LINEARNI MODELI 2

STATISTIČKI PRAKTIKUM 2

4

Kada je opravdano koristiti linearni regresijski model?

Promatramo linearni regresijski model u p varijabli

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

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) *normalna distribuiranost* grešaka

(iv) *homogenost* grešaka (jednakost normalnih distribucija)

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

(i) Linearnost podataka

Pretpostavljamo da Y ovisi linearno o prediktorima.

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* ($\hat{Y} -$ e graf)

U prvom grafu očekujemo simetrično raspršenje podataka oko dijagonale, a u drugom oko apscise (veća odstupanja sugeriraju lošiji model).

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 - promjene u modelu su rezultat drugih faktora, potrebno korištenje vremenskih nizova koji uključuju varijablu vremena)
- čiste koreliranosti (loš odabir modela, koristiti modele koji uključuju koreliranost podataka, npr. ARMA)

Napomena: Nezavisnost dviju normalnih slučajnih varijabli povlači njihovu nekoreliranost.
Budući da se nezavisnost u principu ne može testirati, testiramo nekoreliranost.

Serijsku korelaciju među greškama ε možemo procijeniti koristeći reziduale 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 $4/\sqrt{n}$).

(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).

Ključna pretpostavlja kako bi izbjegli probleme s predviđanjima i omogućili dobre procjene.

Homogenost se može uočiti iz grafičkog prikaza

- reziduala obzirom na vrijeme ($t-$ e graf), ako se radi o vremenskom nizu
- residual-fit plot-a ($\hat{Y}-$ e graf).

Kod oba grafa pozornost treba obratiti na rast reziduala, želimo postići simetrično raspršenje oko osi x.

Moguća rješenja su:

- analiza podataka po dijelovima s jednakom varijancom (više linearnih modela)
- 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, budući da je normalnost grešaka korištena pri određivanju distribucija ostalih varijabli vezanih uz model.

Jedan od uzroka može biti i prisutnost *outliera*, koji mogu stvoriti probleme pri procjeni parametara modela (povećanje kvadratne greške).

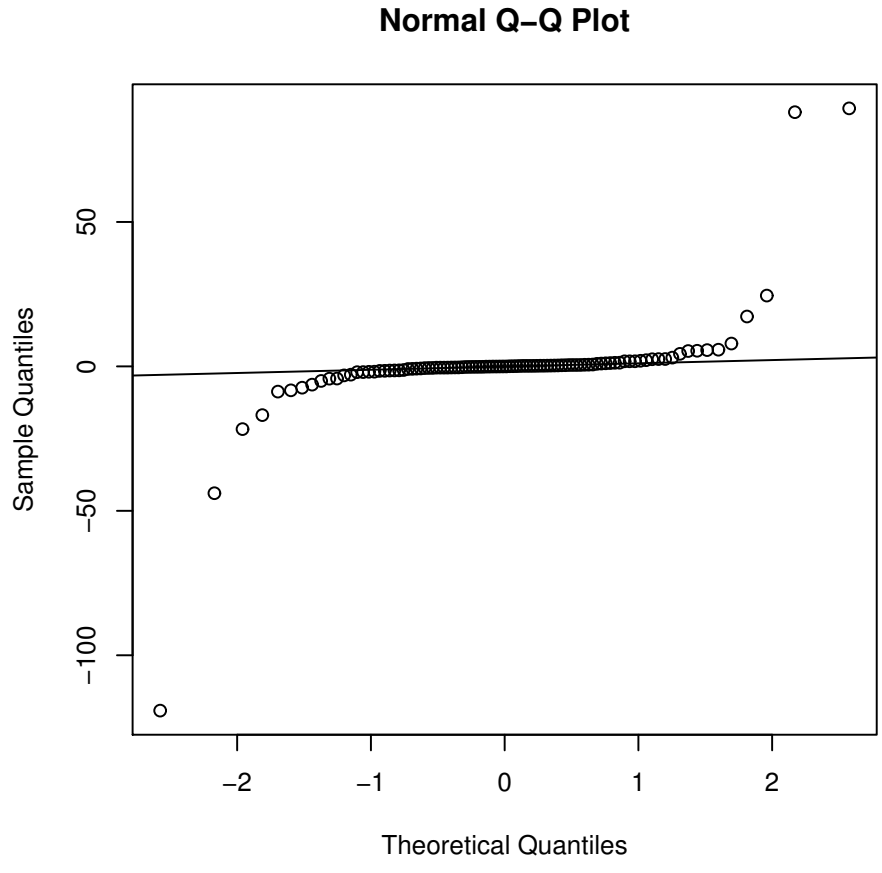
Normalnost grešaka možemo provjeriti preko reziduala, korištenjem

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

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

- Jarque-Bera testa.

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 e

Ako su greške ε_i nezavisne i (normalno) jednako distribuirane, ostaci e ne moraju biti. Naime $\mathbf{e} = (\mathbb{I} - \boldsymbol{H})\varepsilon$,,
$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, pa je $\mathrm{Var}(e) = (\mathbb{I} - \boldsymbol{H})\sigma^2$, odnosno

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}).$$

Problem se rješava uvođenjem studentiziranih ostataka

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

Ako su pretpostavke na grešku modela točne, tada je $\mathrm{Var}(r_i) = 1$, a korelacija $\mathrm{Cor}(r_i, r_j)$ je mala.
U R-u ih dobivamo pomoću naredbe

```
> rstudent(model)
```

i u praksi radimo sa studentiziranim rezidualima.

Poluga

Poluge h_{ii}

- mjere osjetljivost procijenjenog $\hat{y}_i$ s obzirom na promjenu opažene vrijednosti y_i (odnosno, koliko i-ta opservacija utječe na procjenu parametara regresije)
- iz intervala $[0, 1]$ (što je poluga veća, utjecaj pojedine opažene vrijednosti na predviđenu je veći)
- $\sum h_{ii} = p + 1$

Točke visoke poluge - neobično velike ili male vrijednosti prediktora ili neobične vrijednosti s obzirom na vrijednosti ostalih prediktora, u praksi $h_{ii} > 2p/n$

Outlieri - točka čija opažena vrijednost ne prati trend ostalih točaka (opažena vrijednost značajno odstupa od ostalih opaženih vrijednosti točaka sličnih vrijednosti varijabli prediktora)

Utjecajne točke

Utjecajne točke u modelu

- imaju značajan utjecaj na neki dio procijenjenog modela (npr. predviđene vrijednosti, procijenjene koeficijente ili p-vrijednosti testova)
- njihovim uklanjanjem značajno mijenjamo dobivene procjene
- kandidati su točke visoke poluge i outlieri.

Jedan način da otkrivanja je računajući **Cookovu udaljenost** D_i koja kvantificira sveukupan utjecaj točke:

- > 0.5 - moguće je da je točka utjecajna
- > 1 - točka je vrlo vjerojatno utjecajna
- D_i poput palca (poput slova T) odskače od ostalih Cookovih udaljenosti D_j - točka je gotovo sigurno utjecajna

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 $i = 0, 1, \ldots, 100$ 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 i usporedite ga s pravcem $y = x$.

Testiranje (linearnih) hipoteza o parametrima

Možemo li ukloniti neke varijable iz modela, čiji je doprinos mali i bez njih ne dobivamo značajno lošiji model?

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

$$H_0: \quad \boldsymbol{C}b = \mathrm{g}$$

testiramo statistikom

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

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 da je uži model dovoljan, tj.

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

```
> pl2=lm(yy~offset(1*x)+sin(x))
> anova(pl2,pl)
Analysis of Variance Table

Model 1: yy ~ offset(x) + sin(x)
Model 2: yy ~ 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

Želimo izabrati *najbolji* podskup varijabli poticaja (*optimalni* model):

- želimo objasniti podatke na najjednostavniji mogući način (višak varijabli sigurno neće pogoršati model, ali nas zanima hoće li ga *značajno* poboljšati)
- višak varijabli poticaja pojačat će *šumove* kod izračunavanja procjena
- može doći do kolinearnosti - više varijabli pokušava procijeniti isti dio nezavisne varijable (teško se procjenjuje utjecaj pojedine varijable jer se efekt podijeli na njih više)
- veliki modeli su neučinkoviti (invertiranje velikih matrica)
- izbacivanjem prevelikog broja varijabli poticaja gubimo moć predikcije

Procedure za izbor varijabli

1. hijerarhijski izbor modela (anova);
2. procedure korak po korak (dodavanjem ili uklanjanjem jedne po jedne varijable);
3. procedure bazirane na kriterijima (kojeg računamo za sve modele od interesa i na temelju njih odabiremo najbolji).

AIC i BIC kriteriji

Za p mogućih varijabli poticaja, imamo 2^{p+1} mogućih modela od kojih želimo odabrati najbolji prema nekom kriteriju.

1. *Akaike Information Criterion*

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

2. *Bayes Information Criterion*

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

Za linearne modele

$$-2\text{log-likelihood} = n \log(SSE/n).$$

$(SSE = \sum_{k=1}^{n} (\hat{y}_i - y_i)^2)$.

- želimo što veći log-likelihood
- veći modeli $\Longrightarrow$ manji SSE i veći p (detektiramo varijable koje ne smanjuju grešku dovoljno značajno)
- najbolji model je ravnoteža između broja parametara i pristajanja podacima
- BIC strože kažnjava veće modele
- različiti kriteriji mogu donijeti različite odluke

Za određivanje AIC i BIC koristimo naredbu

```
> AIC(ul, k=...),
```

gdje k označava način kažnjavanja velikih modela ($k_{\mathrm{AIC}} = 2$, $k_{\mathrm{BIC}} = \log(n)$)

Ukoliko želimo provesti hijerarhijsku analizu optimalnog modela, da izbjegnemo ručno računanje kriterija za svaki model, koristimo funkciju

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

gdje je `model` najopćenitiji model (sa svim varijablama).

Zadatak

Modelirajmo podatke iz primjera s uniformno distribuiranim greškama i to modelom s funkcijom sinus i polinomom 3. stupnja. Nađimo najpogodniji pod-model za ove podatke

(1) po AIC i BIC kriteriju,
(2) testirajući razliku između modela prigodnim statističkim testom, i to krećući od
    (a) punog modela izbacujući jednu po jednu varijablu
    (b) nul-modela dodajući jednu po jednu varijablu.

Underfitting vs. overfitting

Koliko dobro naš model uči iz podataka?

- *underfitting* - prilagodba modela podacima dobra, ali nedovoljno precizna i prejednostavna (predviđene vrijednosti daleko od stvarnih zbog nedostatka važnih odnosa)
- *overfitting* - model predobro odgovara podacima (prekompleksan model koji opisuje šumove umjesto pravih uzoraka). Model će loše predviđati na novom skupu podataka. Moguća rješenja:
  - *cross-validation* (uzimanje slučajnih poduzoraka od početnog uzorka, na njima se provodi prilagodba te potom uzima neki prosjek rješenja)
  - više podataka
  - jednostavniji podaci
  - dodavanje šuma početnim podacima

Ometajuće varijable

Ometajuće varijable (*confounding variables*)

- ometaju vezu između zavisne i važne nezavisne varijable
- utječu na neku drugu nezavisnu varijablu i na varijablu odziva, čime daje na važnosti toj varijabli odziva koja možda nije potrebna u modelu

Kolinearnost

- statistička povezanost između nezavisnih varijabli
- teško je procijeniti utjecaj pojedine nezavisne varijable na varijablu odziva
- male promjene u podacima značajno mijenjaju procijenjene koeficijente

Mjere kolinearnosti:

- **Tolerancija**
  - mjeri udio varijance u nezavisnoj varijabli koji nije objašnjen ostalim nezavisnim varijablama
  - $T = 1 - R^2$, gdje je R^2 koeficijent determinacije, dobiven regresijom promatrane varijable na sve ostale nezavisne varijable
  - $T < 0.1$ - visok stupanj multikolinearnosti (većinu varijance prediktora objašnjavaju druge prediktorske varijable)
- **VIF** (Variance inflation factor)
  - koliko se varijanca regresijskog koeficijenta povećava zbog kolinearnosti među prediktorima
  - $\text{VIF} = \frac{1}{T} = \frac{1}{1-R^2}$
  - visoka vrijednost VIF-a znači da prediktor dijeli puno varijance s drugim prediktorima, što može dovesti do nestabilnih procjena koeficijenata
  - $\text{VIF} > 5$ potrebno detaljnije promotriti varijable poticaja
  - $\text{VIF} > 10$ značajna multikolinearnost