GENERALIZIRANI LINEARNI MODELI

STATISTIČKI PRAKTIKUM 2

9

Do sada

1. $Y = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$, $x_1, \ldots, x_p$ kvantitativne varijable prediktori
2. prediktori mogu biti i faktorske varijable (paralelni pravci), npr. $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 d_1 + \alpha_3 d_2 + \varepsilon$
3. uvodimo interakciju između prediktora $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 d_1 + \alpha_3 d_2 + \alpha_4 x_1 d_1 + \alpha_5 x_1 d_2 + \varepsilon$
4. dozvoljavamo ne-Gaussovsku strukturu varijable Y

GLM čine široku klasu linearnih modela koja obuhvaća modele sa

- specijalnim strukturama grešaka
- kategorijskim ili uređenim varijablama odaziva
- multinomijalnim varijablama (zavisnim)
- ...

$$Y = g^{-1}(X\beta) + \varepsilon, \quad g(\mathbb{E}Y) = X\beta$$

U standardnom linearnom modelu greške su bile n.j.d. $\sim N(0, \sigma^2)$, a funkcija g je bila identiteta.

Sada Y ne mora više biti neprekidna, ali u `R`-u ne možemo raditi sa bilo kakvim funkcijama g i distribucijama od Y.

Osnove modela

$$Y = g^{-1}(X\beta) + \varepsilon$$

GLM sastoji se od tri elementa:

1. slučajni - vjerojatnosna distribucija F iz eksponencijalne familije distribucija, $Y \sim F$;
2. sistemski - linearni prediktor $X\beta$;
3. veza između slučajne i sistemske komponente - vezna (*link*) funkcija g t.d. $\mu = \mathbb{E}Y = g^{-1}(X\beta)$.

Gustoća iz *eksponencijalne familije* se može zapisati u obliku

$$f(x \mid \theta) = h(x) \exp\left\{\eta(\theta) T(x) - A(\theta)\right\},$$

npr. binomna, Poissonova, eksponencijalna, normalna, geometrijska distribucija.

Procjena parametara

Parametri modela β mogu se procijeniti metodom maksimalne vjerodostojnosti (ML). Procjenitelji se općenito ne mogu dobiti u zatvorenoj formi, ali se uvijek mogu procijeniti iterativnom metodom najmanjih kvadrata s težinama (IWLS).
U `R`-u je implementirana procedura `glm`

```
> glm
glm(formula, family = gaussian, data, weights, subset,
    na.action, ...)
```

Parametrom `family` specificiramo distribuciju i link funkciju modela.

glm

U proceduri `glm` implementirano je 6 distribucija s najčešće korištenim pripadnim link funkcijama:

Distribucija	Link funkcija
normalna	identiteta
binomna	logit, probit
Poissonova	log , identiteta, korijen
...	...

Bernoullijeva razdioba

```
family=binomial(link=logit)
family=binomial(link=probit)
```

Neka Y poprima vrijednosti 0 ili 1, tj. $Y \sim B(p)$, $p = \mathbb{E}Y$. Pri odabiru generaliziranog linearnog modela često se promatraju dvije link funkcije:

- logit $g(y) = \ln\left(\frac{y}{1-y}\right)$, $y \in \langle 0, 1\rangle$
- probit $g(y) = \Phi^{-1}(y)$, $y \in \langle 0, 1\rangle$

Zadatak

U datoteci `binary.csv` nalaze se podaci o uspješnosti upisa 400 studenata na poslijediplomske studije. Za svakog su aplikanta dani rezultati GRE testa, prosjek ocjena (GPA) i rang fakulteta na koji se aplicirao.
Procijenite parametre probit modela za dane podatke, odredite pripadne 95% pouzdane intervale za koeficijente te na temelju danog modela procijenite koja je vjerojatnost da student s GRE rezultatom 750 i prosjekom 3.88 upadne na poslijediplomski program na sveučilištu ranga 1.

Rezultati:

- *devijanca* - mjera odstupanja opažanja od očekivane vrijednosti dane modelom
- *null-deviance* =
  $2(\log L(\text{saturirani model}) - 2\log L(\text{null model}))$ = pokazatelj koliko dobro osnovni model (samo slobodni član) opisuje podatke
- *residual deviance* =
  $2(\log L(\text{saturirani model}) - 2\log L(\text{proposed model}))$ = pokazatelj koliko dobro predloženi model opisuje podatke, odgovara sumi kvadrata reziduala u standardnom linearnom modelu
- AIC (*Akaike information criterion*) $= -2\log L - k \cdot (p+1)$, gdje je $p+1$ broj parametara modela

Odstupanje točaka

Točke u modelu mogu odstupati od ostalih točaka po

- x-osi - točke **visoke poluge** (*eng. high leverage*)
- y-osi - outlieri

Točka (podatak) također može biti **utjecajna** za model ako ima značajan utjecaj na neki dio procijenjenog modela (predviđene vrijednosti varijable odaziva, procijenjene parametre ili procijenjen utjecaj pojedinog prediktora u modelu), tj. izbacivanjem te točke iz modela procijenjene vrijednosti se značajno promijene.

Outlieri i točke visoke poluge mogu i ne moraju biti utjecajne u modelu.

Poluga h_i mjeri odstupanja po x-osi i-tog podatka od prosjeka vrijednosti svih podataka (broj u rasponu od 0 do 1). Predstavlja utjecaj koji opažena vrijednost y_i ima na predviđenu vrijednost $\hat{y}_i$. Što je poluga veća, to je uloga i-tog podatka u formiranju predviđanja $\hat{y}_i$ veća. Vrijedi: $\sum_{i=1}^{n} h_i = p + 1$.

Potencijalno utjecajne točke su točke visoke poluge za koje je

$$h_i > 3\frac{p+1}{n}.$$

Outlieri - podaci čiji su studentizirani reziduali r_i po apsolutnoj vrijednosti veći od 3.

Utjecajne točke

Jedna od mjera kojom možemo identificirati utjecajne točke u modelu je **Cookova udaljenost** (*eng. Cook's distance*). Definirana je s

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p+1)MSE} \frac{h_i}{(1-h_i)^2}.$$

D_i opisuje koliko se sve predviđene vrijednosti promijene kada iz modela izbacimo i-ti podatak, tj. koliko je ta točka utjecajna.

- $D_i > 0.5$ - i-ta točka je možda utjecajna (treba dalje istražiti)
- $D_i > 1$ - i-ta točka je vjerojatno utjecajna
- u usporedbi s ostalim Cookovim udaljenostima, D_i značajno odskače po svojoj vrijednosti (graf izgleda poput slova T) - i-ta točka je gotovo sigurno utjecajna

```
> plot(cooks.distance(model))
> plot(cooks.distance(model), type="h")
```

Što s utjecajnim točkama?

Ekstremnost ne znači nužno da točke trebamo izbaciti iz modela. U principu, točke ćemo izbaciti ukoliko se radi o pogrešci u mjerenju ili prilikom upisivanja podataka. Također, moguće je da podatak nije reprezentativan za populaciju koju promatramo. Točku nećemo izbaciti samo zato što se ne uklapa u model koji indiciraju ostale točke. Svako izbacivanje podataka treba dobro opravdati.

Ukoliko nemamo razloga za odbacivanjem utjecajne točke ili nismo sigurni radi li se o pogrešci možemo:

- provjeriti model koji smo dobili (goodness-of-fit) i napraviti potrebne promjene (uvesti nove prediktore ili interakciju među postojećim prediktorima, proučiti odnos između varijabli i uzeti model koji bi bolje opisao taj odnos),
- analizirati oba modela: s točkom i bez nje i prikazati obje varijante.

Analiza reziduala

Pozivanjem naredbe `plot(model)` dobijemo grafičku analizu reziduala:

- reziduali vs. predviđene vrijednosti $\hat{y}_i$
- normalni vjerojatnosni graf studentiziranih reziduala
- $\sqrt{|r_i|}$ vs. $\hat{y}_i$ (*scale-location plot*, raspored reziduala obzirom na lokaciju podataka unutar modela)
- r_i vs. h_i zajedno s krivuljama Cookovih udaljenosti