



# Korelacja i regresja liniowa

Statystyka i analiza danych 2017/2018

---

Jurek Błaszczński,  
na podstawie slajdów Wojtka Kotłowskiego  
13 maja 2018

- **Populacyjna:**

$$C(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

gdzie  $\mu_X, \mu_Y$  są średnimi zmiennych  $X$  i  $Y$ .

$C(X, Y)$  mierzy **zależność liniową** dwóch zmiennych losowych.

Szczególny przypadek: wariancja  $D^2[X] = C(X, X)$ .

- **Próbkowa** - dla zbioru  $n$  par  $(X_1, Y_1), \dots, (X_n, Y_n)$ :

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

- To unormowana kowariancja – **populacyjna**:

$$\rho(X, Y) = \frac{C(X, Y)}{D[X]D[Y]}, \quad \rho(X, Y) \in [-1, 1]$$

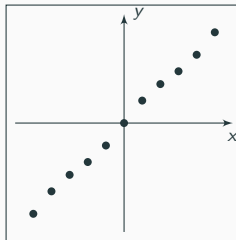
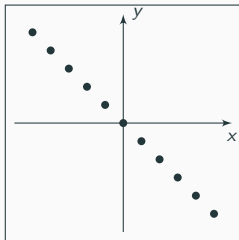
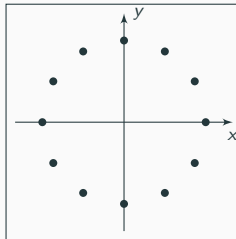
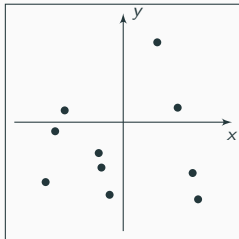
Jeśli  $X, Y$  – niezależne, to  $\rho(X, Y) = 0$  (ale nie odwrotnie!)

- **Próbkowa** - dla zbioru  $n$  par  $(X_1, Y_1), \dots, (X_n, Y_n)$ :

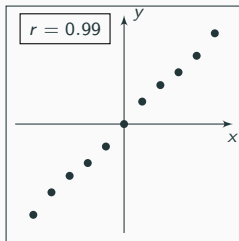
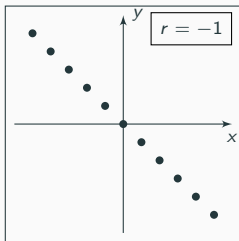
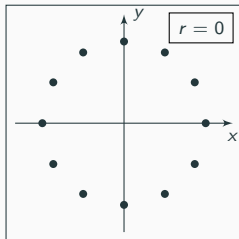
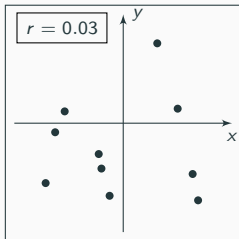
$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(\sum_{i=1}^n (X_i - \bar{X})^2)(\sum_{i=1}^n (Y_i - \bar{Y})^2)}}$$

Zachodzi  $r \in [-1, 1]$ , wartości skrajne  $\{-1, 1\}$  przyjmowane są **wtedy i tylko wtedy** gdy  $Y_i$  są funkcją liniową  $X_i$  (lub odwrotnie)

# Przykłady korelacji



# Przykłady korelacji



# Test na istotność korelacji

- Układ hipotez:

najczęściej			
$H_0 :$	$\rho = 0$	$(\rho \geq 0)$	$(\rho \leq 0)$
$H_1 :$	$\rho \neq 0$	$\rho < 0$	$\rho > 0$

- Statystyka testowa:

$$T = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \sim t(n-2)$$

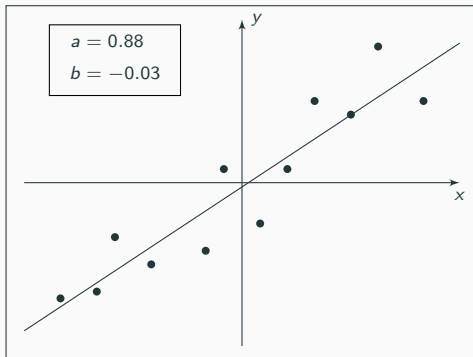
Wartość krytyczną (lub  $p$ -wartość) otrzymujemy z rozkładu t-Studenta z  $n - 2$  stopniami swobody.

# Regresja liniowa

Korelacja mierzy **siłę zależności liniowej**.

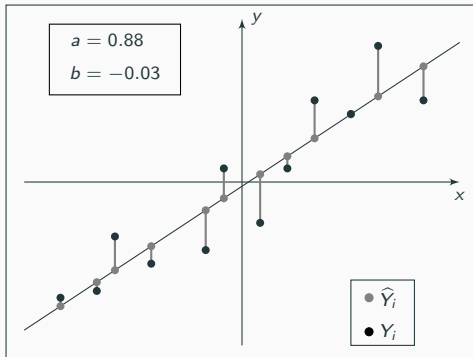
Regresja to wyznaczanie **współczynników zależności liniowej**:

Mając zbiór  $n$  punktów  $(X_1, Y_1), \dots, (X_n, Y_n)$  wyznacz współczynniki  $a, b$  zależności liniowej  $Y = aX + b$ .



# Metoda najmniejszych kwadratów

Dla każdego  $X_i$  błąd modelu liniowego to różnica między wartością odczytaną z prostej  $\hat{Y}_i = aX_i + b$  a prawdziwą wartością  $Y_i$ :



Minimalizujemy **sumę kwadratów błędów**:

$$a, b \leftarrow \min_{a, b} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - aX_i - b)^2$$



$$L(a, b) = \sum_{i=1}^n (Y_i - aX_i - b)^2$$

Przyrównujemy pochodne cząstkowe do zera:

$$\frac{\partial L}{\partial b} = 0 \iff -\sum_{i=1}^n 2(Y_i - aX_i - b) = 0 \iff b = \bar{Y} - a\bar{X}$$

$$\begin{aligned} \frac{\partial L}{\partial a} = 0 &\iff -\sum_{i=1}^n 2(Y_i - aX_i - b)X_i = 0 \\ &\iff \sum_{i=1}^n (Y_i - \bar{Y})X_i = a \sum_{i=1}^n (X_i - \bar{X})X_i \\ &\iff \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) = a \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) \\ &\iff a = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}. \end{aligned}$$

$$a = \frac{s_{XY}}{s_X^2} = r \frac{s_Y}{s_X}$$
$$b = \bar{Y} - a\bar{X}$$

- Linia regresji przechodzi przez punkt  $(\bar{X}, \bar{Y})$
- Współczynnik kierunkowy  $a$  ma **ten sam znak** co współczynnik korelacji  $r$ .