

LINEARNI MODELI 2

STATISTIČKI PRAKTIKUM 2

3. VJEŽBE

Kada je opravdano koristiti linearni regresijski model?

Promatramo linearni regresijski model u p varijabli

$$Y_n = \beta_0 + \beta_1 x_1^{(n)} + \dots + \beta_p x_p^{(n)} + \varepsilon_n.$$

Četiri glavne pretpostavke koje opravdavaju korištenje LRM-a u svrhu analize podataka i predviđanja su:

- (i) *linearni* odnos između varijabli poticaja i odaziva
- (ii) *nezavisnost* grešaka
- (iii) *homogenost* grešaka
- (iv) *normalna distribuiranost* grešaka

Ukoliko neka od pretpostavki nije opravdana, naša previđanja mogu biti nevaljana.

(i) Linearnost podataka

Ukoliko postoji nelinearan odnos jedne ili više varijabli, naša predviđanja (posebno izvan raspona uzorka) mogu biti potpuno netočna.

Detekcija odnosa varijabli i dobar izbor početnog modela ključan je za ispravnu analizu podataka.

Kako prepoznati nelinearnost?

Nelinearnost je najlakše uočiti iz grafičke usporedbe

- ▶ stvarnih i predviđenih vrijednosti varijable odaziva ($Y - \hat{Y}$ graf)
- ▶ reziduala i predviđenih vrijednosti - *residual-fit plot* ($\hat{Y} - \mathbf{e}$ graf)

U prvom grafu očekujemo simetrično raspršenje podataka oko dijagonale, a u drugom oko apscise.

Kako ukloniti nelinearnost?

Potrebno je *transformirati* jednu ili više varijabli poticaja i/ili varijablu odaziva nekom nelinearnom funkcijom. Odabir funkcije ovisi o tipu podataka.

(ii) Nezavisnost grešaka

Problem zavisnosti grešaka javlja se kao moguća posljedica

- ▶ “cluster” podataka (uzorkovanje iz određene grupe umjesto cijele populacije)
- ▶ longitudinalnih podataka (uzorkovanje populacije kroz vrijeme)
- ▶ čiste koreliranosti (loš odabir modela)

Napomena: Nekoreliranost dviju normalnih slučajnih varijabli povlači njihovu nezavisnost.

Serijsku korelaciju među greškama ε možemo procijeniti koristeći rezidualne e . U R-u to čini funkcija `acf`

```
> model=lm(y~x)
> corr=acf(model$res)
> corr$acf
---
> acf(model$res)
```

Output je grafički prikaz vrijednosti *autokorelacijske funkcije*, zajedno s pripadnom 95%-pouzdanom prugom (približne širine $4/\sqrt{n}$).

Problemi uzorkovani “cluster” i longitudinalnim podacima mogu se ukloniti statistički opravdanim skupljanjem podataka. Općenito u slučaju pojave koreliranosti možemo koristiti modele koji uključuju koreliranost podataka (npr. odabirom pravilnog ARIMA modela)

(iii) Homogenost grešaka

Nejednakost varijance grešaka može rezultirati

- ▶ preširokim/preuskim intervalima pouzdanosti za procjene
- ▶ preferiranjem određenog podskupa podataka pri procjeni

Najčešće se pojavljuje kod vremenskih podataka (varijanca raste s vremenom).

Homogenost se može uočiti iz grafičkog prikaza

- ▶ reziduala obzirom na vrijeme ($t - e$ graf)
- ▶ residual-fit plot-a ($\hat{Y} - e$ graf).

Kod oba grafa pozornost treba obratiti na rast reziduala.

Moguća rješenja su:

- ▶ analiza podataka po dijelovima s jednakom varijancom
- ▶ ispitivanje pravilnog odabira modela
- ▶ korištenje alternativnih modela (ARCH)

(iv) Normalnost grešaka

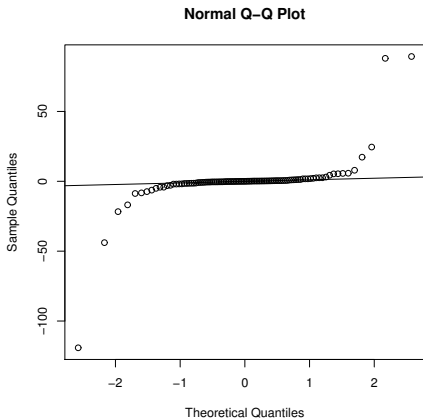
Velika odstupanja od normalnosti mogu prouzročiti brojne probleme pri procjeni vrijednosti i pouzdanih intervala. Jedan od uzroka može biti i prisutnost *outliera*, koji mogu stvoriti probleme pri procjeni parametara modela (povećanje kvadratne greške).

Normalnost grešaka možemo provjeriti preko reziduala, korištenjem

- ▶ normalnog vjerojatnosnog grafa
- ▶ Lillieforsovog testa
- ▶ Shapiro-Wilkovog testa

```
> shapiro.test(model$res)
```

Veliki uzorci iz distribucija *lakog repa* neće prouzročiti probleme u procjeni. Ako su greške kontrolirane i pripadaju distribuciji lakog repa svi rezultati će zbog snage CGT vrijediti asimptotski. Problemi su i dalje mogući kada greške imaju distribuciju *teškog repa*:



Analiza greške ε preko analize ostataka $\mathbf{e}$

Ako su greške ε_i nezavisne i (normalno) jednako distribuirane, ostaci $\mathbf{e}$ ne moraju biti. Naime

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\varepsilon,$$

pa je $\text{Var } \mathbf{e} = (\mathbf{I} - \mathbf{H})\sigma^2$.

Specijalno, uz oznaku $h_i = h_{ii}$ (*leverage*), vrijedi

$$\text{Var } e_i = \sigma^2(1 - h_i).$$

Problem se rješava uvođenjem studentiziranih ostataka

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

Ako su pretpostavke na grešku modela točne, tada je $\text{Var } r_i = 1$, a korelacija $\text{Cor}(r_i, r_j)$ je mala.

U R-u ih dobivamo pomoću naredbe

```
> rstudent(model).
```

Zadatak

Promotrimo model s nenormalno distribuiranim greškama.
Simulirajmo podatke za model

$$Y_i = 1 + x_i + 2 \sin(x_i) + \varepsilon_i,$$

gdje je $x_i = i/10$, za $i = 0, 1, \dots, 100$ i $\varepsilon_i \sim U(-1, 1)$ nezavisne.

- (a) Nacrtajmo stvarni model, procjenu modela i pouzdanu prugu za opažanja.
- (b) Nacrtajte *residual-fit plot* i analizirajte ga.
- (c) Proučite raspon ostataka i nacrtajte pripadni graf *autokorelacijske funkcije*.
- (d) Nacrtajte normalni vjerojatnosni graf ostataka i studentiziranih ostataka.

Izbor modela

Testiranje (linearnih) hipoteza o parametrima

Ukoliko smo opravdali pretpostavke modela, možemo preći na testiranje pouzdanosti modela.

Neka je $\mathbf{C} \in M_{m \times (p+1)}(\mathbb{R})$ i $\mathbf{g} \in M_{m \times 1}$, t.d. $r(\mathbf{C}) = m < p + 1$.

Hipotezu

$$H_0 : \quad \mathbf{C}\mathbf{b} = \mathbf{g}$$

testiramo statistikom

$$F = \frac{(\mathbf{C}\hat{\mathbf{b}} - \mathbf{g})^T [\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T]^{-1} (\mathbf{C}\hat{\mathbf{b}} - \mathbf{g})}{\hat{\sigma}^2} \stackrel{H_0}{\sim} F(m, p + 1)$$

Parametar m predstavlja broj jednadžbi uz koje je vezana nulta hipoteza (nužno manji od broja parametara modela).

Primjer testiranja linearnih hipoteza

Metodom najmanjih kvadrata podatke iz yy modelirajmo kao

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \sin x.$$

```
> pl=lm(yy~x+I(x^2)+sin(x))
```

Testirajmo hipotezu

$$H_0: \beta_2 = 0, \beta_1 = 1.$$

```
> pl2=lm(yy~offset(1*x)+sin(x))
```

```
> anova(pl2,pl)
```

Analysis of Variance Table

Model 1: $yy \sim \text{offset}(x) + \sin(x)$

Model 2: $yy \sim x + \sin(x) + I(x^2)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	99	32.044				
2	97	31.403	2	0.641	0.9896	0.3755

Izbor varijabli

Procedure za izbor varijabli žele izabrati *najbolji* podskup varijabli poticaja. Zašto?

1. Želimo objasniti podatke na najjednostavniji mogući način. Višak varijabli poticaja treba ukloniti.
2. Višak varijabli poticaja pojačat će *šumove* kod izračunavanja procjena.
3. Može doći do kolineranosti. Previše varijabli pokušava procijeniti istu stvar.
4. Troškovi računanja vrijednosti varijabli poticaja mogu se smanjiti.

Procedure za izbor varijabli

Neke od često korištenih procedura su:

1. hijerarhijski izbor modela;
2. procedure korak po korak;
3. procedure bazirane na kriterijima.

AIC i BIC kriterij

$$-2\log\text{-likelihood} + k$$

Ako imamo p mogućih varijabli poticaja, imamo 2^{p+1} mogućih modela. Želimo odabrati najbolji model prema nekom kriteriju. Dva korisna kriterija su *Akaike Information Criterion* (AIC) i *Bayes Information Criterion* (BIC).

$$AIC = -2\log\text{-likelihood} + 2(p + 1)$$

$$BIC = -2\log\text{-likelihood} + p \log n$$

Za linearne modele $-2\log\text{-likelihood}$ je $n \log(SSE/n)$.
($SSE = \sum_{k=1}^n (\hat{y}_i - y_i)^2$) Veći modeli imat će manji SSE , no koristit će više parametara. Najbolji model će biti ravnoteža između veličine i pristajanja podacima. BIC strože kažnjava veće modele.

Za određivanje AIC i BIC koristimo naredbu

```
> AIC(u1)  
> AIC(u1, k=log(101)).
```

Ukoliko želimo provesti hijerarhijsku analizu optimalnog modela, počevši od najopćenitijeg modela `model`, koristimo funkciju

```
> step(model, k=...)
```

Zadatak

Modelirajmo podatke iz primjera s uniformno distribuiranim greškama i to modelom s funkcijom sinus i polinomom 3. stupnja. Nađimo najpogodniji pod-model po AIC i BIC kriteriju.