:

$$\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{b}.$$

Wartości poszukiwanych parametrów $\boldsymbol{b}$ należy wyznaczyć w taki sposób aby jak **najlepiej** dobrać funkcję aproksymującą. Ocenę jakości aproksymacji przeprowadza się licząc odchylenie pomiędzy wartościami wyjść zaobserwowanymi a wyznaczonymi na podstawie przyjętej funkcji. Najczęściej za miarę błędu aproksymacji przyjmuje się:

$$SS_e = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Przedstawiony sposób rozwiązania opiera się na zasadzie najmniejszych kwadratów błędów, metoda rozwiązania oparta na tej zasadzie nazywana jest metodą najmniejszych kwadratów.

Metoda najmniejszych kwadratów

Zadanie znalezienia **najlepszej** funkcji aproksymującej, sprowadza się do znalezienia takich współczynników $\boldsymbol{b}$ aby błąd aproksymacji:

$$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

był jak **najmniejszy**.

Zapisując zaobserwowane wartości wyjść obiektu w postaci wektora $\boldsymbol{y} = [y_1, y_2, \dots, y_n]^T$, błąd aproksymacji można zapisać zależnością:

$$SS_e = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{b})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{b}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\mathbf{X}\boldsymbol{b} - \boldsymbol{b}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{b}^T\mathbf{X}^T\mathbf{X}\boldsymbol{b}.$$

$$SS_e(\boldsymbol{b}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{b}^T\mathbf{X}^T\mathbf{X}\boldsymbol{b}.$$

Wielkość błędu jest więc zależna od wartości współczynników $\boldsymbol{b}$. Zgodnie z teorią rachunku różniczkowego funkcji wielu zmiennych, warunkiem koniecznym istnienia ekstremum funkcji wielu zmiennych w określonym punkcie jest zerowanie się jej wszystkich pochodnych cząstkowych w tym punkcie:

$$\frac{\partial SS_e}{\partial \boldsymbol{b}} = \mathbf{0}.$$

Pochodną błędu SS_e

$$SS_e(\boldsymbol{b}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T\mathbf{X}^T\boldsymbol{y} + \boldsymbol{b}^T\mathbf{X}^T\mathbf{X}\boldsymbol{b}$$

wyznacza się jako:

$$\frac{\partial SS_e(\boldsymbol{b})}{\partial \boldsymbol{b}} = -2\mathbf{X}^T\boldsymbol{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{b}$$

więc po przyrównaniu jej do zera otrzymuje się zależność

$$-2\mathbf{X}^T\boldsymbol{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{b} = \mathbf{0}$$

czyli

$$\mathbf{X}^T\mathbf{X}\boldsymbol{b} = \mathbf{X}^T\boldsymbol{y}$$

Ostatecznie poszukiwane parametry $\boldsymbol{b}$ wyznacza się jako

$$\widehat{\boldsymbol{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

symbol ^ używany jest do oznaczenia estymatorów współczynników wyznaczonych metodą najmniejszych kwadratów.

Z analizy zależności $\widehat{\boldsymbol{b}} = \cdots$ wynika, że istnienie rozwiązania $\widehat{\boldsymbol{b}}$ zależy od postaci macierzy $\mathbf{X}^T\mathbf{X}$ – macierz ta musi być macierzą nieosobliwą.

Interpretacja współczynników

Prognozowane z modelu wartości zmiennej zależnej dla i–tej i j–tej obserwacji wynoszą

$$\hat{y}_i = \hat{b}_0 x_{i0} + \hat{b}_1 x_{i1} + \cdots + b_p x_{ip}$$

$$\hat{y}_j = \hat{b}_0 x_{j0} + \hat{b}_1 x_{j1} + \cdots + b_p x_{jp}$$

gdzie: x_{i0}, x_{j0}, ..., x_{ip}, x_{jp} – wartości z macierzy wejść **X**.

Różnicę w prognozach obliczymy odejmując od siebie obywa równania, tzn.

$$\Delta\hat{y} = \hat{b}_0 \Delta x_0 + \hat{b}_1 \Delta x_1 + \cdots + \hat{b}_p \Delta x_p$$

gdzie: $\Delta\hat{y} = \hat{y}_i - \hat{y}_j$, $\Delta x_0 = xi_0 \_ x_{j0}$, ..., $\Delta x_p = x_{ip} - x_{jp}$.

Stąd, jeżeli założymy, że:

$$\Delta x_k = 1 \text{ oraz } \Delta x_0 = \ldots = \Delta x_{k-1} = \Delta x_{k+1} = \cdots = \Delta x_p = 0$$

to:

$$\Delta\hat{y} = \hat{b}_k$$

Współczynnik $\hat{b}_k$ mierzy zmianę w $\hat{y}$ w przypadku gdy k-ta zmienna niezależna* zmieni swą wartość o 1 jednostkę a pozostałe zmienne niezależne* nie ulegają zmianie.

*W przypadku gdy w macierzy wejść występują funkcje zmiennych niezależnych – zmiana dotyczy wartości funkcji bazowych.

Współczynniki standaryzowane

Współczynniki $\hat{b}_k$ nie mogą być wykorzystywane do oceny, który z nich ma największy wpływ na y. Ocena taka byłaby możliwa, gdyby wszystkie zmienne modelu były standaryzowane:

$$y^* = \frac{y - \overline{y}}{\sigma_y} \qquad x_1^* = \frac{x_1 - \overline{x}_1}{\sigma_{x_1}} \qquad \ldots \qquad x_p^* = \frac{x_p - \overline{x}_p}{\sigma_{x_p}}$$

Po przekształceniu, niestandaryzowane zmienne można zapisać w postaci

$$y = \sigma_y y^* + \overline{y} \qquad x_1 = \sigma_{x_1} x_1^* + \overline{x}_1 \qquad \ldots \qquad x_p = \sigma_{x_p} x_p^* + \overline{x}_p$$

a podstawiając otrzymane zależności do równania regresji

$$y = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p + e$$

otrzymuje się model zawierający standaryzowane zmienne i współczynniki *(slajd ⇨)*

$$y^* = \hat{b}_1^* \, x_1^* + \cdots + \hat{b}_p^* \, x_p^* + e.$$

Standaryzowane współczynniki $\hat{b}_k^* = \hat{b}_k \frac{\sigma_{x_k}}{\sigma_y}$ mogą być wykorzystane do oceny wpływu na k-tej zmiennej niezależnej na zmienną zależną.

Współczynniki standaryzowane

Równanie regresji $y = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p + e$

po podstawieniu $y = \sigma_y y^* + \overline{y}$, $x_1 = \sigma_{x_1} x_1^* + \overline{x}_1$, ..., $x_p = \sigma_{x_p} x_p^* + \overline{x}_p$

zapisuje się kolejno

$$\sigma_y y^* + \overline{y} = \hat{b}_0 + \hat{b}_1\left(\sigma_{x_1} x_1^* + \overline{x}_1\right) + \cdots + \hat{b}_p\left(\sigma_{x_p} x_p^* + \overline{x}_p\right) + e$$

$$\sigma_y y^* = \hat{b}_1 \sigma_{x_1} x_1^* + \cdots + \hat{b}_p \sigma_{x_p} x_p^* + e + \underbrace{\hat{b}_0 + \hat{b}_1 \overline{x}_1 + \cdots + \hat{b}_p \overline{x}_p - \overline{y}}_{=0}$$

a uwzględniając, że

$$\overline{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum\left(\hat{b}_0 + \hat{b}_1 x_{i1} + \cdots + \hat{b}_p x_{ip} + e_i\right) = \hat{b}_0 + \hat{b}_1 \overline{x}_1 + \cdots + \hat{b}_p \overline{x}_p + \underbrace{\overline{e}}_{=0}$$

tzn. $\hat{b}_0 + \hat{b}_1 \overline{x}_1 + \cdots + \hat{b}_p \overline{x}_p - \overline{y} = 0$, równanie można zapisać w postaci

$$\sigma_y y^* = \hat{b}_1 \sigma_{x_1} x_1^* + \cdots + \hat{b}_p \sigma_{x_p} x_p^* + e$$

po wprowadzeniu współczynników standaryzowanych $\hat{b}_k^* = \hat{b}_k \frac{\sigma_{x_k}}{\sigma_y}$, ostatecznie jako

$$y^* = \hat{b}_1^* x_1^* + \cdots + \hat{b}_p^* x_p^* + e.$$

Metoda najmniejszych kwadratów

Celem badań doświadczalnych jest określenie funkcji obiektu badań $y = f(x_1, x_2)$ przy założeniu, że:

$$y = b_0 + b_1 x_1 + b_2 x_2 \qquad 1 \le x_1 \le 3, \qquad 0 \le x_2 \le 4.$$

Planując eksperyment zdecydowano o przeprowadzeniu 4 doświadczeń: po jednym na każdym poziomie zmiennej wejściowej.

Lp.	x_1	x_2	y
1	1	0	
2	1	4	
3	3	0	
4	3	4	

Wyniki otrzymanych obserwacji można zapisać w postaci wektora $\boldsymbol{y} = \begin{bmatrix} 12 \\ 0 \\ -2 \\ 26 \end{bmatrix}$.

Lp.	x_1	x_2	y
1	1	0	12
2	1	4	0
3	3	0	-2
4	3	4	26

Z przyjętej postaci funkcji $y = b_0 + b_1 x_1 + b_2 x_2$ wynika, że funkcje bazowe należy w tym przypadku przyjąć jako

$$\phi_0(\boldsymbol{x}) = 1, \quad \phi_1(\boldsymbol{x}) = x_1, \quad \phi_2(\boldsymbol{x}) = x_2.$$

Dla takich założeń i przyjętych w eksperymencie wartości zmiennych wejściowych macierz wejść należy zapisać w postaci

$$\mathbf{X} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \phi_2(\boldsymbol{x}_2) \\ \phi_0(\boldsymbol{x}_3) & \phi_1(\boldsymbol{x}_3) & \phi_2(\boldsymbol{x}_3) \\ \phi_0(\boldsymbol{x}_4) & \phi_1(\boldsymbol{x}_4) & \phi_2(\boldsymbol{x}_4) \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 4 \\ 1 & 3 & 0 \\ 1 & 3 & 4 \end{bmatrix}.$$

Współczynniki funkcji regresji wyznacza się jako

$$\widehat{\boldsymbol{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 0 & 4 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 4 \\ 1 & 3 & 0 \\ 1 & 3 & 4 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 0 & 4 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 12 \\ 0 \\ -2 \\ 26 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix}$$

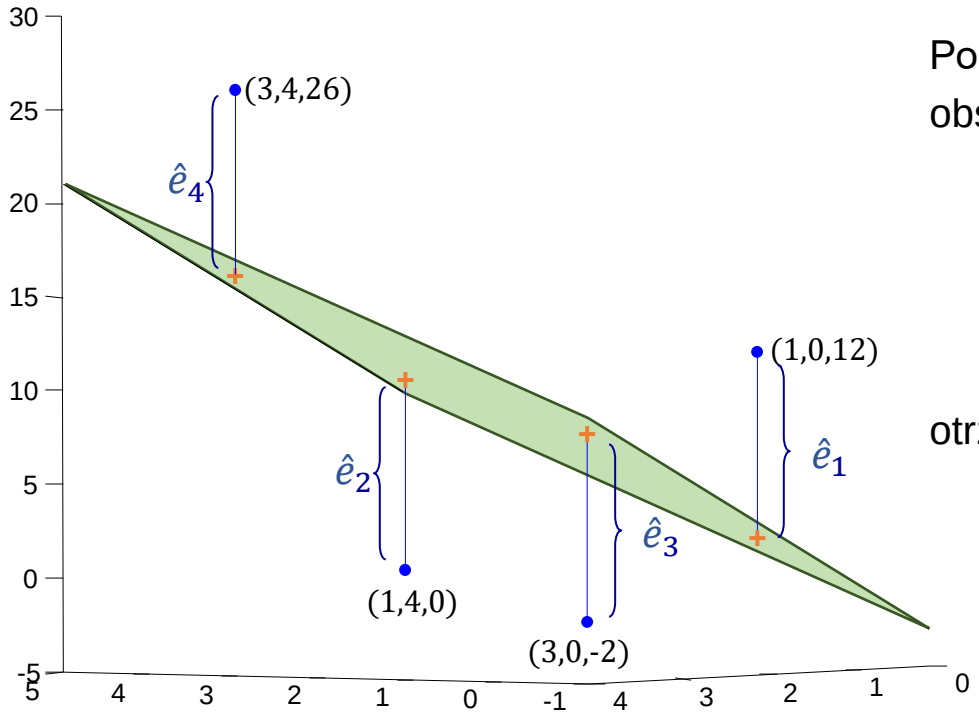
więc znaleziona funkcja ma postać

$$\hat{y} = -1 + 3x_1 + 2x_2.$$

Metoda najmniejszych kwadratów

Aproksymowane wartości zmiennej wyjściowej można otrzymać podstawiając do powyższego wzoru wartości zmiennych wejściowych lub korzystając z wzoru

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{b}} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 4 \\ 1 & 3 & 0 \\ 1 & 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 10 \\ 8 \\ 16 \end{bmatrix}.$$



Porównując wartości aproksymowane z obserwowanymi

$$\boldsymbol{y} = \begin{bmatrix} 12 \\ 0 \\ -2 \\ 26 \end{bmatrix}$$

otrzymuje się wektor reszt

$$\hat{\boldsymbol{e}} = \begin{bmatrix} 10 \\ -10 \\ -10 \\ 10 \end{bmatrix}$$

Interpretacja współczynników

Współczynniki wyznaczonego modelu $\hat{y} = -1 + 3x_1 + 2x_2$ oznaczają, że:

- zmiana wartości zmiennej x_1 o 1 jednostkę spowoduje wzrost zmiennej $\hat{y}$ o 3 jednostki,
- zmiana wartości zmiennej x_2 o 1 jednostkę spowoduje wzrost zmiennej $\hat{y}$ o 2 jednostki.

Do oceny, która ze zmiennych niezależnych ma największy wpływ na y można wykorzystać współczynniki standaryzowane $\hat{b}_1^*$ i $\hat{b}_2^*$. Obliczając kolejno:

$$\overline{x}_1 = 2, \quad \overline{x}_2 = 2, \quad \overline{y} = 9, \qquad \sigma_y = \sqrt{\tfrac{1}{3}\left((12-9)^2 + \cdots + (26-9)^2\right)} \approx 12{,}91$$

$$\sigma_{x_1} = \sqrt{\tfrac{1}{3}\left((1-2)^2 + \cdots + (3-2)^2\right)} \approx 1{,}16, \qquad \sigma_{x_2} = \sqrt{\tfrac{1}{3}\left((0-2)^2 + \cdots + (4-2)^2\right)} \approx 2{,}31,$$

otrzymuje się:

$$\hat{b}_1^* = \hat{b}_1 \frac{\sigma_{x_1}}{\sigma_y} = 3 \cdot \frac{1{,}16}{12{,}91} \approx 0{,}27, \qquad \hat{b}_2^* = \hat{b}_2 \frac{\sigma_{x_2}}{\sigma_y} = 2 \cdot \frac{2{,}31}{12{,}91} \approx 0{,}36$$

co oznacza, że zmienna x_2 ma większy wpływ na zmienną $\hat{y}$.

Lp.	x_1	x_2	y
1	1	0	12
2	1	4	0
3	3	0	-2
4	3	4	26

Metoda najmniejszych kwadratów

Załóżmy, że celem badań jest określenie funkcji obiektu w postaci $y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$. Funkcje bazowe należy w tym przypadku przyjąć jako

$$\phi_0(\boldsymbol{x}) = 1, \quad \phi_1(\boldsymbol{x}) = x_1, \quad \phi_2(\boldsymbol{x}) = x_2, \quad \phi_{12}(\boldsymbol{x}) = x_1 x_2.$$

Dla takich założeń i przyjętych w eksperymencie wartości zmiennych wejściowych macierz wejść należy zapisać w postaci

$$\mathbf{X} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_1) & \phi_3(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \phi_2(\boldsymbol{x}_2) & \phi_3(\boldsymbol{x}_2) \\ \phi_0(\boldsymbol{x}_3) & \phi_1(\boldsymbol{x}_3) & \phi_2(\boldsymbol{x}_3) & \phi_3(\boldsymbol{x}_3) \\ \phi_0(\boldsymbol{x}_4) & \phi_1(\boldsymbol{x}_4) & \phi_2(\boldsymbol{x}_4) & \phi_3(\boldsymbol{x}_4) \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ 1 & x_{31} & x_{32} & x_{31}x_{32} \\ 1 & x_{41} & x_{42} & x_{41}x_{42} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 4 & 4 \\ 1 & 3 & 0 & 0 \\ 1 & 3 & 4 & 12 \end{bmatrix}.$$

Współczynniki funkcji regresji wyznacza się jako

$$\hat{\boldsymbol{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \begin{bmatrix} 19 \\ -7 \\ -8 \\ 5 \end{bmatrix}$$

więc znaleziona funkcja ma postać

$$\hat{y} = 19 - 7x_1 - 8x_2 + 5x_1 x_2.$$

Wartości aproksymowane odpowiadają obserwowanym, wszystkie reszty są zerowe – znaleziona funkcja idealnie opisuje wpływ zmiennych wejściowych na zmienną wyjściową.

Lp.	x_1	x_2	y	$\hat{y}$
1	1	0	12	12
2	1	4	0	0
3	3	0	-2	-2
4	3	4	26	26

Analiza regresji – założenia

Błędy losowe e_i:

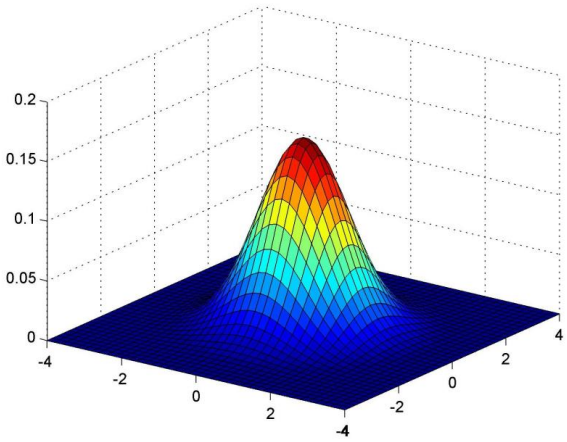
- są od siebie niezależne, tzn.: dla $i \neq j$ kowariancja $\sigma_{e_i e_j} = 0$,
- wartość oczekiwana $E[e_i] = 0$,
- wariancja $D^2[e_i] = \sigma^2$,
- oraz $e_i \sim \mathcal{N}(0, \sigma)$,

tzn.

- reszty $\boldsymbol{e}$ mają n – wymiarowy rozkład normalny o parametrach $\mu = \mathbf{0}$ i $\Sigma = \sigma^2\mathbf{I}$

$$\boldsymbol{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

($\mathbf{I}$ to macierz jednostkowa)



Własności wektora obserwacji $\boldsymbol{y}$

Z przyjętego modelu funkcji regresji: $\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{e}$, wynika, że:

- wartość oczekiwana $E[\boldsymbol{y}] = E[\mathbf{X}\boldsymbol{b}]$

$$E[\boldsymbol{y}] = E[\mathbf{X}\boldsymbol{b} + \boldsymbol{e}] = E[\mathbf{X}\boldsymbol{b}] + E[\boldsymbol{e}] = E[\mathbf{X}\boldsymbol{b}]$$

- macierz kowariancji $D^2[\boldsymbol{y}] = \sigma^2\mathbf{I}$

$$D^2[\boldsymbol{y}] = D^2[\mathbf{X}\boldsymbol{b} + \boldsymbol{e}] = D^2[\boldsymbol{e}] = \sigma^2\mathbf{I}$$

tzn.

- wektor obserwacji $\boldsymbol{y}$ ma n – wymiarowy rozkład normalny o $\mu = \mathbf{X}\boldsymbol{b}$ i $\Sigma = \sigma^2\mathbf{I}$

$$\boldsymbol{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{b}, \sigma^2\mathbf{I})$$

Własności parametrów $\hat{b}$

- wartość oczekiwana $E[\hat{\boldsymbol{b}}] = \boldsymbol{b}$

$$E[\hat{\boldsymbol{b}}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{b}+\boldsymbol{e})] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{b} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{e}] =$$
$$= E[\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}_{=\mathbf{I}}\boldsymbol{b}] + E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{e}] = E[\boldsymbol{b}] + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underbrace{E[\boldsymbol{e}]}_{=\mathbf{0}} = E[\boldsymbol{b}]$$

- macierz kowariancji $D^2[\hat{\boldsymbol{b}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

$$D^2[\hat{\boldsymbol{b}}] = D^2[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underbrace{D^2[\boldsymbol{y}]}_{=\sigma^2\mathbf{I}}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T =$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \sigma^2\mathbf{I} \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}_{=\mathbf{I}}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Własności parametrów $\hat{b}$

- parametry $\hat{\boldsymbol{b}} \sim \mathcal{N}_{p+1}(\boldsymbol{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$

1. $\hat{\boldsymbol{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$
2. $\hat{\boldsymbol{b}}$ jest liniowo zależne od wektora obserwacji $\boldsymbol{y}$
3. obserwacje $\boldsymbol{y}$ mają rozkład normalny
4. $E[\hat{\boldsymbol{b}}] = \boldsymbol{b}$ i $D^2[\hat{\boldsymbol{b}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

} parametry $\hat{\boldsymbol{b}}$ mają rozkład normalny

$$\hat{\boldsymbol{b}} \sim \mathcal{N}_{p+1}(\boldsymbol{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

wprowadzając $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$ oraz c_{ii} (i-ty element diagonalny macierzy $\mathbf{C}$)

$$\hat{\boldsymbol{b}} \sim \mathcal{N}_{p+1}(\boldsymbol{b}, \sigma^2\,\mathbf{C})$$

$$\hat{b}_i \sim \mathcal{N}\left(b_i, \sigma\sqrt{c_{ii}}\right)$$

wariancja $\hat{b}_i$ wynosi $\sigma^2 c_{ii}$
kowariancja pomiędzy $\hat{b}_i$ i $\hat{b}_j$ wynosi $\sigma^2 c_{ij}$

Własności wektora reszt e

1) macierz wejść jest nieskorelowana z e, tzn.: $\mathbf{X}^T\boldsymbol{e} = \mathbf{0}$

$$\mathbf{X}^T\boldsymbol{e} = \mathbf{X}^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\hat{\boldsymbol{y}} = \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{b}} = \mathbf{X}^T\boldsymbol{y} - \underbrace{\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}}_{=\mathbf{I}}\mathbf{X}^T\boldsymbol{y}$$

$$= \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\boldsymbol{y} = \mathbf{0}$$

2) suma elementów wektora $\boldsymbol{e}$ wynosi 0 (tylko jeżeli funkcja regresji zawiera stałą)

jeśli funkcja regresji zawiera stałą to pierwsza kolumna macierzy $\mathbf{X}$ jest zbudowana z jedynek i w konsekwencji pierwszy wiersz macierzy $\mathbf{X}^T$ jest również zbudowany z jedynek, na mocy własności 1) otrzymuje się więc:

$$\sum e_i = 0$$

3) wektor $\hat{\boldsymbol{y}}$ jest nieskorelowany z wektorem reszt, tzn.: $\hat{\boldsymbol{y}}^T\boldsymbol{e} = \mathbf{0}$

$$\hat{\boldsymbol{y}}^T\boldsymbol{e} = \left(\mathbf{X}\hat{\boldsymbol{b}}\right)^T\boldsymbol{e} = \hat{\boldsymbol{b}}^T\underbrace{\mathbf{X}^T\boldsymbol{e}}_{=\mathbf{0}} = \mathbf{0}$$

Dekompozycja zmienności zmiennej zależnej

Jeżeli funkcja regresji zawiera stałą to całkowitą zmienność zmiennej zależnej SS_T można zdekomponować na zmienność wyjaśnioną równaniem regresji SS_r i zmienność niewyjaśnioną przyjętym modelem SS_e, tzn.:

$$SS_T = SS_r + SS_e$$

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad SS_r = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad SS_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y} + y_i - \hat{y}_i)^2 = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{= SS_r} + 2\underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i)}_{= 0} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{= SS_e}$$

$$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})e_i = \sum_{i=1}^{n}\hat{y}_i\, e_i - \sum_{i=1}^{n}\overline{y}\, e_i = \underbrace{\boldsymbol{y}^T\boldsymbol{e}}_{\text{(własność 3)} = 0} - \overline{y}\underbrace{\sum_{i=1}^{n} e_i}_{= 0\ \text{(własność 2)}} = 0$$

Jakość dopasowania równania regresji

Współczynnik determinacji

Współczynnik opisuje siłę związku pomiędzy rzeczywistym wyjściem obiektu $\boldsymbol{y}$ a prognozowanym wyjściem $\widehat{\boldsymbol{y}}$.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SS_r}{SS_T}$$

Współczynnik przyjmuje on wartości z przedziału $[0, 1]$.

Jeśli funkcja regresji jest:

- słabo dopasowana to wskaźnik R^2 ma wartości bliskie 0 (słaby związek $\boldsymbol{y}$ z $\widehat{\boldsymbol{y}}$)
- im lepiej dopasowana funkcja (silniejszy związek $\boldsymbol{y}$ z $\widehat{\boldsymbol{y}}$) tym wyższa wartość R^2.

Jakość dopasowania równania regresji

Współczynnik determinacji

Dla modeli zawierającej stałą, współczynnik determinacji zapisywany jest również jako:

$$R^2 = \frac{SS_r}{SS_T} = \frac{SS_T - SS_e}{SS_T} = 1 - \frac{SS_e}{SS_T}.$$

W przypadku regresji liniowej współczynnik determinacji jest równy kwadratowi współczynnika korelacji liniowej Paersona: $R^2 = r^2$.

Skorygowany współczynnik determinacji

W celu uniezależnienia wartości wskaźnika od liczby stopni swobody wprowadzany jest współczynnik skorygowany:

$$\overline{R}^2 = 1 - \frac{SS_e/n-p-1}{SS_T/n-1} = 1 - \frac{n-1}{n-p-1}\frac{SS_e}{SS_T} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Jakość dopasowania równania regresji

Współczynnik determinacji dla modelu

$$\hat{y} = -1 + 3x_1 + 2x_2$$

wyznacza się obliczając kolejno

Lp.	x_1	x_2	y	$\hat{y}$	$\hat{e}$
1	1	0	12	2	10
2	1	4	0	10	-10
3	3	0	-2	8	-10
4	3	4	26	16	10

$$\overline{y} = 9$$

$$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2 = 400$$

$$SS_r = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = (2-9)^2 + (10-9)^2 + (8-9)^2 + (16-9)^2 = 100$$

$$SS_T = SS_r + SS_e = 400 + 100 = 500$$

$$R^2 = 1 - \frac{SS_e}{SS_T} = 1 - \frac{400}{500} = 0{,}2$$

$$\overline{R}^2 = 1 - \frac{n-1}{n-p-1} \cdot \frac{SS_e}{SS_T} = 1 - \frac{4-1}{4-2-1} \cdot \frac{400}{500} = -1{,}4$$

dopasowanie jest bardzo słabe, zmienność niewyjaśniona modelem znacznie przewyższa zmienność wyjaśnioną

Szacowanie σ^2

Po wprowadzeniu średnich kwadratów odchyleń:

MS_T	MS_r	MS_e
$\frac{1}{n-1} SS_T$	$\frac{1}{p} SS_r$	$\frac{1}{n-p-1} SS_e$

i obliczeniu ich wartości oczekiwanych:

MS_r	MS_e
$\sigma^2 + \frac{1}{p\sigma^2} \tilde{\boldsymbol{b}}^T \mathbf{X}_c^T \mathbf{X}_c \tilde{\boldsymbol{b}}$	σ^2

okazuje się, że wariancję σ^2 można oszacować w oparciu o:

- średnią MS_r – jeżeli dla każdego i prawdziwa jest hipoteza $H_0: b_i = 0$
- średnią MS_e – niezależnie od prawdziwości stawianych hipotez zerowych $H_0: b_i = 0$

*$\tilde{\boldsymbol{b}} = [b_1, \ldots, b_p]^T$, $\mathbf{X}_c$ – wycentrowana macierz wejść

D.C. Montgomery, E.A. Peck, , G.G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, 2015

Istotność funkcji regresji

Funkcję regresji uznaje się za istotną jeżeli przynajmniej jeden ze współczynników funkcji jest istotnie różny od zera. Istotność bada się stawiając hipotezę zerową o braku wpływu zmiennych niezależnych na zmienną zależną, tzn.

$$H_0: \; b_1 = b_2 = \ldots = b_p = 0$$

wobec hipotezy alternatywnej

$$H_1: \; b_i \neq 0.$$

Do przeprowadzenia tego testu wykorzystuje się statystykę

$$F_n = \frac{MS_r}{MS_e} = \frac{SS_r}{p} \Big/ \frac{SS_e}{n-p-1}.$$

Statystyka F_n ma rozkład F Fishera-Snedecora o $v_1 = p$ i $v_2 = (n - p - 1)$ stopniach swobody. Jeśli hipoteza H_0 jest prawdziwa rozkład jest centralny, jeśli prawdziwa jest hipoteza H_1 rozkład jest niecentralny o stopniu niecentralności:

$$\delta = \frac{1}{p\sigma^2} \widetilde{\boldsymbol{b}}^T \mathbf{X}_c^T \mathbf{X}_c \widetilde{\boldsymbol{b}}.$$

Istotność współczynnika regresji b_i

Współczynnik uznaje się za istotny jeżeli jest istotnie różny o zera. Istotność współczynnika bada się więc testując hipotezę o braku istotności

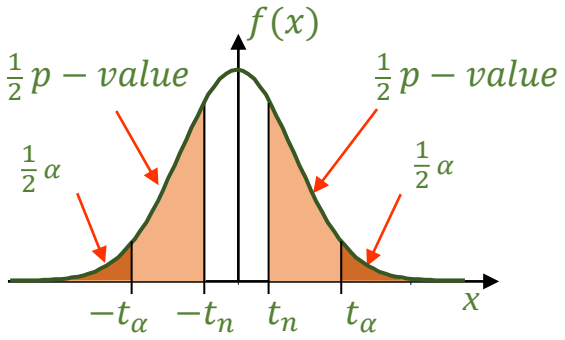
$$H_0: \; b_i = 0$$

wobec hipotezy alternatywnej

$$H_1: \; b_i \neq 0.$$

Współczynniki równania regresji mają rozkład normalny $\hat{b}_i \sim \mathcal{N}\left(b_i, \sigma\sqrt{c_{ii}}\right)$, do przeprowadzenia testu wykorzystywana jest statystyka t_n o rozkładzie *t–Studenta* o $(n-p-1)$ stopniach swobody

$$t_n = \frac{\hat{b}_i - 0}{s\sqrt{c_{ii}}} = \frac{\hat{b}_i}{s\sqrt{c_{ii}}}, \quad s = \sqrt{MS_e}.$$



Istotność funkcji regresji

W teście istotności dla modelu

$$\hat{y} = -1 + 3x_1 + 2x_2$$

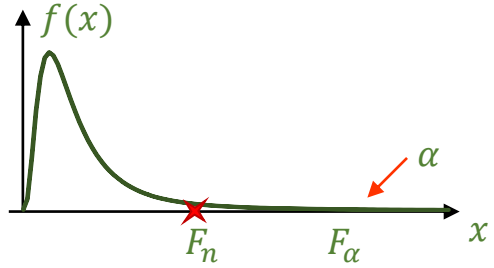
oblicza się kolejno

Lp.	x_1	x_2	y	$\hat{y}$	$\hat{e}$
1	1	0	12	2	10
2	1	4	0	10	-10
3	3	0	-2	8	-10
4	3	4	26	16	10

$$MS_e = \frac{SS_e}{n-p-1} = \frac{400}{4-2-1} = 400$$

$$MS_r = \frac{SS_r}{p} = \frac{100}{2} = 50$$

$$F_n = \frac{MS_r}{MS_e} = \frac{50}{400} = 0{,}125$$



$$F_\alpha = F^{-1}_{F(2,4-2-1)}(1-0{,}05) \approx 199{,}5$$

$$p-value = 1 - F_{F(p,n-p-1)}(F_n) = 1 - F_{F(2,4-2-1)}(0{,}125) \approx 1 - 0{,}1056 \approx 0{,}894$$

wartość statystyki testowej poza obszarem krytycznym
poziom istotności α jest mniejszy od $p-value$

nie ma podstaw do odrzucenia
hipotezy o braku istotności

Istotność współczynników funkcji regresji

Test istotności wykazał, że znaleziona funkcja regresji jest statystycznie nieistotna – oznacza to, że współczynniki funkcji b_1, b_2 nie różnią się w sposób istotny od zera. Brak istotności współczynników można również pokazać przeprowadzając testy istotności dla każdego współczynnika oddzielnie.

Do przeprowadzenia testu konieczne jest oszacowanie wariancji współczynników $\hat{b}_i$

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 0 & 4 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 4 \\ 1 & 3 & 0 \\ 1 & 3 & 4 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 1{,}5 & -0{,}5 & -0{,}125 \\ -0{,}5 & 0{,}25 & 0 \\ -0{,}125 & 0 & 0{,}0625 \end{bmatrix}, \quad s = \sqrt{MS_e} = \sqrt{400} = 20$$

.

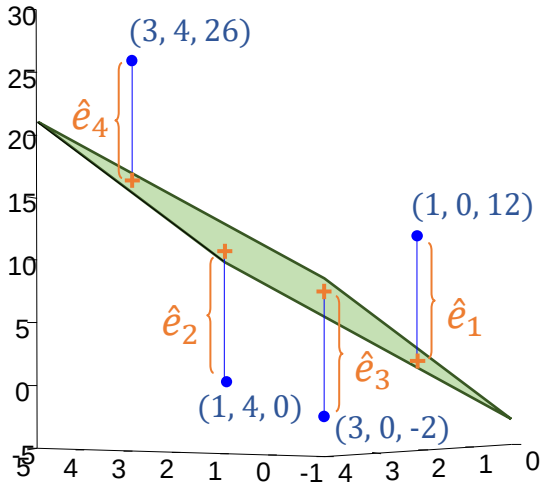
i	$\hat{b}_i$	$t_n = \frac{\hat{b}_i}{s\sqrt{c_{ii}}}$	$p - value = 2F_{t(n-p-1)}(t_n)$
0	-1	$\frac{-1}{20\sqrt{1{,}5}} \approx -0{,}04$	$2F_{t(4-2-1)}(-0{,}04) \approx 0{,}9740$
1	3	$\frac{3}{20\sqrt{0{,}25}} = 0{,}3$	$2F_{t(4-2-1)}(-0{,}3) \approx 0{,}8145$
2	2	$\frac{2}{20\sqrt{0{,}0625}} = 0{,}4$	$2F_{t(4-2-1)}(-0{,}4) \approx 0{,}7578$

wartości statystyk testowych oraz $p - value$ wskazują, że **nie można odrzucić** hipotez o braku istotności współczynników regresji

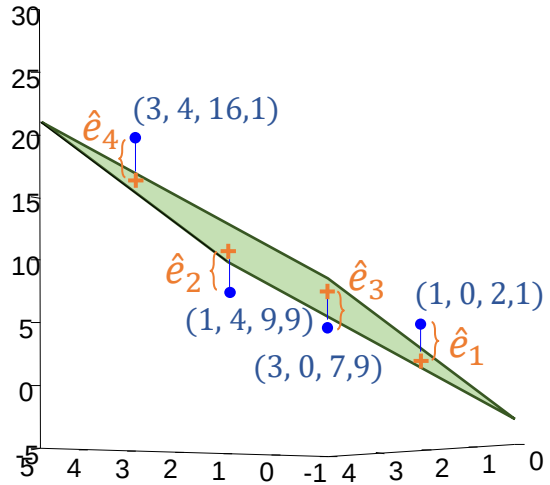
$$t_\alpha = -F^{-1}_{t(4-2-1)}(0{,}05/2) \approx 12{,}7062$$

Zbadana zostanie istotność funkcji regresji $\hat{y} = -1 + 3x_1 + 2x_2$ przy założeniu, że reszty $\hat{e}_i = \pm 0{,}1$.

Lp.	x_1	x_2	y	$\hat{y}$	$\hat{e}$
1	1	0	12	2	10
2	1	4	0	10	-10
3	3	0	-2	8	-10
4	3	4	26	16	10



Lp.	x_1	x_2	y	$\hat{y}$	$\hat{e}$
1	1	0	2,1	2	0,1
2	1	4	9,9	10	-0,1
3	3	0	7,9	8	-0,1
4	3	4	16,1	16	0,1



$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \left(\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & 3 & 3\\ 0 & 4 & 0 & 4\end{bmatrix}\cdot\begin{bmatrix}1 & 1 & 0\\ 1 & 1 & 4\\ 1 & 3 & 0\\ 1 & 3 & 4\end{bmatrix}\right)^{-1}\cdot\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & 3 & 3\\ 0 & 4 & 0 & 4\end{bmatrix}\cdot\begin{bmatrix}2{,}1\\ 9{,}9\\ 7{,}9\\ 16{,}1\end{bmatrix} = \begin{bmatrix}-1\\ 3\\ 2\end{bmatrix}$$

Jakość dopasowania równania regresji

Współczynnik determinacji dla modelu

$$\hat{y} = -1 + 3x_1 + 2x_2$$

wyznacza się obliczając kolejno

Lp.	x_1	x_2	y	$\hat{y}$	$\hat{e}$
1	1	0	2,1	2	0,1
2	1	4	9,9	10	-0,1
3	3	0	7,9	8	-0,1
4	3	4	16,1	16	0,1

$$\overline{y} = 9$$

$$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2 = 0{,}04$$

$$SS_r = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = (2-9)^2 + (10-9)^2 + (8-9)^2 + (16-9)^2 = 100$$

$$SS_T = SS_r + SS_e = 0{,}04 + 100 = 100{,}04$$

$$R^2 = 1 - \frac{SS_e}{SS_T} = 1 - \frac{0{,}04}{100{,}04} = 0{,}9996$$

$$\overline{R}^2 = 1 - \frac{n-1}{n-p-1} \cdot \frac{SS_e}{SS_T} = 1 - \frac{4-1}{4-2-1} \cdot \frac{0{,}04}{100{,}04} = 0{,}9988$$

dopasowanie jest bardzo silne, zmienność wyjaśniona modelem znacznie przewyższa zmienność niewyjaśnioną

Istotność funkcji regresji

W teście istotności dla modelu

$$\hat{y} = -1 + 3x_1 + 2x_2$$

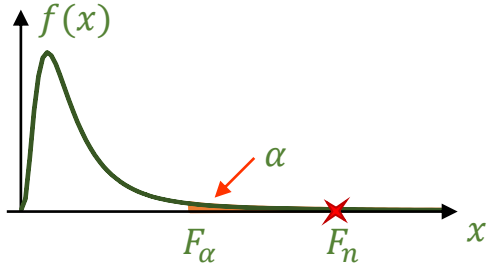
oblicza się kolejno

Lp.	x_1	x_2	y	$\hat{y}$	$\hat{e}$
1	1	0	2,1	2	0,1
2	1	4	9,9	10	-0,1
3	3	0	7,9	8	-0,1
4	3	4	16,1	16	0,1

$$MS_e = \frac{SS_e}{n-p-1} = \frac{0{,}04}{4-2-1} = 0{,}04$$

$$MS_r = \frac{SS_r}{p} = \frac{100}{2} = 50$$

$$F_n = \frac{MS_r}{MS_e} = \frac{50}{0{,}04} = 1250$$



$$F_\alpha = F^{-1}_{F(2,4-2-1)}(1-0{,}05) \approx 199{,}5$$

$$p-value = 1 - F_{F(p,n-p-1)}(F_n) = 1 - F_{F(2,4-2-1)}(1250) \approx 0{,}02$$

Wartość statystyki testowej w obszarze krytycznym poziom istotności α jest większy od $p-value$ → hipotezę o braku istotności należy odrzucić

Istotność współczynników funkcji regresji

Test istotności wykazał, że znaleziona funkcja regresji jest statystycznie istotna – testy istotności dla poszczególnych współczynników pozwolą zweryfikować ich istotność.

Do przeprowadzenia testu konieczne jest oszacowanie wariancji współczynników $\hat{b}_i$

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 1{,}5 & -0{,}5 & -0{,}125 \\ -0{,}5 & 0{,}25 & 0 \\ -0{,}125 & 0 & 0{,}0625 \end{bmatrix}, \quad s = \sqrt{MS_e} = \sqrt{0{,}04} = 0{,}2$$

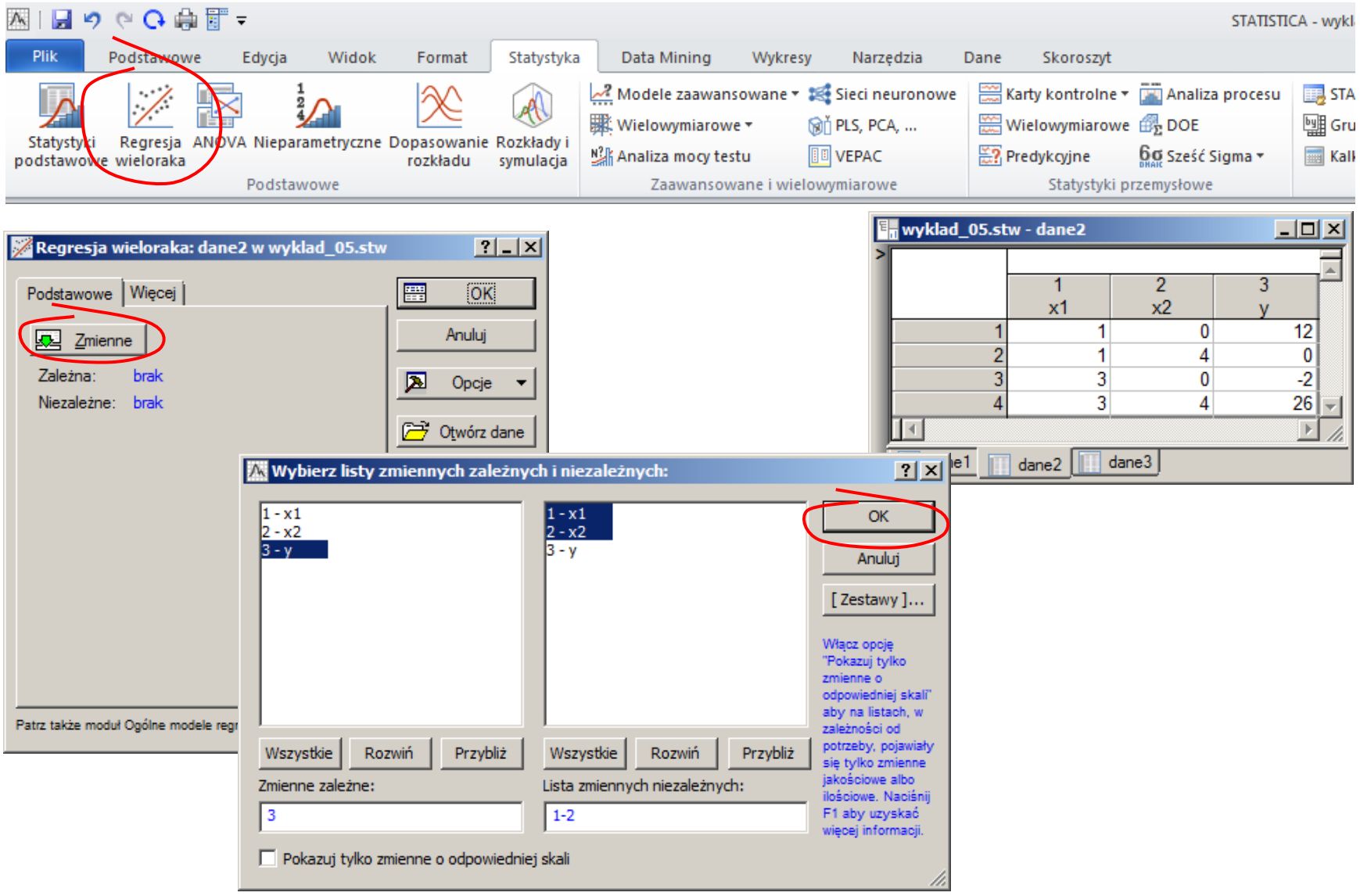
i	$\hat{b}_i$	$t_n = \frac{\hat{b}_i}{s\sqrt{c_{ii}}}$	$p - value = 2F_{t(n-p-1)}(t_n)$
0	-1	$\frac{-1}{0{,}2\sqrt{1{,}5}} \approx -4{,}08$	$2F_{t(4-2-1)}(-4{,}08) \approx 0{,}1529$
1	3	$\frac{3}{0{,}2\sqrt{0{,}25}} = 30$	$2F_{t(4-2-1)}(-30) \approx 0{,}0212$
2	2	$\frac{2}{0{,}2\sqrt{0{,}0625}} = 40$	$2F_{t(4-2-1)}(-40) \approx 0{,}0159$

wartości statystyk testowych oraz $p - value$ wskazują, że:

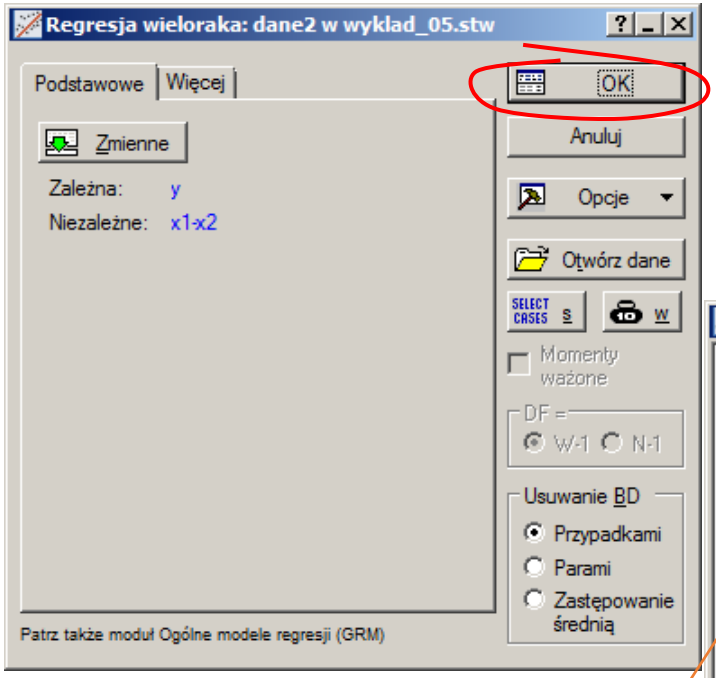
- **nie można odrzucić** hipotezy o braku istotności wsp. $\hat{b}_0$
- **należy odrzucić** hipotezy o braku istotności wsp. $\hat{b}_1$ i $\hat{b}_2$

$$t_\alpha = -F^{-1}_{t(4-2-1)}(0{,}05/2) \approx 12{,}7062$$

STATISTICA – analiza regresji



STATISTICA – analiza regresji



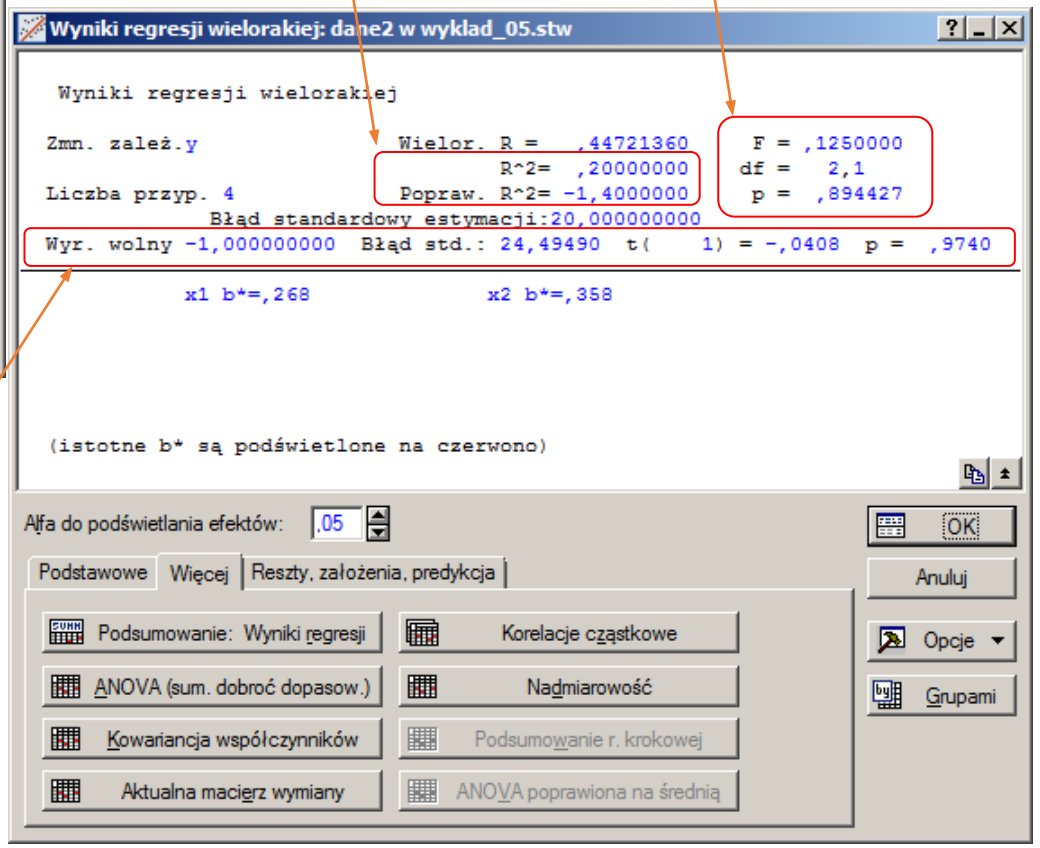
test istotności funkcji regresji

F – statystyka testowa

df – liczba stopni swobody F

p – $p-value$

R^2, $\overline{R}^2$



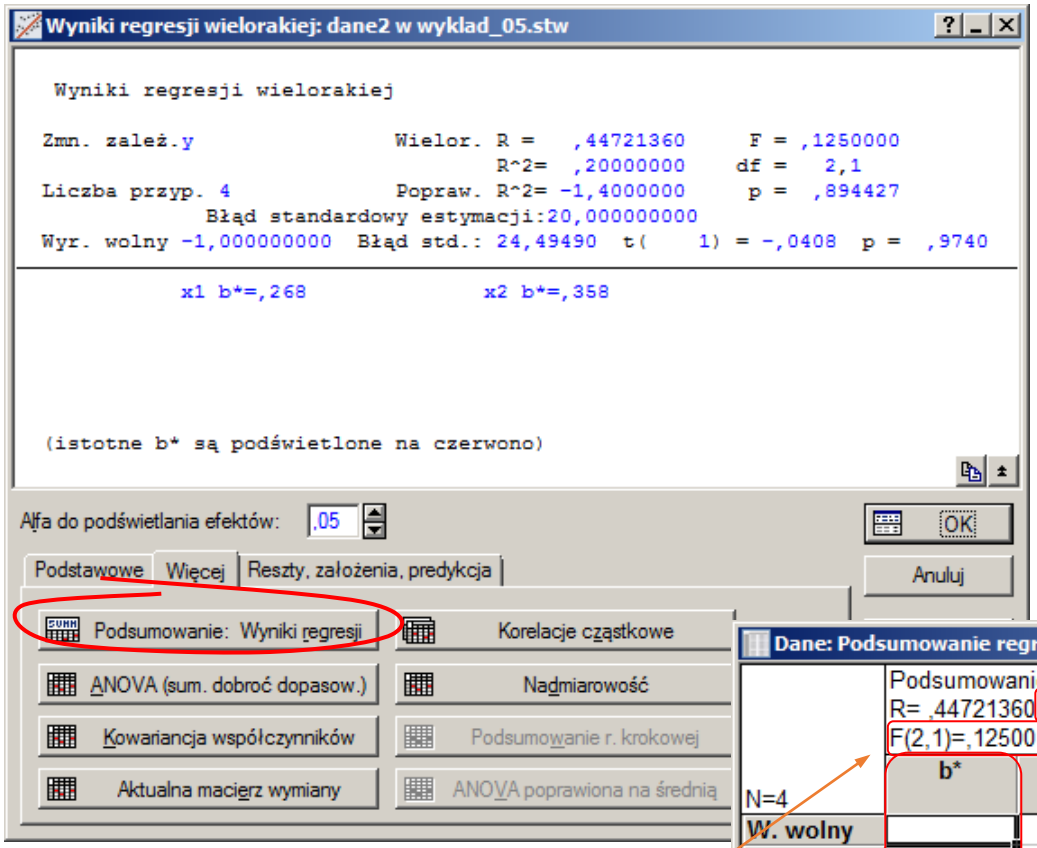
test istotności dla wyrazu wolnego

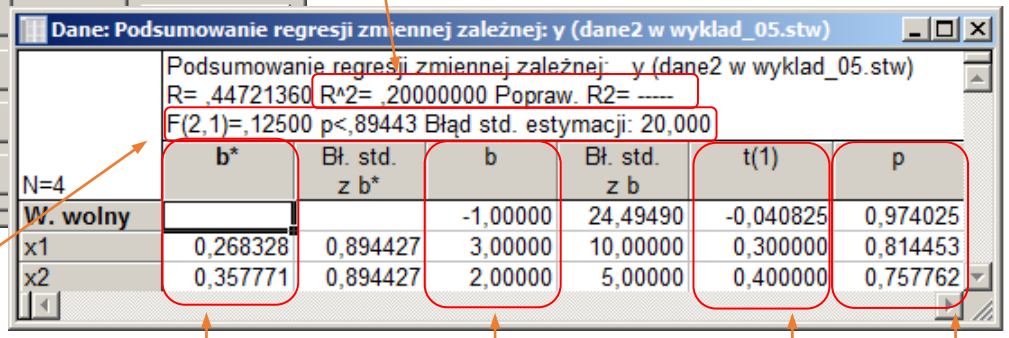
Wyr. wolny – $\hat{b}_0$

t – statystyka testowa (w nawiasie liczba stopni swobody)

p – $p-value$

STATISTICA – analiza regresji





R^2, $\overline{R}^2$

test istotności funkcji regresji

$\hat{b}_i^*$ $\hat{b}_i$ t_n $p-value$

test istotności współczynników funkcji regresji:

STATISTICA – analiza regresji

funkcja regresji nieistotna

Dane: Podsumowanie regresji zmiennej zależnej: y (dane2 w wyklad_05.stw)

Podsumowanie regresji zmiennej zależnej: y (dane2 w wyklad_05.stw)
R= ,44721360 R^2= ,20000000 Popraw. R2= -----
F(2,1)=,12500 p<,89443 Błąd std. estymacji: 20,000

N=4	b*	Bł. std. z b*	b	Bł. std. z b	t(1)	p
W. wolny			-1,00000	24,49490	-0,040825	0,974025
x1	0,268328	0,894427	3,00000	10,00000	0,300000	0,814453
x2	0,357771	0,894427	2,00000	5,00000	0,400000	0,757762

współczynniki funkcji regresji nieistotne

$$\hat{y} = -1 + 3x_1 + 2x_2$$

Dane: Podsumowanie regresji zmiennej zależnej: y (dane3 w wyklad_05.stw)

Podsumowanie regresji zmiennej zależnej: y (dane3 w wyklad_05.stw)
R= ,99980006 R^2= ,99960016 Popraw. R2= ,99880048
F(2,1)=1250,0 p<,02000 Błąd std. estymacji: ,20000

N=4	b*	Bł. std. z b*	b	Bł. std. z b	t(1)	p
W. wolny			-1,00000	0,244949	-4,08248	0,152928
x1	0,599880	0,019996	3,00000	0,100000	30,00000	0,021213
x2	0,799840	0,019996	2,00000	0,050000	40,00000	0,015912

wyraz wolny nieistotny
współczynniki $\hat{b}_1$, $\hat{b}_2$ istotne

funkcja regresji istotna