Laboratorijska vježba 4 - Statističko zaključivanje dvije varijable - linearna regresija

Metode statističke analize podataka

2025./2026.

Cilj zadatka

Primijeniti metode statističke analize podataka na datasetu `AirQualityUCI.csv` kako biste:

- učitali i očistili relevantne varijable,
- izgradili i analizirali jedan linearan regresijski model,
- procijenili parametre modela, intervale povjerenja i adekvatnost modela,
- izračunali koeficijent korelacije između odabranih varijabli i interpretirali rezultate.

Dataset: Koristite dataset `AirQualityUCI.csv` koji sadrži mjerenja kvalitete zraka u gradu (koncentracije plinova, temperaturne i vlažne uvjete, itd.). Primjeri varijabli:

- `CO(GT)` – koncentracija ugljičnog monoksida (mg/m^3),
- **`NO2(GT)`** – koncentracija dušikova dioksida (µg/m^3),
- **`C6H6(GT)`** – benzene (µg/m^3),
- `T` – temperatura (°C),
- `RH` – relativna vlažnost (%),
- `AH` – apsolutna vlažnost.

Koraci za rješavanje zadatka

1. **Učitavanje i pregled podataka**

   a) Učitajte dataset u Python (npr. `pandas.read_csv("AirQualityUCI.csv")`).

   b) Ispišite nekoliko prvih redaka tablice i ispitajte strukturu podataka:

   - struktura: `data.info()`,
   - osnovne statistike: `data.describe()`,
   - tipovi podataka: `data.dtypes`.

c) Provjerite nedostajuće vrijednosti i posebne kodove nedostajućih vrijednosti:
   - `data.isna().sum()`,
   - obratite pozornost na vrijednosti -200 koje u ovom datasetu označavaju nedostajuće mjere određenih senzora.

d) Očistite podatke:
   - zamijenite vrijednosti -200 sa `NaN`,
   - uklonite retke s `NaN` *za dvije varijable koje ćete koristiti u regresijskom modelu*.

e) Kratko komentirajte odabir načina čišćenja (zadržati, odbaciti, zamijeniti).

2. **Definiranje hipoteza i odabir varijabli za regresiju**

a) Odaberite jednu zavisnu varijablu Y i jednu nezavisnu varijablu X za koju pretpostavljate linearnu vezu.

b) Zapišite model:
$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

c) Za svaki model odaberite:
   - ciljnu varijablu Y (npr. `NO2(GT)` ili `CO(GT)`),
   - jednu nezavisnu varijablu X za koju pretpostavljate linearnu vezu s Y (npr. `C6H6(GT)`, `T`, `RH`, ...).

d) Zapišite hipoteze:
   - $H_0 : \beta_1 = 0$ (nema linearne veze između X i Y),
   - $H_1 : \beta_1 \neq 0$ (postoji linearna veza).

3. **Izgradnja linearnog regresijskog modela**

a) Procijenite parametre modela:
   - `scipy.stats.linregress(X, Y)` i/ili
   - `statsmodels.OLS`.
   - ako su skale vrlo različite (npr. koncentracija plina vs. temperatura), po potrebi standardizirajte ili skalirajte varijable (npr. `StandardScaler` iz `sklearn`).

b) Prikažite regresijski model na dijagramu raspršenja (scatter + regresijska linija).

c) Pazite na ispravno obilježavanje osi (nazivi varijabli i jedinice).

d) Izračunajte i interpretirajte:
   - nagib $\hat{\beta}_1$,
   - odsječak $\hat{\beta}_0$,
   - koeficijent determinacije R^2.

4. **Intervali povjerenja za parametre**

a) Izračunajte 95% intervale povjerenja za β_0 i β_1.

b) Možete koristiti:
   - rezultate iz `statsmodels` (`results.conf_int()`),
   - ili ručni izračun pomoću standardne pogreške i kritične vrijednosti t-distribucije.

c) Protumačite uključuje li interval za β_1 nulu i što to znači.

5. **Analiza adekvatnosti modela**

a) Izračunajte reziduale i grafički ih prikažite:
   - reziduali naspram predviđenih vrijednosti,
   - histogram reziduala.

b) Analizirajte reziduale kako biste procijenili je li linearni model primjeren:
   - nacrtajte graf *reziduali naspram predviđenih vrijednosti*,
   - komentirajte: postoji li vidljiv uzorak? (trend, zakrivljenost, oblik lijevka ...),
   - procijenite postoji li heteroskedastičnost (nejednaka varijanca reziduala).

c) Provjerite normalnost reziduala:
   - histogram reziduala,
   - QQ-plot,
   - kratka interpretacija koliko odstupaju od normalne razdiobe.

d) Komentirajte homoskedastičnost:
   - jesu li reziduali otprilike jednako raspršeni oko 0 za sve predviđene vrijednosti?

e) Još jednom izračunajte i interpretirajte R^2 kao mjeru kvalitete modela:
   - koliko varijance zavisne varijable je objašnjeno modelom,
   - je li takav R^2 u ovom kontekstu zadovoljavajući?

6. **Koeficijent korelacije**

a) Izračunajte Pearsonov koeficijent korelacije između ciljane i nezavisne varijable za svaki model.

b) Interpretirajte:
   - smjer povezanosti (pozitivan/negativan),
   - jačinu povezanosti (slaba/umjerena/snažna),
   - povezanost s R^2 (za jednostavnu linearnu regresiju vrijedi $R^2 = r^2$).

7. **Testiranje adekvatnosti modela – F-test**

a) U okviru jednostavne linearne regresije potrebno je provesti F-test kako bi se ispitalo ima li regresijski model uopće objašnjavačku moć (odnosno je li barem jedan regresijski koeficijent različit od nule).

b) Izračunajte F-statistiku i pripadnu p-vrijednost (npr. iz `statsmodels` `summary()`).

c) Interpretirajte rezultate:
   - za odabranu razinu značajnosti α (npr. 0.05), odlučite odbacujete li nul-hipotezu da model nema objašnjavačku moć,

- objasnite u kontekstu problema (predviđanje kvalitete zraka).

8. **Zaključak**

   a) Sažmite najvažnije rezultate analize:
      - značajnost β_1,
      - jačina veze,
      - adekvatnost modela prema rezidualima i F-testu.

9. **Deskriptivna statistika** (opcionalno)

   a) Odaberite **četiri** kvantitativne varijable (npr. `CO(GT)`, `NO2(GT)`, `T`, `RH`) i izračunajte:
      - apsolutne srednje vrijednosti, standardne devijacije,
      - relativne mjere (npr. koeficijent varijacije) uz kratku interpretaciju.

   b) Vizualizirajte odabrane varijable:
      - histogrami,
      - box-plotovi,
      - dijagrami raspršenja za parove varijabli za koje očekujete povezanost (npr. `NO2(GT)`–`C6H6(GT)`, `T`–`RH`).

   c) Napišite kratku interpretaciju:
      i) **Varijabilnost:** Koji od odabranih parametara imaju najveću varijabilnost (relativno na njihovu srednju vrijednost)?
      ii) **Iznimne vrijednosti:** Postoje li iznimno visoke ili niske vrijednosti (outlieri) koje su ostale u datasetu nakon čišćenja?
      iii) **Obrazloženje odabira:** Objasnite zašto ste odabrali baš te četiri varijable za analizu (npr. važnost za kvalitetu zraka, logična fizička povezanost).