Statističko učenje 23./24.
Treća domaća zadaća

Rok za predaju: **12. siječnja, 2023.**

Broj bodova: 15

Teorijski dio

Napomena: *Dovoljno je riješiti jedan teorijski zadatak po izboru. U tom slučaju preostali zadatak služi kao vježba za završni ispit.*

Zadatak 1 (Eksponencijalne familije)

Gama razdiobu $\Gamma(\alpha, \beta)$ s parametrima $\alpha, \beta > 0$ definiramo kao neprekidnu razdiobu s gustoćom[1]

$$f_{\alpha,\beta}(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}} 1_{(0,\infty)}(y)\,,\ y \in \mathbb{R}\,.$$

(a) Pokažite da je familija gama razdioba eksponencijalna familija distribucija uz prirodni parametar $\theta = \frac{-1}{\alpha\beta} < 0$ i parametar disperzije $\phi = \frac{1}{\alpha} > 0$.

(b) Koristeći dio (a), izvedite formule za $\mu := \mathbb{E}[Y]$ i $\mathrm{Var}(Y)$ u ovisnosti o α, β, odnosno θ, ϕ, ako je $Y \sim \Gamma(\alpha, \beta)$, te napišite kako izgleda funkcija varijance $V(\mu)$.

(c) Odredite koja je kanonska funkcija veze $g_c(\mu)$ za GLM za gama familiju razdioba, te napišite formulom kako u takvom GLM-u očekivanje $\mu_i = \mathbb{E}[Y \mid X = x^{(i)}]$ ovisi o linearnom prediktoru $\eta_i = (x^{(i)})^\tau \beta$ za $\beta \in \mathbb{R}^p$.[2]

Zadatak 2 (Devijanca GLM-a)

Za danu eksponencijalnu familiju $\{P_{\theta,\phi}\}$, GLM pretpostavlja da su $Y_1, \ldots, Y_n$ nezavisne i takve da

$$Y_i \sim P_{\theta_i,\phi_i}\,,\ i = 1, \ldots, n$$

pri čemu je $\theta_i = \theta(\mu_i)$ za očekivanje $\mu_i = \mathbb{E}[Y_i \mid X^{(i)} = x^{(i)}]$, te

[1] Postoje i drugačije parametrizacije.

[2] Ukoliko nema nikakvih restrikcija na koeficijente β, ako umjesto funkcije veze $g(\mu)$ koristimo $cg(\mu)$ za proizvoljnu konstatnu $c \neq 0$ dobivamo ekvivalentan GLM (razlika je samo u procijenjenim koeficijentima). U slučaju gama razdiobe se tipično $-g_c(\mu)$ navodi i koristi kao kanonska funkcija veze.

1. za neki $\beta \in \mathbb{R}^p$, μ_i i linearni prediktor $\eta_i = (x^{(i)})^\tau \beta$ su povezani preko strogo monotone funkcije veze g t.d. $g(\mu_i) = \eta_i$, tj. $\mu_i = g^{-1}(\eta_i)$.

2. parametar disperzije je oblika $\phi_i = \frac{\phi}{w_i}$ pri čemu su "težine" $w_1, \ldots, w_n$ i parametar $\phi > 0$ poznati.

Neka su $y_1, \ldots, y_n$ realizacije od $Y_1, \ldots, Y_n$ te

$$\ell(\beta) := \sum_{i=1}^{n} \log f_{\theta_i, \phi_i}(y_i)\,,\ \beta \in \mathbb{R}^p$$

log-vjerodostojnost.

(i) Pokažite da vrijedi

$$\ell(\beta) := \sum_{i=1}^{n} \log(h(y_i, \phi/w_i)) + \frac{1}{\phi} \sum_{i=1}^{n} w_i(\theta(\mu_i)y_i - b(\theta(\mu_i)))\,.$$

(ii) Ako su $\widehat{\beta}$ koeficijenti koji maksimiziraju log-vjerodostojnost, a $\hat{\mu}_i = g^{-1}((x^{(i)})^\tau \widehat{\beta})$ procijenjena očekivanja, pokažite da je devijanca oblika

$$D = 2 \sum_{i=1}^{n} w_i \left[y_i(\theta(y_i) - \theta(\widehat{\mu}_i)) - (b(\theta(y_i)) - b(\theta(\widehat{\mu}_i))) \right]\,,$$

te specijalno ne ovisi o ϕ. [3]

(iii) Pokažite da u slučaju Poissonovog GLM-a (dakle $\phi = w_i = 1$) uz kanonsku funkciju veze, vrijedi

$$D = 2 \sum_{i=1}^{n} \left[y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i) \right]\,.$$

[3] To je zapravo bitno u slučajevima kada je ϕ nepoznat.

Praktični dio

Zadatak 1 (Poissonov GLM)

U `ceb.rda` nalaze se podaci o broju rođene djece ("childern ever born") ovisno o određenim grupama žena, u državi Fiji. Dostupne varijable su:

- `nwom`: broj žena u svakoj grupi
- `nceb`: broj rođene djece u grupi (odziv)
- `dur`: vrijeme (u godinama) od vjenčanja; ovo je kategorijalna varijabla, a kategorije su 0-4 (1), 5–9 (2), 10–14 (3), 15–19 (4), 20–24 (5) i više od 25 (6)
- `res`: kateogrijalna varijabla o mjestu stanovanja; kategorije su Suva (1), urbana (2), ruralna (3)
- `educ`: kategorijalna (ili preciznije, ordinalna) varijabla o razini obrazovanja; kategorija su bez obrazovanja (1), niže osnovno (2), više osnovno (3), srednja škola ili više (4)
- `var`: procijenjena varijanca za broj djece unutar svake grupe

Želimo koristeći Poissonovu regresiju ispitati ovisnost broj rođene djece o dostupnim kovarijatama.

(i) Treba li u ovom modelu uključiti `offset` i zašto? Ako da, koju varijablu treba uključiti kao `offset`?

(ii) Prilagodite Poissonov GLM (uz kanonsku funkciju veze) koji uključuje `offset` i kovarijate `dur`, `res` i `educ`. Provjerite prilagodbu modela koristeći njegovu devijancu. Napišite precizno koji je točno model pretpostavljen te interpretirajte sve koeficijente (uključujući i slobodni član ako je to moguće). Koristeći test omjera vjerodostojnosti ispitajte statističku značajnost svake od kovarijata u ovom modelu. Koja je kovarijata najznačajnija?

(iii) U model uključite interakciju između kovarijata `dur` i `educ`. Objasnite kako se u ovom modelu mijenja očekivanje odziva ako uspoređujemo ženu kojoj je od vjenčanja prošlo 5-9 godina u odnosu na onu kojoj je prošlo 0-4, ukoliko imaju istu vrijednost kovarijata `res` i `educ`; ovisi li ta promjena o konkretnoj zajedničkoj vrijednosti kovarijate `res` i/ili `educ`? Koristeći test omjera vjerodostojnosti provjerite je li potrebno u model uključiti ovu interakciju.

Zadatak 2 (Poissonov GAM)

U ovom zadatku primijenit ćete generalizirani aditivan model (GAM) za Poissonovu razdiobu na podatke `chicago` iz paketa `gamair`. Jedna od poruka je da GAM-ovima možemo neparametarski modelirati trendove i sezonalnost u vremenskim nizovima.

Podaci sadrže informacije o dnevnom broju umrlih osoba u gradu Chicagu (`death`) (odziv), a potencijalne kovarijate su: razine ozona (`o3median`), sumporovog dioksida (`so2median`) i određenih štetnih čestica (`pm10median`) u zraku, te prosječna dnevna temperatura (`tmpd`) i dan (`time`); dan 0 je 31. prosinca 1993. U GAM-ovima u nastavku zadatka vezu između odziva i svake od kovarijate modelirajte koristeći *smoothing spline*; ako broj stupnjeva slobode nije eksplicitno zadan, koristite zadanu (*default*) vrijednost (to je 4). Prije svega:

- Temperatura je zadana u stupnjevima Fahrenheita – pretvorite ih u stupnjeve Celzijeve.
- Izbacite stupac koji odgovara kovarijati `pm25median`.
- Izbacite sve retke koji sadrže barem jednu `NA` vrijednost.

(a) Nacrtajte podatke o broju dnevnih smrti u ovisnosti o vremenu. Trebali biste primijetiti 4 uzastopna dana s neobično visokim brojem smrti. Kojim datumima odgovaraju te vrijednosti (koristite funkciju `as.Date`)?

(b) Prilagodite GAM za Poissonovu razdiobu uz kanonsku funkciju veze, za odziv `death` u ovisnosti o svim dostupnim kovarijatama; za kovarijatu `time`, ovdje i u nastavku, koristite 150 kao broj stupnjeva slobode.

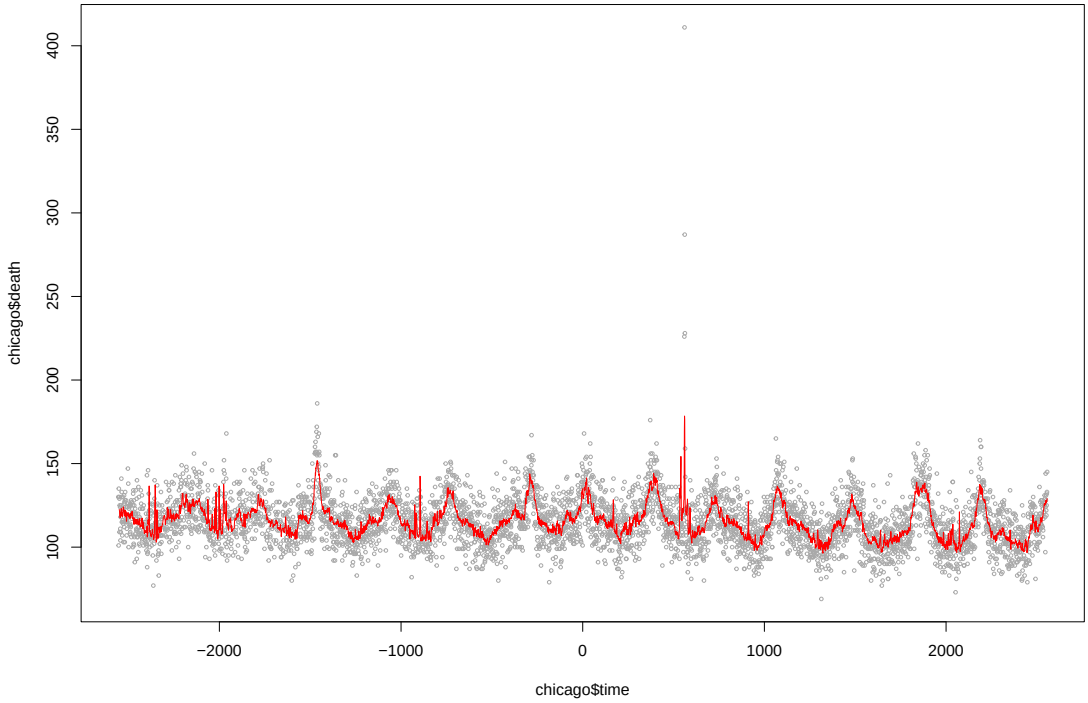
(b1) Precizno napišite koji ste model zapravo pretpostavili.

(b2) Nacrtajte procijenjene funkcije za svaku kovarijatu (tzv. parcijalne funkcije). Interpretirajte grafove (npr. kako svaka od kovarijata utječe na odziv, jesu li veze približno linearne, koja se čini da ima veći/manji utjecaj na odziv, što znači kada je funkcija pozitivna/negativna itd.).

(b3) Nacrtajte vrijednosti dnevnih smrti dobivene iz modela u ovisnosti o kovarijati `time` zajedno sa stvarnim podacima. Koliko dobro ovaj model predviđa 4 "outliera"?

(b4) Koliko je ukupan broj stupnjeva slobode?

(c) Sada ćete umjesto kovarijate o dnevnoj temperaturi te kovarijata koje opisuju količinu plinova/čestica u zraku u jednom danu, koristiti prosjek vrijednosti trenutnog i 3 prethodna dana. Zašto to ima smisla? Ponovite korake (b2)-(b4) s transformiranim kovarijatama i komentirajte rezultate. Što se promijenilo? Jesu li "outlieri" još uvijek tu?



Slika 1: Predikcije dnevnih smrti (crveno) iz modela u Zadatku 1(d) u ovisnosti o vremenu zajedno sa stvarnim podacima.

(d) *(*dodatni zadatak*) Sada u model dodajemo interakcije, tj. utjecaj (transformiranih) kovarijata `tmpd` i `o3median` modelirat ćemo zajedno umjesto zasebno. Ponovite korake te odgovorite na pitanja kao u (b2)-(b4) dijelu zadatka. Čini li se efekt ove interakcije značajan? Je li dodavanje ove interakcije značajno utjecalo na predviđanje naših "outliera"?[4] (*Napomena:* Ukoliko ćete rješavati ovaj dio zadatka, trebat ćete koristiti paket `mgcv` te malo proučiti kako se on koristi. Ovdje nije potrebno fiksno zadati broj stupnjeva slobode već algoritam to radi sam; možete za kovarijatu `time` zadati gornju ogradu od npr. 200).

[4]Općenito, znamo da koliko se dobro koja metoda prilagođava podacima nije dobar indikator koliko je ta metoda zaista dobra. Ipak, za modele koji koriste podjednak broj stupnjeva slobode, smisleno je izabarati onu koje sa najbolje prilagođava podacima.