

Analiza ZALEŻNOŚCI pomiędzy CECHAMI (Analiza KORELACJI i REGRESJI)

- korelacyjny wykres rozrzutu (korelogram)
- rodzaje zależności (brak, nieliniowa, liniowa)
- pomiar siły zależności liniowej (współczynnik korelacji Pearsona, współczynnik korelacji rang Spearmana)
- liniowa funkcja regresji

Badamy jednostki statystyczne pod kątem dwóch różnych cech - cechy X oraz cechy Y.

Pytanie jakie sobie stawiamy to:

czy istnieje zależność pomiędzy cechą X i cechą Y ?

Jeżeli taka zależność istnieje, to poszukujemy odpowiedzi na kolejne pytania:

- jaki jest charakter tej zależności oraz
- jaka jest jej siła ?

Zależność korelacyjna pomiędzy cechami X i Y charakteryzuje się tym, że wartościom jednej cechy są przyporządkowane ściśle określone wartości średnie drugiej cechy.

Informacja statystyczna niezbędna do zbadania zależności pomiędzy cechami X i Y przyjmuje najczęściej 2 formy:

- szereg(i) szczegółowy par informacji o cechach X oraz Y; ma on postać ciągu par $\{ (x_i, y_i) \}$,
- szereg rozdzielczy w postaci tzw. tablicy korelacyjnej.

Korelacyjny wykres rozrzutu KORELOGRAM

Jeżeli obie cechy X i Y są mierzalne, to analizę zależności rozpoczynamy od sporządzenia korelogramu.

Korelogram jest to wykres punktowy par $\{ (x_i, y_i) \}$.

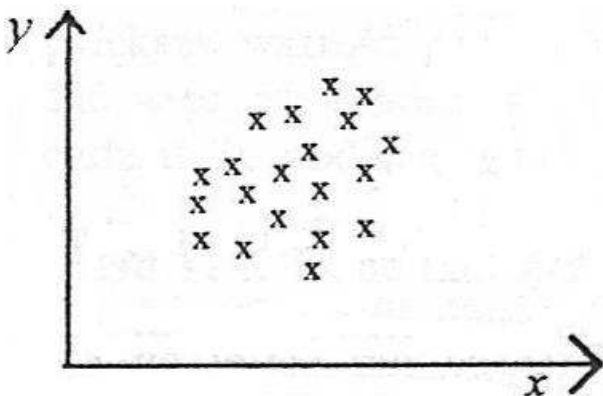
(Excel nazywa taki wykres: „wykresem XY”).

W kartezjańskim układzie współrzędnych xOy pary te odpowiadają punktom o współrzędnych

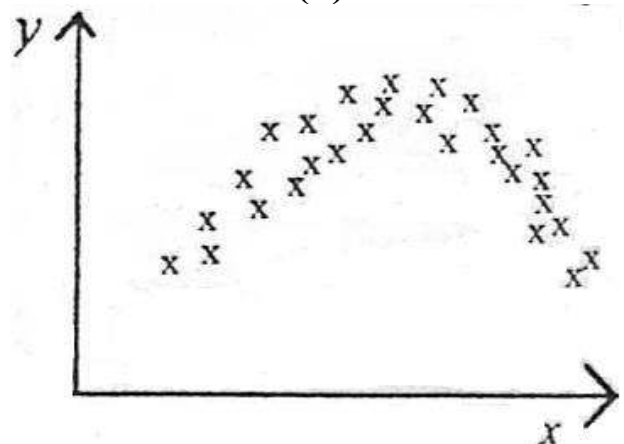
$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

PRZYKŁADY korelogramów (każdy punkt oznaczono x)

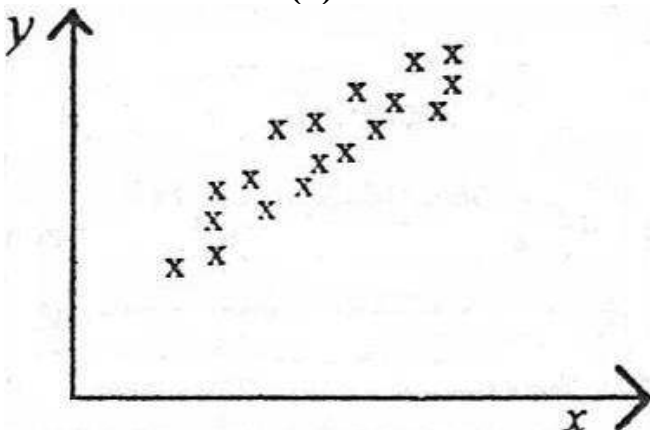
(a)



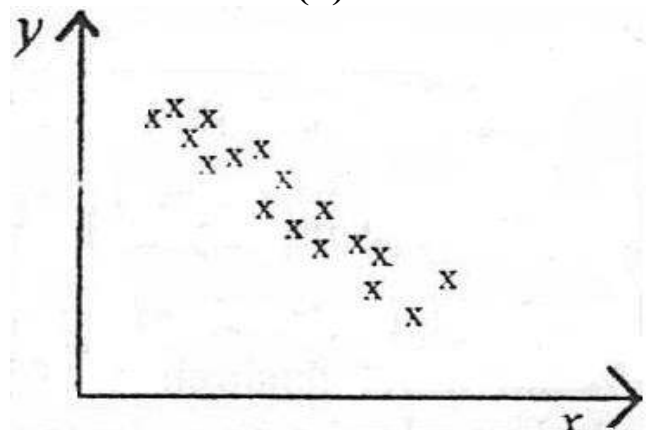
(b)



(c)



(d)



Jeżeli otrzymamy bezlądny zbiór punktów, który nie przypomina kształtem wykresu znanego związku funkcyjnego, to powiemy że pomiędzy cechami X i Y nie ma zależności. Ilustruje to rysunek (a).

Na rysunku (b) widać, że smuga punktów układa się w kształt paraboli. Powiemy zatem, że istnieje zależność pomiędzy cechami X i Y i jest to związek nieliniowy; zależność nieliniowa.

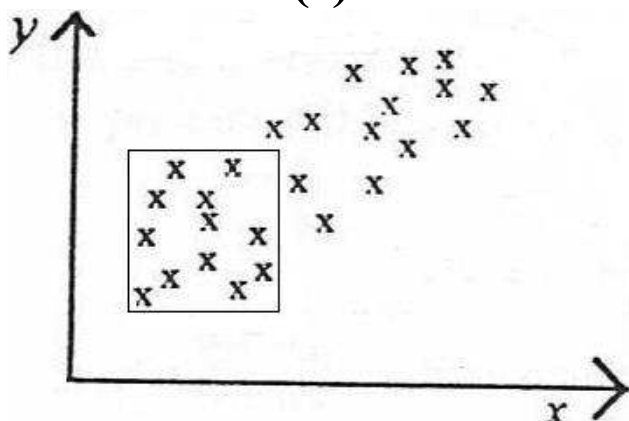
Na rysunkach (c) i (d) smuga punktów układa się wzdłuż linii prostej. Powiemy zatem, że istnieje zależność pomiędzy cechami X i Y i jest to związek liniowy; zależność liniowa.

Rysunki (e) i (f) ilustrują przypadki błędów we wnioskowaniu o zależności cech X i Y na podstawie korelogramu.

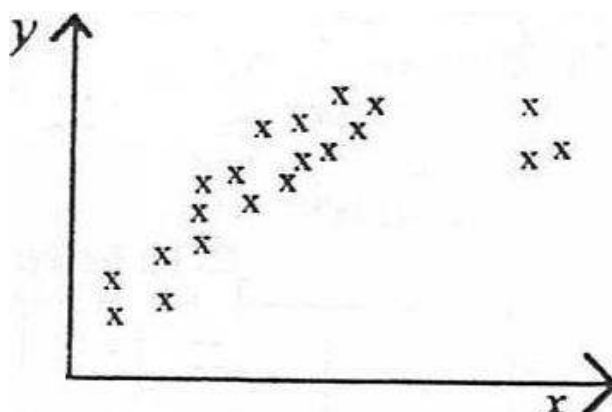
Rysunek (e) – za mało danych. Zebrano dane (punkty obwiedzione kwadratem) i z korelogramu wynika brak zależności. W rzeczywistości jest zależność liniowa.

Rysunek (f) – nietypowe dane. Trzy ostatnie punkty (odseparowane) to dane nietypowe. Sugerują zależność nieliniową (parabola). Po odrzuceniu tych nietypowych informacji widać, że jest wyraźna zależność liniowa.

(e)



(f)



Pomiar KIERUNKU i SIŁY zależności liniowej Szeregi szczegółowe

WSPÓŁCZYNNIK KORELACJI (Pearsona)

Współczynnik korelacji (Pearsona) r_{xy} obliczamy dla cech ilościowych wg następującego wzoru:

$$r_{xy} = \frac{C(X, Y)}{S_x S_y}$$

gdzie:

$C(X, Y)$ – kowariancja pomiędzy cechami X i Y

s_x (s_y) – odchylenie standardowe cechy X (cechy Y)

Kowariancja jest kluczowym parametrem rozkładu dwóch cech w badaniu zależności cech ilościowych X i Y. Wylicza się ją wg następującego wzoru (dla szeregu(ów) szczegółowego):

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

Współczynnik korelacji (Pearsona) r_{xy} spełnia zawsze warunek:

$$-1 \leq r_{xy} \leq 1$$

Współczynnik korelacji (Pearsona) jest miarą symetryczną, tzn.

$$r_{xy} = r_{yx}$$

INTERPRETACJA współczynnika korelacji r_{xy}

Znak współczynnika r_{xy} mówi nam o kierunku zależności. I tak:

- znak plus – zależność liniowa dodatnia, tzn. wraz ze wzrostem wartości jednej cechy rosną średnie wartości drugiej z cech,
- znak minus – zależność liniowa ujemna, tzn. wraz ze wzrostem wartości jednej cechy maleją średnie wartości drugiej z cech.

Wartość bezwzględna współczynnika korelacji, czyli $|r_{xy}|$,

mówi nam o sile zależności. Jeżeli wartość bezwzględna $|r_{xy}|$:

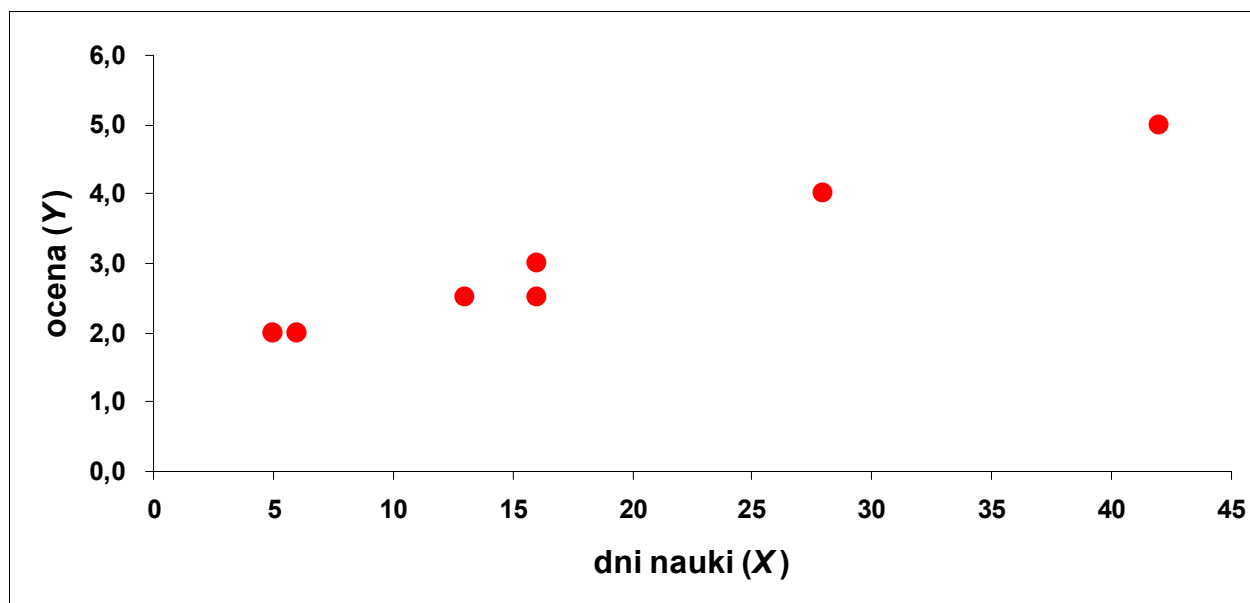
- jest mniejsza od 0,2, to praktycznie brak związku liniowego pomiędzy badanymi cechami,
- 0,2 – 0,4 - zależność liniowa wyraźna, lecz niska,
- 0,4 – 0,7 - zależność liniowa umiarkowana,
- 0,7 – 0,9 - zależność liniowa znacząca,
- powyżej 0,9 - zależność liniowa bardzo silna.

PRZYKŁAD 1

W grupie 7 studentów badano zależność pomiędzy oceną z egzaminu ze statystyki (Y), a liczbą dni poświęconych na naukę (X).

<i>nr studenta</i>	<i>ocena z egzaminu (Y)</i>	<i>liczba dni nauki (X)</i>
<i>i</i>	<i>y_i</i>	<i>x_i</i>
1	2,0	5
2	2,5	13
3	2,5	16
4	4,0	28
5	5,0	42
6	3,0	16
7	2,0	6

Sporządzamy korelogram.



Widać tutaj wyraźną zależność liniową (dodatnią).

Obliczamy współczynnik korelacji (Pearsona).

UWAGA ! Liczebność populacji jest mała ($n=7$). Użyjemy tak małego przykładu tylko dlatego, aby sprawnie zilustrować procedurę liczenia.

Obliczanie średnich, wariancji oraz kowariancji.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i	y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	(4)*(4)	(5)*(5)	(4)*(5)
1	2,0	5					
2	2,5	13					
3	2,5	16					
4	4,0	28	1,0	10	1,00	100	10,0
5	5,0	42	2,0	24	4,00	576	48,0
6	3,0	16	0,0	-2	0,00	4	0,0
7	2,0	6	-1,0	-12	1,00	144	12,0
<i>razem</i>			X	X			

$$n = 7 \quad \bar{x} = \frac{126}{7} = 18 \quad \bar{y} = \frac{21}{7} = 3$$

$$s_x^2 = \frac{1022}{7} = 146$$

$$s_y^2 = \frac{7,5}{7} = 1,07$$

$$s_x = \sqrt{146} = 12,08$$

$$s_y = \sqrt{1,07} = 1,03$$

$$C(X, Y) = \frac{86,5}{7} = 12,36$$

Współczynnik korelacji (Pearsona) wynosi dla danych z przykładu 1:

$$r_{xy} = \frac{C(X, Y)}{s_x s_y} = \frac{12,36}{12,08 \times 1,03} = +0,993$$

INTERPRETACJA

W badanej grupie studentów wystąpiła bardzo silna dodatnia (znak *plus*) zależność liniowa pomiędzy czasem nauki (cecha X), a uzyskaną oceną z egzaminu (cecha Y).

Oznacza to, że wraz ze wzrostem czasu poświęconego na naukę rosła w tej grupie uzyskiwana ocena.

WSPÓŁCZYNNIK KORELACJI RANG (Spearmana)

Współczynnik korelacji rang (Spearmana) r_S używamy w przypadku gdy:

1. choć jedna z badanych cech jest cechą jakościową (niemierzalną), ale istnieje możliwość uporządkowania (ponumerowania) wariantów każdej z cech;
2. cechy mają charakter ilościowy (mierzalny), ale liczebność zbiorowości jest mała ($n < 30$).

Numery jakie nadajemy wariantom cech noszą nazwę rang.

UWAGA ! W procesie nadawania rang stymulanty porządkujemy malejąco, a destymulanty rosnąco.

UWAGA ! W procesie nadawania rang może zdarzyć się więcej niż 1 jednostka o takiej samej wartości cechy (np. k jednostek).

Wówczas należy na chwilę nadać tym jednostkom kolejne rangi.

Następnie należy zsumować takie rangi i podzielić przez k (otrzymamy w ten sposób średnią rangę dla tej grupy k jednostek).

W ostateczności każda jednostka z tych k jednostek otrzyma identyczną rangę (średnią dla danej grupy k jednostek).

Współczynnik korelacji rang (Spearmana) r_S wyznaczamy wg następującego wzoru:

$$r_S = 1 - \frac{6 \times \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

d_i – różnica pomiędzy rangami dla cechy X i cechy Y

Współczynnik korelacji rang (Spearmana) r_S spełnia zawsze warunek:

$$-1 \leq r_S \leq 1$$

INTERPRETACJA

Analogiczna jak dla współczynnika korelacji (Pearsona).

PRZYKŁAD 2

Dla danych z przykładu 1 obliczenia współczynnika korelacji rang (Spearmana) są następujące:

(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	y_i	x_i	<i>rangi cechy Y</i>	<i>rangi cechy X</i>	d_i	d_i^2
1	2,0	5				
2	2,5	13				
3	2,5	16				
4	4,0	28			0,0	0,00
5	5,0	42			0,0	0,00
6	3,0	16			0,5	0,25
7	2,0	6			-0,5	0,25
<i>razem</i>	X	X	X	X	X	

$$r_S = 1 - \frac{6 \times \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 2}{7(7^2 - 1)} = +0,964$$

Wartość współczynnika korelacji rang (Spearmana) potwierdza bardzo silną, dodatnią (znak *plus*) zależność pomiędzy czasem nauki (X), a uzyskaną oceną (Y).

Pomiar KIERUNKU i SIŁY zależności liniowej Szeregi rozdzielcze

TABLICA KORELACYJNA

Schemat tablicy korelacyjnej

Warianty cechy X (x_i)	Warianty cechy Y (y_j)				(razem) $n_{i\bullet}$
	y_1	y_1	\dots	y_s	
x_1	n_{11}	n_{12}	\dots	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	\dots	n_{2s}	$n_{2\bullet}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_r	n_{r1}	n_{r2}	\dots	n_{rs}	$n_{r\bullet}$
(razem) $n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet s}$	n

Oznaczenia:

n_{ij} - liczba jednostek, która charakteryzuje się wartością x_i cechy X oraz wartością y_j cechy Y

$n_{i\bullet}$ - liczba jednostek, która charakteryzuje się wartością x_i cechy X

$$n_{i\bullet} = \sum_{j=1}^s n_{ij}$$

$n_{\bullet j}$ - liczba jednostek, która charakteryzuje się wartością y_j cechy Y

$$n_{\bullet j} = \sum_{i=1}^r n_{ij}$$

n - liczebność populacji

$$n = \sum_{i=1}^r \sum_{j=1}^s n_{ij} = \sum_{i=1}^r n_{i\bullet} = \sum_{j=1}^s n_{\bullet j}$$

PRZYKŁAD 3

Podobnie jak w przykładzie 1 zbadamy zależność pomiędzy czasem nauki (X), a uzyskaną oceną (Y).

W tablicy korelacyjnej zestawiono informację o 400 studentach ($n=400$).

Ocena (Y)	Czas nauki (X) w dniach				$n_{i\bullet}$
	0 - 7	7 - 14	14 - 21	21 - 28	
2	80				
3	10	80			
3,5		60	10		
4		20	30		
4,5			50	10	
5				50	
$n_{\bullet j}$					

Obliczamy osobno dla każdej z cech: średnie, wariancje i odchylenia standardowe.

Ocena (Y)	Czas nauki (X)				(a) $n_{i\bullet}$	$y_i n_{i\bullet}$	(b) $y_i - \bar{y}$	(c) (b)*(b)	(d) (c)*(a)
	0 - 7	7 - 14	14 - 21	21 - 28					
2	80				80				
3	10	80			90				
3,5		60	10		70				
4		20	30		50	200	0,5	0,25	12,5
4,5			50	10	60	270	1	1	60
5				50	50	250	1,5	2,25	112,5
$n_{\bullet j}$	90	160	90	60	400		x	x	
\dot{x}_j			17,5	24,5	x	x	x	x	x
$\dot{x}_j n_{\bullet j}$			1575	1470		x	x	x	x
$\dot{x}_j - \bar{x}$			4,9	11,9	x	x	x	x	x
$\left(\dot{x}_j - \bar{x}\right)^2$			24,01	141,61	x	x	x	x	x
$\left(\dot{x}_j - \bar{x}\right)^2 n_{\bullet j}$			2160,9	8496,6		x	x	x	x

$$n = 400$$

$$\bar{x} = \frac{5040}{400} = 12,6$$

$$\bar{y} = \frac{1395}{400} = 3,5$$

$$s_x^2 = \frac{18816}{400} = 47,04$$

$$s_y^2 = \frac{387,5}{400} = 0,97$$

$$s_x = \sqrt{47,04} = 6,86$$

$$s_y = \sqrt{0,97} = 0,98$$

Przechodzimy do obliczania kowariancji $C(X, Y)$.

Na początek policzymy wszystkie iloczyny $(\dot{x}_j - \bar{x})(y_i - \bar{y})$:

$y_i - \bar{y}$	$\dot{x}_j - \bar{x}$			
	-9,1	-2,1	4,9	11,9
-1,5				
-0,5				
0	0	0	0	0
0,5	-4,55	-1,05	2,45	5,95
1	-9,1	-2,1	4,9	11,9
1,5	-13,65	-3,15	7,35	17,85

Wykorzystamy tabelę początkową:

Ocena (Y)	Czas nauki (X) w dniach			
	0 - 7	7 - 14	14 - 21	21 - 28
2	80	0	0	0
3	10	80	0	0
3,5	0	60	10	0
4	0	20	30	0
4,5	0	0	50	10
5	0	0	0	50

i policzymy wszystkie iloczyny $(\dot{x}_j - \bar{x})(y_i - \bar{y})n_{ij}$

$y_i - \bar{y}$	$\dot{x}_j - \bar{x}$				<i>razem</i>
	-9,1	-2,1	4,9	11,9	
-1,5					
-0,5					
0	0	0	0	0	
0,5	0	-21	73,5	0	
1	0	0	245	119	
1,5	0	0	0	892,5	
<i>razem</i>					

Zatem kowariancja wynosi:

$$C(X, Y) = \frac{2530,5}{400} = 6,33$$

Współczynnik korelacji (Pearsona) wynosi dla danych z przykładu 3:

$$r_{xy} = \frac{C(X, Y)}{s_x s_y} = \frac{6,33}{6,86 \times 0,98} = +0,942$$

INTERPRETACJA

W badanej grupie 400 studentów wystąpiła **bardzo silna dodatnia** (znak *plus*) **zależność liniowa** pomiędzy czasem nauki (cecha X), a uzyskaną oceną z egzaminu (cecha Y).

Inne miary zależności wyliczalne na podstawie tablicy korelacyjnej

Obok współczynnika korelacji Personna stosowane są inne miary zależności pomiędzy cechą Y i cechą X. Są to:

- Stosunek korelacji (e_{yx})
- Miary oparte na chi-kwadrat (χ^2)

Stosunek korelacji

- Miara ta jest oparta na spostrzeżeniu, że przy braku zależności średnie poziomy cechy Y wewnątrz grup (klas) pokrywają się ze średnią ogólną cechy Y.
- Miara ta spełnia warunki

$$0 < e_{yx} < 1$$

$$|r_{yx}| \leq e_{yx}$$

- Warunkiem policzenia stosunku korelacji jest mierzalność cechy Y.
- Jest to miara zalecana w przypadku badania zależności dla związków nieliniowych.

Miary oparte na chi-kwadrat

- Miary te oparte są na badaniu różnic pomiędzy liczebnościami empirycznymi a liczebnościami teoretycznymi, które wyliczane są przy założeniu niezależności cechy Y i cechy X.
- Do tej grupy należą współczynniki (por. wykład 10):
 - C – Personna
 - Q – Yule’a
 - T – Czuprowa
- V – Cramera

REGRESJA PROSTA

Ważnym uzupełnieniem zagadnienia badania kierunku i siły zależności pomiędzy cechami X i Y jest analiza regresji.

Przez analizę regresji rozumiemy metodę badania wpływu zmiennych uznanych za niezależne (przyczyny) na zmienną uznana za zależną (skutek).

Jeżeli w analizie uwzględnimy tylko 1 zmienną niezależną, to mówimy o REGRESJI PROSTEJ.

Cecha X (zmienna niezależna) - przyczyna.

Cecha Y (zmienna zależna) - skutek.

Przypadek większej liczby zmiennych niezależnych będzie rozwinięty w przedmiocie „Ekonometria” (dla słuchaczy kierunku Zarządzanie).

Podstawowym narzędziem badania jest tutaj funkcja regresji.

Rozważymy tylko przypadek zależności liniowej dla regresji prostej. Narzędziem będzie zatem funkcja regresji postaci:

$$\hat{y}_i = ax_i + b$$

\hat{y}_i - teoretyczna wartość zmiennej zależnej (Y)

x_i - empiryczna wartość zmiennej niezależnej (X)

a – współczynnik regresji (współczynnik kierunkowy)

INTERPRETACJA: jeżeli wartość zmiennej niezależnej X wzrośnie o jednostkę, to wartość zmiennej zależnej Y :

- wzrośnie (jeżeli $a > 0$) o $|a|$ jednostek lub
- spadnie (jeżeli $a < 0$) o $|a|$ jednostek.

b – wyraz wolny

INTERPRETACJA: stały poziom wartości zmiennej zależnej Y niezależny od zmian wartości zmiennej niezależnej X.

Uwaga ! Interpretacja wyrazu wolnego nie zawsze ma sens ekonomiczny.

Zauważmy, że liniowa funkcja trendu (omówiona w wykładzie 6)

$$\hat{y}_t = at + b$$

może być również traktowana jako liniowa funkcja regresji prostej.

Zmienna zależna Y opisuje tam poziom badanego zjawiska Y .

Zmienną niezależną X jest tam czas (zmienna czasowa t).

W efekcie podstawiając X zamiast t oraz zmieniając wskaźnik t

na wskaźnik i otrzymamy funkcję regresji

$$\hat{y}_i = ax_i + b$$

W nowym układzie funkcja trendu może być traktowana jako funkcja regresji Y względem czasu t .

Szacowanie parametrów a i b funkcji regresji

$$a = \frac{C(X, Y)}{s_x^2}$$

$$b = \bar{y} - a\bar{x}$$

PRZYKŁAD 4

Dla danych z przykładu 1 szacowanie parametrów funkcji regresji przebiega następująco:

$$\bar{x} = 18 \quad \bar{y} = 3 \quad s_x^2 = 146 \quad C(X, Y) = 12,36$$

$$a = \frac{C(X, Y)}{s_x^2} = \frac{12,36}{146} = 0,085$$

$$b = \bar{y} - a\bar{x} = 3 - 0,085 \times 18 = 1,47$$

Funkcja regresji w przykładzie 1 ma więc postać:

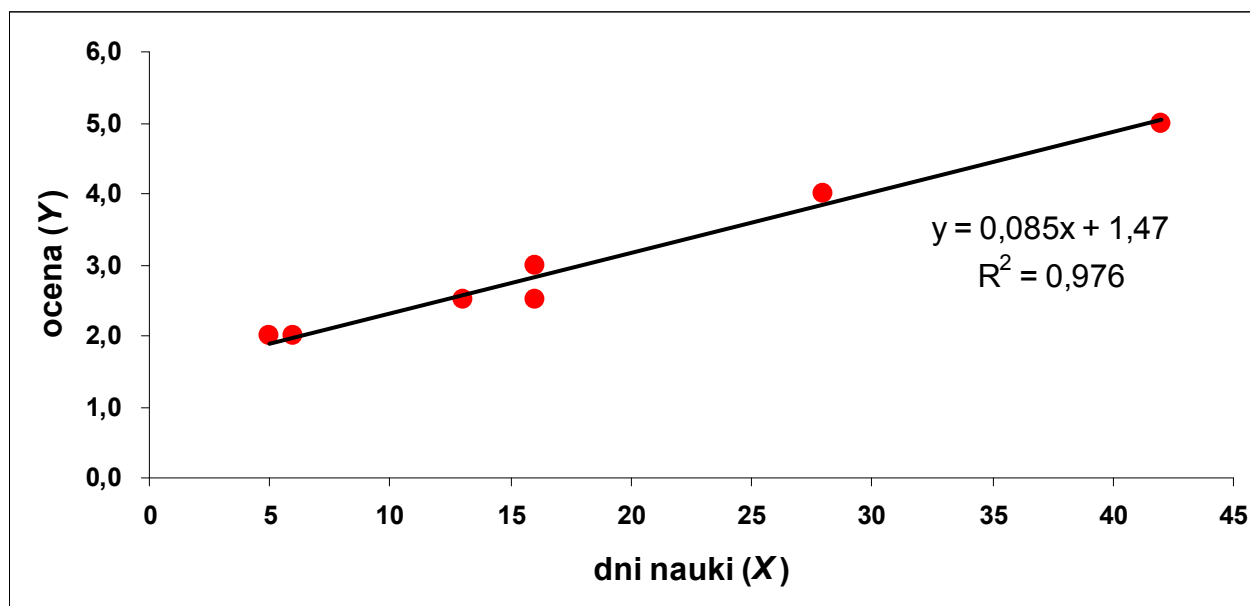
$$\hat{y}_i = 0,085 \cdot x_i + 1,47$$

INTERPRETACJA:

współczynnik regresji ($a=0,085 > 0$) - jeżeli liczba dni nauki wzrośnie o jednostkę (o 1 dzień), to ocena z egzaminu wzrośnie o 0,085 (inaczej: każdy dzień nauki podnosi średnio ocenę o 0,085)

wyraz wolny ($b=1,47$) - stały, niezależny od liczby dni nauki ($x=0$) poziom uzyskanej oceny z egzaminu to 1,47 (poniżej niedostatecznej)

Otrzymaną funkcję regresji, wykreśloną na korelogramie pokazano na rysunku:

**Wykorzystanie funkcji regresji do prognozowania**

Słuchacz o numerze 8 (przypomnijmy, że badanie przeprowadzono dla $n=7$ studentów) poświęcił na naukę 20 dni ($x_8=20$).

Jakiej oceny może spodziewać się (średnio) przy takim nakładzie czasu na naukę ?

$$\hat{y}_8 = 0,085 \cdot x_8 + 1,47 = 0,085 \times 20 + 1,47 = 3,17$$

Poświęcając 20 dni na naukę słuchacz może spodziewać się (średnio !!!) oceny 3,17 czyli „dst+”.

Ocena dopasowania funkcji regresji do danych empirycznych

Problem oceny dopasowania był już częściowo omawiany (wykład 6) przy okazji analitycznego wygładzania szeregu czasowego za pomocą liniowej funkcji trendu.

Podstawowymi miarami „dobroci” dopasowania linii regresji do danych empirycznych są:

- współczynnik zbieżności (φ^2)
- współczynnik determinacji (R^2)
- średni błąd szacunku (pierwiastek z tzw. wariancji resztowej)

Współczynnik zbieżności (φ^2):

$$\varphi^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{gdzie} \quad 0 \leq \varphi^2 \leq 1$$

Im φ^2 jest bliższy 0, tym dopasowanie jest lepsze.

Współczynnik determinacji (R^2):

$$R^2 = 1 - \varphi^2 \quad \text{gdzie} \quad 0 \leq R^2 \leq 1$$

Przy zależności liniowej można go wyznaczyć również jako:

$$R^2 = r_{xy}^2 \quad \text{lub} \quad R^2 = r_{yx}^2$$

Im R^2 jest bliższy 1, tym dopasowanie jest lepsze.

Średni błąd szacunku (S_e):

$$S_e = \sqrt{S_e^2} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k}}$$

gdzie:

k – liczba szacowanych parametrów funkcji regresji

(tutaj $k=2$; szacujemy dwa parametry: a i b)

Jest to pierwiastek z wariancji resztowej (S_e^2).

Nazwa bierze się od reszty (e_i), którą definiuje się jako:

różnicę pomiędzy wartością empiryczną, a wartością teoretyczną cechy zależnej Y :

$$e_i = y_i - \hat{y}_i$$

PRZYKŁAD 5

Ocena dopasowania funkcji regresji dla danych z przykładu 1.

$$\hat{y}_i = 0,085 \cdot x_i + 1,47$$

$$\bar{y} = 3$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i	y_i	x_i	\hat{y}_i	$(y_i - \bar{y})$	$(y_i - \hat{y}_i)$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	2,0	5					
2	2,5	13					
3	2,5	16	2,83	-0,5	-0,33	0,25	0,1089
4	4,0	28	3,85	1,0	0,15	1,00	0,0225
5	5,0	42	5,04	2,0	-0,04	4,00	0,0016
6	3,0	16	2,83	0,0	0,17	0,00	0,0289
7	2,0	6	1,98	-1,0	0,02	1,00	0,0004
razem	X	X	X	X	X		

Współczynnik zbieżności

$$\varphi^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{0,1787}{7,5} = 0,024$$

Współczynnik determinacji

$$R^2 = 1 - \varphi^2 = 1 - 0,024 = 0,976$$

lub wg innego wzoru

$$R^2 = r_{xy}^2 = (0,993)^2 = 0,986$$

Uwaga! Różnice w wartości współczynnika determinacji wynikają z błędów zaokrągleń na etapie liczenia współczynników: zbieżności i korelacji

Średni błąd szacunku

$$S_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k}} = \sqrt{\frac{0,1787}{7 - 2}} = 0,189$$

W celu wyrobienia sobie poglądu nt. wielkości tego błędu odniesiemy go średniego poziomu cechy Y:

$$S_e / \bar{y} \times 100\% = 0,189 / 3 \times 100\% = 6,3\%$$

Uwaga! Nie można użyć znanego współczynnika zmienności (V_x) ponieważ średnia wartość reszt jest teoretycznie równa 0. Wystąpiłoby zatem dzielenie przez zero.

PODSUMOWANIE (przykład 5)

Wszystkie policzone miary dopasowania potwierdzają bardzo dobre dopasowanie funkcji regresji do danych empirycznych.

PRZYKŁAD 6

Na zakończenie wyznaczmy funkcję regresji dla danych z przykładu 3. Badaniu poddano tam 400 studentów. Wcześniej otrzymaliśmy tam:

$$n = 400 \quad \bar{x} = 12,6 \quad \bar{y} = 3,5 \quad s_x^2 = 47,04$$

$$C(X, Y) = 6,33 \quad r_{xy} = 0,942$$

Parametry funkcji regresji wynoszą:

$$a = \frac{C(X, Y)}{s_x^2} = \frac{6,33}{47,04} = 0,135$$

$$b = \bar{y} - a\bar{x} = 3,5 - 0,135 \times 12,6 = 1,799$$

Funkcja regresji w przykładzie 3 ma postać:

$$\hat{y}_i = 0,135 \cdot x_i + 1,799$$

Dobroć dopasowania do danych empirycznych mierzona współczynnikiem determinacji wynosi:

$$R^2 = r_{xy}^2 = (0,942)^2 = 0,887$$

Powyższa funkcja regresji w 88,7% objaśnia kształtowanie się oceny z egzaminu (Y) w zależności od czasu nauki (X).

WYKORZYSTANIE funkcji regresji do PROGNOZY oceny.

Słuchacz o numerze 401 poświęcił na naukę 20 dni ($x_{401}=20$).

Jakiej oceny może się spodziewać (średnio) ?

$$\hat{y}_{401} = 0,135 \cdot x_{401} + 1,799 = 0,135 \times 20 + 1,799 = 4,499$$

Poświęcając 20 dni na naukę słuchacz może spodziewać się (średnio !!!) oceny 4,499 czyli „db+”.