

# SKRIPTA - LINEARNA REGRESIJA I INTERPOLACIJA

## deo 1: LINEARNA REGRESIJA

### 1.1 Model Višestruke Linearne Regresije

**Model:**

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n + \epsilon$$

**Objašnjenje oznaka:**

- **y** - zavisna promenljiva (predviđana vrednost)
- **b<sub>0</sub>** - intercept (odsečak) - vrednost y kada su sve x<sub>i</sub> = 0
- **b<sub>i</sub>** - parametri regresije koji predstavljaju uticaj svake nezavisne promenljive
- **x<sub>i</sub>** - nezavisne promenljive koje se koriste za predviđanje
- **ε** - greška modela (slučajna komponenta)

### 1.2 Koeficijent Determinacije (R<sup>2</sup>)

**Definicija:** Mera koliko se dobro regresioni model uklapa u podatke

**Karakteristike:**

- Kreće se između 0 i 1
- R<sup>2</sup> = 0 → nikakvo uklapanje
- R<sup>2</sup> = 1 → savršeno uklapanje
- **Interpretacija:** R<sup>2</sup> = 0.72 znači da 72% varijacije u y možemo objasniti pomoću x, dok je 28% varijacije šum koji nije obuhvaćen modelom

### 1.3 Metoda Najmanjih Kvadrata

**Cilj:** Minimizujemo grešku, odnosno razliku između predikcija i stvarnih vrednosti sa ciljem da dobijemo optimalne parametre modela.

**Zbir kvadrata grešaka:**

$$SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

gde su:

- $y_i$  - stvarna vrednost
- $\hat{y}_i$  - predviđena vrednost

## 1.4 Linearnost po Parametrima

**Važno:** Regresija pomoću polinoma trećeg stepena (npr.  $y = b_0 + b_1x + b_2x^2 + b_3x^3$ ) je takođe linearna regresija jer je **linearna po parametrima**, ne po nezavisnoj promenljivoj.

## 1.5 Statističko Testiranje - t-test

**Svrha:** Testiranje statističke hipoteze o značajnosti parametara

**Postupak:**

1. Postavljamo hipotezu lošeg slučaja:  $H_0: b_i = 0$
2. Ako je p-vrednost  $< 0.05 \rightarrow$  odbacujemo hipotezu
3. To znači da je parametar različit od 0 i da postoji linearна веза између  $y$  и  $x_i$

**Interpretacija p-vrednosti (95% pouzdanost):**

- $p < 0.05 \rightarrow$  odbacujemo hipotezu da ne postoji linearна веза
- Zaključak: postoji statistički značajna linearна веза

## 1.6 Interpretacija Koeficijenata u Višestrukoj Regresiji

**Pravilo:** Ako povećam neko  $x_i$  za 1 vrednost, srednja vrednost  $y$  će se povećati za  $b_i$ , **pod uslovom da su ostali x fiksirani**.

**Primer negativnog koeficijenta:** Model za predikciju cena kuća uključuje površinu kuće i broj kupatila. Ako koeficijent ispred površine ima negativnu vrednost, to znači da cena kuće opada sa povećanjem površine (ako je fiksan broj kupatila) - npr. ogromna kuća sa samo 1 kupatilom.

## 1.7 Prilagođeni Koeficijent Determinacije

**Problem sa običnim  $R^2$ :** Kako dodajemo promenljive,  $R^2$  će biti isti ili rasti, ne znamo da li smo overfitovali i da li su dodate promenljive besmislene.

**Rešenje:** Prilagođeni  $R^2$  uzima u obzir broj nezavisnih promenljivih i kažnjava dodavanje nepotrebnih varijabli.

## 1.8 Overfitting (Preprilagođavanje)

**Definicija:** Model je loše uslovljen - za male promene u  $x$  dobijam velike promene u  $y$ . Model počinje da se prilagođava ne samo signalu nego i šumu.

**Indikatori:** Model odlično radi na training podacima, ali loše na novim podacima.

## 1.9 Procena Korisnosti Nezavisnih Promenljivih

**Kriterijum:** Kod promenljivih gde je p-vrednost  $< 0.05$ , one su korisne jer odbacujemo hipotezu da je parametar = 0, što znači da postoji linearna veza.

## 1.10 Kada Ne Uklanjati Nezavisnu Promenljivu

**Situacije:**

- **Cilj - predikcija:** Možemo da uklonimo iz modela ako nam nije korisna
- **Cilj - istraživanje:** Hoćemo da vidimo baš vezu između te promenljive i y, čak i ako nije statistički značajna

## 1.11 Prepostavke Linearne Regresije (LINE)

### L - Linearity (Linearnost)

- Postoji linearna veza između x i y
- **Ako je narušena:** Radimo transformaciju (npr.  $x^2$ )

### I - Independence of Errors (Nezavisnost grešaka)

- $y_i$  ne zavisi od  $y_{i-1}$
- **Ako zavisi:** Koristi se model vremenskih serija
- **Na grafiku reziduala:** Greške ravnomerno osciluju

### N - Normality of Errors (Normalnost grešaka)

- Greške prate normalnu raspodelu
- **Interpretacija:** Najverovatniji je podatak na samom regresionom modelu, sa manjom verovatnićom će biti udaljen

### E - Equal Variance (Konstantna varijansa grešaka)

- **Problem:** Na početnom grafiku su greške manje pa kasnije veće
- **Rešenje:** Logaritamska ili korenska transformacija y
- **Na grafiku reziduala:** Reziduali ne smeju da imaju šablon

## 1.12 Multikolinearnost

**Problem:** npr.  $x_1 = 2x_2 + x_3$

## Posledice:

- Rešavamo sistem sa d jednačina i d nepoznatih
- Ako su neki u međusobnoj vezi, sistem može nemati rešenja
- Približna multikolinearnost → model nestabilan, sklon overfittingu

## 1.13 Reziduali

**Definicija:** Rezidual =  $y_i$  (stvarno) -  $\hat{y}_i$  (predviđeno)

**Svrha:** Analiza reziduala pomaže u proveri prepostavki modela.

---

## DEO 2: INTERPOLACIJA

### 2.1 Problem Interpolacije

**Cilj:** Imamo skup tačaka i hoćemo da vidimo šta se dešava između tih tačaka.

**Osnovna ideja:** Prepostavka da će to biti interpolacioni polinom koji **precizno prolazi kroz svaku od tačaka**.

### 2.2 Interpolacija vs Regresija

Aspekt	Interpolacija	Regresija
Kada koristiti	Merenja su precizna (zvuk, slika)	Tražimo trend u podacima
Primer	Izmereni podaci ostaju takvi kakvi jesu	Cena nekretnine - za istu kvadraturu različite cene
Cilj	Šta je između podataka	Opšti trend

### 2.3 Interpolacioni Polinom

Za N tačaka:

- **Oblik:**  $y = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1}$
- **Određujemo:** n parametara od  $a_0$  do  $a_{n-1}$
- **Stepen polinoma:** N-1

**Ako je stepen polinoma prevelik:** Koristimo splajn - na svakom intervalu imamo poseban splajn (linearni, kvadratni, kubni).

### 2.4 Predstava Polinoma u Python-u

**Format:** Kao vektori koeficijenata

**Primer:**  $5.5x^3 + x - 3 \rightarrow [5.5, 0, 1, -3]$

## 2.5 Zapisi Interpolacionog Polinoma

### 1. Standardni Zapis

$$g(x) = a_1 + a_2x + a_3x^2 + \dots + a_nx^{n-1}$$

**Problem:** n jednačina sa n nepoznatih - loše uslovljen sistem jer za male promene dobijamo veliko rešenje (skoro paralelne prave).

### 2. Njutnov Zapis (za polinom 2. stepena)

$$g_n(x) = b_1 + b_2(x-x_1) + b_3(x-x_1)(x-x_2)$$

**Prednost:** Ako dodamo još jednu tačku, niži koeficijenti ostaju nepromenjeni - računamo samo jedan dodatni koeficijent.

### 3. Lagranžov Zapis (za polinom 2. stepena)

$$\begin{aligned} g_1(x) &= [(x-x_2)(x-x_3)]/[(x_1-x_2)(x_1-x_3)] \cdot f(x_1) + \\ &[(x-x_1)(x-x_3)]/[(x_2-x_1)(x_2-x_3)] \cdot f(x_2) + \\ &[(x-x_1)(x-x_2)]/[(x_3-x_1)(x_3-x_2)] \cdot f(x_3) \end{aligned}$$

## 2.6 Faktori koji Utiču na Tačnost Interpolacije

- Razdaljina tačaka** - što je manja, biće preciznije
- Broj tačaka** - sa povećanjem stepena polinoma dobijamo preciznije rezultate

## 2.7 Ekstrapolacija

**Definicija:** Koristimo isti polinom kao kod interpolacije, ali za predviđanje šta se dešava **daleko od tačaka**.

**Napomena:** Trudimo se da ekstrapolaciju izbegnemo jer je manje pouzdana.

## 2.8 Interpolacija Splajnom

### Osnovna Ideja

- Podelimo interval na N-1 podintervala (ako je N broj tačaka)
- Na svakom intervalu imamo poseban splajn

- Na kraju dobijemo krivu koja prolazi kroz sve tačke

### Problem Linearnog Splajna

- Koristi linearni polinom
- Kriva može da ne bude dovoljno glatka
- Ne prati baš precizno oblik podataka

### Kvadratni Splajn za 4 Tačke

- **3 podintervala** (između 4 tačke)
  - **Oblik polinoma za svaki podinterval:**  $a_i x^2 + b_i x + c_i$
  - **Ukupno nepoznatih koeficijenata:** 9 (3 koeficijenta  $\times$  3 podintervala)
- 

## KLJUČNE FORMULE ZA PAMĆENJE

1. **Model višestruke linearne regresije:**  $y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n + \varepsilon$
2. **Zbir kvadrata grešaka:**  $SSE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$
3. **Rezidual:**  $r_i = y_i - \hat{y}_i$
4. **Interpolacioni polinom:**  $y = a_0 + a_1 x + a_2 x^2 + \dots + a_{n-1} x^{n-1}$
5. **Kriterijum statističke značajnosti:**  $p < 0.05$

## PRAKTIČNI SAVETI ZA ISPIT

1. **Uvek objasniti sve oznake** u modelima
2. **Pamtiti LINE pretpostavke** i kada se krše
3. **Razlikovati interpolaciju od regresije** - ključ je u tome da li su merenja precizna
4. **Interpretacija koeficijenata** - uvek spomenuti "ostali faktori fiksni"
5. **p-vrednost** - uvek povezati sa 0.05 i hipotezom testiranja