Korelacja i regresja

Na podstawie materiałów WK

Miary populacyjne

- **Kowariancja**:

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

Mierzy **zależność liniową** dwóch zmiennych losowych.
Szczególny przypadek: wariancja $D^2[X] = C(X,X)$.

- **Korelacja** – unormowana kowariancja:

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

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

Miary próbkowe

Dla zbioru n par $(X_1, Y_1), \ldots, (X_n, Y_n)$:

- **Kowariancja**:

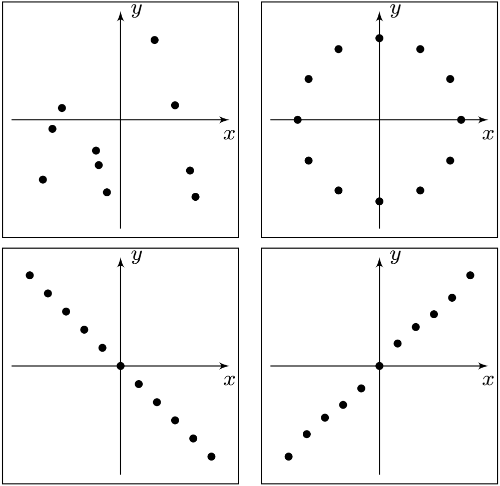
$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

- **Korelacja**:

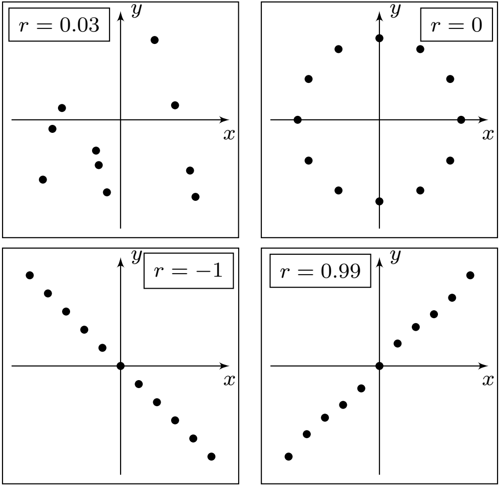
$$r \;=\; \frac{s_{XY}}{s_X s_Y} \;=\; \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^{n} (X_i - \overline{X})^2\right)\left(\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right)}}$$

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**:

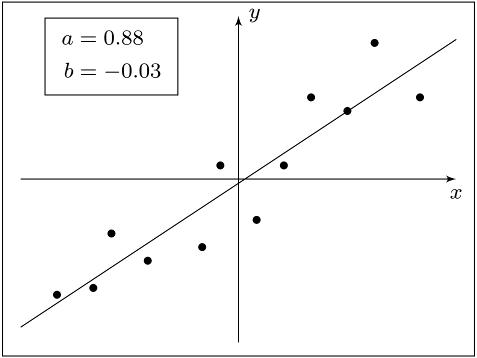
$$T = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2} \;\sim\; t(n-2)$$

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

Regresja liniowa

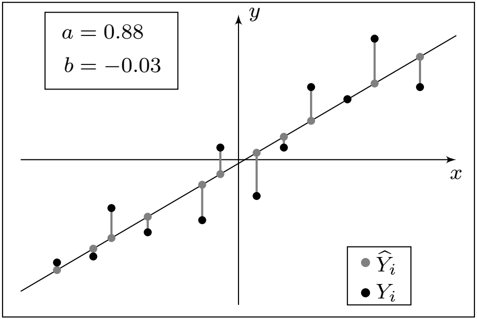
Korelacja mierzy **siłę zależności liniowej**.
Regresja to wyznaczanie **współczynników zależności liniowej**:

> Mając zbiór n punktów $(X_1, Y_1), \ldots (X_n, Y_n)$ 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 $\widehat{Y}_i = aX_i + b$ a prawdziwą wartością Y_i:



Minimalizujemy **sumę kwadratów błędów**:

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

Współczynniki regresji

$$\begin{aligned} a &= \frac{s_{XY}}{s_X^2} = r\frac{s_Y}{s_X} \\ b &= \overline{Y} - a\overline{X} \end{aligned}$$

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