

Generalizirani linearni modeli

Bojan Basrak, PMF–MO Zagreb

Financijski praktikum

23. svibnja 2016.

U mnogim znanstvenim istraživanjima, ali i u financijama odn. osiguranju, važno je utvrditi da li neka odabrana veličina (npr. duljina života) ovisi o drugim mjerenim veličinama (npr. spolu, potrošnji duhanskih proizvoda, visini u odrasloj dobi, itd.). Veza između takvih mjerenja je vrlo rijetko jasna i deterministička, pa je najčešće predstavljamo koristeći vjerojatnosne modele. Veličinu od interesa modeliramo kao slučajnu varijablu koju nazivamo **ovisnom varijablom ili odzivom** (eng. response), a sva ostala mjerenja zovemo **neovisnim varijablama, predviđateljima ili kovarijatama** (eng. predictors).

Podatke koje želimo opisati tipično reprezentiramo kao niz parova

$$(y_i, \mathbf{x}_i), \quad i = 1, \dots, n,$$

gdje je y_i realizacija sl. varijable Y_i čija razdioba ovisi o kovarijatama \mathbf{x}_i . Odmah želimo primjetiti da kovarijate mogu primiti numeričke vrijednosti ili kategorijalne vrijednosti. U potonjem slučaju zovemo ih faktorima (npr. spol ili kategorija vozila). Takodjer, kovarijate mogu biti proizvoljno velike dimenzije.

Primjer (jednostavna linearna regresija)

Najpoznatiji ovakav model je svakako jednostavna linearna regresija. Kod nje pretpostavljamo da sl. varijabla Y_i na linearan način ovisi o numeričkoj kovarijati x_i , koja ima oblik

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

gdje je (ε_i) čine niz n.j.d. normalnih sl. varijabli s očekivanjem 0 i varijancom σ^2 . Ekivalentno bismo mogli pisati da je

$$EY_i = \beta_0 + \beta_1 x_i.$$

Primjetite da ovaj model ima tri parametra β_0 , β_1 i σ^2

Linearna zavisnost izmedju dvije varijable ne znači kauzalnost. Kontroverznih primjera je puno (v. spurious correlations)

- članak iz JL, ljeta 2006.: seks. eksplicitne pjesme i spolno ponašanje tinejdžera
- rode i djeca u Švedskoj u 19. st.
- današnji i sutrašnji povrati na cijene dionica
- pušenje (R.A. Fisher)
- itd.

Primjer (analiza varijance - ANOVA)

Ovdje je x_i jednodimenzionalna kategorijalna varijabla s recimo $m + 1$ različitih kategorija npr. $0, 1, 2, \dots, m$, sl. kao gore

$$Y_i = h(x_i) + \varepsilon_i,$$

gdje su ε_i kao i gore, a funkcija h zadovoljava npr.

$$h(0) = \mu$$

$$h(1) = \mu + \alpha_1$$

$$\vdots$$

$$h(m) = \mu + \alpha_m$$

Iako funkcija h , pa ni model ne izgledaju linearno to se može riješiti uvođenjem tzv. "dummy" varijabli

$$\begin{aligned} z_{i_1} &= \mathbb{I}\{x_i = 1\} \\ &\vdots \\ z_{i_m} &= \mathbb{I}\{x_i = m\}. \end{aligned}$$

Sad je model moguće zapisati i kao

$$Y_i = \mu + \sum_{j=1}^m \alpha_j z_{i_j} + \varepsilon_i,$$

pa je specijalno zbog normalnosti od ε_i

$$Y_i \sim N(h(x_i), \sigma^2)$$

Kategorijalne varijable dakle nisu nužno problem za linearne modele, no mnoge varijable koje susrećemo u aktuarskoj praksi nisu normalno distribuirane. Ponekad se i to može riješiti transformacijom ili promjenom razdiobe od ε_i .

Ostaje problem kako modelirati podatke s kategorijalnim ili npr. cjelobrojnim vrijednostima. Jedan od najvažnijih primjera su sl. varijable koje mogu imati dva stanja, npr. 1 ili 0, recimo ako je osiguranik prijavio štetu u ugovorenom periodu ili ne. U ovakvom slučaju bismo mogli prepostaviti da Y_i imaju Bernoullijevu razdiobu, ali tako da vjerojatnost uspjeha odn. $p_i = P(Y_i = 1)$ ovisi o kovarijati x_i , npr. dobi osiguranika. Pri tom dakako moramo voditi računa da p_i mora biti u intervalu $[0, 1]$, što neće biti uvijek zadovoljeno ako prepostavimo da $p_i = EY_i$ na linearan način ovisi o kovarijati x_i .

Ponekad bismo htjeli da težina repa ili varijanca podataka također ovise o kovarijatama ili o tzv. izloženosti (eng. exposure). Uglavnom, potrebna nam je bolja ideja.

Tri komponente generaliziranih linearnih modela

Osnovna ideja linearnih modela je pretpostavka da postoji linearna veza između očekivanja odziva i kovarijata tj.

$$EY_i = \sum_{j=1}^d \beta_j x_{ij},$$

gdje je d broj mogućih kovarijata (konstanta može biti uključena). Često čak pretpostavljamo i normalnost

$$Y_i \sim N\left(\sum_{j=1}^d \beta_j x_{ij}, \sigma^2\right).$$

Kod generaliziranih linearnih modela prepostavljamo

$$EY_i = g^{-1} \left(\sum_{j=1}^d \beta_j x_{ij} \right),$$

gdje je

- ▷ g^{-1} inverz funkcije veze g
- ▷ $\sum_{j=1}^d \beta_j x_{ij}$ tzv. linearni prediktor
- ▷ za zadano očekivanje Y_i ima unaprijed određenu razdiobu iz tzv. eksponencijalne familije.

Općenito o GLM

Teorija GLM omogućuje simultano modeliranje ovisnosti varijable odziva o numeričkim i kategorijalnim varijablama. Iako iz teorijskih razloga, slučajna komponenta odziva mora imati razdiobu iz neke od eksponencijalnih familija, ova restrikcija je u praksi tipično prihvatljiva jer ove familije uključuju najčešće korištene razdiobe.

Postoji etablirana statistička teorija procjene parametara, ali i veliki broj radova i knjiga koje zagovaraju primjenu GLMa.

Procjena parametara, kao i razni testovi za GLM-e uključeni su u komercijalne statističke pakete kao što su npr. SAS, SPSS, Splus, Statistica, kao i u lako dostupni R.

Eksponencijalne familije

Kažemo da sl. varijabla Y pripada nekoj eksponencijalnoj familiji ako joj gustoća (neprekidna ili diskretna) ima oblik

$$f(y; \theta, \varphi) = \exp \left[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi) \right] \quad (1)$$

za neke funkcije a, b i c . Primjetite da familija ima dva parametra: θ , tzv. **prirodni parametar** i φ tzv. **parametar disperzije** ili **skaliranja**. Pokazat ćemo da očekivanje od Y ovisi isključivo o parametru θ . Kasnije, u generaliziranom linearnom modelu, dopustit ćemo da θ ovisi o linearnoj kombinaciji kovarijata.

- ▷ Funkcija b u gornjoj definiciji je uvijek dvaput neprekidno diferencijabilna i t.d. je b' invertibilna.
- ▷ Funkcija a parametra φ se zove **funkcija disperzije** omogućuje dodatnu fleksibilnost u modelu, tako da ne moraju svi odzivi imati istu varijancu npr (inače bi bilo dosta pretpostaviti da je a identiteta). Često pretpostavljamo $\varphi > 0$ jer predznak ne mijenja ništa bitno u obliku razdiobe.
- ▷ Funkciju c tipično ignoriramo jer nema utjecaja u procesu procjene parametara GLMa.
- ▷ Parametar θ ima vrijednosti u otvorenom skupu.

Česti oblik funkcije disperzije je

$$a(\varphi) = \frac{\varphi}{w},$$

gdje je u praksi $w = w_i$ faktor težine odn. izloženosti i tog opažanja.

Ovako na relativno jednostavan način u model uvodimo nejednake varijance.

Promotrimo funkciju log-vjerodostojnosti $l(y; \theta, \varphi) = \log(f(y; \theta, \varphi))$ unutar neke eksponencijalne familije. Ona će nam trebati kasnije pri procjeni GLM. Trenutno trebamo dva vrlo dobro znana rezultata iz statističke teorije:

$$E \left[\frac{\partial l}{\partial \theta} \right] = 0 \quad \text{i} \quad E \left[\frac{\partial^2 l}{\partial \theta^2} \right] + E \left[\left(\frac{\partial l}{\partial \theta} \right)^2 \right] = 0. \quad (2)$$

Da bismo pokazali prvu od gornjih jednakosti, pretpostavimo da možemo diferencirati

$$\int f(y; \theta, \varphi) dx,$$

po θ jednostavno uvođenjem znaka diferenciranja pod integral (to je uvijek moguće unutar eksponencijalnih familija). Kako je gornji integral jednak 1 za sve θ , diferenciranjem ćemo dobiti 0.

Dakle

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial \theta} f(y; \theta, \varphi) \mathbf{d}x = \int \frac{\partial}{\partial \theta} f(y; \theta, \varphi) \frac{f(y; \theta, \varphi)}{f(y; \theta, \varphi)} \mathbf{d}x \\ &= \int \frac{\partial}{\partial \theta} l(y; \theta, \varphi) f(y; \theta, \varphi) \mathbf{d}x = E \left[\frac{\partial l}{\partial \theta} \right] \end{aligned}$$

Slično ako po θ diferenciramo jednakost

$$\int \frac{\partial}{\partial \theta} l(y; \theta, \varphi) f(y; \theta, \varphi) \mathbf{d}x = 0$$

dobijemo i drugu jednakost u (2).

Za log-vjerodostojnost eksponencijalnih familija vrijedi

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\varphi)}.$$

Stoga iz prve od jednakosti u (2) slijedi

$$\mu = E[Y] = b'(\theta) \iff \theta = b'^{-1}(\mu).$$

Kako vrijedi i

$$\frac{\partial^2 l}{\partial \theta^2} = \frac{-b''(\theta)}{a(\varphi)}$$

takodjer iz (2) imamo

$$\begin{aligned} \text{var}(Y) &= a(\varphi)^2 E \left(\frac{Y - b'(\theta)}{a(\varphi)} \right)^2 \\ &= a(\varphi)^2 E \left[\left(\frac{\partial l}{\partial \theta} \right)^2 \right] = -a(\varphi)^2 E \left[\frac{\partial^2 l}{\partial \theta^2} \right] = a(\varphi) b''(\theta). \end{aligned}$$

Dakle varijanca od Y iznosi

$$\text{var}(Y) = a(\varphi)b''(\theta),$$

gdje crtica označava derivaciju s obzirom na θ . Dakle, očekivanje ne ovisi o φ , dok varijanca općenito ovisi o oba parametra.

Može se pokazati da je b' neprekidna i invertibilna (čak striktno rastuća funkcija) osim u trivijalnim egzotičnim slučajevima. Stoga stavljajući $\mu = b'(\theta)$ zapravo uvodimo novi parametar, tzv. parametar srednje vrijednosti (mean value parameter), naime sad je $\theta = b'^{-1}(\mu)$, pa je dobro definirana **funkcija varijance** relacijom $\mu \mapsto V(\mu) = b''(\theta) = b''(b'^{-1}(\mu))$.

Uočimo, varijanca podataka ima dvije komponente: jednu koja uključuje parametar skaliranja, i drugu koja određuje način na koji varijanca ovisi o očekivanju. Da bi naglasili utjecaj očekivanja na varijancu izrazimo je kao

$$\text{var}(Y) = a(\varphi)V(\mu).$$

Normalna distribucija

$$\begin{aligned} f_Y(y; \theta, \varphi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(y - \mu)^2}{2\sigma^2} \right] \\ &= \exp \left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2} \left(\frac{y^2}{\sigma^2} + \log 2\pi\sigma^2 \right) \right] \end{aligned}$$

što je oblika (1) sa

$$\theta = \mu$$

$$\varphi = \sigma^2$$

$$a(\varphi) = \varphi$$

$$b(\theta) = \theta^2/2$$

$$c(y, \varphi) = -\frac{1}{2} \left(\frac{y^2}{\sigma^2} + \log 2\pi\sigma^2 \right) .$$

Stoga je prirodni parametar normalne distribucije jednak μ , a parametar skaliranja je σ^2 .

Prvo, promatramo li normalnu distribuciju, iz (2) možemo izvesti očekivanje i varijancu

$$b(\theta) = \theta^2/2 \quad \text{te stoga} \quad E[Y] = b'(\theta) = \theta = \mu$$
$$a(\varphi) = \varphi \quad \text{te stoga} \quad \text{var}[Y] = a(\varphi)b''(\theta) = \varphi = \sigma^2 .$$

Za normalnu distribuciju, varijanca ne ovisi o očekivanju (zbog $b''(\theta) = 1$), ali ćemo kod drugih distribucija vidjeti da to nije uvijek slučaj.

Poissonova distribucija

$$f_Y(y; \theta, \varphi) = \frac{\mu^y e^{-\mu}}{y!} = \exp[y \log \mu - \mu - \log y!]$$

što je oblika (1) sa

$$\theta = \log \mu$$

$$\varphi = 1, \text{ te zato } a(\varphi) = 1$$

$$b(\theta) = e^\theta$$

$$c(y, \varphi) = -\log y!$$

Zato je prirodni parametar Poissonove distribucije $\log \mu$, očekivanje je $E[Y] = b'(\theta) = e^\theta = \mu$, a funkcija varijance je $V(\mu) = b''(\theta) = e^\theta = \mu$. Funkcija varijance nam kaže da je varijanca proporcionalna očekivanju. Vidimo da je varijanca u stvari *jednaka* očekivanju, jer je $a(\varphi) = 1$.

Binomna distribucija Binomnu slučajnu varijablu moramo prvo podijeliti s n . Pretpostavimo, dakle, da je $Z \sim$ binomna (n, μ) . Stavimo $Y = \frac{Z}{n}$, tako da je $Z = nY$. Diskretna gustoća od Z je $f_Z(z; \theta, \varphi) = \binom{n}{z} \mu^z (1 - \mu)^{n-z}$, te supstituirajući za z , distribucija od Y je (za $y = k/n$, $k = 0, \dots, n$) zadana gustoćom

$$\begin{aligned}
 f_Y(y; \theta, \varphi) &= \binom{n}{ny} \mu^{ny} (1 - \mu)^{n-ny} \\
 &= \exp \left[n(y \log \mu + (1 - y) \log(1 - \mu)) + \log \binom{n}{ny} \right] \\
 &= \exp \left[n \left(y \log \left(\frac{\mu}{1 - \mu} \right) + \log(1 - \mu) \right) + \log \binom{n}{ny} \right]
 \end{aligned}$$

Dakle opet dobijemo izraz (1) sa

$$\theta = \log \left(\frac{\mu}{1 - \mu} \right)$$

(uočite da je inverzno preslikavanje $\theta \mapsto \mu = \frac{e^\theta}{1 + e^\theta}$),

$$\varphi = n$$

$$a(\varphi) = \frac{1}{\varphi}$$

$$b(\theta) = \log(1 + e^\theta)$$

$$c(y, \varphi) = \log \binom{n}{ny}.$$

Zato je prirodni parametar binomne (specijalno i Bernoullijeve) distribucije $\log\left(\frac{\mu}{1-\mu}\right)$, očekivanje je

$$E[Y] = b'(\theta) = \frac{e^\theta}{1 + e^\theta},$$

a funkcija varijance je

$$V(\mu) = b''(\theta) = \frac{e^\theta}{(1 + e^\theta)^2} = \mu(1 - \mu).$$

Gama distribucija Kod gama distribucije korisno je zamijeniti parametre iz α i λ u α i $\mu = \frac{\alpha}{\lambda}$, t.j., $\lambda = \frac{\alpha}{\mu}$.

$$\begin{aligned} f_Y(y; \theta, \varphi) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} = \frac{\alpha^\alpha}{\mu^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y\alpha}{\mu}} \\ &= \exp \left[\left(-\frac{y}{\mu} - \log \mu \right) \alpha + (\alpha - 1) \log y + \alpha \log \alpha - \log \Gamma(\alpha) \right] \end{aligned}$$

Ponovo je gustoća oblika (1) sa

$$\theta = -\frac{1}{\mu}$$

$$\varphi = \alpha$$

$$a(\varphi) = \frac{1}{\varphi}$$

$$b(\theta) = -\log(-\theta)$$

$$c(y, \varphi) = (\varphi - 1) \log y + \varphi \log \varphi - \log \Gamma(\varphi).$$

Stoga je, ignorirajući negativan predznak (zapravo prebacujući ga na funkciju odziva), prirodni parametar gama distribucije jednak $1/\mu$. Očekivanje je $E[Y] = b'(\theta) = -1/\theta = \mu$. Funkcija varijance je $V(\mu) = b''(\theta) = 1/\theta^2 = \mu^2$, te je varijanca jednaka μ^2/α .

Konačno, lognormalna distribucija koja se često upotrebljava u neživotnom osiguranju za modeliranje distribucija veličine šteta, također se može uključiti u okvir GLM, jer ako je $Y \sim \text{lognormalna}$, tada je $\log Y \sim \text{normalna}$. Dakle, ako želimo koristiti lognormalnu distribuciju, podatke prvo trebamo logaritmirati, te se tada može primjeniti modeliranje normalnom distribucijom.

Generalizirani linearni model

Generalizirani linearni model pretpostavlja da varijable odziva Y opažamo na nezavisan način za različite vrijednosti kovarijate x . Model pri tom ima sljedeće komponente:

- Kovarijate x koje mogu biti višedimenzionalne i na linearan način utječu na razdiobu od Y , preko tzv. linearnog predviđatelja $\eta = \beta_1 x_1 + \dots + \beta_r x_r$, gdje je $x = (x_1, \dots, x_r)$.
- Razdioba sl. varijable Y za dane kovarijate pripada uvijek istoj eksponencijalnoj familiji razdioba.
- Očekivanje od Y je glatka i invertibilna funkcija linearnog predviđatelja η oblika $b' \circ h$ za neku funkciju h , tj.

$$\theta = h(\eta) \quad \text{i} \quad \mu = EY = b'(\theta) = b'(h(\eta))$$

pa je

$$\mu = g^{-1}(\eta) \quad \text{odn} \quad \eta = g(\mu),$$

za funkciju $g = h^{-1} \circ b'^{-1}$ koju zovemo **funkcija veze**.

I prirodni parametar θ je u ovom slučaju glatka funkcija od η

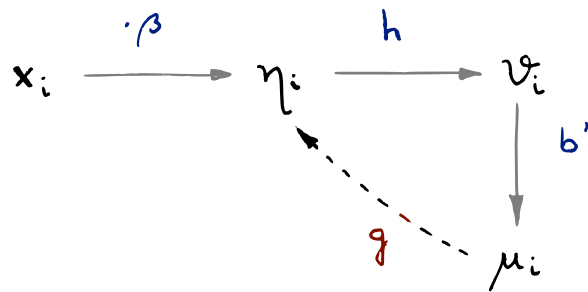
$$\theta = b'^{-1}(\mu) = h(\eta).$$

Ako $h \equiv id$, tada je $\theta \equiv \eta$, a

$$g = b'^{-1}$$

se naziva kanonska funkcija veze. Općenito utjecaj kovarijata na očekivanje odziva prikazuje dijagram

$$(\mathbf{x}_i) \longmapsto \eta_i = \beta_1 x_{i,1} + \cdots + \beta_r x_{i,r} \xrightarrow{h} \theta_i \xrightarrow{b'} \mu_i$$



funkcija veze $g = h^{-1} \circ b'^{-1}$

Primjer (linearna regresija s normalnim šumom)

Kako smo već naveli ovaj model specificira da odziv Y ima razdiobu kao $\mu + \varepsilon$, gdje je $\varepsilon \sim N(0, \sigma^2)$, a broj μ predstavlja očekivanje od Y i ovisi o kovarijatama $x = (x_1, \dots, x_p)$ na sljedeći način

$$\mu = \eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

Primjetite da je ovo specijalni slučaj GLM, kod kojeg je eksponencijalna familija familija normalnih razdioba, a funkcija veze je jednostavno identiteta $g(\mu) = \mu$. Razmislite – zašto uvijek možemo dopustiti konstantni član?

Primjer (logistička regresija)

Ako želimo modelirati binarni odziv kao Bernoullijevu sl. varijablu, moramo odrediti vjerojatnosti $\mu = P(Y = 1) = EY$ kao funkciju od kovarijata. Naravno, ta funkcija mora primiti vrijednosti u intervalu $[0, 1]$. Jedan često korišten model specificira ovisnost oblika

$$\log \frac{\mu}{1 - \mu} = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r, \text{ tj. } \mu = \frac{e^\eta}{1 + e^\eta}.$$

Funkcija veze je tzv. **logit funkcija** $g(\mu) = \log \mu / (1 - \mu)$, dakle upravo kanonska funkcija veze za binomnu razdiobu.

Primjer (probit veza)

Mogli smo razumno pretpostaviti i da je

$$\mu = \Phi(\eta) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r)$$

sada je funkcija veze tzv. **probit funkcija** $g(\mu) = \Phi^{-1}(\mu)$.

Primjer (Poissonova razdioba)

Pretpostavimo cjelobrojni odziv modeliramo kao Poissonovu sl. varijablu s parametrom $\mu = EY$ koji ovisi o kovarijatima. Ako koristimo kanonsku funkciju veze vrijedi $g(\mu) = \log \mu$ stoga je

$$\mu = e^\eta = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r) .$$

Primjer (gama razdioba)

Razmislite koje vrijednosti linearnog prediktora daju nedopuštene vrijednosti prirodnog parametra ako se koristi kanonska funkcija veze (prema Ohlssen & Johansson ovo je razlog nepopularnosti kanonske funkcije veze u aktuarskoj praksi, premda se i to može dopustiti ako pazimo kod procjene parametara β_i .)

Za linearnu i logističku regresiju je $\theta = g(\mu) = \eta$, odn. $g = b'^{-1}$, a linearni predviđatelj jednak je prirodnom parametru modela. Takve funkcije veze zovemo također **prirodnim ili kanonskim**. Sljedeća tablica sadrži neke važnije familije i pripadne kanonske funkcije veze.

Razdioba	Kanonska veza	Naziv
Binomna	$\log(\mu/(1 - \mu))$	logit funkcija
Poissonova	$\log \mu$	logaritam
Normalna	μ	identiteta
Gama	μ^{-1}	inverz

Nije nužno uvijek koristiti kanonske funkcije veze. One su izabrane zbog jednostavnosti. Općenito moramo samo voditi računa da vrijednost parametra $\theta = g^{-1}(\eta)$ mora pripadati u skup prihvatljivih parametara.

Napomenimo još jednom da kategorijalne varijable ili faktore u linearni predivitelj možemo uvesti preko tzv. "dummy variables".

Linearni predviditelji

Kako smo istaknuli kovarijate x_i utiču na razdiobu od Y preko linearnog predviditelja. Ako imamo jednu numeričku kovarijatu (ili kontrolnu varijablu, npr. dob) x linearni predviditelj ima oblik

$$\beta_0 + \beta_1 x.$$

Ako uz nju imamo još jedan faktor (ili kategorijalnu kovarijatu, npr. spol) s dvije kategorije onda će linearni predviditelj imati oblik

$$\alpha_i + \beta x, \quad i = 1, 2$$

gdje konstantu α_1 koristimo za prvu, a α_2 za drugu od dvije kategorije.

Ponekad će i utjecaj ostalih kovarijata ovisiti o faktoru, npr. dob može imati različiti utjecaj na varijablu Y za osobe ženskog u odn. na osobe muškog spol. Tada kažemo da su te dvije kovarijate u **interakciji**. Linearni predviditelj bismo mogli zapisali u obliku

$$\alpha_i + \beta_i x, \quad i = 1, 2$$

gdje se konstante α_i, β_i razlikuju za dvije kategorije.

Ako imamo dvije kovarijate, obje kategorijalne, dva faktora dakle, i tada je moguće za između njih postoji interakcija, pa bi u tom slučaju linearni predviditelj pisali u obliku

$$\alpha_i + \beta_j + \gamma_{ij}, \quad i, j = 1, 2, \dots$$

Postoji uobičajena notacija za specificiranje tih modela, koja je za gornje modele kako slijedi:

model	linearni predviditelj
dob	$\beta_0 + \beta_1 x$
spol	α_i
dob + spol	$\alpha_i + \beta x$
dob + spol + dob. spol	$\alpha_i + \beta_i x$
dob* spol	$\alpha_i + \beta_i x$

Zadnja dva modela su ekvivalentna, a prikazana su odvojeno za ilustraciju oznaka koje se koriste za modele. Slična notacija se koristi i u standardnim statističkim paketima.

Linearni predviđitelji su vrlo korisni u praksi, no ponekad bismo htjeli dopustiti mogućnost da kovarijate utiču na odziv na način koji nije nužno linearan. I ovaj zahtjev možemo uključiti u GLM (v. pritom Nelder's dictum, cf. D. Williams)

model

linearni predviđitelj

dob

$$\beta_0 + \beta_1 x$$

dob + dob²

$$\beta_0 + \beta_1 x + \beta_2 x^2$$

dob + trajanje

$$\beta_0 + \beta_1 x + \beta_2 x_2$$

log(dob)

$$\beta_0 + \beta_1 \log x$$

spol

$$\alpha_i$$

grupa za određivanje premije vozila

$$\beta_j$$

spol + grupa za određivanje premije vozila

$$\alpha_i + \beta_j$$

spol* grupa za određivanje premije vozila

$$\alpha_i + \beta_j + \gamma_{ij}$$

Interpretacija parametara

U jednostavnoj linearnoj regresiji je jasno

$$EY_i = \beta_0 + \beta_1 x_i,$$

pa je parametar β_1 lako interpretirati. Općenito to nije slučaj u GLM modelima.

Primjer (logistička regresija)

Prisjetimo se odziv Y ima Bernoullijevu razdiobu s parametrom $\mu = P(Y = 1) = EY$, a zbog jednostavnosti pretpostavimo da imamo samo jednu binarnu numeričku kovarijatu $x = 0, 1$, tj. neka je $\eta = \beta_0 + \beta_1 x$. Jasno

$$\mu = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

Sad je

$$\mu(x = 1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \quad \text{odn.} \quad \mu(x = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

Za Bernoullijeve sl. varijable nekad računamo tzv. izgleda (engl. odds)

$$\text{odds}(Y) = \frac{\mu}{1 - \mu}.$$

Omjer izgleda za Y, Y' Bernoullijeve s vjerojatnostima uspjeha μ, μ' je

$$\text{odds ratio}(Y, Y') = \frac{\frac{\mu}{1 - \mu}}{\frac{\mu'}{1 - \mu'}}.$$

Uočite, omjer izgleda nije relativni rizik

$$\text{relative risk}(Y, Y') = \frac{\mu}{\mu'}.$$

Posebno ako je $Y_1 \sim \text{Ber}(\mu(x = 1))$ i $Y_0 \sim \text{Ber}(\mu(x = 0))$

$$\text{odds ratio}(Y_1, Y_0) = e^{\beta_1},$$

tj. izgledi odziva Y su za ovaj faktor veći (manji) ako je $x = 1$.

Primjer (Poissonova razdioba)

Pretpostavimo cjelobrojni odziv modeliramo kao Poissonovu sl. varijablu s parametrom $\mu = EY$ koji ovisi o numeričkim kovarijatama. Ako koristimo kanonsku funkciju veze lako je interpretirati parametre u relaciji

$$\mu = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_r x_r) .$$

Procjena i odabir modela

Procjena parametara

Procjena parametara u GLM može biti poprilično složena. Standardna procedura za njihovu procjenu je MLE procedura tj. korištenje maksimuma funkcije vjerodostojnosti. U praksi osim za jednostavne modele pronaći maksimume nije jednostavno i koriste se numerički algoritmi (Newton-Raphson ili iterative weighted least squares). Za izradu intervala pouzdanosti i testove potrebni su nam i elementi Fisherove informacijske matrice. Parametar φ se obično može neovisno procijeniti. U osnovnom primjeru linearne regresije procjena je relativno jednostavan zadatak.

Primjer (linearna regresija) Uz pretpostavku da podaci $\mathbf{y} = (y_i)$ slijede model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

njihova je funkcija log-vjerodostojnosti oblika

$$l(\mathbf{y}; \beta_0, \beta_1, \sigma^2) = \sum_{i=1}^n -\frac{\log \sigma^2}{2} - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

Da bismo pronašli procjenitelje za β_0 i β_1 možemo derivirati gornju funkciju po ovim parametrima. Iz oblika funkcije l , jasno je da MLE procjenitelji $\hat{\beta}_0$ i $\hat{\beta}_1$ neće ovisiti o σ^2 .

Procjenitelj dobijemo minimizirajući izraz

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Dakle naši procjenitelji su jednostavno rezultat metode najmanjih kvadrata. A sumu

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

zovemo **residual sum of squares** (RSS). On je mjera kvalitete podudaranja podataka s modelom i našim procjeniteljima.

Jasno je da bi model s proizvoljno parametara mogao u mnogim slučajevima postići savršenu predikciju očekivanih opaženih odziva tj. riješiti sustav

$$y_i = g^{-1}(\eta_i) = g^{-1}(\beta_1 x_{i,1} + \dots + \beta_r x_{i,r}).$$

Model u kojem je to moguće naziva se **zasićeni** i tipično nije jako koristan (razmislite zašto). Ipak usporedimo li naš procjenjeni model sa zasićenim mogli bismo saznati koliko je on zaista dobar.

Kako model s većim brojem parametara uključuje model s manjim brojem parametara kao svoj (restringirani) podmodel, ako njihove maksimizirane log-vjerodostojnosti označimo s \hat{l} odn. sa \hat{l}_0 vrijedi

$$\hat{l} \geq \hat{l}_0,$$

a obje vrijednosti su manje od \hat{l}_F odn. log-vjerodostojnosti zasićenog modela.

Devijanca i odabir modela

Općenito ako sa \hat{l} označimo maksimiziranu log-vjerodostojnost koju smo postigli s našim procjeniteljima, a sa \hat{l}_F maksimiziranu log-vjerodostojnost u zasićenom modelu, (skaliranu) **devijancu** našeg modela definiramo kao

$$d_M = 2(\hat{l}_F - \hat{l}).$$

Neskalirana devijanca za odabrani model, D_M , definira se kao

$$D_M = d_M \varphi.$$

U slučaju linearne regresije iz našeg primjera, devijanca D_M je točno jednaka RSS (provjerite). Neskalirana devijanca ne ovisi o mjernoj jedinici.

Devijanca je općenito oblika

$$D_M = \sum w_i d(y_i, \mu_i),$$

a funkcija d ovisi o eksponencijalnoj familiji razdioba. Odredite je za normalne, Poissonove i gama razdiobe.

Nakon što smo odabrali model i procjenili njegove parametre važno je znati koje od kovarijata su zaista bitne u modeliranju odziva, a koje možemo ispustiti. Da bismo to napravili pretpostavimo da imamo p kontroliranih varijabli, tako da je linearni procjenitelj oblika

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

Prepostavimo da smo maksimiziranjem funkcije log-vjerodostojnosti pronašli procjenitelje $\hat{\beta}_i$ i dosegli maksimum \hat{l} .

Prepostavite da želimo testirati nul hipotezu $H_0 : \beta_{q+1} = \cdots = \beta_p = 0$.

Procedura testa ima tri koraka

- Koristeći MLE procjenimo parametre u restringiranom modelu
- Nadjemo maksimalnu log-vjerodostojnost u restringiranom modelu, recimo \hat{l}_0
- Izračunamo **log-likelihood ratio statistics ili statistiku omjera vjerodostojnosti**

$$2(\hat{l} - \hat{l}_0).$$

Ako je ona iznad kritične vrijednosti za izabrani nivo značajnosti i razdiobu χ^2 sa $p - q$ stupnjeva slobode odbacujemo nul hipotezu.

Uočite da je testna statistika razlika devijanci, tj.

$$2(\hat{l} - \hat{l}_0) = d_{M_0} - d_M$$

Test u posljednjem koraku je egzaktno samo za normalnu razdiobu, inače vrijedi asimptotski. Lako se vidi da je test moguće bazirati i na različitim skaliranim devijancijama, te da se dobije potpuno isti izraz. Pri tome smo ignorirali činjenicu da je za određivanje skaliranih devijanci (kao i \hat{l} i \hat{l}_0) potrebno znati φ (ili σ^2 u slučaju linearne regresije npr.) Ako je φ nužno prvo procijeniti test se ponekad obavlja u odnosu na F razdiobu. U praksi se testovi često provode uz nivo značajnosti od 5%.

Alternativno možemo promatrati z -statistike $\hat{\beta}_k/\text{s.e.}(\hat{\beta}_k)$, koje bi uz nul-hipotezu $\beta_k = 0$ trebale imati (aproksimativno) standardnu normalnu razdiobu, pa bismo nul-hipotezu odbacivali kad je taj izraz po apsolutnoj vrijednosti veći od $1.96 \approx 2$.

Primjer (Štete motornih vozila. Iz P. McCullagh i J. Nelder, “Generalized Linear Models”).

Skalirane devijacije za potpun raspon mogućih modela pokazane su u donjoj tablici.

Model	Devijanca	Stupnjevi slobode	Razlike	
1	638.32	122		
do			81.13	7
do + gv	557.19	115	226.5	3
do + gv + sv	330.65	112	204.1	3
do + gv + sv + do.gv	126.51	109	34.32	21
do + gv + sv + do.gv + do.sv	92.191	88	18.77	21
do + gv + sv + do.gv + do.sv + gv.sv	73.416	67	3.89	9
do + gv + sv + do.gv + do.sv + gv.sv + do.gv.sv	69.524	58	69.524	58
do + gv + sv + do.gv + do.sv + gv.sv + do.gv.sv	0	0		

Uočite da se prvi model sastoji jednostavno od konstante (1 parametar je $\eta = \beta_0$, d.f. = broj podataka -1), a zadnji je zasićen, te je stoga prilagodba savršena. Zadnji model također sadrži i član s trostrukom interakcijom. U praksi, kod prilagodjivanja modela određivanja premija takve je modele najbolje izbjegavati.

Pri odlučivanju koji model(i) adekvatno objašnjava podatke, treba ispitivati razlike u devijancama i stupnjevima slobode. Redoslijed kojim se članovi dodaju modelu utječe na rezultate, te se u praksi može gledati nekoliko redoslijeda tako da se ne propusti ništa važno. Na primjer, svaki glavni faktor može se prilagoditi sam za sebe, umjesto da se svaki dodaje modelu kao što je napravljeno gore.

Primjer Binarni odziv.

```
X <- read.table("ex12.8.1.dat", header = TRUE)
names(X)
```

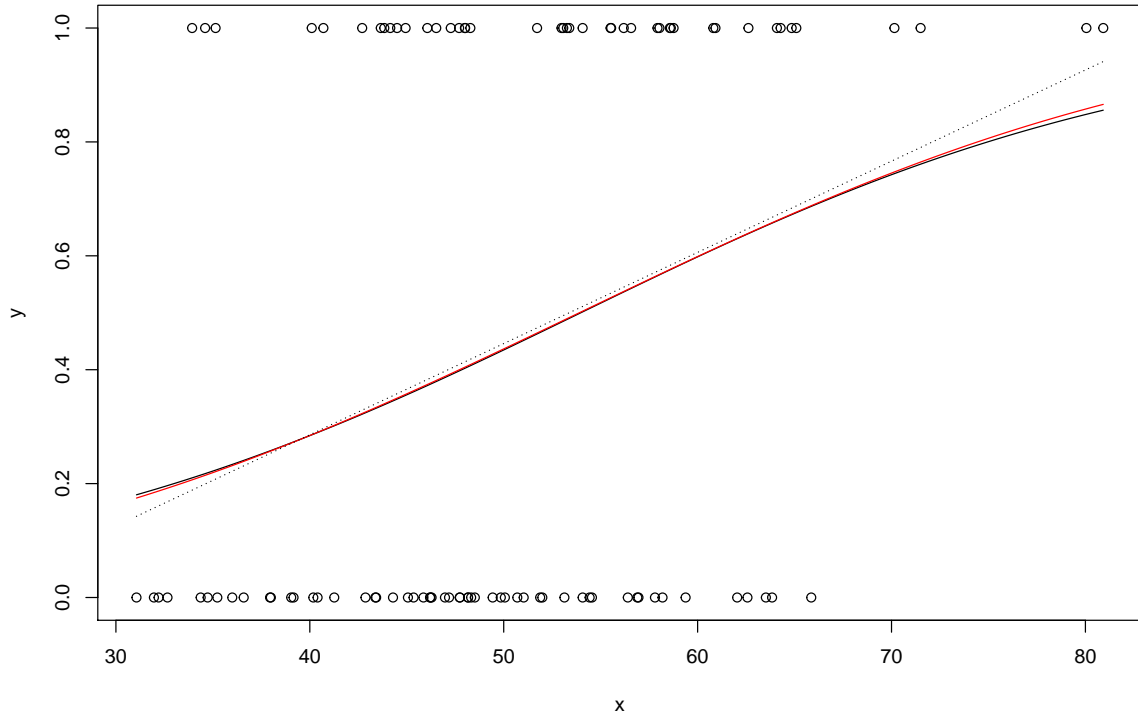
```
## [1] "y" "x" "z"
```

```
attach(X)
X[1:12,]
```

```
##      y      x      z
## 1  0 49.85 36.82
## 2  1 44.50 41.55
## 3  0 50.06 71.37
## 4  1 58.76 48.47
## 5  1 64.85 40.98
## 6  0 62.04 59.00
## 7  0 46.22 36.12
## 8  0 46.97 42.62
## 9  0 56.40 59.54
## 10 0 45.35 47.34
## 11 1 56.57 56.45
## 12 1 42.70 26.75
```

```
glm(y ~ x, family = binomial);  
glm(y ~ x, family = binomial(link = "probit"));  
glm(y ~ x+I(x^2)+z, family = binomial)
```

Linearni, logit i probit model



Primjer Binarni odziv - podaci o štetama u osiguranju. Poredak je važan uočite.

```
ydf=read.table("PodaciGLM2013.csv", header = TRUE, sep = ",")
attach(ydf)
names(ydf)
```

```
## [1] "Steta"          "spolovi"        "regija"         "tip"
## [5] "voz.iskustvo"  "cijena.voz"    "dob"
```

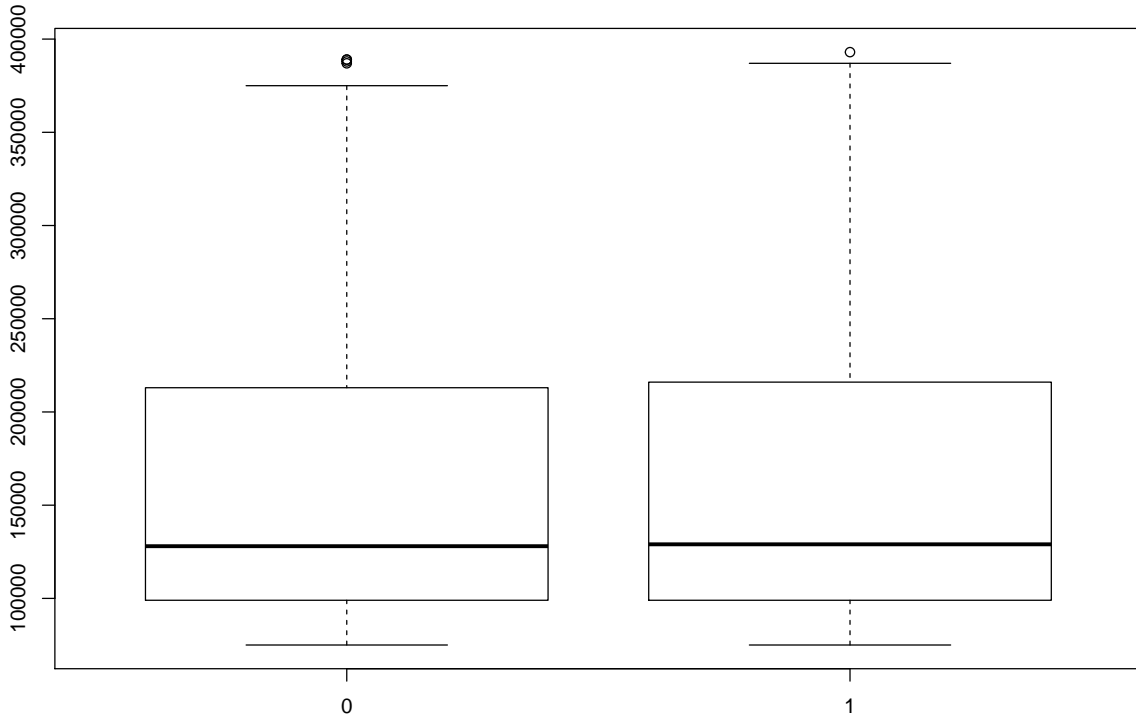
```
ydf[1:10,]
```

```
##      Steta spolovi regija tip voz.iskustvo cijena.voz dob
## 1      1      M      2     1          18      103000  37
## 2      1      Z      2     2          42       85000  60
## 3      1      Z      2     2          25       86000  43
## 4      1      M      2     2          47       88000  66
## 5      0      Z      2     2          47       99000  65
## 6      1      M      3     3          20      106000  38
## 7      1      M      1     3           6      323000  27
## 8      1      M      1     3          11      106000  29
## 9      0      Z      1     3          16      113000  35
## 10     0      Z      2     5          27      114000  49
```

```
t(table(Steta,regija))/colSums(table(Steta,regija))
```

```
##           Steta
## regija         0         1
##    1 0.2963863 0.7036137
##    2 0.3480019 0.6519981
##    3 0.3172193 0.6827807
```

```
boxplot(cijena.voz~Steta)
```



```
outA.logit <- glm(Steta ~ cijena.voz+ dob+
as.factor(spolovi)+as.factor(regija), family = binomial)
anova(outA.logit)
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model: binomial, link: logit
```

```
##
```

```
## Response: Steta
```

```
##
```

```
## Terms added sequentially (first to last)
```

```
##
```

```
##
```

##		Df	Deviance	Resid. Df	Resid. Dev
##	NULL			9999	12586.6
##	cijena.voz	1	1.69	9998	12584.9
##	dob	1	3117.29	9997	9467.7
##	as.factor(spolovi)	1	54.17	9996	9413.5
##	as.factor(regija)	2	42.97	9994	9370.5

Primjer Podaci o broju policijskih zaustavljanja po okruzima i rasi. Uočite *offset*.

##	eth	precinct	stops	past.arrests
## 1	1	1	202	980
## 2	2	1	102	295
## 3	3	1	81	381
## 4	1	2	132	753
## 5	2	2	144	557
## 6	3	2	71	431
## 7	1	3	752	2188
## 8	2	3	441	627
## 9	3	3	410	1238
## 10	1	4	385	471

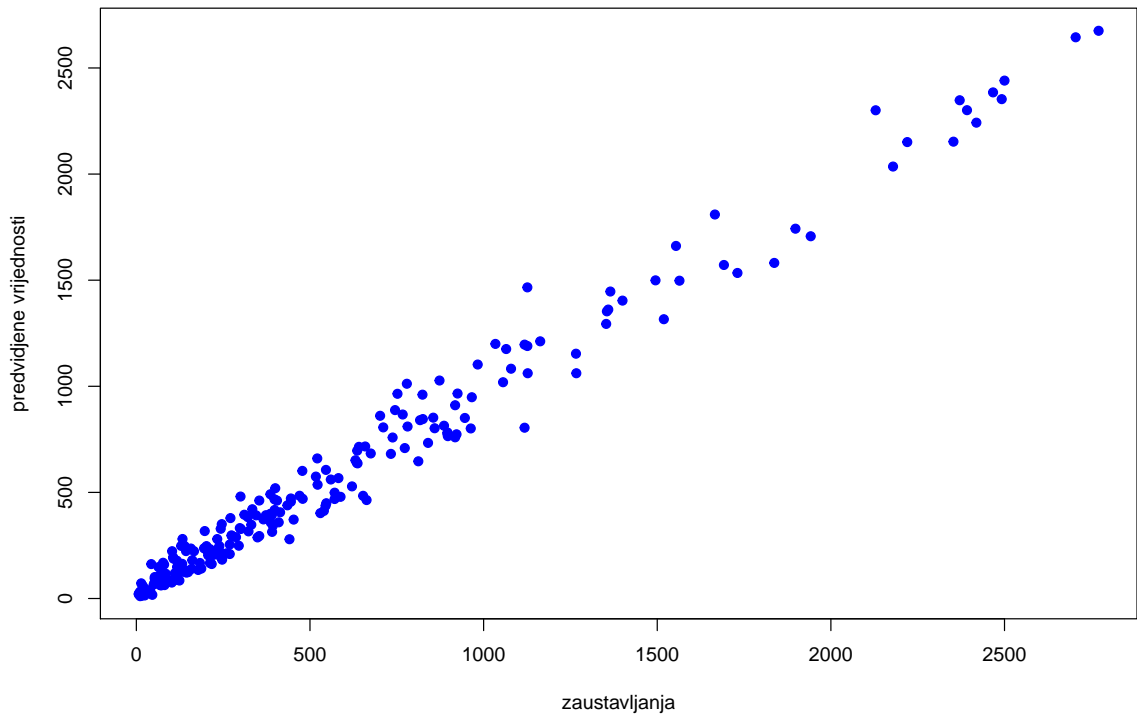
```
fit.2 <- glm (stops ~ factor(eth), data=stops,  
family=poisson, offset=log(past.arrests))  
summary(fit.2)
```

```
##  
## Call:  
## glm(formula = stops ~ factor(eth), family = poisson, data = stops,  
##      offset = log(past.arrests))  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q        Max  
## -47.327   -7.740   -0.182   10.241   39.140  
##  
## Coefficients:  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  -0.588086   0.003784 -155.40 <2e-16 ***  
## factor(eth)2  0.070208   0.006061  11.58 <2e-16 ***  
## factor(eth)3 -0.161581   0.008558 -18.88 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for poisson family taken to be 1)  
##  
##      Null deviance: 46120  on 224  degrees of freedom  
... - - - - -
```

```
fit.2 <- glm (stops ~ factor(eth), data=stops,  
family=poisson, offset=log(past.arrests))  
summary(fit.2)  
fit.3 <- glm (stops ~ factor(eth) + factor(precinct),  
data=stops, family=poisson, offset=log(past.arrests))  
summary(fit.3)
```

```
anova(fit.3)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: stops
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                224      46120
## factor(eth)         2      683      222      45437
## factor(precinct) 74    42010      148      3427
```

Ako je razlika u devijancama veća od kritične vrijednosti, tada je dodani član značajan u objašnjenju varijacije u odzivu. Stoga vidimo da se svaki od glavnih faktora čini značajnim i treba biti upotrebljen u modelu. Međutim, niti jedna interakcija ne čini se naročito značajnom.

Oprez: redosljed kojim smo uvodili nove komponente u model je bitan. Moguće je ponekad da drugi redosljed da i drugačije rezultate, tj. drugačiji konačni model. Postoje i automatizirane procedure zasnovane na AIC i sličnim kriterijima za odabir modela, no uloga aktuara/statističara ostaje presudna.

U praksi treba pokušati izbjeći "overfitting".

Analiza ostataka i ocjena prilagodbe modela

Nakon što je upotrebom svih devijanci nađen mogući model, treba ga provjeriti gledajući ostatke i značajnost parametara. Ostaci se zasnivaju na razlikama između opaženih i predviđenih odziva

$$y_i - \hat{\mu}_i.$$

U primjeru regresije ostaci su bili $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Ako imamo dva faktora na primjer, i upotrijebimo log funkciju veze, predviđeni odzivi mogu se dobiti iz

$$\hat{\mu} = e^{\alpha_j + \beta_j}.$$

Da bismo usporedili ostatke moramo ih prethodno standardizirati. U softverskim paketima najčešće se koriste dva oblika standardizacije:

Pearsonovi ostaci su

$$\frac{y - \hat{\mu}}{\sqrt{\text{var}(\hat{\mu})}}$$

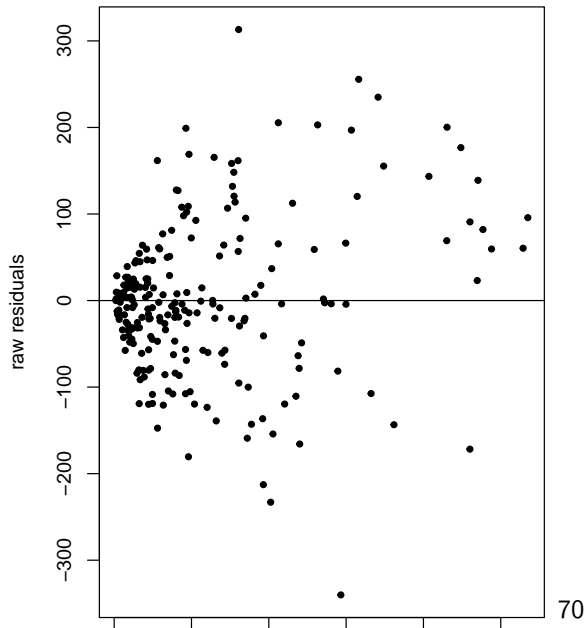
Ostaci u devijanci su definirani kao produkt predznaka od $y - \hat{\mu}$ i drugog korijena doprinosa od y devijanci. I oni kao i Pearsonovi ostaci (kad je model točan) imaju približno i asimptotski standardnu normalnu razdiobu.

U praksi bismo analizirali ostatke koristeći histogram ili *qq*-plot da vidimo jesu li naše prepostavke o razdiobi opravdane. Također bismo tražili skrivenu strukturu u ostacima, plotajući ih nasuprot pojedinih kovarijata. Ako grafovi sugeriraju još neki oblik zavisnosti preostaje nam pogledati i proširiti naš model.

```

par(mfrow=c(1,2))
pv <- fitted(fit.3)
r <- (stops$stops - fitted(fit.3))
plot(pv, r, pch=20, ylab="raw residuals", xlab="predicted value")
abline(h=0)
sr <- (stops$stops - fitted(fit.3))/sqrt(fitted(fit.3))
plot(pv, sr, pch=20, ylab="standardized residuals", xlab="predicted val")
abline(h=c(-2,0,2),lty=c(2,1,2))

```



70

