UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Leon Horvat

Posplošeni linearni modeli

Delo diplomskega seminarja

Mentor: prof. dr. Jaka Smrekar

Ljubljana, 2018

Kazalo

Posplošeni linearni modeli

Povzetek

V povzetku na kratko opišite vsebinske rezultate dela. Sem ne sodi razlaga organizacije dela – v katerem poglavju/razdelku je kaj, pač pa le opis vsebine.

Angleški naslov dela

Abstract

Prevod zgornjega povzetka v angleščino.

Math. Subj. Class. (2010): navedite vsaj eno klasifikacijsko oznako – dostopne so na `www.ams.org/mathscinet/msc/msc2010.html`

Ključne besede: navedite nekaj ključnih pojmov, ki nastopajo v delu

Keywords: angleški prevod ključnih besed

1. Uvod

Kako na pričakovano življenjsko dobo vplivajo hrana, kraj rojstva, izobrazba, športne aktivnosti posameznika in kajenje? Kakšna je verjetnost otrokovega uspeha v šoli glede na šolski uspeh staršev, socialno in finančno stanje družine? Ali lahko na podlagi starosti voznika, starosti in modela vozila predvidimo višino povzročene materialne škode, ki jo bo zavarovalnica morala poravnati? Modeliranje povezav med opazovanimi spremenljivkami je ključno za statistično raziskovanje in analizo. Konstruiranje teh modelov za povezovanje pojasnjevalnih spremenljivk in proučevanih spremenljivk poda globlji vpogled v razmerja med njimi, sploh v večjih količinah podatkov, kjer so ta razmerja manj vidna na prvi pogled. Pri modeliranju teh povezav je zaželena enostavnost modela z manj pojasnjevalnimi spremenljivkami, ki daje intuitiven vpogled in lažjo interpretacijo. Pri kompleksnem modelu z več pojasnjevalnimi spremenljivkami je to lahko težje, hkrati pa mogoče niti ne poda boljših rezultatov.

V tem delu bom predstavil posplošene linearne modele. Najprej bom na kratko opisal linearno regresijo in jo postopoma nadgradil. Pri posameznih poglavjih bom skušal teorijo ponazoriti s primeri, prav tako pa bom celotno teorijo na koncu apliciral na praktičen primer.

2. Linearna regresija

V enostavnem modelu linearne regresije je proučevana slučajna spremenljivka Y podana kot

$$Y = x^T\beta + \epsilon,$$

kjer je β vektor neznanih parametrov, x je vektor pojasnjevalnih slučajnih spremenljivk, ϵ pa je slučajna spremenljivka, ki meri slučajna odstopanja; to so lahko meritvene, zaokrožitvene napake ali pa druga odstopanja. Dva pomembna vidika tega modela sta linearna odvisnost proučevane slučajne spremenljivke Y od neznanih parametrov β in struktura napake ϵ. Privzamemo, da je $E(\epsilon) = 0$; iz tega sledi $\mu = E(Y) = x^T\beta$. Torej je v tem modelu pričakovana vrednost preučevane slučajne spremenljivke linearna funkcija pojasnjevalnih slučajih spremenljivk. Neznane parametre β ocenimo po metodi najmanjših kvadratov, torej da minimiziramo vsoto kvadratov residualov ali pa po metodi največjega verjetja. Če gledamo Y_i kot i-to proučevano slučajno spremenljivko in x_i^T vektor znanih spremenljivk, potem model linearne regresije oceni Y_i kot

$$\hat{Y_i} = \hat{\beta}_1 x_i 1 + \hat{\beta}_2 x_i 2 + \cdots + \hat{\beta}_p x_i p \equiv \hat{\eta}_i.$$

Taka linearna povezava med Y_i in linearno cenilko $\hat{\eta}_i$ je hitro berljiva in matematično enostavna, vendar ni vedno najboljša, lahko je celo nepravilna ali nemogoča. Če dovolimo, da parametri β zavzamejo poljubno vrednost, lahko linearna cenilka η_i pade izven obsega vrednosti, ki jih lahko zavzame preučevana slučajna spremenljivka Y. To se lahko na primer zgodi, ko je Y spremenljivka, ki opiše verjetnost ali razmerje. Y lahko zavzame vrednosti le na zaprtem intervalu med 0 in 1, η_i pa zavzame vrednost izven tega intervala.

Ena metoda, ki reši ta problem, je transformacija slučajne spremenljivke Y z neko funkcijo g, torej naredimo linearno regresijo na slučajni spremenljivki $g(Y)$. To pomeni, da za slučajno spremenljivko Y_i velja $E(g(Y_i))4 = \sum_{j=1}^{p} \beta_j x_{ij}$. Na primer, velikokrat se kot povezovalno funkcijo g uporabi logaritem. V aktuarstvu

je logaritemska transformacija uporabljena na višini odškodninskih zahtevkov, ki so pozitivni.

Druga metoda, ki reši zgornji problem, je uporaba linearne regresije na pričakovani vrednosti μ slučajne spremenljivke Y, transformirane s funkcijo g. Torej

$$g(E(Y_i)) = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Ta pristop je temelj posplošenega linearnega modeliranja. Predstavil bom tri primere posplošenih linearnih modelov: logistično regresijo, Poissonovo (linearno) regresijo in 'probit' model.

Recimo, da nas zanima delež moških voznikov, ki bodo v prihodnjem letu podali zavarovalnici odškodninski zahtevek. Označimo iskan delež za posamezno starost x s π_x, Y_x pa naj bo delež voznikov starosti x v danem vzorcu, ki so podali odškodninski zahtevek. Ker je Y_x delež, torej je element intervala $(0, 1)$, je tudi pričakovana vrednost $E(Y_x) = \pi_x$ element tega intervala. Če za model vzamemo $\pi_x = \beta_0 + \beta_1 x$ nam lahko vrednosti π_x padejo izven intervala $(0, 1)$, takih vrednosti pa ne moremo interpretirati kot deleže. Zato model izboljšamo z uporabo logistične funkcije. Logistična funkcija je definirana kot $logit(x) = \log(\frac{x}{1-x})$. Opazimo, da ta funkcija slika iz intervala $(0, 1)$ na celotno realno os. V modelu logistične regresije bomo torej modelirali $logit(\pi)$ kot linearno kombinacijo pojasnjevalnih spremenljivk, v našem primeru starosti voznika. Model se torej glasi:

$$g(\pi_x) = logit(\pi_x) = \log \frac{\pi_x}{1 - \pi_x} = \beta_0 + \beta_1 x = \eta_x$$

$$\pi_x = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} = \frac{exp(\eta_x)}{1 + exp(\eta_x)}$$

Sedaj nimamo več problema, da π_x ne bi bil vsebovan v intervalu $(0, 1)$. Logistična regresija se velikokrat uporablja, če nas zanimajo deleži ali verjetnosti. Lahko bi nas zanimala verjetnost razvoja raka glede na spol, starost in druge zdravstvene značilnosti. Ali pa bi nas zanimala verjetnost, da študent opravi vse izpite pred poletjem.

Iščemo lahko tudi modele za slučajne spremenljivke, ki so porazdeljene Poissonovo. To je lahko na primer število nesreč na odseku avtoceste, ki jih lahko pojasnimo z vremenom, letnim časom, dnevom v tednu, uro. Ker je pričakovana vrednost slučajne spremenljivke Y porazdeljene kot $Poiss(\lambda)$ enaka λ, pri tem pa je ta pozitivna, je model linearne regresije neprimeren, ocena za λ bi bila ob neki izbiri podatkov negativna. Lahko pa vzamemo log-linearni model oz. Poissonovo regresijo. To pomeni, da modeliramo

$g(\lambda) = \log \lambda = x^T \beta = \eta$ oziroma $\lambda = e^\eta$ V tem modelu $E(Y) = \lambda$ in linearno kombinacijo pojasnjevalnih spremenljivk povezuje logaritemska funkcija. Logaritem slika iz intervala $(0, \inf)$ na realno os; to pomeni da bo napoved za λ Poissonove slučajne spremenljivke vedno pozitivna in ne bomo imeli problema iz enostavne linearne regresije, ko bi bila lahko λ tudi negativna. Probit model je eden prvih primerov posplošenih linearnih modelov, ki ni linearen v klasičnem smislu. Rešuje podobno problematiko kot logistični model, torej se uporablja, ko želimo oceniti neki delež ali verjetnost. Recimo, da nas zanima verjetnost preživetja osebe Y_x ob zaužitju nekega zdravila količine x. S π_x označimo pričakovano vrednost te slučajne spremenljivke. V probit modelu vzamemo za povezovalno funkcijo inverz

kumulativne porazdelitvene funkcije standardne normalne porazdelitve Φ^{-1}. Ta slika iz intervala $(0, 1)$ na realno os. π_x bomo torej ocenili na naslednji način:

$$g(\pi_x) = \Phi^{-1}(\pi_x) = \beta_0 + \beta_1 x = \eta_x$$

oziroma

$$\pi_x = \Phi(\beta_0 + \beta_1 x) = \Phi(\eta_x).$$

S posplošenimi linearnimi modeli lahko modeliramo vse slučajne spremenljivke, ki so v družini eksponentnih porazdelitev. To so na primer normalna porazdelitev, Poissonova porazdelitev, binomska porazdelitev in porazdelitev gama.

3. Bolj podroben pregled linerane regresije

Pri klasični linearni regresiji privzamemo, da je $Y = (Y_1, \ldots, Y_n)$ slučajen vektor, komponente pa so neodvisne, normalno porazdeljene spremenljivke, $Y_i \sim N(\mu_i, \sigma^2)$, in velja $E(Y_i) = \mu_i = x_i^T \beta$. Model lahko zapišemo v matrični obliki $Y = X\beta + \epsilon$, pri čemer je Y slučajen proučevan vektor, X je $n \times p$ matrika pojasnjevalnih spremenljivk, torej imamo p pojasnjevalnih spremenljivk, ϵ pa je vektor slučajnih napak.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vektor slučajnih napak je porazdeljen $\epsilon \sim N(0, \sigma^2)$. Proučevano slučajno spremenljivko Y_i torej razdelimo na sistematično komponento $x_i^T \beta$ in slučajno komponento ϵ_i. Sistematična komponenta je linearna cenilka za $E(Y_i)$. Pojasnjevalne slučajne spremenljivke so v tem modelu lahko kategorične, numerične ali pa teh kombinacija. Temu modelu pravimo splošen linearen model. Če je v modelu konstanta, potem je prvi stolpec v matriki pojasnjevalnih spremenljivk X sestavljen iz enic.

Izraz ANOVA (analysis of variance), analiza variance, je pogosto uporabljen za model, ki ima kategorične pojasnjevalne spremenljivke. V takem primeru si želimo primerjati različne skupine, ki so definirane s kategorijami teh spremenljivk. ANCOVA (analysis of covariance), analiza kovariance, pa opisuje model, v katerem nastopajo tako kategorične kot tudi numerične pojasnjevalne slučajne spremenljivke. V tem primeru primerjamo različne skupine, definirane s kategorijami, na napovedvsake skupine pa vpliva tudi numerična spremenljivka.

Neznane parametre β lahko izračunamo z metodo najmanjših kvadratov. Če je $y = (y_1 \ldots y_n)$ vzorec proučevanih spremenljivk in so $x_i = (x_{i1} \ldots x_{ip}), i = 1, \ldots, n$ vzorci pripadajočih pojasnjevalnih slučajnih spremenljivk, torej minimiziramo

$$\sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2.$$

Vektor β lahko izračunamo tudi prek metode največjega verjetja. V tem primeru iščemo maksimum funkcije verjetja

$$L(y, x, \sigma^2, \beta) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} exp\left(-\frac{(y_i - x_i^T \beta)^2}{2\sigma^2}\right)$$

oziroma logaritma funkcije verjetja

$$l(y, x, \sigma^2, \beta) = -n \log \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T)^2$$

Opazimo, da je cenilka pridobljena z metodo največjega verjetja enaka cenilki po metodi najmanjših kvadratov. Še dodatno nam Gauss-Markov teorem pove, da je cenilka za parametre β pridobljena po katerikoli od obeh metod (sta enaki) najboljša linearna nepristranska cenilka za β.

(poravnava enačb!)

4. Eksponentna družina

Najprej si bomo pogledali, kako je lahko opazovana slučajna spremenljivka Y porazdeljena, da jo lahko modeliramo s posplošenimi linearnimi modeli. Pri grajenju modela je izbira porazdelitvene družine prva naloga. V tem delu si bomo pogledali enoparameterske verjetnostne porazdelitve, ki pripadajo eksponentni družini. Te so tudi najpogosteje uporabljene pri modeliranju posplošenih linearnih modelov.

Definicija 4.1. Slučajna spremenljivka Y pripada enoparametrski eksponentni družini, če ima gostoto f_Y, ki je oblike

$$f_Y(y; \theta, \phi) = exp\left(\frac{A(y \cdot \theta - \gamma(\theta))}{\phi} + \tau(y, \frac{\phi}{A}) \right),$$

pri čemer je θ naravni (ali kanonični) parameter, ϕ je disperzijski parameter, γ in τ sta funkciji ter A je utež.

V nadaljevanju bom izpeljal nekaj lastnosti porazdelitev iz eksponentne družine. Logaritemska funkcija verjetja je enaka naravnemu logaritmu porazdelitve, torej

$$l(\theta) = \log f_Y(y; \theta, \phi) = A\frac{y \cdot \theta - \gamma(\theta)}{\phi} + \tau(y, \frac{\phi}{A})$$

?(slovenski prevod - score function, kaj nam pove U) funkcija U je parcialni odvod logaritemske funkcije verjetja po θ in jo označimo z $U(\theta) = \frac{\partial}{\partial\theta} l = \frac{A}{\phi}(y - \gamma'(\theta))$

Trditev 4.2. *Naj bo Y slučajna spremenljivka iz eksponentne družine. Potem velja*

$$E(Y) = \mu = \gamma'(\theta), \;\; Var(Y) = \frac{\phi}{A}\gamma''(\theta)$$

Dokaz. Recimo, da je Y slučajna spremenljivka z gostoto, ki pripada eksponentni družini. Za dokaz prve enakosti si bomo pomagali z matematičnim upanjem funkcije U slučajne spremenljivke Y.

$$E(U(\theta)) = E\left(\frac{\partial}{\partial\theta} l\right) = \int f_Y \frac{\partial}{\partial\theta} l dy = \int f_Y \frac{\partial}{\partial\theta} \log f_Y dy =$$
$$\int f_Y \frac{1}{f_Y} \frac{\partial f_Y}{\partial\theta} dy = \int \frac{\partial f_Y}{\partial\theta} dy = \frac{\partial}{\partial\theta} \int f_Y dy = \frac{\partial}{\partial\theta} 1 = 0$$

Iz $E(U(\theta)) = E\left(\frac{A}{\phi}(Y - \gamma'(\theta))\right) = \frac{A}{\phi}(E(Y) - \gamma'(\theta))$ sledi, da je $E(Y) = \gamma'(\theta)$

Za dokaz druge enakosti pa si oglejmo naslednje:

$$E\left(\frac{\partial^2}{\partial\theta^2}l+\left(\frac{\partial}{\partial\theta}l\right)^2\right)=\int f_Y\left(\frac{\partial}{\partial\theta}\left(\frac{1}{f_Y}\frac{\partial f_Y}{\partial\theta}\right)+\left(\frac{1}{f_Y}\frac{\partial f_Y}{\partial\theta}\right)^2\right)dy=$$

$$=\int f_Y\left(-\frac{1}{f_Y^2}\left(\frac{\partial f_Y}{\partial\theta}\right)^2+\frac{1}{f_Y}\frac{\partial^2 f_Y}{\partial\theta^2}+\frac{1}{f_Y^2}\left(\frac{\partial f_Y}{\partial\theta}\right)^2\right)dy=$$

$$=\int\frac{\partial^2 f_Y}{\partial\theta^2}dy=\frac{\partial^2}{\partial\theta^2}\int f_Y dy=\frac{\partial^2}{\partial\theta^2}1=0$$

Ker velja naslednje

$$\frac{\partial^2}{\partial\theta^2}l=\frac{\partial}{\partial\theta}\left(\frac{A}{\phi}\left(y-\gamma'(\theta)\right)\right)=-\frac{A}{\phi}\gamma''(\theta),\quad E\left(\frac{\partial^2}{\partial\theta^2}l\right)=-\frac{A}{\phi}\gamma''(\theta)$$

$$E\left(\left(\frac{\partial}{\partial\theta}l\right)^2\right)=E\left(\frac{A^2}{\phi^2}\left(Y-\gamma'(\theta)\right)^2\right)=\frac{A^2}{\phi^2}E\left((Y-\mu)^2\right)=\frac{A^2}{\phi^2}Var(Y),$$

lahko varianco izračunamo:

$$-E\left(\frac{\partial^2}{\partial\theta^2}l\right)=E\left(\left(\frac{\partial}{\partial\theta}l\right)^2\right)$$

$$\frac{A}{\phi}\gamma''(\theta)=\frac{A^2}{\phi^2}Var(Y)$$

$$Var(Y)=\frac{\phi}{A}\gamma''(\theta)$$

□

Zgornja trditev nam poleg formule za izračun pričakovane vrednosti in variance neke porazdelitve iz eksponentne družine pokaže tudi, da sta pričakovana vrednost in varianca neodvisni. Ta lastnost je pomembna za posplošene linearne modele.

Lahko označimo funkcijo variance $V(\mu)=\gamma''(\theta)$.

Iz prejšnjega dokaza lahko opazimo tudi, kaj velja za Fisherjevo informacijo $I(\theta)$ v slučajni spremenljivki Y za θ, ki nam pove spodnjo mejo za varianco (nepristranske cenilke?) θ. Fisherjeva informacija za θ je definirana z $I(\theta)=E((\frac{\partial}{\partial\theta}l)^2)$, po zgoraj izračunanem pa velja tudi $I(\theta)=-E(\frac{\partial^2}{\partial\theta^2}l)$.

V nadaljevanju bom podal nekaj primerov porazdelitev iz eksponentne družine.

4.1. **Normalna porazdelitev.** Normalno porazdeljena slučajna spremenljivka $Y\sim N(\mu,\sigma^2)$ ima gostoto oblike

$$f_Y(y)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

za vsak $y\in\mathbb{R}$.

Log-verjetnostno funkcijo lahko zapišemo tako:

$$l(\mu,\sigma^2;y)=\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)-\frac{(y-\mu)^2}{2\sigma^2}=-\frac{1}{2}\log(2\pi\sigma^2)-\frac{y^2-2y\mu+\mu^2}{2\sigma^2}=$$

$$=\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right)$$

Iz zgornjega zapisa in iz definicije gostote porazdelitve iz eksponentne družine sledi:

$$\theta = \mu,\ \phi = \sigma^2,\ \gamma(\theta) = \frac{\theta^2}{2},\ A = 1 \text{ in } \tau(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)$$

Vidimo, da normalna porazdelitev res spada v eksponentno družino, ker smo uspeli preoblikovati njeno gostoto oz. log-verjetnostno funkcijo v primerno obliko. Opazimo še:

$$E(Y) = \gamma'(\theta) = \theta \ \text{ in } \ Var(Y) = \frac{\phi}{A}\gamma''(\theta) = \sigma^2 \cdot 1 = \sigma^2$$

4.2. **Poissonova porazdelitev.** Poissonovo porazdeljena slučajna spremenljivka $Y \sim Poiss(\mu)$ ima porazdelitev

$$f_Y(y) = P(Y = y) = \mu^y \frac{e^{-\mu}}{y!} = exp(y \log \mu - \mu - \log y!)$$

za $y \in \mathbb{N}_0$. Takoj lahko opazimo, da se oblika gostote ujema z gostoto eksponentne družine, pri čemer velja:

$$\theta = \log \mu,\ A = \phi = 1,\ \gamma(\theta) = e^\theta \text{ in } \tau(y, \phi) = -\log y!$$

Matematično upanje in varianca pa sta v tem primeru:

$$E(Y) = \gamma'(\theta) = e^\theta = \mu \ \text{ in } \ Var(Y) = \frac{\phi}{A}\gamma''(\theta) = 1 \cdot e^\theta = \mu$$

V Poissonovi porazdelitvi je naravni parameter $\log \mu$ in ne samo μ, zato je velikokrat v posplošenem linearnem modelu s Poissonovo porazdeljeno slučajno spremenljivko povezovalna funkcija logaritem.

4.3. **Binomska porazdelitev.** Naj bo S binomsko porazdeljena slučajna spremenljivka $S \sim B(n, p)$. Vemo, da je opazovano razmerje uspehov $Y = \frac{S}{n}$ nepristranska cenilka za verjetnost uspeha $p = \mu = E(Y)$. Verjetnostno porazdelitev slučajne spremenljivke Y lahko izrazimo kot:

$$P(Y = y) = P(S = ny) = \binom{n}{ny} p^{ny}(1-p)^{n-ny} =$$

$$= exp\left(ny \log \frac{p}{1-p} + n \log(1-p) + \log \binom{n}{ny}\right) =$$

$$= exp\left(n(y \log \frac{p}{1-p} + \log(1-p)) + \log \binom{n}{ny}\right)$$

Iz tega razberemo:

$$\theta = logit(p),\ \gamma(\theta) = \log(1 + e^\theta),\ \phi = 1,\ A = n \text{ in } \tau(y, \frac{\phi}{A}) = \log \binom{n}{ny}$$

To pomeni, da slučajna spremenljivka Y, ki predstavlja delež uspehov v n Bernoullijevih poskusih z verjetnostjo p, pripada eksponentni družini porazdelitev. Naravni parameter za binomsko porazdelitev je torej $logit(p)$ in ne le p. Pričakovana vrednost in varianca sta:

$$E(Y) = \gamma'(\theta) = \frac{e^\theta}{1 + e^\theta} = \frac{\frac{p}{1-p}}{1 + \frac{p}{1-p}} = p = \mu$$

$$Var(Y) = \frac{\phi}{A}\gamma''(\theta) = \frac{1}{n}\frac{e^\theta}{(1+e^\theta)^2} = \frac{1}{n}p(1-p) = \frac{1}{n}\mu(1-\mu)$$

4.4. **Gama porazdelitev.** Slučajna spremenljivka ima porazdelitev gama, $Y \sim \Gamma(\alpha, \lambda)$, s parametroma α in λ, če ima gostoto oblike

$$f_Y(y) = \frac{\lambda^\alpha y^{\alpha-1} e^{-\lambda y}}{\Gamma(\alpha)} \quad \text{za } y > 0.$$

Da lahko pokažemo, da gama porazdelitev tudi spada v družino eksponentnih porazdelitev, gostoto reparametriziramo. S slučajno spremenljivko Y je upanje enako $E(Y) = \frac{\alpha}{\lambda} = \mu$, nova parametra pa sta α in $\mu = \frac{\alpha}{\lambda}$. Če gostoto ponovno izrazimo z novima parametroma dobimo

$$f_Y(y) = \frac{(\frac{\alpha}{\mu})^2 y^{\alpha-1} e^{\frac{\alpha}{\mu} y}}{\Gamma(\alpha)} = exp\left(-\frac{\alpha}{\mu} y + \alpha \log \alpha - \alpha \log \mu + (\alpha - 1) \log y - \log \Gamma(\alpha)\right) =$$

$$= exp\left(\frac{y(-\frac{1}{\mu}) - \log \mu}{\frac{1}{\alpha}} y + \alpha \log \alpha + (\alpha - 1) \log y - \log \Gamma(\alpha)\right)$$

Vidimo, da je to oblika gostote porazdelitve iz družine eksponentnih porazdelitev, pri čemer je

$$\theta = -\frac{1}{\mu},\ \gamma(\theta) = \log(-\frac{1}{\theta}),\ \phi = \frac{1}{\alpha},\ A = 1$$

in

$$\tau(y, \frac{\phi}{A}) = \alpha \log \alpha + (\alpha - 1) \log y - \log \Gamma(\alpha)$$

Napišimo še

$$E(Y) = \gamma'(\theta) = -\frac{1}{\theta} = \mu \ \text{ in } \ Var(Y) = \frac{\phi}{A} \gamma''(\theta) = \frac{1}{\alpha} \frac{1}{\theta^2} = \frac{\mu^2}{\alpha}$$

Pri uporabi posplošenih linearnih modelov bomo predpostavljali, da podatki pripadajo eni izmed porazdelitev ekponentne družine. Z metodo največjega verjetja bomo dobili ocene za relevantne parametre neke porazdelitve. Za n opazovanj $y = (y_1, y_2, \ldots, y_n)$ iz eksponentne družine, pri čemer je ϕ znan, je log-verjetnostna funkcija oblike

$$l(\theta, \phi; y) = \sum_{i=1}^{n} \left(A_i(y_i \theta_i - \gamma(\theta_i))/\phi + \tau(y_i, \phi/A_i)\right),$$

s parcialnim odvajanjem po posameznem parametru θ in enačenju tega z 0 dobimo cenilko največjega verjetja za θ.

5. Povezovalne funkcije

Kot sem omenil že na začetku, sta v posplošenem linearnem modelu pričakovana vrednost preučevane slučajne spremenljivke $E(Y) = \mu$ in pojasnjevalne slučajne spremenljivke x povezane prek povezovalne funkcije g, torej

$$g(E(Y)) = x^T \beta = \eta.$$

Naslednji korak v gradnji modela je izbira povezovalne funkcije, ki ustreza opazovani spremenljivki Y. Čeprav imamo na izbiro več možnosti, med katerimi izberemo povezovalno funkcijo, obstajajo kanonične oz. naravne povezovalne funkcije za porazdelitve iz eksponentne družine. V prejšnjem poglavju smo definirali kanoničen oz. naravni parameter θ v družini eksponentnih porazdelitev. Izkaže se, da za ta

parameter velja $g_N(\mu) = \theta$, pri čemer je g_N naravna povezovalna funkcija, μ pa pričakovana vrednost Y. To funkcijo lahko dobimo še na drugačen način. Za eksponentne družine velja $E(Y) = \mu = \gamma'(\theta)$. Opazimo, da g_N zadošča $g_N = (\gamma')^{-1}$. V prejšnjem poglavju smo za primere porazdelitev iz eksponentne družine izračunali θ v odvisnosti od μ ter funkcijo variance, iz tega dobimo naslednjo tabelo:

Tabela 1. Tabela naravnih povezovalnih funkcij

Družina	$g_N(\mu)$	$V(\mu)$	ϕ
Normalna	μ	1	σ^2
Poissonova	$\log \mu$	μ	1
Binomska	$\text{logit}\mu$	$\mu(1-\mu)$	1
Gamma	$-1/\mu$	μ^2	$1/\alpha$

Zgoraj navedenih naravnih povezovalnih funkcij ni nujno uporabljati pri določeni porazdelitvi iz eksponentne družine. V posameznem primeru se odločimo, ali bomo uporabljali naravno povezovalno funkcijo ali pa katero drugo. Za binomsko porazdelitev lahko na primer izberemo *probit* funkcijo ($g(\mu) = \Phi^{-1}(\mu)$), pri porazdelitvi gamma pa običajno vzamemo funkcijo $1/\mu$ namesto $-1/\mu$, čeprav seveda obe funkciji predstavljata isti model. V splošnem je lahko g povezovalna funkcija, če je injektivna in veljajo dodatne tehnične omejitve (dovoljkrat zvezno odvedljiva).

6. Pojasnjevalne spremenljivke

V tem razdelku bomo pregledali različne pojasnjevalne spremenljivke: številske pojasnjevalne spremenljivke, kategorične spremenljivke in interakcije. Ko ugotovimo, kakšna je porazdelitev podatkov in izberemo povezovalno funkcijo, moramo izbrati pojasnjevalne spremenljivke, ki imajo vpliv na opazovano slučajno spremenljivko. Za izbran nabor pojasnjevalnih spremenljivk lahko izračunamo $\eta = x^T\beta$, pri čemer parametre β_i izračunamo z metodo največjega verjetja. Najprej pa podrobneje opišimo različne pojasnjevalne spremenljivke.

6.1. **Številske in kategorične slučajne spremenljivke.** Številske pojasnjevalne slučajne spremenljivke so spremenljivke, ki lahko zavzamejo poljubno številsko vrednost, to so na primer starost, zavarovalna vsota, dohodek. Pojasnjevalne slučajne spremenljivke pa so lahko tudi kategorične. Te zavzamejo nek končen nabor vrednosti, ki jim rečemo kategorije. Take spremenljivke so recimo spol (moški/ženska), raven izobrazbe (npr. osnovna, srednja, visokošolska, univerzitetna), status kajenja (da/ne). V modelu z eno kategorično spremenljivko moramo oceniti parameter za vsako kategorijo, ki jo spremenljivka lahko zavzame. Če dodamo v model kategorično spremenljivko z recimo b kategorijami, potrebujemo oceniti dodatnih $b-1$ parametrov. Kategorična spremenljivka z b kategorijami spremenimo v $b-1$ dihotomnih spremenljivk. Dihotomna spremenljivka je spremenljivka, ki lahko pokaže le vrednost 0 in 1. Za vsako kategorijo prvotne slučajne spremenljivke naredimo novo dihotomno slučajno spremenljivko, razen za eno. Vrednost tiste je določena z vrednostjo ostalih. Če vse dihotomne spremenljivke, ki nadomestijo kategorično spremenljivko pokažejo 0, potem je spremenljivka dosegla vrednost v kategoriji, ki nima lastne dihotomne spremenljivke.

Za primer vzemimo okus sladoleda. To je kategorična slučajna spremenljivka (S), ki zavzame 3 vrednosti: vanilija (V), čokolada (C), jagoda (J). Dihotomni spremenljivki, ki ju dobimo, sta (S_V), ki pokaže 1, če je okus sladoleda vanilija, in 0 sicer,

in (S_C), ki pokaže 1, če je okus sladoleda čokolada, in 0 sicer.

$$S = \begin{bmatrix} V \\ C \\ V \\ J \end{bmatrix} \longrightarrow \quad S\_V = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad S\_C = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Spremenljivke S_J ne potrebujemo, ker je okus sladoleda jagoda enolično določen s spremenljivkama (S_V) in (S_C). Če obe pokažeta 0, potem je okus sladoleda jagoda, sicer pa ne.

6.2. **Interakcije.** Interakcija med dvema slučajnima spremenljivkama obstaja, ko vrednost ene vpliva na efekt na linearno cenilko η druge spremenljivke. Interakcija lahko obstaja med dvema kategoričnima spremenljivkama, recimo, da sta to spremenljivki W z w kategorijami in U z u kategorijami. Torej za vsako kategorijo, ki jo zavzame W, vrednost U vpliva drugače na linearno cenilko η. Model, ki vključuje samo kategorični spremenljivki W in U, brez njune interakcije ima $w + u - 1$ parametrov, če pa vključimo še interakcijo, moramo oceniti $w \cdot u$ parametrov. Interakcija nam torej doda $w \cdot u - (w + u - 1) = (w - 1)(u - 1)$ parametrov.

Lahko imamo tudi interakcijo med kategorično in številsko slučajno spremenljivko. Vpliv številske slučajne spremenljivke X na linearno cenilko η se spreminja v odvisnosti od kategorije, ki jo zavzame kategorična spremenljivka W z w kategorijami. Če v tak model vključimo interakcijo, moramo za vsako kategorijo w oceniti parameter za X, dodamo torej $w - 1$ parametrov.

Interakcija lahko obstaja tudi med dvema številskima slučanjima spremenljivkama, recimo X_1, X_2. Tako interakcijo lahko vključimo v model na način, da je produkt $X_1 \cdot X_2$ nova dodatna številska pojasnjevalna spremenljivka.

6.3. **Nelinearen vpliv pojasnjevalnih spremenljivk.** Številska slučajna spremenljivka lahko na linearno cenilko η vpliva tudi nelinearno. Tak vpliv vključimo v model tako, da pojasnjevalno slučajno spremenljivko preoblikujemo s funkcijo, za katero mislimo, da bo bolje pojasnila zvezo med spremenljivko in linearno cenilko η. Taka funkcija je na primer kvadriranje, logaritemska funkcija, koren ... Tako interakcije, kot tudi nelinearne funkcije pojasnjevalnih spremenljivk so pomembne za manjšanje pristraskosti linearnega posplošenega modela.

7. Minimalna zadostna statistika

Minimalna zadostna statistika je funkcija vzorca, v kateri so zbrane vse informacije, ki jih lahko dobimo iz vzorca samega. Torej, če poznamo minimalno zadostno statistiko nam vzorčne vrednosti ne podajo ničesar novega. Pri ocenjevanju parametrov uporabljamo samo funkcije teh statistik. V nadaljevanju bomo grobo spoznali minimalno zadostno statistiko vzorca porazdelitve iz eksponentne družine.

Naj bo $y_1, \ldots, y_n$ neodvisen vzorec opazovanj, pri čemer ima y_i porazdelitev iz dane eksponentne družine s parametroma (θ_i, ϕ). Potem je log-verjetnostna funkcija za ta vzorec enaka

$$l(\theta, y) = \sum_{i=1}^{n} \left(A_i \left(\frac{y_i \theta_i - \gamma(\theta_i)}{\phi} \right) + \tau \left(y_i, \frac{\phi}{A_i} \right) \right).$$

Če je $g(\mu_i) = \theta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \eta$ (naravna povezovalna funkcija), potem je log-verjetnostna funkcija za parametre β_i

$$l(\beta, y) = \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} A_i \frac{y_i x_{ij}}{\phi} - \sum_{i=1}^{n} \left(A_i \frac{\gamma(\theta_i)}{\phi} + \tau\left(y_i, \frac{\phi}{A_i}\right)\right).$$

Če privzamemo, da so ϕ_i znani, se po Fisher-Neymanovem faktorizacijskem izreku izkaže, da je množica minimalnih zadostnih statistik za parametre β_i

$$\left\{ \sum_{i=1}^{n} A_i \frac{y_i x_{ij}}{\phi}; j = 1, \ldots, p \right\}.$$

Ta množica statistik nam poda vse zadostne informacije za oceno β, poznavanje vrednosti y_i nam ne razkrije nobenih dodatnih informacij. Pri računanju minimalnih zadostnih statistik za porazdelitev iz eksponente družine spet vidimo, zakaj funkcijo $g(\mu) = \theta = \eta$ imenujemo naravna povezovalna funkcija.

Prametre β ocenjujemo prek metode največjega verjetja. (mogoče kaj več?)

8. Izbira pojasnjevalnih spremenljivk

Ko imamo nabor slučajnih spremenljivk, ki pojasnjujejo opazovano spremenljivko, se moramo odločiti, katere vključiti v končen model, tako da bo ta kar se da točen. Kriterija, na podlagi katerih sprejmemo to odločitev, sta prileganje podatkom in skopost/enostavnost, želimo torej model z dobrim prileganjem podatkov in majhnim številom parametrov, ker je takšne modele lažje uporabljati in razumeti. Ocenjevanje parametrov za pojasnjevalne slučajne spremenljivke, ki imajo zelo majhen vpliv na opazovano spremenljivko, je škodljivo, ker poveča varianco ocen parametrov in nam posledično zmanjša natančnost rezulatatov. Na žalost pa tukaj tiči konflikt interesov: Če želimo izboljšati prileganje podatkom, potrebujemo več parametrov, s tem pa zmanjšujemo enostavnost modela.

Prileganje podatkov je pogosto merjeno z devianco, vendar je njena največja uporaba v primerjavi gnezdenih modelov, torej če je množica pojasnjevalnih slučajnih spremenljivk v enem modelu podmnožica spremenljivk v drugem.

8.1. **Devianca in nasičen/poln(saturated) model.** Devianca je mera prileganja modela podatkom, vendar je samo relativna. Uporabljamo jo torej lahko samo za primerjavo med posameznimi modeli. Poln model (tudi nasičen ali maksimalen) je model, ki se popolnoma prilega danim podatkom. Število parametrov v takem modelu (označimo jih z n_S) je lahko kar enako številu opazovanih podatkov. Tak model se najbolje prilega podatkom, vendar je to prileganje pretirano, saj je v praksi tak model neuporaben.

Primernost modela M s $p < n_S$ regresijskimi parametri ocenimo tako, da primerjamo maksimalno vrednost verjetnostne funkcije modela M z maksimumom verjetnostne funkcije polnega modela. (dopolni) Recimo, da sta L_M in L_S maksimalni vrednosti verjetnostne funkcije modela M in polnega modela S za dan vzorec podatkov. Potem je razmerje največjih verjetij $\lambda = L_S/L_M$ in logaritem razmerja $\log \lambda = l_S - l_M$, pri čemer velja $l_S = \log L_S$ in $l_M = \log L_M$. Opazimo, da je $L_S \geq L_M$, ker se poln model popolnoma prilega podatkom, bolje od modela M, in potem velja tudi $l_S - l_M \geq 0$. Če se tudi model M dobro prilega podatkom, potem je $L_S \doteq L_M$ in $\log \lambda$ majhen. Velika vrednost $\log \lambda$ govori obratno, model M se slabo prilega podatkom.

Zdaj predpostavimo, da je v modelu, ki ga preučujemo, disperzijski parameter $\phi = 1$. Ocena največjega verjetja v polnem modelu za $\mu_i = \gamma'(\theta_i)$ je kar y_i. Če obrnemo, je ocena za θ_i v polnem modelu $\theta(y_i) \equiv (\gamma')^{-1}(y_i)$. Označimo še oceno največjega verjetja za θ_i v modelu M s $\hat{\theta}_i$. Potem je devianca (oz. $2\log\lambda$) modela M za vzorec podatkov y

$$D_M = 2(l_S - l_M) = 2\sum_{i=1}^{n_S} A_i \left([y_i\theta(y_i) - \gamma(\theta(y_i))] - [y_i\hat{\theta}_i - \gamma(\hat{\theta}_i)] \right),$$

kjer je posamezen prispevek opazovanja y_i k celotni devianci D_M

$$D_M(y_i) = 2A_i \left([y_i\theta(y_i) - \gamma(\theta(y_i))] - [y_i\hat{\theta}_i - \gamma(\hat{\theta}_i)] \right).$$

Devianco D_M lahko interpretiramo kot razdaljo med polnim modelom in modelom M. Najenostavnejši model je konstanten model, kjer je $E(Y) = \mu$ konstanta za vse i, takemu modelu pravimo tudi ničti model M_0. Ta model ima samo en parameter - μ. Devianci ničtega modela pravimo ničta devianca.

Če je disperzijski parameter ϕ različen od ena, potem devianca D_M ni več enaka $2\log\lambda$. V takem primeru opazujemo D_M/ϕ. Pravimo ji skalirana devianca.

Poglejmo si primer deviance v normalnem modelu. Najprej predpostavimo, da imamo vzorec n neodvisnih opazovanj normalne slučajne spremenljivke in da je $\phi = 1 = \sigma^2$. Log-verjetnostna funkcija za normalno porazdelitev, ki smo jo izpeljali v poglavju o eksponentnih družinah, je

$$l(\mu, \sigma^2; y) = \sum_{i=1}^{n} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - \mu)^2}{2\sigma^2} \right)$$

V polnem modelu S velja $\theta(y_i) \equiv (\gamma')^{-1}(y_i)$, to pa je pri normalni porazdelitvi enako y_i. Maksimalna vrednost funkcije l_S v polnem modelu je torej

$$l_S = \sum_{i=1}^{n} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - y_i)^2}{2\sigma^2} \right) = n\log\left(\frac{1}{\sqrt{2\pi}}\right).$$

V modelu M pa označimo oceno največjega verjetja za θ_i s $\hat{\theta}_i = \hat{\mu}_i$. Maksimalna vrednost funkcije l_M je enaka

$$l_M = \sum_{i=1}^{n} \left(\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(y_i - \hat{\mu}_i)^2}{2} \right) = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{2}.$$

Zdaj lahko izračunamo devianco D_M

$$D_M = 2(l_S - l_M) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2.$$

Če ϕ ni enak 1, potem imamo skalirano devianco modela M

$$\frac{D_M}{\phi} = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}{\sigma^2},$$

še več, izkaže se, da je porazdeljena enako kot χ^2_{n-p} (n je število parametrov v polnem modelu in p število parametrov v modelu M). Če naši podatki niso porazdeljeni normalno, potem je D_M/ϕ približno enako porazdeljena kot χ^2_{n-p}.

8.2. **Primerjava modelov z devianco.** Recimo, da vemo, da mora naš model vsebovati spremenljivke $x_1, \ldots, x_p$, nismo pa prepričani, ali naj bi vseboval tudi spremenljivke $x_{p+1}, \ldots, x_{p+q}$. Označimo model s p pojasnjevalnimi spremenljivkami z M_1, večji model s $p+q$ spremenljivkami pa z M_2. H_0 naj bo hipoteza, da je pravilen model M_1, povedano drugače, da so parametri $\beta_{p+1} = \cdots = \beta_{p+q} = 0$. Testna statistika na podlagi razmerja verjetij je tako

$$2(l_{M_2} - l_{M_1}) = 2(l_S - l_{M_1}) - 2(l_S - l_{M_1}) = D_{M_1} - D_{M_2} \equiv \Delta D \dot{\sim} \chi^2_q.$$

Torej, če dodatne pojasnjevalne spremenljivke res nimajo nobenega efekta v modelu, potem velja zgornje, hipotezo H_0 pa zavrnemo, če velja nasprotno. Če je H_0 napačna, potem se model M_2 bolje prilega podatkom in ima manjšo devianco, razlika devianc ΔD pa je posledično večja. Dodatne spremenljivke je smiselno vključiti v model, če ob stopnji zaupanja α velja $\Delta D > \chi^2_{1-\alpha,q}$. Takšno vključevanje dodatnih spremenljivk po navadi poteka za vsako spremenljivko posebej. Torej najprej preverimo ali naj bi model vseboval tudi spremenljivko x_{p+1}. Če velja $\Delta D > \chi^2_{1-\alpha,1}$, potem to spremenljivko vključimo v model in nadaljujemo z x_{p+2}.

8.3. **Analiza residualov.** Ko izberemo model s pomočjo deviance, kot je pokazano v prejšnjem razdelku, moramo preveriti, da so vse začetne predpostavke pravilne in če se model dobro prilega podatkom. Na začetku predpostavljamo, da so pojasnjevalne spremenljivke neodvisne in porazdeljene z gostoto iz iste eksponentne družine. Preveriti moramo tudi, če smo vključili vse razpoložljive pojasnjevalne spremenljivke in uporabili primerno povezovalno funkcijo. Pogosto lahko napake v modelu opazimo iz grafov residualov.

Residuali temeljijo na razliki med opazovanimi podatki in napovedanimi vrednostmi modela. Pearsonov residual (i-ti) za dan model je

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}},$$

kjer je $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ in $Var(\hat{\mu}_i) = (\phi/A_i)V(\hat{\mu}_i)$. Residual deviance ($i$-ti) pa je enak

$$r_{d_i} = sgn(y_i - \hat{\mu}_i)\sqrt{d_i},$$

kjer je d_i doprinos i-tega podatka celotni devianci.

Pearsonovi residuali so pogosto izkrivljeni (skewed) pri podaatkih, ki niso normalno porazdeljeni. zato je interpretacija težja. Po drugi strani pa so residuali deviance porazdeljeni normalno in so pogosto bolj priljubljeni. Za normalno porazdeljene podatke se izkaže, da so Pearsonovi residuali enaki residualom devianc.

Če vse predpostavke za posplošene linearne modele držijo, potem pričakujemo, da residuali ne kažejo nobenih vzorcev. Preverimo torej grafe residualov v odvisnosti od posameznih pojasnjevlnih spremenljivk. Če opazimo kakršenkoli trend, potem smo izpustili vpliv neke spremenljivke iz modela, ki bi ga radi zajeli. Histogram residualov pa uporabimo za testiranje predpostavke o porazdelitvi v našem modelu.