KORELACJA I REGRESJA.

KORELACJA

X , Y - cechy badane równocześnie.
Dane statystyczne zapisujemy w szeregu statystycznym dwóch cech

x_i	x_1	x_2	.....	x_n
y_i	y_1	y_2	.....	y_n

lub w **tablicy korelacyjnej**.

Y / X	y_1	y_2	.....	y_l	$n_{i.}$
x_1	n_{11}	n_{12}	.....	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	.....	n_{2l}	$n_{2.}$
....	....	....	....	....	.....
x_k	n_{k1}	n_{k2}	.....	n_{kl}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	.....	$n_{.l}$	n

gdzie

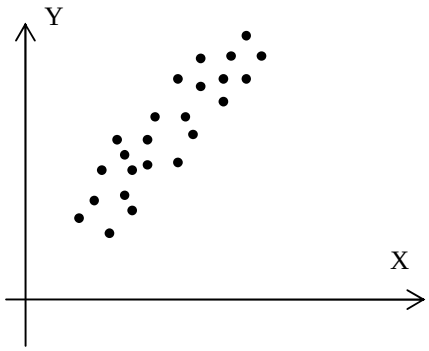
$x_1, x_2, \ldots, x_k$ - warianty lub środki klas dla cechy X,

$y_1, y_2, \ldots, y_l$ - warianty lub środki klas dla cechy Y,

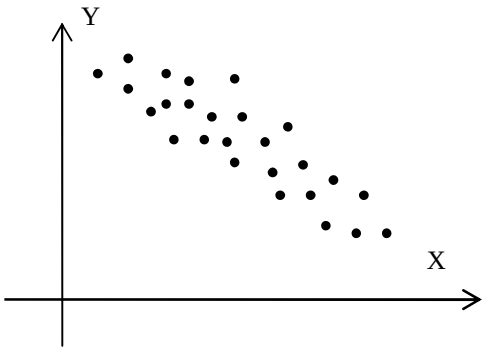
$n_{.j}$ - sumy liczebności kolumn,

$n_{i.}$ - sumy liczebności wierszy.

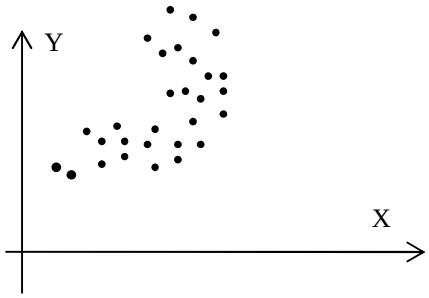
Wstępnie siłę i kształt zależności między cechami możemy ocenić na podstawie diagramu korelacyjnego:



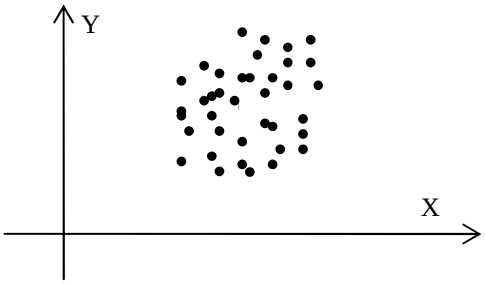
korelacja liniowa dodatnia



korelacja liniowa ujemna



korelacja krzywoliniowa



brak korelacji

Siłę zależności między cechami mierzymy **współczynnikiem korelacji liniowej Pearsona**

$$r = \frac{\operatorname{cov}\ (X, Y)}{S_X S_Y}$$

Uwaga.

$$r \in \langle -1;\ 1 \rangle$$

gdzie

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

lub (gdy dane w tablicy korelacyjnej)

$$\operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right) n_{ij} =$$

$$= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j n_{ij} - \bar{x}\,\bar{y}$$

jest **kowariancją** między cechami X i Y (kowariancja też mierzy siłę zależności między cechami, jej znak określa kierunek zależności lecz jest wielkością nieunormowaną)

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2}$$

$$S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2 - (\bar{y})^2}$$

lub (gdy dane w tablicy korelacyjnej)

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_{i.}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_{i.} - \left(\bar{x}\right)^2}$$

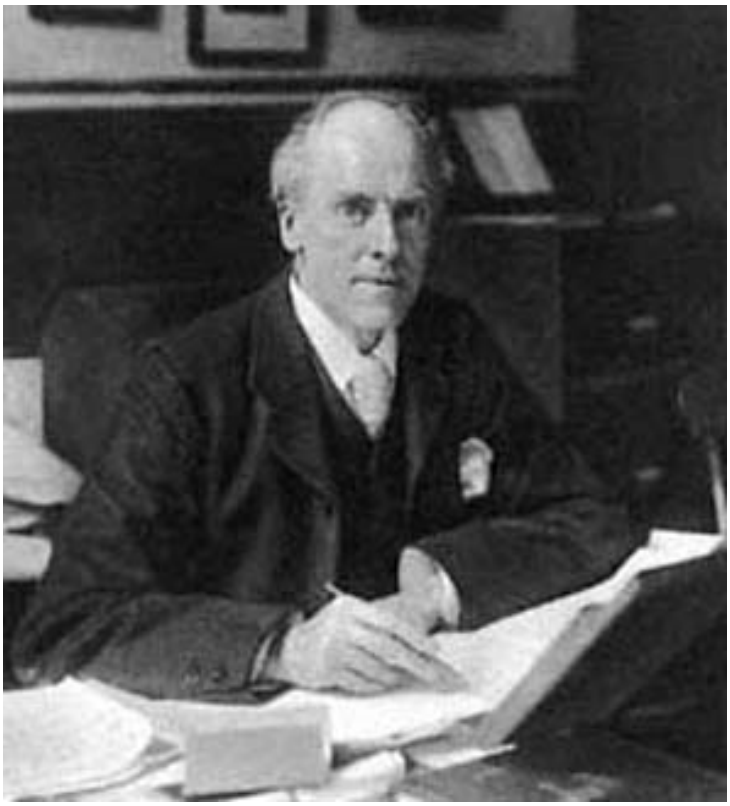
$$s_Y = \sqrt{\frac{1}{n}\sum_{j=1}^{l} n_{.j}\left(y_i - \bar{y}\right)^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{l} y_i^2 n_{.j} - \left(\bar{y}\right)^2}$$

są odchyleniami standardowymi dla cech X i Y.

Uwaga:

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

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



Karl Pearson (1857 - 1936), angielski matematyk, prekursor statystyki matematycznej

Jeśli r > 0 to mówimy, że cechy są **skorelowane dodatnio** (wzrostowi cechy X towarzyszy wzrost cechy Y),

Jeśli r < 0 to mówimy, że cechy są **skorelowane ujemnie**,(wzrostowi cechy X towarzyszy spadek cechy Y),

Jeśli r = 0 to mówimy, że cechy są **nieskorelowane**, (zmiany wartości cechy X nie powodują zmian wartości cechy Y),

Jeśli $0 < |r| < 0{,}3$ to mówimy, że cechy są skorelowane słabo,

Jeśli $0{,}3 \leq |r| < 0{,}5$ to mówimy, że cechy są skorelowane średnio,

Jeśli $0{,}5 \leq |r| < 0{,}7$ to mówimy, że cechy są skorelowane mocno,

Jeśli $0{,}7 \leq |r|$ to mówimy, że cechy są skorelowane bardzo mocno.

Powyższe przedziały mają zakres umowny.

Interpretując powyższy współczynnik korelacji należy pamiętać, że jego wartość bliska zera nie zawsze oznacza brak zależności a jedynie brak zależności liniowej. W tym przypadku należy skorzystać z wykresu lub skorzystać z innych miar zależności np. policzyć tzw. stosunki korelacyjne.

Wartość współczynnika korelacji zależy od zakresu zmienności badanych cech, podobnie jak średnia arytmetyczna podlega wpływom skrajnych wartości.

Przykład

Badano zależność wartości zużytych surowców (w tys. zł.) Y od wielkości produkcji (tys. szt.) X w 6-ciu zakładach produkcyjnych.

x_t	1	2	1,5	1	3	0,5
y_t	2	5	4	4	7	2

Wyznaczamy wartość współczynnika korelacji.

Obliczenia wykonamy w tabeli

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	2	-0,5	-2	1	0,25	4
2	5	0,5	1	0,5	0,25	1
1,5	4	0	0	0	0	0
1	4	-0,5	0	0	0,25	0
3	7	1,5	3	4,5	2,25	9
0,5	2	-1	-2	2	1	4
9	24	0	0	8	4	18

$$\bar{x} = \frac{9}{6} = 1{,}5; \qquad \bar{y} = \frac{24}{6} = 4\,; \quad r = \frac{8}{\sqrt{4}\sqrt{18}} = 0{,}9428$$

zatem związek pomiędzy wartością zużytych surowców a wielkością produkcji jest bardzo silny (korelacja dodatnia).

Przykład.
Badano zależność liczby błędów na stronie maszynopisu Y od stażu pracy X (podano środek przedziału stażu pracy) w grupie 50 sekretarek.

Y X	0	1	2	3	$n_{i.}$
4			5	10	15
12			10		10
20		10	5		15
28	5	5			10
$n_{.j}$	5	15	20	10	50

$$\bar{x} = \frac{4 \cdot 15 + 12 \cdot 10 + ... + 28 \cdot 10}{50} = \frac{760}{50} = 15{,}2;$$

$$\bar{y} = \frac{0 \cdot 5 + 1 \cdot 15 + ... + 3 \cdot 10}{50} = \frac{85}{50} = 1{,}7$$

$$S_X^2 = \frac{4^2 \cdot 15 + 12^2 \cdot 10 + \ldots + 28^2 \cdot 10}{50} - 15{,}2^2 = \frac{15520}{50} - 15{,}2^2 = 79{,}36$$

$$S_Y^2 = \frac{0^2 \cdot 5 + 1^2 \cdot 15 + \ldots + 3^2 \cdot 10}{50} - 1{,}7^2 = \frac{185}{50} - 1{,}7^2 = 0{,}81$$

$$\operatorname{cov}(X,Y) = \frac{4 \cdot 2 \cdot 5 + 4 \cdot 3 \cdot 10 + 12 \cdot 2 \cdot 10 \ldots + 28 \cdot 1 \cdot 5}{50} - 15{,}2 \cdot 1{,}7 = -7{,}04$$

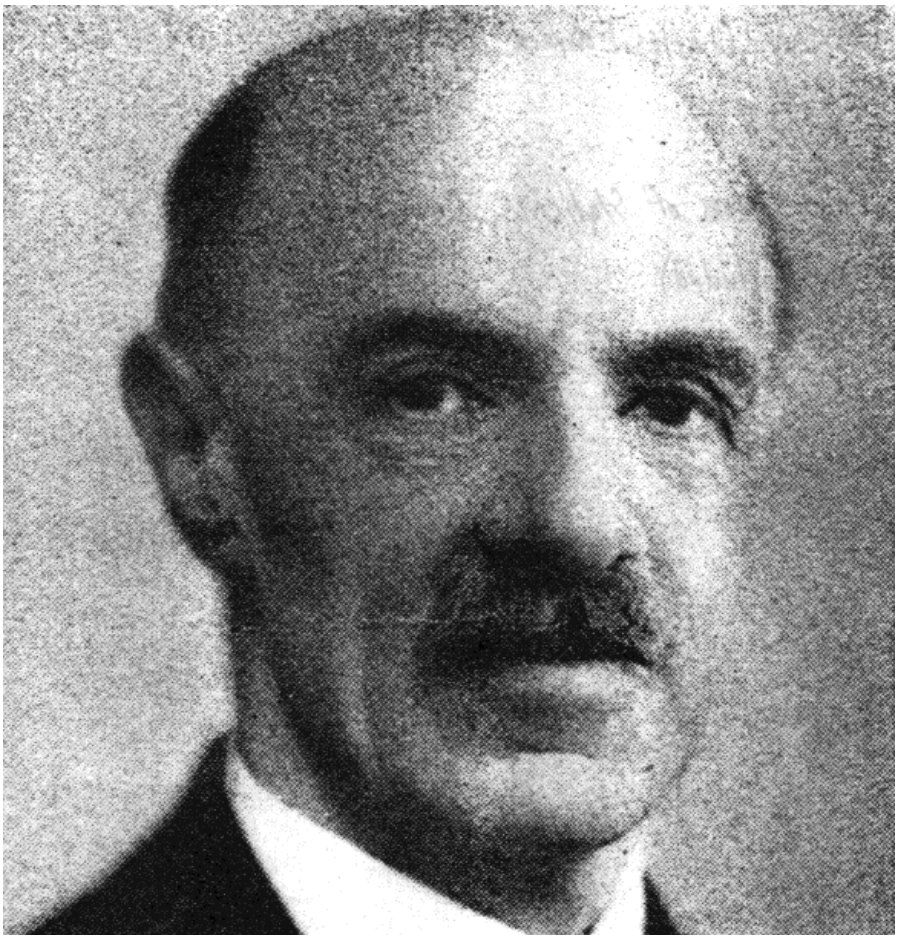
$$r = \frac{-7{,}04}{\sqrt{79{,}36}\sqrt{0{,}81}} = -0{,}878$$

zatem związek pomiędzy stażem a ilością błędów jest bardzo silny (korelacja ujemna).

Siłę zależności możemy również mierzyć **współczynnikiem korelacji rang Spearmana**:
Obserwacje numerujemy od najmniejszej do największej (nadajemy rangi). Jeśli dane powtarzają się to przypisujemy im jednakowe rangi równe średniej arytmetycznej z kolejnych numerów.

$$Q = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

gdzie d_i - różnice rang.



Charles Edward Spearman (1863 - 1945)
angielski psycholog i statystyk

Współczynnik ten stosujemy w przypadku małej liczby danych lub w przypadku cech niemierzalnych, których wartości można uporządkować.

W przypadku cech niemierzalnych można mierzyć siłę zależności współczynnikiem Cramera lub Czuprowa (definicja będzie podana przy teście niezależności chi kwadrat).

Przykład.

Dwóch członków komisji przetargowej A i B oceniało nadesłane oferty. Członek A oceniał jakość ofert opisowo natomiast członek B przydzielał im punkty od 0 do 100.

Oferta	Ocena A	Ocena B	Ranga oceny A	Ranga oceny B	d_i	d_i^2
I	**mniej niż przeciętna**	50				
II	**słaba**	45				
III	**dobra**	25				
IV	**przeciętna**	30				
V	**bardzo dobra**	25				
VI	**bardzo słaba**	42				
VII	**przeciętna**	40				
Razem	x	x				

Oferta	Ocena A	Ocena B	Ranga oceny A	Ranga oceny B	d_i	d_i^2
I	mniej niż przeciętna	50	3	7		
II	słaba	45	2	6		
III	dobra	25	6	1,5		
IV	przeciętna	30	4,5	3		
V	bardzo dobra	25	7	1,5		
VI	bardzo słaba	42	1	5		
VII	przeciętna	40	4,5	4		
Razem	x	x	x	x		

Oferta	Ocena A	Ocena B	Ranga oceny A	Ranga oceny B	d_i	d_i^2
I	mniej niż przeciętna	50	3	7	-4	16
II	słaba	45	2	6	-4	16
III	dobra	25	6	1,5	4,5	20,25
IV	przeciętna	30	4,5	3	1,5	2,25
V	bardzo dobra	25	7	1,5	5,5	30,25
VI	bardzo słaba	42	1	5	-4	16
VII	przeciętna	40	4,5	4	0,5	0,25
Razem	x	x	x	x	0	101

$$Q = 1 - \frac{6 \cdot 101}{7^3 - 7} = -0{,}8$$

Wynika stąd zupełny brak zgodności ocen obu członków komisji (bardzo silna korelacja ujemna).

REGRESJA LINIOWA

Regresja to kształt zależności między badanymi cechami. Interesuje nas najprostsza zależność w postaci funkcji liniowej.

Wyznaczymy prostą

$$\hat{Y} = b_0 + b_1 X$$

Najlepiej „dopasowaną” do danych (x_i, y_i)

Y - zmienna objaśniana, y_i - wartości (obserwacje) zmiennej Y; $i = 1,...,n$ - numer obserwacji,

X - zmienna objaśniająca, x_i - wartości zmiennej X,

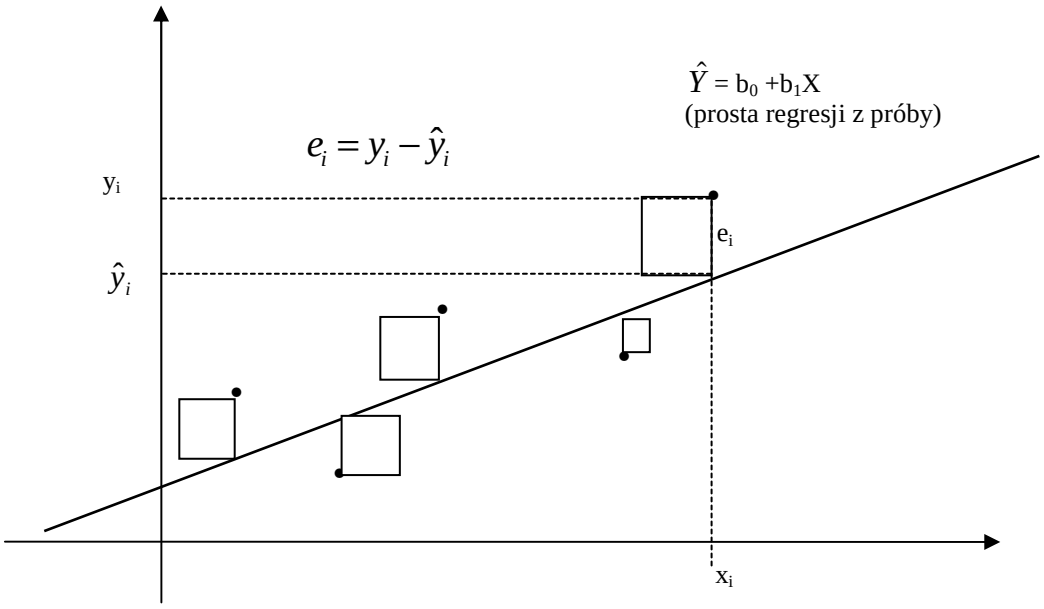
b_0, b_1 - parametry strukturalne (ich wartość wyznacza się na podstawie obserwacji (x_i, y_i))

Aby wyznaczyć wartość parametrów strukturalnych b_0, b_1 na podstawie próby stosujemy metodę najmniejszych kwadratów (MNK).

MNK polega na wyznaczeniu takich

$$b_0, b_1$$

aby dla danych obserwacji (x_i, y_i) suma kwadratów odchyleń zaobserwowanych wartości y_i od wartości $\hat{Y} = b_0 + b_1 X$ była minimalna, tzn. chcemy wyznaczyć minimum funkcji:


$\hat{Y}$ = b0 +b1X
(prosta regresji z próby)
$e_i = y_i - \hat{y}_i$
yi
$\hat{y}_i$
ei
xi

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

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

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

Uwaga.

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

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

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

$$\frac{\partial S}{\partial b_0} = 2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)(-1) = -2\left(\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i - nb_0\right) = 0$$

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

rozwiązując otrzymany układ równań otrzymamy wzory na przybliżone wartości parametrów strukturalnych

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} =$$

$$= \frac{\frac{1}{n}\sum x_i y_i - \overline{xy}}{\frac{1}{n}\sum x_i^2 - (\overline{x})^2} = \frac{\sum \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum \left(x_i - \overline{x}\right)^2} =$$

$$= \frac{S_Y}{S_X} r = \frac{\mathrm{cov}(X,Y)}{S_X^2}$$

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

Prostą

$$\hat{Y} = b_0 + b_1 X$$

nazywamy **prostą regresji z próby**.

Miary dopasowania.

Wariancja resztowa:

Wariancja resztowa to uśrednienie pól kwadratów zbudowanych na resztach
i odzwierciedla 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

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

czyli

$$S_e^2 = \frac{\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i}{n-2} =$$

$$= \frac{n}{n-2}\left(1 - r^2\right) S_Y^2$$

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

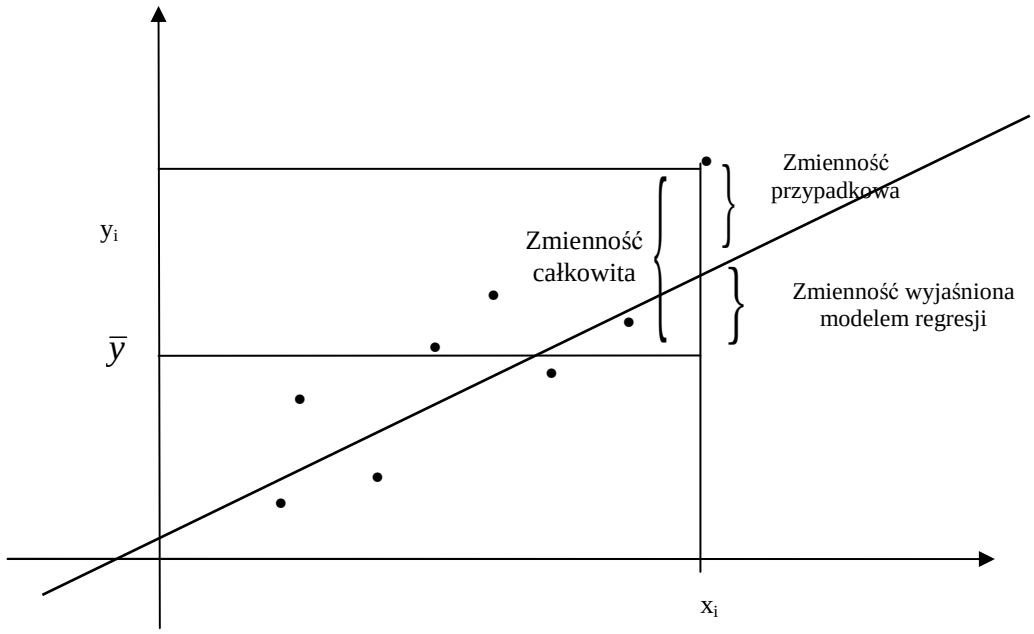
Dopasowanie modelu do danych empirycznych można oceniać odchyleniem standardowym reszt lecz jest to miara bezwzględna i nieunormowana, dlatego do porównań lepsze są miary względne lub unormowane.

Najprostszą względną miarą dopasowania jest **współczynnik zmienności resztowej**:

$$V_e = \frac{S_e}{\overline{Y}} 100\%$$

Współczynnik ten informuje jaką część średniej wartości badanego zjawiska stanowi odchylenie standardowe reszt.

Mniejsze wartości tego współczynnika wskazują na lepsze dopasowanie modelu do danych empirycznych, niekiedy żąda się aby np. $V_e < 0{,}2$.



Wprowadzamy oznaczenia:

Całkowita suma kwadratów (zmienność całkowita): $CSK = \sum (y_i - \bar{y})^2$

Wyjaśniona suma kwadratów (zmienność wyjaśniona): $WSK = \sum (\hat{y}_i - \bar{y})^2$

Niewyjaśniona suma kwadratów (zmienność przypadkowa): $NSK = \sum e_i^2$

$gdzie: \hat{y}_i = b_0 + b_1 x_i$

Własność: $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2$

Czyli CSK = WSK + NSK

Miarą dopasowania modelu do rzeczywistości (wartości zaobserwowanych) jest również współczynnik determinacji R^2

Współczynnik determinacji:

$$\boxed{R^2 = \frac{WSK}{CSK}} \qquad R^2 \in \langle 0, 1 \rangle$$

<u>współczynnik ten określa jaka część całkowitej zmienności zmiennej objaśnianej została wyjaśniona przez model regresji liniowej.</u>

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

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

$$= \frac{b_1\left(\sum x_i y_i - n\bar{x}\bar{y}\right)}{\sum y_i^2 - n(\bar{y})^2} = \frac{\mathrm{cov}^2(X,Y)}{S_X^2 S_Y^2} = r^2$$

Przykład

Badano zależności kosztów całkowitych (w tys. zł.) Y od wielkości produkcji (tys. szt.) X w 6-ciu zakładach produkcyjnych.

x_i	4	8	6	4	12	2
y_i	2	5	4	4	7	2

Dla $\hat{Y} = b_0 + b_1 x$ wyznaczamy przybliżone wartości parametrów strukturalnych i współczynnik determinacji.

Obliczenia wykonamy w tabeli

x_i	y_i	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
4	2			
8	5			
6	4			
4	4			
12	7			
2	2			
36	24			

x_i	y_i	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
4	2	4	4	4
8	5	2	4	1
6	4	0	0	0
4	4	0	4	0
12	7	18	36	9
2	2	8	16	4
36	24	32	64	18

$$\bar{x} = \frac{36}{6} = 6; \qquad \bar{y} = \frac{24}{6} = 4;$$

$$b_1 = \frac{32}{64} = 0{,}5; \quad b_0 = 4 - 0{,}5 * 6 = 1$$

zatem związek pomiędzy kosztami całkowitymi a wielkością produkcji wyraża się zależnością liniową w postaci

$$\hat{Y} = 1 + 0{,}5X$$

Współczynnik determinacji

$$R^2 = \frac{16}{18} = 0{,}89$$

należy oczekiwać, że rozpatrywany model wyjaśnia 89% całkowitej zmienności badanego zjawiska.

Standardowe błędy oszacowania parametrów strukturalnych.

$$S(b_1) = \frac{S_e}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{S_e}{\sqrt{n}\, S_X}$$

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

$$= S(b_1) \cdot \sqrt{S_X^2 + (\bar{x})^2} = \frac{S_e}{\sqrt{n} \cdot \sqrt{1 + \frac{(\bar{x})^2}{S_X^2}}}$$

Stosujemy niekiedy zapis

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

Uwaga.
W celu dokładniejszego zbadania kształtu zależności między cechami można wykonać wykresy **empirycznych linii regresji**.
Są to łamane wyznaczone przez średnie warunkowe:

$$\bar{x}_j = \frac{\sum_{i=1}^{k} x_i n_{ij}}{n_{.j}}$$

(tzn. obliczamy średnią wartość X przy ustalonej wartości y_j)

$$\bar{y}_i = \frac{\sum_{j=1}^{l} y_j n_{ij}}{n_{i.}}$$

(tzn. obliczamy średnią wartość Y przy ustalonej wartości x_i)

Regresja Y względem X

$$(x_1, \bar{y}_1); (x_2, \bar{y}_2); .....; (x_k, \bar{y}_k)$$

Regresja X względem Y

$$(\bar{x}_1, y_1); (\bar{x}_2, y_2); .....; (\bar{x}_l, y_l)$$

Łamane te przecinają się w punkcie $(\bar{x}, \bar{y})$. Im bliżej siebie są położone tym silniejszy jest związek miedzy cechami.

Przykład.
Badano zależność wartości sprzedaży Y (mln zł) od wydatków na reklamę X (tys. zł) w grupie 100 firm.

Y / X	3-5	5-7	7-9	9-11	11-13	13-15	$n_{i.}$
50-100	10	13					23
100-150		10	27	11			48
150-200			1	9	8	2	20
200-250					4	5	9
$n_{.j}$	10	23	28	20	12	7	100

$$\bar{x} = \frac{13250}{100} = 132{,}5; \quad \bar{y} = \frac{844}{100} = 8{,}44$$

Zestawienie średnich warunkowych:

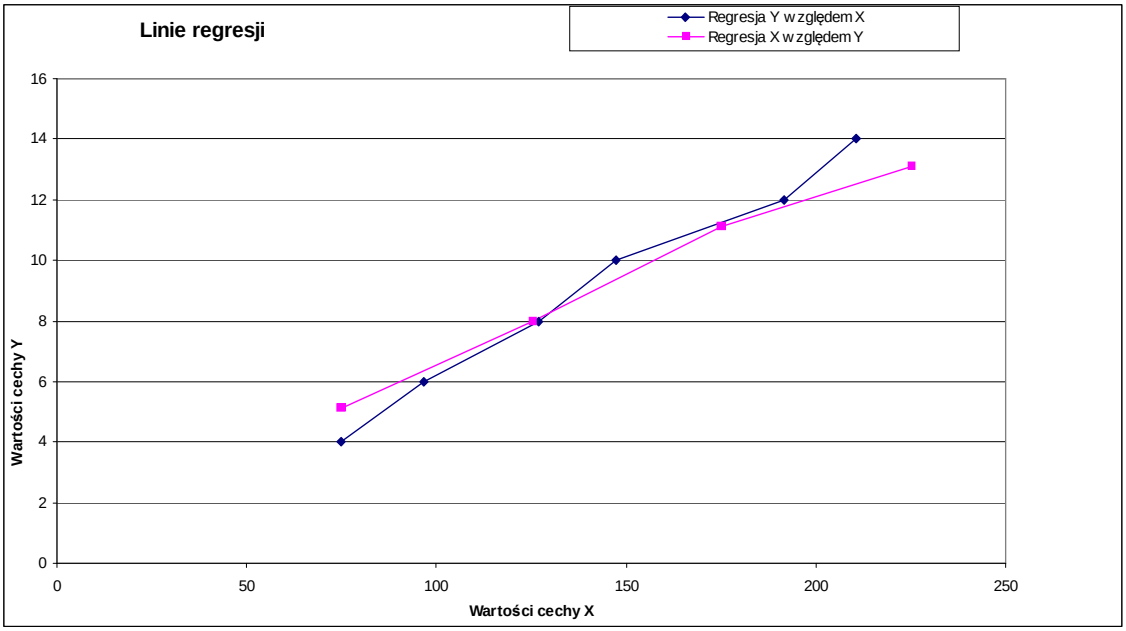
$(x_j, \bar{y}_j)$

75	4
96,7	6
126,8	8
147,5	10
191,7	12
210,7	14

$(x_j, \bar{y}_j)$

75	5,1
125	8
175	11,1
225	13,1

Wykres empirycznych linii regresji.



W przypadku gdy wykres danych w układzie współrzędnych wskazuje na brak zależności liniowej możemy próbować dobrać funkcję nieliniową do opisu zależności między cechami.

Równość wariancyjna.

$$S^2(y) = S^2(\bar{y}_i) + \overline{S_i^2(y)}$$

gdzie

$S^2(y)$ - wariancja cechy Y

$S^2(\bar{y}_i)$ - wariancja międzygrupowa $\quad S^2(\bar{y}_i) = \dfrac{\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_{i.}}{n}$

mierzy zróżnicowanie cechy Y wywołane oddziaływaniem cechy X. Jest to wariancja średnich warunkowych Y(X = x_i).

$\overline{S_i^2(y)}$ - wariancja wewnątrzgrupowa $\quad \overline{S_i^2(y)} = \dfrac{\sum_{i=1}^{k} s_i^2(y) n_{i.}}{n}$

mierzy zróżnicowanie cechy Y wywołane oddziaływaniem czynników poza cechą X. Jest to średnia ważona rozkładów warunkowych Y(X = x_i).

Stosunek korelacyjny

$$e_{yx} = \frac{S(\bar{y}_i)}{S(y)}$$

mierzy siłę zależności cechy Y względem cechy X.

Analogicznie stosunek korelacyjny

$$e_{xy} = \frac{S(\bar{x}_i)}{S(x)}$$

mierzy siłę zależności cechy X względem cechy Y.

Stosunki korelacyjne pokazują siłę związku, lecz nie informują o jego kierunku.

Przyjmują wartości z przedziału [0, 1]. Wartości e_{yx} i e_{xy} są na ogół różne. Różnica między kwadratem stosunku korelacyjnego a kwadratem współczynnika korelacji Pearsona (zwany **wskaźnikiem krzywoliniowości**) mierzy stopień krzywoliniowości regresji:

$m_{yx} = e_{yx}^2 - r^2$ zmiennej Y względem X,

$m_{xy} = e_{xy}^2 - r^2$ zmiennej X względem Y,

Niekiedy przyjmuje się, że jeśli wskaźnik krzywoliniowości jest nie większy niż 0,2 to wpływ jednej cechy na drugą jest liniowy i można stosować regresję liniową, w przeciwnym przypadku lepiej stosować regresję nieliniową.

Prognoza.

Prognoza punktowa

τ - moment (okres prognozy)

x_τ - wartość cechy X w okresie prognozy

$$\boxed{y_\tau^* = b_0 + b_1 x_\tau}$$

Standardowy błąd prognozy

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

Uwaga

1) $S_\tau > S_e$

2) S_τ jest minimalne dla $x_\tau = \bar{x}$

błąd względny prognozy:

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

Model tendencji rozwojowej

Gdy X jest zmienną czasową $x_i = t$ (t = 1, 2, ..., n) tzn. model regresji ma postać

$$\hat{Y} = b_0 + b_1 t$$

wówczas taki model nazywamy **modelem tendencji rozwojowej** lub **modelem trendu liniowego.**

Wtedy korzystając z własności:

(*) $\sum_{t=1}^{n} t = \frac{n(n+1)}{2}$, $\sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}$,

$\bar{t} = \frac{n+1}{2}$ $\sum (t-\bar{t})^2 = \sum t^2 - n(\bar{t})^2$

mamy

$$b_1 = \frac{n\sum ty_t - \sum t \sum y_t}{n\sum t^2 - \left(\sum t\right)^2} =$$

$$= \frac{12\sum (t-\bar{t})y_t}{n(n^2-1)} = \frac{12\left(\sum ty_t - \bar{t}\sum y_t\right)}{n(n^2-1)}$$

$$b_0 = \bar{y} - b_1 \bar{t} = \bar{y} - b_1 \frac{n+1}{2}$$

Wariancja resztowa

Niech $e_i = y_i - \hat{y}_i$, (gdzie $\hat{y}_i = b_0 + b_1 t$) to reszty modelu, wtedy

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

czyli $$\boxed{s_e^2 = \frac{\sum_{t=1}^{n} y_t^2 - b_0 \sum_{t=1}^{n} y_t - b_1 \sum_{t=1}^{n} t y_t}{n-2}}$$

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

Dopasowanie modelu do danych empirycznych oceniamy też współczynnikiem determinacji

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

$$= \frac{b_0 \sum y_i + b_1 \sum t y_t - n\bar{y}^2}{\sum y_t^2 - n(\bar{y})^2} = \frac{b_1\left(\sum t y_t - n\bar{t}\,\bar{y}\right)}{\sum y_t^2 - n(\bar{y})^2} = r^2$$

Prognoza dla modelu trendu

Niech t_τ – okres prognozy.

Prognoza punktowa y_τ^* to przewidywana wartość cechy Y w okresie t_τ.

$$\boxed{y_\tau^* = b_0 + b_1 t_\tau}$$

Standardowy błąd prognozy punktowej

$$s_\tau = s_e \sqrt{1 + \frac{1}{n} + \frac{(t_\tau - \bar{t})^2}{\sum_{t=1}^{n} (t - \bar{t})^2}} = s_e \sqrt{1 + \frac{\sum_{t=1}^{n} t^2 + n t_\tau^2 - 2 t_\tau \sum_{t=1}^{n} t}{n \sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t \right)^2}}$$

Wzór ten można uprościć korzystając z własności (*).

$$s_\tau = s_e \sqrt{1 + \frac{\frac{n(n+1)(2n+1)}{6} + nt_\tau^2 - 2t_\tau \frac{n(n+1)}{2}}{\frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}}} =$$

$$= s_e \sqrt{1 + \frac{2(2n+1) + \frac{12t_\tau^2}{n+1} - 12t_\tau}{n^2 - n}}$$

Zatem należy traktować wartość prognozy jako

$$y_\tau^* \pm s_\tau$$

Jakość prognozy punktowej możemy ocenić względnym błędem prognozy punktowej

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

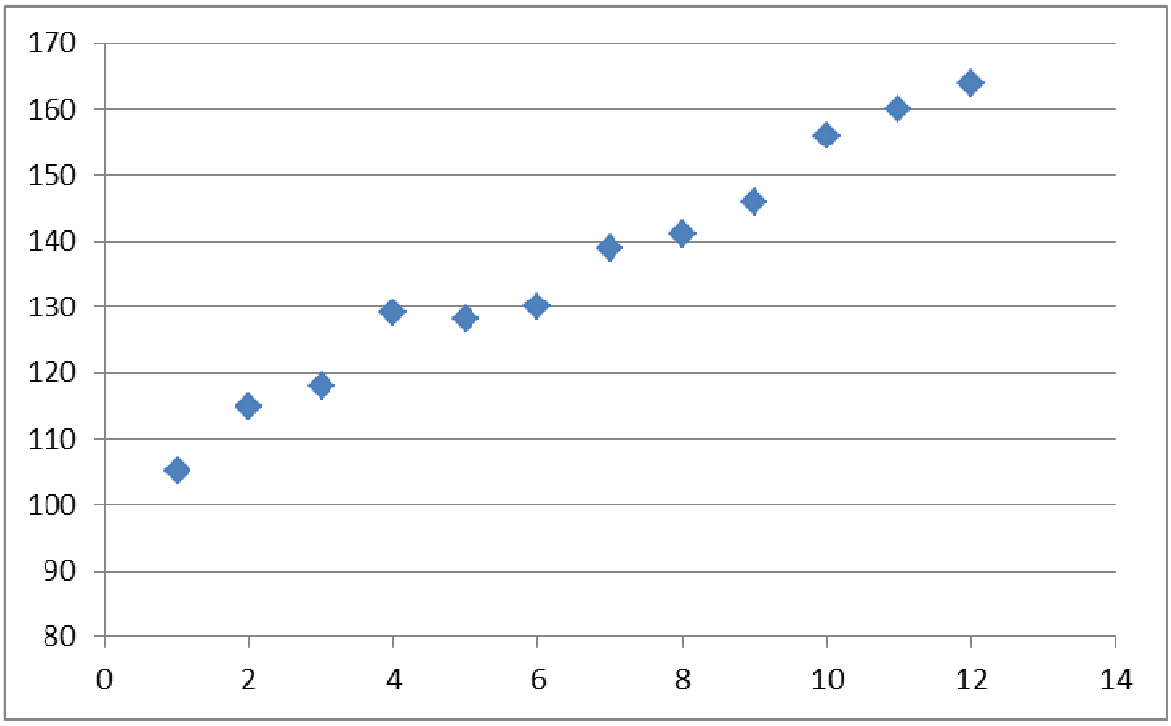
Przykład

Y – wielkość sprzedaży (tys. szt.).

Dane z kolejnych półroczy 2003-2008:

105, 115, 118, 129, 128, 130, 139, 141, 146, 156, 160, 164.

Wyznaczyć prognozę na pierwsze półrocze 2010 roku i ocenić jej dokładność.


170
160
150
140
130
120
110
100
90
80
0
2
4
6
8
10
12
14

t	y_t	ty_t	y_t^2
1	**105**		
2	**115**		
3	**118**		
4	**129**		
5	**128**		
6	**130**		
7	**139**		
8	**141**		
9	**146**		
10	**156**		
11	**160**		
12	**164**		

t	y_t	ty_t	y_t^2
1	105	105	11025
2	115	230	13225
3	118	354	13924
4	129	516	16641
5	128	640	16384
6	130	780	16900
7	139	973	19321
8	141	1128	19881
9	146	1314	21316
10	156	1560	24336
11	160	1760	25600
12	164	1968	26896
78	**1631**	**11328**	**225449**

tśr	6,5
yśr	135,9167
b1	5,08042
b0	102,8939

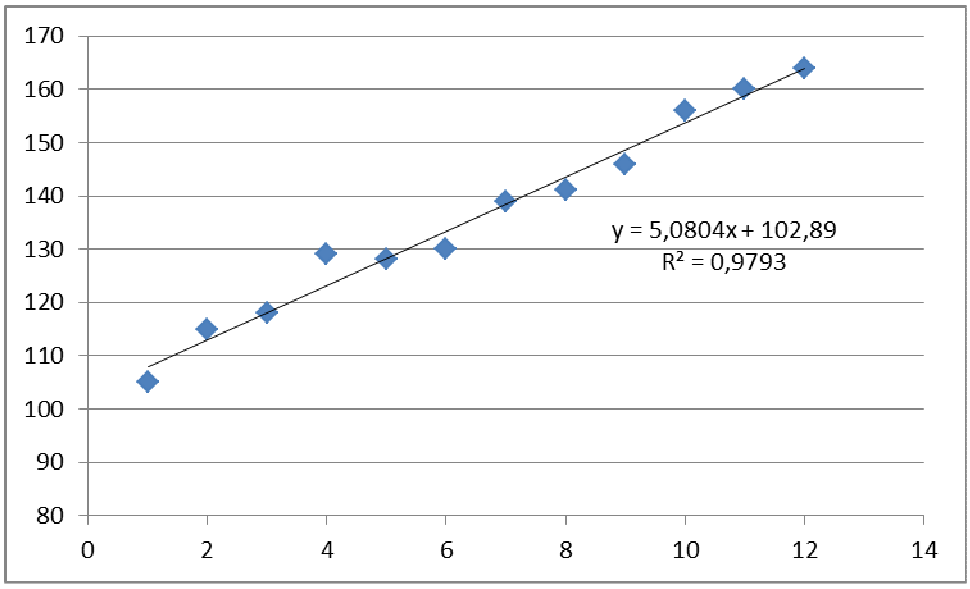
Se^2	7,799184
Se	2,792702

tt	15

yt*	179,1002

St	3,662272

d_{pkt}	2,04%


170
160
150
140
130
120
110
100
90
80
0
2
4
6
8
10
12
14
y = 5,0804x + 102,89
R² = 0,9793