Linearna regresija

Matija Bašić
PMF-MO, Zagreb
25. ožujka 2022.

Teachers' Inquiry in
Mathematics Education

Co-funded by the
Erasmus+ Programme
of the European Union

VAŠA ISKUSTVA

- Kad ste se prvi put susreli s linearnom regresijom?
- Jeste li ikad održali nastavni sat na ovu temu? Kakvi su dojmovi?
- Što smatrate važnim istaknuti?

- Na koji način matematički istražujete novu temu?

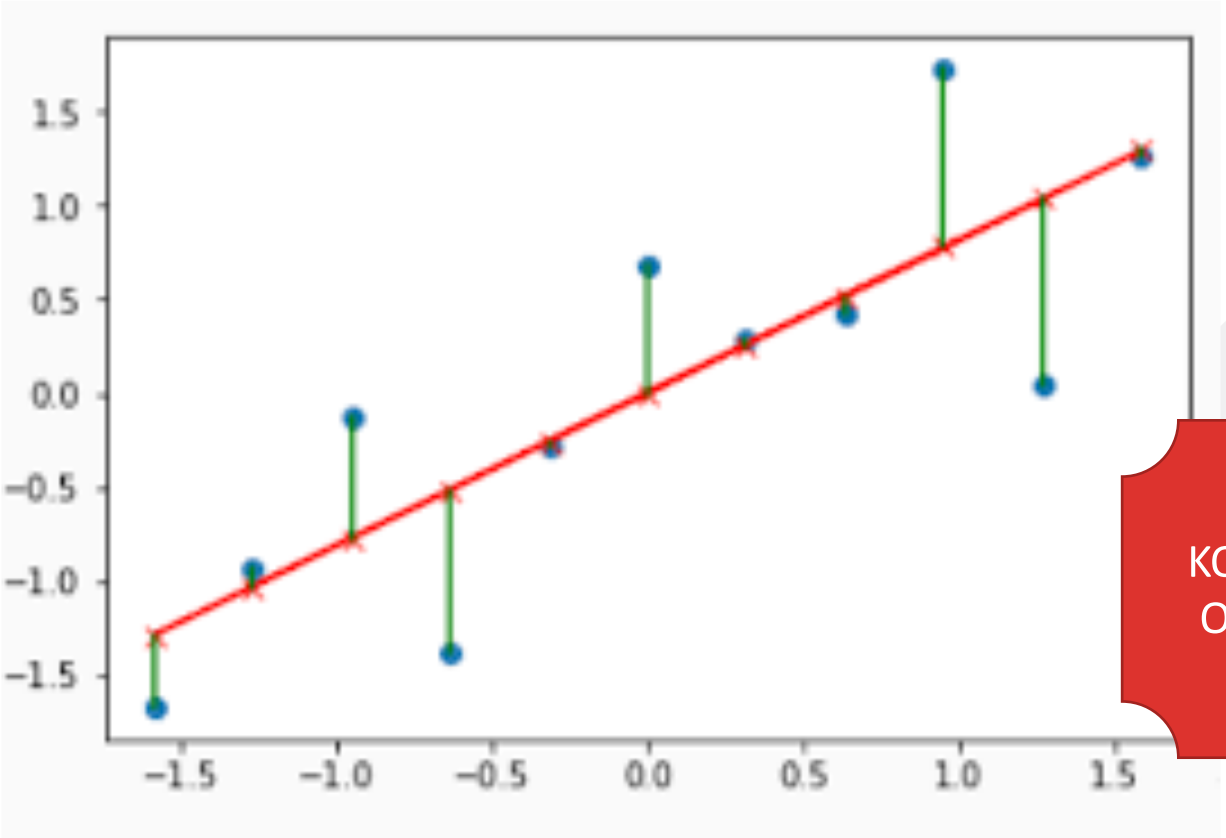
Co-funded by the Erasmus+ Programme of the European Union

PROBLEM LINEARNE REGRESIJE

x	y
x_1	y_1
x_2	y_2
…	…
x_n	y_n

LINEARNI MODEL

$$f(x) = ax + b$$



KOJI PRAVAC ODABRATI?

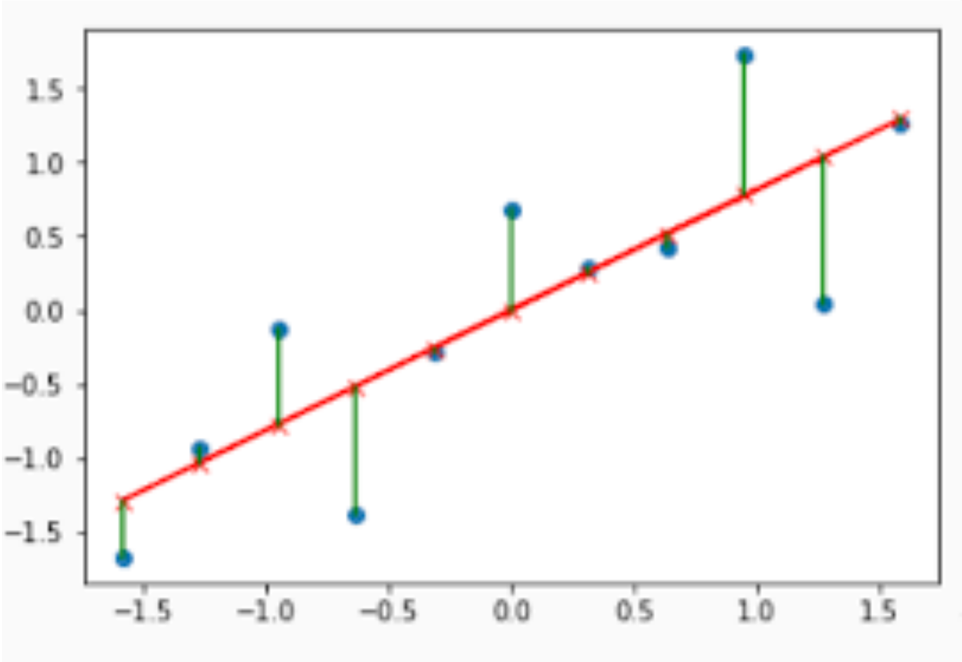
Co-funded by the Erasmus+ Programme of the European Union

PRAVILO RAVNOTEŽE

$$ax_1 + b - y_1 + \cdots + ax_n + b - y_n = 0.$$

$$f(\bar{x}) = \bar{y}, \qquad a\bar{x} + b = \bar{y}$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Co-funded by the Erasmus+ Programme of the European Union

METODA NAJMANJIH KVADRATA

Tražena funkcija $f(x) = ax + b$

takva da je izraz

$$(ax_1 + b - y_1)^2 + \cdots + (ax_n + b - y_n)^2$$

najmanji mogući.

Minimum tražimo izjednačavajući parcijalne derivacije s nulom

ili

primjenom pravila ravnoteže i određivanjem tjemena kvadratne funkcije u varijabli a.

METODA NAJMANJIH KVADRATA

Uvrstimo li $b = \bar{y} - a\bar{x}$, dobivamo funkciju

$$F(a) = (ax_1 + \bar{y} - a\bar{x} - y_1)^2 + \cdots + (ax_n + \bar{y} - a\bar{x} - y_n)^2$$

$$F(a) = a^2 \sum_i (x_i - \bar{x})^2 - 2a \sum_i (x_i - \bar{x})(y_i - \bar{y}) + \sum_i (y_i - \bar{y})^2$$

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{s_{x,y}}{s_x^2}$$

ZAŠTO METODA NAJMANJIH KVADRATA?

Zašto promatramo funkciju

$$F(a,b) = (ax_1 + b - y_1)^2 + \cdots + (ax_n + b - y_n)^2$$

umjesto funkcije

$$G(a,b) = |ax_1 + b - y_1| + \cdots + |ax_n + b - y_n|\ ?$$

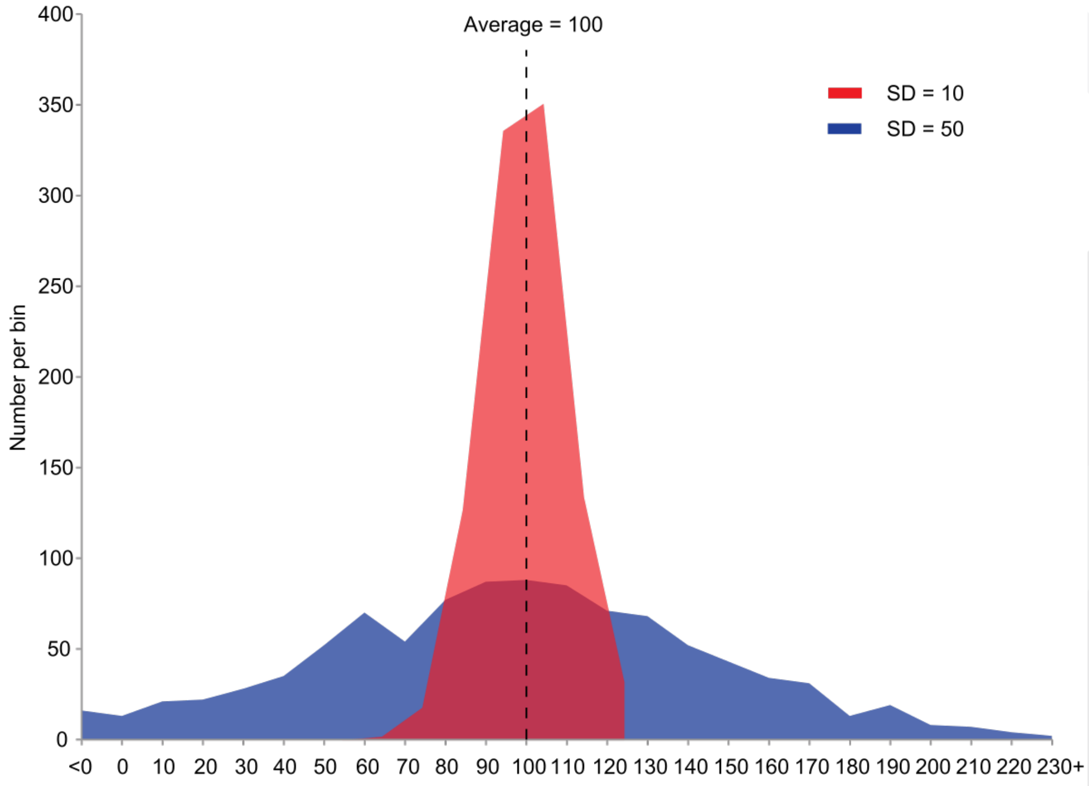
ZAŠTO METODA NAJMANJIH KVADRATA?

Uočite da je varijanca normalno distribuirane populacije

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \cdots (x_n - \mu)^2}{n}$$

što možemo izraziti kao prosječno kvadratno odstupanje od prosjeka.

Pitanje je zapravo usko vezano uz razloge zašto kao mjeru raspršenosti promatramo **varijancu**.



ZAŠTO METODA NAJMANJIH KVADRATA (MNK)?

- Kvadratna funkcija je **derivabilna** u svim točkama domene, a apsolutna vrijednost nije
- *Prosjek* minimizira funkciju $f(x) = (x - x_1)^2 + \cdots + (x - x_n)^2$

dok *medijan* minimizira funkciju $f(x) = |x - x_1| + \cdots + |x - x_n|$

- Ako su X i Y nezavisne, vrijedi Var(X+Y)=Var X + Var Y (posljedica Pitagore)
- Među raznim normama na $\mathbb{R}^n$, euklidska norma je jedina inducirana skalarnim produktom
- Prostor slučajnih varijabli s konačnom varijancom ima skalarni produkt koji je dan kovarijancom, a varijanca je norma

ZAŠTO METODA NAJMANJIH KVADRATA (MNK)?

- Ako su greške (tj. Y-aX-b) normalno distribuirane, onda će parametri linearnog modela *dobivenog MNK* također biti **normalno distribuirani** što omogućuje testiranje hipoteza o tim parametrima