

GENERALIZIRANI LINEARNI MODELI. PROPENSITY SCORE MATCHING.

STATISTIČKI PRAKTIKUM 2

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.

Osnove modela

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. $\mu = \mathbb{E}Y = g^{-1}(X\beta)$.

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.

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 odskoče po svojoj vrijednosti (graf izgleda poput slova T) - i -ta točka je gotovo sigurno utjecajna

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*)
- ▶ r_i vs. h_i zajedno s krivuljama Cookovih udaljenosti

Propensity Score Matching

PSM je metoda procjenjivanja efekata tretmana na temelju modela koji ocjenjuje vjerojatnost primanja tretmana.

Neka su X_1, \dots, X_k varijable poticaja koje utječu na primanje tretmana, X varijabla koja mjeri efekt tretmana te Y binarna varijabla koja označava tretman ($Y \in \{0, 1\}$).

Ideja:

1. Pomoću binarnog modela procijeniti vjerojatnost ulaska u tretman.
2. Svakoj tretiranoj jedinki naći (jedan ili više) par iz netretirane skupini s najsličnijom vjerojatnosti ulaska u tretman.
3. Efekt tretmana uspoređivati na temelju dvaju dobivenih grupa.

Uparivanje u PSM-u

Traženje para u PSM-u može se odvijati sa i bez vraćanja, najčešće korištene metode su

- ▶ metoda najbližeg susjeda
- ▶ caliper metoda (najbliži susjed s radijusom)
- ▶ metoda radijusa
- ▶ Mahalanobis metoda

Primjer

U datoteci `lalonde.txt` nalaze se podaci za analizu efekta programa usavršavanja u SAD-u 70-tih godina.

(LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review* 76:604-620.).

```
> names(lalonde)
[1] "age"      "educ"     "black"    "hispanic" "married"  "nodegr"   "re74"
[8] "re75"    "re78"     "u74"      "u75"      "treat"
```


Prvo procijenimo pripadni probit model:

```
> model=glm(treat~age+educ+black+hisp+married+
  nodegr+re74+re75,family=binomial(link=probit),data=lalonde)
> summary(model)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	7.300e-01	6.536e-01	1.117	0.26406
age	2.876e-03	8.868e-03	0.324	0.74571
educ	-4.403e-02	4.444e-02	-0.991	0.32175
black	-1.403e-01	2.273e-01	-0.617	0.53696
hisp	-5.232e-01	3.087e-01	-1.694	0.09018 .
married	1.029e-01	1.717e-01	0.599	0.54917
nodegr	-5.622e-01	1.943e-01	-2.893	0.00381 **
re74	-1.969e-05	1.568e-05	-1.255	0.20931
re75	3.864e-05	2.679e-05	1.442	0.14916

Procijenimo vrijednosti varijable odaziva Y , tj. vjerojatnost ulaska u program usavršavanja \hat{Y} :

```
> p=predict(model, type="response", se.fit = TRUE)[1]
```

Promotrimo stvarne i procijenjene vrijednosti,

```
> x=cbind(lalonde$treat,p)
```

te za svaku osobu koja je ušla u program ($Y = 1$) nađemo osobu koja nije bila u programu ($Y = 0$) s najbližom vjerojatnošću ulaska u program (\hat{Y}).

```
> par=matrix(0,nrow=185,ncol=2)
> indeks=186:445

> for(i in 1:185){
+ m=min(abs(p[186:445]-p[i]))
+ par[i,]=c(indeks[abs(p[186:445]-p[i])==m][1],m)
+ }
```

Sada možemo testirati efekt tretmana X =zarada 1978 godine (nakon ulaska u program osposobljavanja).

```
> x=lalonde$re78
```

Podijelimo x u dvije skupine - skupinu tretiranih i skupinu njihovih parova.

```
> x1=x[1:185]
> x0=x[par[,1]]
```

Prigodnim testom provjerimo postoji li razlika u mjerenoj varijabli X između tretirane i netretirane skupine (efekt tretmana)

```
> wilcox.test(x1,x0,alternative="greater",paired=T)
```

Wilcoxon signed rank test with continuity correction

data: x1 and x0

$V = 9873$, $p\text{-value} = 0.0001871$

alternative hypothesis: true location shift is greater than 0