Poglavlje 4

Metoda najmanjih kvadrata

Primjeri u Excelu vezani za ovu cjelinu nalaze se u dokumentu *MNK.xlsx*.

Pretpostavimo da su zadana dva skupa podataka,

$$x = \{x_1, x_2, \ldots, x_n\} \quad \text{i} \quad y = \{y_1, y_2, \ldots, y_n\}.$$

Zanima nas jesu li ta dva skupa podataka međusobno povezana (korelirana) i ako da, na koji način. U slučaju da su veličine x i y korelirane, onda ako znamo jednu od njih, x_i, možemo procijeniti drugu, y_i. Najjednostavnija veza među podatcima je linearna. To znači da su vrijednosti x i y povezane linearnom funkcijom,

$$y := f(x) = ax + b.$$

Podatke prikazujemo kao točke u koordinatnom sustavu pri čemu je

$$T_1 = (x_1, y_1), T_2 = (x_2, y_2), \ldots, T_n = (x_n, y_n).$$

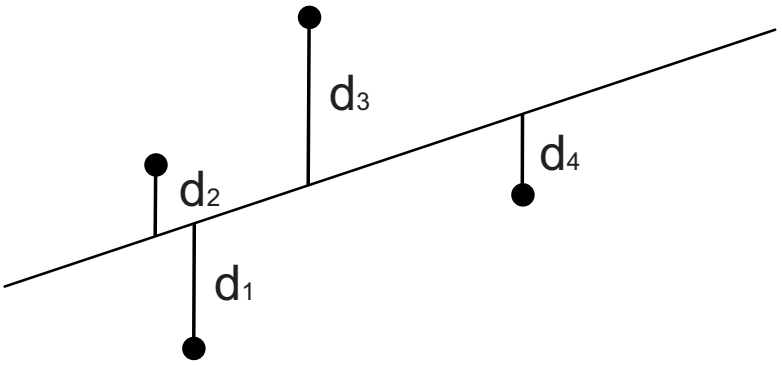
Imajući u vidu da je graf linearne funkcije pravac, vezu među podatcima grafički prikazujemo kao pravac. Taj se pravac naziva **regresijski pravac** ili pravac linearne regresije.

U općem slučaju nije moguće povući jedan pravac koji će prolaziti kroz sve točke. Stoga, tražimo pravac koji prolazi "dovoljno blizu" svim točkama. Biramo ga tako da promatramo udaljenosti d_i, (slika 4.1)

$$d_i = y_i - (ax_i + b),$$

gdje su a i b nepoznati parametri. Potrebno je da suma kvadrata ovih udaljenosti bude što manja,

$$\sum_{i=1}^{n} d_i^2 = d_1^2 + d_2^2 + \ldots + d_n^2 \to \min.$$



Slika 4.1: Metoda najmanjih kvadrata

Iz gornjeg zahtjeva dobiju se parametri a i b koji se potom uvrste u jednadžbu $y = ax + b$. Ovdje nećemo raditi izvod tih parametara, nego ćemo za dobivanje jednadžbe regresijskog pravca korisiti mogućnosti Excela.

Odgovor na pitanje koliko dobro regresijski pravac aproksimira zadane podatke daje **koeficijent linearne korelacije** R. Njegova vrijednost je između -1 i 1. Koeficijent R je pozitivan ako je regresijska funkcija rastuća, a negativan ako je padajuća. Što je R po apsolutnoj vrijednosti bliže 1, to je aproksimacija bolja. Ako bi imali situaciju u kojoj sve zadane točke leže na regresijskom pravcu, vrijednost koeficijenta R bila bi 1 ili -1. Što ćemo uzeti za dovoljno dobru aproksimaciju ovisi o problemu. Uglavnom možemo smatrati da su podatci jako linearno korelirani ako je

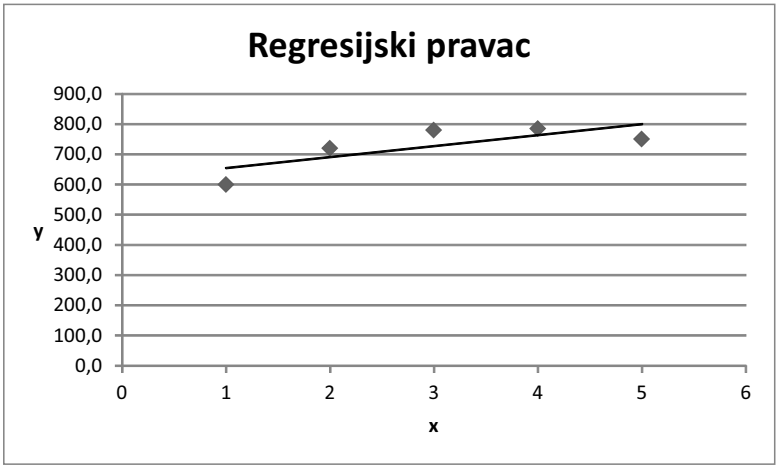
$$|R| \geq 0.9.$$

U Excelu se koeficijent R računa naredbom

$$\text{CORREL(vrijednosti } x\text{; vrijednosti } y\text{)}.$$

Kada znamo vezu podataka x i y, za dodatne vrijednosti x_k pripadne vrijednosti y_k računamo jednostavnim uvrštavanjem u jednadžbu regresijskog pravca. Ovdje treba imati na umu da su dobivene vrijednosti približne, te više vjerodostojne što je bolji koeficijent R.

Također, potrebno je znati da koreliranost podataka ne uvjetuje nužno i njihovu uzročnost. Odnosno, moguće je da su dva skupa podataka dobro korelirana, ali su u stvarnosti neovisni. Stoga uvijek treba biti pažljiv pri interpretaciji rezultata.



Slika 4.2: Primjer regresijskog pravca