Generalizirani linearni modeli II

(Poissonova regresija, prekomjerna disperzija, interakcije, modeliranje stopa)

H. Planinić

Prosinac 2023.

Poissonova regresija

- Ukoliko imamo odzive $y_i \in \{0, 1, 2, \dots, \}$ (counts), možemo koristiti GLM uz Poissonovu razdiobu kao model za distribuciju odziva, tj. pretpostavka je $Y_i \sim \text{Poiss}(\mu_i)$, gdje je $\mu_i = \mathbb{E}[Y_i \mid X^{(i)} = x^{(i)}]$.
- kanonska funkcija veze je $g(\mu) = \log(\mu)$ – u tom slučaju očekivanje i linearni prediktor povezani su kao

$$\mu(x) := \mathbb{E}[Y \mid X = x] = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\,.$$

- u ovom slučaju je, kao i kod logističke regresija, parametar disperzije $\phi = 1$, te su devijanca i skalirana devijanca modela jednake.
- za razliku od logističke regresije, devijanca za velike n i ako je naš model točan ima približno χ^2 razdiobu s brojem stupnjeva slobode jednakim $n - p$, što omogućava testiranje hipoteze da je naš model (približno) točan.[1]

[1] Ipak, to ne vrijedi baš uvijek, ali diskusija o tome izlazi van okvira ovog kolegija. Sličan slučaj je zapravo i kod binomnog modela.

Interpretacija parametara

– pretp. da je X_j kvantitativna kovarijata te da je skup kovarijata $x' \in \mathbb{R}^p$ dobiven iz x tako da samo x_j povećamo za 1, a sve ostale kovarijate ostavimo nepromijenjene, imamo da je

$$\frac{\mu(x')}{\mu(x)} = \frac{\mathbb{E}[Y \mid X = x']}{\mathbb{E}[Y \mid X = x]} = \exp\{\beta_j\}$$

$\rightsquigarrow$ dakle, kada se X_j poveća za 1 i sve ostale kovarijate ostanu iste, očekivanje se mijenja za faktor $\exp\{\beta_j\}$.

– analogna interpretacija je i u slučaju kategorijalne kovarijate

– za slobodni član imamo da je

$$\mathbb{E}[Y \mid x = (0, \ldots, 0)] = \exp(\beta_0)\,.$$

Zadatak 1

Učitajte podatke `intention.rda`:

"Subjects navigated a website that contained, among other things, an advertisement for candies. During the site navigation, an “eye-tracker” measured the location on the screen on which the subject's eyes were fixated. The tracker also recorded whether the subject saw the ad and for how long it was in sight. Additionally, facial expression analysis software (FaceReader) can be used to guess the subject's emotions when the ad was in sight. At the end of the study, a questionnaire measured the subject's intention to buy this type of candy and socio-demographic variables. Only the 120 subjects that had seen the ad in question are included in the data."

Zadatak 1

Zanimat će nas kako broj kupljenih paketića slatkiša `nitem` ovisi o sljedećim kovarijatama:

- `fixation` : the total duration of fixation on the ad (in seconds).
- `emotion` : a measure of reaction during fixation; the ratio of the probability of showing a positive emotion to the probability of showing a negative emotion.
- `sex` : sex of subject, either man (0) or woman (1).
- `age` : age (in years).
- `revenue` : categorical variable indicating the subject's annual income; one of (1) [0, 20k); (2) [20k, 60k); (3) 60k and above.
- `educ` : categorical variable indicating the highest educational achievement, either (1) high school or lower; (2) college or (3) university degree.
- `marital` : civil status, either single (0) or in a relationship (1).

Zadatak 1

(a) Prikažite marginalne ovisnosti odziva o svim gorespomenutim kovarijatama (za kategorijalne kovarijate koristite npr. `boxplot`). Postoji li naznaka da neke od kovarijata utječu na broj kupljenih paketića?

(b) Prilagodite Poissonovu regresiju ovim podacima (uz kanonsku funkciju veze) te napišite koji je točno model pretpostavljen. Koristeći test omjera vjerodostojnosti (uspoređujući puni model s modelima u kojima je izbačena po jedna kovarijata) testirajte statističku značajnost svake od kovarijata. Interpretirajte koeficijente uz statistički značajne kovarijate.

(c) Provjerite prilagodbu modela uspoređujući devijancu s odgovarajućom χ^2 razdiobom. Čini li se prilagodba dobra?

Prekomjerna disperzija

- U Poissonovom modelu pretpostavljamo da je

$$\mathrm{Var}(Y \mid X = x) = \mathbb{E}[Y \mid X = x] = \mu(x)\,.$$

- Ukolko se prilagodba Poissonovg modela ne čini dobra (kao što je to npr. slučaju Zadatku 1), često je problem u tome što je zapravo

$$\mathrm{Var}(Y \mid X = x) > \mu(x)\,,$$

  fenomen koji nazivamo prekomjerna disperzija (engl. overdispersion).

- U tom slučaju možemo probati negativni binomni model.

Negativna binomna razdioba

- Koristit ćemo sljedeću parametrizaciju negativne binomne razdiobe: za parametre $\mu > 0$ i $\alpha > 0$, definiramo je kao Poissonovu razdiobu čiji je parametar slučajan te ima gama razdiobu s parametrima μ/α i $1/\alpha$.
- Očekivanje i varijanca su

$$\mathbb{E}[Y] = \mu\,, \quad \mathrm{Var}(Y) = \mu + \alpha\mu^2\ (> \mu)$$

- Dakle, imamo dodatni parametar koji dopušta fleksibilniju vezu između očekivanja i varijance.
- Kada $\alpha \to 0$, dobivamo Poissonovu razdiobu s parametrom μ, kao specijalan slučaj.

Negativna binomna regresija

– u negativnoj binomnoj regresiji pretpostavljamo da Y_i ima negativnu binomnu razdiobe te da je funkcija veze $g(\mu) = \mu$, to jest da opet imamo

$$\mu(x) := \mathbb{E}[Y \mid X = x] = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\,.$$

pa je interpretacija parametara ista kao i za Poissonov slučaj.

– pretpostavljamo da je parametar α isti za sve Y_i-eve te se također procjenjuje iz podataka (nećemo ulaziti u detalje).[2]

– u `R`-u možemo koristiti funkciju `glm.nb` iz paketa `MASS` pri čemu je parametar θ zapravo $\theta = 1/\alpha$.

[2] Također, nije jasno, tj. ja nisam uspio naći rezultate u literaturi, ima li i pod kojim uvjetima devijanca negativnog binomnog modela približno χ^2 razdiobu (kao što je to nekad slučaj s binomnim i Poissonovim modelom).

Je li nam potreban NB u odnosu na Poissonov model?

– ukoliko želimo testirati $H_0 : \alpha = 0$ (tj. dovoljan je Poissonov model) u odnosu na $H_1 : \alpha > 0$, možemo koristiti činjenicu da uz H_0, za velike n,

$$T = 2(\widehat{\ell}_{\mathrm{NB}} - \widehat{\ell}_{\mathrm{Poiss}})$$

ima miješanu razdiobu $\frac{1}{2}\chi^2_1 + \frac{1}{2}\delta_0$; χ^2 ima jedan stupanj slobode jer smo dodali jedan parametar, to je α.

– intuitivno, velike vrijednosti od T sugeriraju da nam je potreban NB model, a formalno, ako je $t > 0$ realizacija od T, p-vrijednost testa računamo kao $\frac{1}{2}\mathbb{P}(\chi^2_1 > t)$.[3]

[3] Zašto?

Zadatak 2

(a) Podacima iz Zadatka 1 prilagodite NB model te usporedite dobivene koeficijente. Koristeći test omjera vjerodostojnosti testiratjte značajnost svake od kovarijata te usporedite s rezultatom iz Poissonovog modela.

(b) Koristeći test omjera vjerodostojnosti testiraje ima li dovoljno dokaza da nam je potreban NB model u odnosu na Poissonov model; za vrijednost log-vjerodostojnosti koristite funkciju `logLik`. Što vam to govori o zaključcima o značajnosti parametara iz (a) dijela u odnosu na one iz Zad1(b) dijela?

(c) Na istom grafu prikažite točke $(\widehat{\mu}_i, y_i)$, $i = 1, \ldots, n$ gdje je $\widehat{\mu}_i$ procjena za $\mathbb{E}[Y \mid X = x^{(i)}]$ iz NB, odnosno Poissonovog modela.

Interakcije

- pretpostavimo da je odziv $Y \in \{0, 1, 2, \dots\}$ te da imamo jednu binarnu kovarijatu $X_1 \in \{\text{musko}, \text{zensko}\}$ te jednu kvantitativnu kovarijatu $X_2 \in \mathbb{R}$, te da koristimo model

$$\mathbb{E}[Y \mid X = x] = \exp\{\beta_0 + \beta_1 1_{\{x_1 = \text{zensko}\}} + \beta_2 x_2\}$$

- u ovom slučaju, $\exp\{\beta_2\}$ je faktor promjene očekivanja ako se x_2 poveca za 1 za dvije osobe istog spola – dakle, faktor promjene ne ovisi o spolu.

- često ta pretpostavka nije opravdana pa u model trebamo uključiti interakciju između spola i X_2

Interakcije

– preciznije, promatramo model

$$\mathbb{E}[Y \mid X = x] = \exp\{\beta_0 + \beta_1 1_{\{x_1 = \text{zensko}\}} + \beta_2 x_2 + \beta_3 x_2 1_{\{x_1 = \text{zensko}\}}\}$$

– u ovom modelu imamo

$$\mathbb{E}[Y \mid X = x] = \begin{cases} \exp\{\beta_0 + \beta_2 x_2\}, & x_1 = \text{musko}, \\ \exp\{\beta_0 + \beta_1 + (\beta_2 + \beta_3) x_2\}, & x_1 = \text{zensko}. \end{cases}$$

$\rightsquigarrow$ sada je faktor promjene za muškarce jednak $\exp\{\beta_2\}$, a za žene $\exp\{\beta_2 + \beta_3\}$
$\rightsquigarrow$ dakle, $\exp\{\beta_3\}$ faktor faktora promjene za žene u odnosu na muškarce.

Interakcije

- uočimo da i koeficijent β_1 više nema istu interpretaciju: u modelu bez interakcija $\exp\{\beta_1\}$ je faktor promjene očekivanja za žene u odnosu na muškarce s istom vrijednosti kovarijate x_2 – dakle, faktor promjene ne ovisi o x_2.
- u modelu s interakcijom, faktor promjene je $\exp\{\beta_1 + \beta_3 x_2\}$.

Napomene

- u `R`-u interakciju između varijabli `a` i `b` dodajmo koristeći naredbu `a:b`, a naredba `a*b` je isto što i `a+b+a:b`, dakle dodaje interakciju te svaku od varijabla zasebno.
- Ukoliko u model uključujemo interakciju između dvije varijable, model bi trebao uključivati i obje kovarijate zasebno (tzv. main effects), bez obzira na to bile one statistički značajne ili ne.

Zadatak 3

Ponovno promatramo podatke iz `intention.rda` uz varijablu `nitem` kao odziv.

(a) Prikažite na istom grafu broj kupljeni paketića slatkiša u odnosu na vrijednost varijablu `fixation` posebno za žene te posebno za muškarce; npr. koristite drugačiju boju za točke, možete dodati i legendu na graf tako da bude jasno koje su koje točke. Prilagodite Poissonov GLM ovim podacima koristeći samo kovarijate `sex` i `fixation`. Nacrtajtu ne prethodnom grafu procijenjene očekivane vrijednosti u modelu u ovisnosti o varijabli `fixation`, posebno za žene te posebno za muškarce.

(b) Prilagodite novi model u kojem ćete dodati i interakciju između spola i varijable `fixation`. Napišite koji je točno model te interpretirajte koeficijent u varijablu interakcije (to može uključivati i druge koeficijente iz modela). Nacrtaje isti graf s predikcijama kao i u (a) dijelu zadatka, te ih usporedite.

(c) Koristeći test omjera vjerodostojnosti testirajte je li varijabla interakcija statistički značajna u odnosu na model bez nje.

Interakcije - dvije kategorijalne kovarijate

– pretpostavimo da je odziv $Y \in \{0, 1, 2, \dots\}$ te da imamo dvije binarne kovarijate $X_1 \in \{\text{musko}, \text{zensko}\}$, $X_2 \in \{\text{dijete}, \text{odrasli}\}$, te jednu kvantitativnu kovarijatu $X_3 \in \mathbb{R}$, model bez interakcija je npr.

$$\mathbb{E}[Y \mid X = x] = \exp\{\beta_0 + \beta_1 1_{\{x_1 = \text{zensko}\}} + \beta_2 1_{\{x_2 = \text{odrasli}\}} + \beta_3 x_3\}$$

– ako u model dodamo interakciju između spola i dobi, dobivamo model s jednim dodatnim parametrom

$$\mathbb{E}[Y \mid X = x] = \exp\{\beta_0 + \beta_1 1_{\{x_1 = \text{zensko}\}} + \beta_2 1_{\{x_2 = \text{odrasli}\}} + \beta_3 x_3 + \beta_4 1_{\{x_1 = \text{zensko}, x_2 = \text{odrasli}\}}\}$$

Interakcije - dvije kategorijalne kovarijate

– dakle, imamo

$$\mathbb{E}[Y \mid X = x] = \begin{cases} \exp\{\beta_0 + \beta_3 x_3\}, & \text{muško, dijete} \\ \exp\{\beta_0 + \beta_1 + \beta_3 x_3\}, & \text{žensko, dijete} \\ \exp\{\beta_0 + \beta_2 + \beta_3 x_3\}, & \text{muško, odrasli} \\ \exp\{\beta_0 + \beta_1 + \beta_2 + \beta_3 x_3 + \beta_4\}, & \text{žensko, odrasli} \end{cases}$$

– npr.

(i) $\exp\{\beta_1\}$ je faktor promjene očekivanja za žensko u odnosu na muško dijete s istom vrijednosti za x_3

(ii) $\exp\{\beta_1 + \beta_4\}$ je faktor promjene očekivanja za žensku u odnosu na mušku odraslu osobu s istom vrijednosti za x_3

Modeliranje stopa

- do sada u Poissonovoj ili NB regresiji implicitno pretpostavljali da su Y_i-evi za različite i-eve usporedivi.
- to često nije slučaj, npr. broj smrti u regiji ovisi o broju stanovnika
- ako Y_i-evi nisu uspredivi, možemo modelirati stope, u gornjem primjeru to je broj smrti po stanovniku

Primjer 1

Promatramo podatke iz `crash.rda`:

"The National Highway Traffic Safety Administration (NHTSA) compiles statistics about road traffic deaths in the Fatality Analysis Reporting System. The yearly mortality counts for 2010 and 2018 are given in `crash` according to whether the accident occured during daytime or nighttime (`time`), and according to the NHTSA-defined geographic area (`region`)."

- Neka je odziv Y_i broj smrti (varijabla `ndeath`) u danoj godini (2010. ili 2018., kovarijata je `year`), po danu ili po noći, za jednu od regija; dakle, ukupan broj podataka je 4 puta broj regija.
- Neka je n_i broj stanovnika u danoj regiji (varijabla `popn`).

Primjer 1

- Standardni Poissonov model bio bi

$$\mathbb{E}[Y_i \mid X^{(i)} = x^{(i)}] = \exp\{\beta_0 + \beta_1 1_{\{\text{year}=2018\}} + \beta_2 1_{\{\text{time}=\text{noc}\}}\}$$

- Ipak, smislenije je modelirati stopu smrti po stanovniku Y_i/n_i, tj. pretpostaviti model

$$\mathbb{E}[Y_i/n_i \mid X^{(i)} = x^{(i)}] = \exp\{\beta_0 + \beta_1 1_{\{\text{year}=2018\}} + \beta_2 1_{\{\text{time}=\text{noc}\}}\}$$

što je ekvivalentno s

$$\begin{aligned}\mathbb{E}[Y_i \mid X^{(i)} = x^{(i)}] &= n_i \exp\{\beta_0 + \beta_1 1_{\{\text{year}=2018\}} + \beta_2 1_{\{\text{time}=\text{noc}\}}\} \\ &= \exp\{\beta_0 + \beta_1 1_{\{\text{year}=2018\}} + \beta_2 1_{\{\text{time}=\text{noc}\}} + \log(n_i)\}\,.\end{aligned}$$

- član $\log(n_i)$ je dakle uključen kao kovarijata i naziva se `offset`, s tim da se njen koeficijent ne procjenjuje već je postavljen kao $\beta = 1$.

Zadatak 4

Promatramo podatke iz `crash.rda` i broj smrti u regiji `ndeath` kao odziv.

(a) Grafički prikažite (npr. koristeći `boxplot`-ove) kako broj smrti po stanovniku ovisi o kovarijatama `year` i `time`. Uočavate li neke povezanosti između kovarijata i odziva?

(b) Prilagodite Poissonov ili negativni binomni model podacima koristeći kovarijate `year` i `time` te koristeći `log(popn)` kao `offset`; u formulu za model jednostavno dodajte `offset(log(popn))` kao kovarijatu. Čini li se potreban NB model? Ako da, u nastavku koristite njega.

(c) Napišite koji je pretpostavljeni model, testirajte značajnost svake od kovarijata (ne i `offset`-a) koristeći test omjera vjerodostojnosti, te interpretirajte sve koeficijente.

Zadatak 4

(d) U model dodajte interakciju između `time` i `year`, napišite koji je sada pretpostavljeni model, te interpretirajte sve koefcijente. Koristeći test omjera vjerodostojnosti testirajte je li interakcija statistički značajna, tj. testirajte je li potreban model s interakcijom (u odnosu na model bez interakcije, tj. samo sa zasebnim kovarijatama).

Završne napomene

- naći dobar model za podatke je puno teži problem od problema predikcije
- svi testovi koji smo koristili (značajnost kovarijata, usporedba modela, devijanca i χ^2 razdioba) bazirani su na pretpostavi da je model koji smo pretpostavili (približno) točan, tj. da su naši podaci simulirani iz modela kojeg smo pretpostavili
- ne postoji testovi koji mogu dokazati da je naš model točan, već je najviše što možemo provjeriti ima li snažnih dokaza da ne vrijede pretpostavke našeg model (to je uostalom osnovna filozofija statističkih testova)
- u principu, donosimo zaključke na temelju našeg modela, tek kad smo se uvjerili da nema snažnih indikacija da naš model nije dobar.
- problem provjere pretpostavki našeg modela (model check) je netrivijalan problem i detaljno bavljenje njime izlazi van okvira ovog kolegija

Završne napomene – odabir kovarijata

- kada je broj kovarijata malen, možemo npr. koristiti testove omjera vjerodostojnosti i usporediti nekoliko različith modela
- ipak, u slučaju većeg broja takvih testova dolazimo da problema višestrukog testiranja
- u tom slučaju, koriste se unakrsna validacija i/ili AIC/BIC kriteriji gdje ulogu sume kvadradata rezidula preuzima log-vjerodostojnost $\ell(\hat{\beta})$, tj.
  (i) $\mathrm{AIC}(\hat{\beta}) = -2\ell(\hat{\beta}) + 2p$;
  (ii) $\mathrm{BIC}(\hat{\beta}) = -2\ell(\hat{\beta}) + \log(n)p$.

Završne napomene – regularizacija

- ponovno, ako je broj kovarijata p velik u odnosu na n, možemo koristiti regularizaciju, te vrijede slična svojstva kao kod linearne regresije
- ridge procjenitelj je

$$\hat{\beta}_\lambda^r := \arg\min_\beta \{-\ell(\beta) + \lambda \|\beta\|_2^2\}$$

- lasso procjenitelj je

$$\hat{\beta}_\lambda^{las} := \arg\min_\beta \{-\ell(\beta) + \lambda \|\beta\|_1\}$$

Završne napomene – GAMovi

– u Generaliziranom Aditivnom Modelu (GAM) pretpostavljamo da je za funkciju veze g,

$$g(\mu_i) = \alpha + \sum_{j=1}^{p} f_j(x_{ij}), \quad i = 1, \ldots, n,$$

pri čemu funkcije f_j procjenjujemo neparametarski.