
# Analiza podataka upotrebom linearne regresije

Milena Jelić

Fakultet tehničkih nauka
Univerzitet u Novom Sadu

13. maj 2024.

# Jednostruka linearna regresija

- Proučavanje veze između dve kontinualne promenljive
- Zavisna (y) i nezavisna (x) promenljiva
- Deterministička (funkcionalna) zavisnost nam nije od interesa jer možemo odrediti tačnu vrednost

# Jednostruka linearna regresija

- Proučavanje veze između dve kontinualne promenljive
- Zavisna (y) i nezavisna (x) promenljiva
- Deterministička (funkcionalna) zavisnost nam nije od interesa jer možemo odrediti tačnu vrednost
- $O = 2r\pi$

# Statistička zavisnost

- Posmatramo probleme gde imamo neku statističku zavisnost između podataka, nemamo "formulu"
- Primer takve zavisnosti: visina i težina ljudi

# Najpogodnija linija

- Linearna zavisnosnost znači da je regresiona funkcija linearna po koeficijentima
- Ako želimo pravu, imamo dva koeficijenta: $\hat{y} = ax + b$
- Određivanje dva koeficijenta na osnovu mnogo više od 2 para vrednosti daje preodređen sistem
- Moguća je situacija da dva podatka imaju istu vrednost $x$, a različito $y$

# Greška

- Greška predstavlja razliku između prave vrednosti i predviđene vrednosti
- Kvadratna funkcija greške: $SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

# Linearna regresija kao polinom proizvoljnog stepena

- Polinom stepena $m$: $p(x) = \sum_{j=0}^{m} a_j x^j$
- Imamo $m + 1$ parametar, a ne 2
- Opet dobijemo preodređen sistem

# Metod najmanjih kvadrata

- Najpribližnije rešenje kod preodređenih sistema
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - ax_i - b)^2$
- $a = \frac{\frac{\mathbf{1}}{n}\sum_{i=\mathbf{1}}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{\mathbf{1}}{n}\sum_{i=\mathbf{1}}^{n}(x_i - \bar{x})^{\mathbf{2}}} = \frac{Cov(x,y)}{Var(x)}$
- $b = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i) = \bar{y} - a\bar{x}$

# Pretpostavke linearne regresije

- Linearnost
- Nezavisnost grešaka
- Normalnost grešaka
- Jednake varijanse

# Višestruka linearna regresija

- Na osnovu 2 ili više nezavisne promenljive predviđa se zavisna
- Primer: $y = 5x_1 + 7x_2 + 3$

# Pretpostavke višestruke regresije

- Sve prethodno pomenute pretpostavke
- Dodatana pretpostavka o kolinearnosti - ne postoji savršena kolinearnost

# Savršena kolinearnost

- Koeficijent korelacije - broj u intervalu od $-1$ do 1
- Savršena pozitivna kolinearnost - koeficijent korelacije je 1
- Savršena pozitivna kolinearnost - koeficijent korelacije je -1
- Ne želimo da imamo 2 kolone koje su u savršenoj kolinearnosti, jednu od te dve kolone treba ukloniti
- $x_i = ax_j + b$

# Jaka korelacija

- Korelacija u intervalu $(-1, -0.8] \cup [0.8, 1)$
- Pretpostavka nije narušena, ali vredi obratiti pažnju
- Matrica korelacije

# Evaluacija modela

- Koeficijent determinacije - uzima vrednosti u opsegu $[0, 1]$. Što je $r^2$ bliže 1, to je model bolji
- $r^2 = \frac{SSR}{SST}$
- $SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$
- $SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$

# Prilagođeni koeficijent determinacije

- Koeficijent determinacije ne uzima u obzir količinu podataka i broj nezavisnih promenljivih
- vrednost $r^2$ nikad ne opada sa dodavanjem novih kolona
- $\bar{r}^2 = 1 - (1 - r^2) \cdot \frac{n-1}{n-p-1}$
- n - broj podataka u skupu
- p - broj nezavisnih promenljivih

# Hvala na pažnji!

Pitanja?