UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Leon Horvat

Posplošeni linearni modeli

Delo diplomskega seminarja

Mentor: izred. prof. dr. Jaka Smrekar

Ljubljana, 2018

Kazalo

1. Uvod 4
2. Linearna regresija 4
3. Eksponentna družina 7
3.1. Normalna porazdelitev 9
3.2. Poissonova porazdelitev 9
3.3. Binomska porazdelitev 9
3.4. Gama porazdelitev 10
4. Povezovalne funkcije 11
5. Pojasnjevalne spremenljivke 11
5.1. Številske in kategorične slučajne spremenljivke 12
5.2. Interakcije 12
5.3. Nelinearen vpliv pojasnjevalnih spremenljivk 12
6. Minimalna zadostna statistika 13
7. Izbira pojasnjevalnih spremenljivk 13
7.1. Devianca in poln model 14
7.2. Primerjava modelov z devianco 15
7.3. Analiza residualov 15
8. Analiza prekinitev pogodb življenjskega zavarovanja 16
8.1. Oblike življenjskega zvarovanja 16
8.2. Izbira povezovalne funkcije 17
8.3. Podatki in pojasnjevalne spremenljivke 17
8.4. Gradnja modelov 19
9. Zaključek 30
Slovar strokovnih izrazov 31
Literatura 31

Posplošeni linearni modeli

Povzetek

Analiziranje in modeliranje povezav med spremenljivkami postaja vedno bolj pomembno. V delu so predstavljeni linearni posplošeni modeli in njihov razvoj iz modelov linearne regresije. Diskutirane so teoretične zahteve modelov, podrobno je predstavljena eksponentna družina porazdelitev slučajnih spremenljivk in pripadajoče naravne povezovalne funkcije. Definirane so osnovne oblike pojasnjevalnih slučajnih spremenljivk, za lažje razumevanje so podani njihovi primeri. Razloženi so tudi postopki za preverjanje prileganja gnezdenih modelov z devianco.

Vse našteto je uporabljeno za gradnjo modelov prekinitev, kapitalizacij in odkupov polic življenjskega zavarovanja. Za bolj jasno sliko so opisane oblike življenjskih zavarovanj in njihova povezava s prekinitvami, kapitalizacijami in odkupi. Razčlenjeno je čiščenje in preoblikovanje podatkov, ki so bili na voljo za modeliranje. Podrobno so raziskani vplivi posameznih pojasnjevalnih slučajnih spremenljivk na proučevano spremenljivko in ugotovitve, katere spremenljivke so pomembne za napovedovanje in katere ne. Vse to omogoča zavarovalnici globlji vpogled v kompleksna razmerja v njenem portfoliu in boljšo pripravljenost na dejavnike tveganja v prihodnosti.

Generalized linear models

Abstract

Analysing and modelling relationships between variables are getting more and more important. In this work, we introduce generalized linear models, and develop them from linear regression models. We discuss theoretical assumptions for these models, and give an in-depth explanation of exponential families of distributions and the associated canonical link functions. We classify the most standard types of explanatory variables, and provide several examples for easier understanding. We explain the procedure for comparing nested models with deviance.

We apply the theory described above to constructing models for lapsed, paid up, and surrendered life insurance policies. For a clearer picture, different forms of life insurance and their relationships with lapsed, paid up and surrendered policies are presented. We analyse the influences of individual explanatory variables on the response variable, and determine which explanatory observations are essential and which are not. With this, an insurance company may gain insight into complex relationships in its portfolio and better readiness for risk factors in the future.

Math. Subj. Class. (2010): 62J12

Ključne besede: posplošeni linearni modeli, linearna regresija, logistična regresija, eksponentna družina, povezovalna funkcija, devianca, življenjsko zavarovanje, prekinitev, kapitalizacija, odkup

Keywords: generalized linear models, linear regression, logistic regression, exponential family, link function, deviance, life insurance, lapse, paid up insurance, surrender

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 proučevane slučajne spremenljivke linearna funkcija neznanih parametrov β. Neznane parametre β ocenimo po metodi najmanjših kvadratov, torej da minimiziramo vsoto kvadratov residualov ali pa po metodi največjega verjetja. S tem postopkom dobimo cenilko za parametre β, ki se glasi $\hat{\beta} = (X^TX)^{-1}X^TY$, pri čemer je X matrika, sestavljena iz stolpcev pojasnjevalnih spremenljivk. Privzamemo, da ima matrika X poln rang. Kot opazimo je cenilka $\hat{\beta}$ linearna cenilka. Če gledamo Y_i kot i-to proučevano slučajno spremenljivko in x_i^T vektor znanih spremenljivk, potem model linearne regresije oceni Y_i z

$$\hat{Y_i} = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} \equiv \hat{\eta}_i.$$

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 proučevana 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 $E(g(Y_i)) = \sum_{j=1}^{p} \beta_j x_{ij}$. 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 pa naj bo slučajna spremenljivka, ki zavzame vrednost 1, če je zavarovanec podal odškodninski zahtevek, in 0 sicer. $E(Y|X = x)$ je tedaj verjetnost odškodninskega zahtevka pogojno na to, da je starost zavarovanca enaka x, to pa je cenilka za π_x. Ker je π_x delež, mora biti element intervala (0,1). Če za model vzamemo $\pi_x = \beta_0 + \beta_1 x$ 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 $\text{logit}(x) = \log(\frac{x}{1-x})$. Opazimo, da ta funkcija slika interval $(0, 1)$ bijektivno na celotno realno os. V modelu logistične regresije bomo torej modelirali $\text{logit}(\pi)$ kot linearno kombinacijo pojasnjevalnih spremenljivk, v našem primeru starosti voznika. Model se torej glasi:

$$g(\pi_x) = \text{logit}(\pi_x) = \text{logit}(E(Y|X = x)) = \log \frac{\pi_x}{1 - \pi_x} = \beta_0 + \beta_1 x = \eta_x$$

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

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 zanimale 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 $g(\lambda_x) = \log \lambda_x = x^T \beta = \eta_x$ oziroma $\lambda_x = e^{\eta_x}$ V tem modelu $\lambda_x = E(Y|X = x)$ in linearno kombinacijo pojasnjevalnih spremenljivk povezuje logaritemska funkcija. Logaritem slika iz intervala $(0, \infty)$ 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.

Pri klasični linearni regresiji privzamemo, da je $Y = (Y_1, \ldots, Y_n)$ slučajni vektor, komponente pa so neodvisne, normalno porazdeljene spremenljivke, $Y_i \sim N(\mu_i, \sigma^2)$, in velja $E(Y_i) = \mu_i = x_i^T \beta$. Tu razumemo vektorje x_i ne več kot slučajne vektorje, ampak kot konstantne vekotorje. 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 polnega ranga, 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 $\epsilon \sim N(0, \sigma^2 I)$. 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 kombinacija obojih. 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 napoved vsake 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, ki jih razumemo kot konstante, torej minimiziramo

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

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

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

Opazimo, da je cenilka pridobljena z metodo največjega verjetja enaka cenilki po metodi najmanjših kvadratov in ta se glasi $\hat{\beta} = (X^TX)^{-1}X^TY$. To formulo dobimo na sledeč način. Poiskati moramo minimum izraza

$$\begin{aligned}(Y - X\beta)^T(Y - X\beta) &= Y^TY - \beta^TX^TY - Y^TX\beta + \beta^TX^TX\beta = \\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta\end{aligned}$$

Z matričnim odvajanjem izraza po β in enačenjem z 0 dobimo enačbo $-2X^TY + 2X^TX\beta = 0$. Če izrazimo β, dobimo iskano cenilko. Ta β je minimum zgornjega izraza, saj je drug odvod izraza pozitiven. Še dodatno nam izrek Gaussa in Markova pove, da je cenilka za parametre β pridobljena po katerikoli od obeh metod, (cenilki sta enaki), najboljša linearna nepristranska cenilka za β.

3. Eksponentna družina

Najprej si bomo pogledali, kako mora biti 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 enoparametrične verjetnostne porazdelitve, ki pripadajo eksponentni družini. Te so tudi najpogosteje uporabljene pri modeliranju posplošenih linearnih modelov.

Definicija 3.1. Slučajna spremenljivka Y pripada enoparametrični 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 gostote porazdelitve, torej

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

z U pa označimo parcialni odvod logaritemske funkcije verjetja po θ in je torej $U(\theta) = \frac{\partial}{\partial\theta} l = \frac{A}{\phi}(y - \gamma'(\theta))$

Trditev 3.2. *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

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

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

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 je varianca funkcija pričakovane vrednosti. 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 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.

3.1. **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:

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

3.2. **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.

3.3. **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:

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

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 $\text{logit}(p)$ in ne le p. Pričakovana vrednost in varianca sta:

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

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

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

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

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

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

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.

4. Povezovalne funkcije

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

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

Prvi 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	$\operatorname{logit} \mu$	$\mu(1-\mu)$	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).

5. 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 ocenimo z metodo največjega verjetja. Najprej pa podrobneje opišimo različne pojasnjevalne spremenljivke.

5.1. **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 (neki) 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. Kategorično 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.

5.2. **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.

5.3. **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.

6. 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 o porazdelitvi proučevane slučajne spremenljivke. 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.

7. 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.

7.1. **Devianca in poln model.** Devianca je mera prileganja modela podatkom, vendar je samo relativna. Uporabljamo jo torej lahko samo za primerjavo med posameznimi, gnezdenimi 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. 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 enako $\lambda = L_S/L_M$ logaritem razmerja pa $\log \lambda = l_S - l_M$, 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 v modelu, ki ga preučujemo, za disperzijski parameter velja $\phi = 1$. 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 enaka

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

kjer posamezen prispevek opazovanja y_i k celotni devianci D_M znaša

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

Devianco D_M lahko interpretiramo kot razdaljo med polnim modelom in modelom M. Najenostavnejši model je konstantni model, kjer je $E(Y) = \mu$ konstanta za vse i, takemu modelu pravimo tudi ničti model M_0. Ta model ima samo en parameter, in sicer 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 porazdeljena kot χ^2_{n-p}.

7.2. **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 velja $\beta_{p+1} = \cdots = \beta_{p+q} = 0$. Testna statistika na podlagi razmerja verjetij je tako

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

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}.

7.3. **Analiza residualov.** Ko izberemo model s pomočjo deviance, kot je pokazano v prejšnjem razdelku, moramo preveriti, ali so vse začetne predpostavke pravilne in ali se model dobro prilega podatkom. Na začetku predpostavljamo, da so pojasnjevalne spremenljivke neodvisne in porazdeljene z gostoto iz iste eksponentne družine. Preveriti moramo tudi, ali 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 dani 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 asimetrični pri podatkih, 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 diagrami residualov ne kažejo nobenih pravilnosti vzorcev, ki bi lahko pomenili sistematično odstopanje in ne samo slučajnega odstopanja. Preverimo torej grafe residualov v odvisnosti od posameznih pojasnjevalnih 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.

8. Analiza prekinitev pogodb življenjskega zavarovanja

Pri življenjskih zavarovanjih je tveganje za zavarovalnico poleg smrti tudi prekinjen dotok premije stranke. To se lahko zgodi zaradi prekinitve pogodbe, kapitalizacije zavarovanja ali odkupa.

Zavarovalnice želijo natančne podatke o številu prekinitev pogodb, kapitaliziranih in odkupljenih polic za različne namene. Potrebujejo jih za določanje cen zavarovanj, rezervacije, oceno vgrajene vrednosti in ostalo. Dognanja, specifična za vsako zavarovalnico, so razvita iz analiziranja zgodovinskih podatkov zavarovalnice, da lahko ocenimo in napovemo vrednosti v prihodnosti. Ti rezultati so uporabljeni tudi za oceno, ali so historični podatki dober pokazatelj prihodnosti ali pa je potrebno zaradi prevelike spremembe okolja (ekonomskega, pravnega, fiskalnega) oceno pridobiti drugače oziroma jo prilagoditi. Pomembno je, da zavarovalnice ocenjujejo in analizirajo prekinitve, kapitalizacije in odkupe zavarovanj pogosto, saj samo tako dobro poznajo tveganja, ki prihajajo iz tega naslova.

V tej analizi sem poskušal poiskati in raziskati spremenljivke, ki vplivajo na prekinitev zavarovalne police z uporabo posplošenih linearnih modelov. S tem pa tudi možnost napovedovanja prihodnjih trendov verjetnosti prekinitev, kapitalizacij in odkupov zavarovalnih polic.

Ta različna tveganja nastopijo pri različnih oblikah življenjskega zavarovanja. Poznamo življenjsko zavarovanje za primer smrti, zavarovanje za primer doživetja, mešano življenjsko zavarovanje in naložbeno življenjsko zavarovanje. Za boljše razumevanje je podana razlaga posameznih oblik življenjskega zavarovanja.

8.1. **Oblike življenjskega zvarovanja.**

Življenjsko zavarovanje za primer smrti oz. riziko je zavarovanje, pri katerem zavarovalnica izplača dogovorjeno zavarovalno vsoto za primer smrti, to pomeni, če zavarovanec umre v času trajanja zavarovanja. Zavarovanje je lahko sklenjeno za določeno dobo trajanja ali pa je sklenjeno za primer smrti za vse življenje. Riziko zavarovanje je lahko sklenjeno tudi s padajočo zavarovalno vsoto, pri katerem se zavarovalna vsota znižuje s pretečenim trajanjem zavarovanja.

Življenjsko zavarovanje za primer doživetja je zavarovanje, ki izplača zavarovalno vsoto v primeru, da zavarovanec preživi dobo trajanja zavarovanja. Ob morebitni smrti zavarovanca zavarovalnica ne izplača nič.

Mešano življenjsko zavarovanje je zavarovanje za primer smrti in doživetja, torej se ob izteku zavarovanja izplača dogovorjena zavarovalna vsota z dobičkom, ki je bil ustvarjen do konca zavarovanja, zavarovalna vsota pa se izplača tudi v primeru smrti. Tako zavarovanje ima tudi varčevalno komponento, zato je to hkrati

zavarovanje in varčevanje. Varčevalni del premije se nalaga na poseben kritni sklad zavarovalnice, ki ima več regulatornih omejitev za zagotavljanje čim večje varnosti sredstev. Zavarovalci (tisti, ki zavarujejo zavarovane osebe) nimajo vpliva na naložbeno politiko sklada, prav tako pa tudi ne prevzemajo nobenega naložbenega tveganja, višina izplačila ob izteku zavarovanja je zajamčena.

Naložbeno življenjsko zavarovanje je podobno mešanemu zavarovanju, prav tako je to zavarovanje za primer smrti in doživetja. Razlika je v tem, da pri naložbenem zavarovanju zavarovalec izbira, v katere investicijske sklade bo varčevalni del premije naložen. S tem prevzema nase tudi naložbeno tveganje, vložena sredstva so podvržena nihanju kapitalskih trgov.

Prekinitev zaradi plačevanja se zgodi lahko pri vseh zavarovanjih, vendar pri življenjskih zavarovanjih z varčevalno komponento le v prvih dveh ali treh letih - odvisno od oblike in trajanja zavarovalnega produkta. V tem primeru zavarovalnica nima več nobene obveze do stranke, zavarovanje postane ničnо. V primeru, da stranka preneha plačevati premijo po preteku dobe, v kateri lahko nastane prekinitev, pride do kapitalizacije. To pomeni, da zavarovanje ne preneha, ampak se glede na dosedanjo plačano premijo oblikuje nova zavarovalna vsota, za katero je zavarovanec zavarovan do konca trajanja police. Pri mešanem ali naložbenem zavarovanju pa je po preteku določenega obdobja možen še odkup. Odkup je prekinitev zavarovanja na prošnjo zavarovalca, ob tem pa mu zavarovalnica izplača odkupno vrednost police, torej varčevalni del premije, zmanjšan za morebitne stroške.

8.2. **Izbira povezovalne funkcije.** V svojem primeru sem obravnaval verjetnost posamezne prekinitve pogodbe življenjskega zavarovanja kot proučevano slučajno spremenljivko in skušal narediti čim boljši posplošen linearni model z danimi podatki. Za povezovalno funkcijo sem izbral funkcijo logit, ker opazujemo razmerje prekinitev pogodb. V razdelku o porazdelitvah iz eksponentne družine smo spoznali, da je to binomsko porazdeljena slučajna spremenljivka, logit pa je naravna povezovalna funkcija za binomsko porazdelitev.

8.3. **Podatki in pojasnjevalne spremenljivke.** V svoji analizi sem zaobjel vsa zavarovanja dotične zavarovalnice od leta 2002 dalje. To leto sem izbral, ker so v svojo ponudbo dodali naložbeno zavarovanje, leta 2005 pa še življenjsko zavarovanje za primer smrti oziroma riziko. Ti dve obliki življenjskih zavarovanj sta tudi edini v sedanji ponudbi zavarovalnice, zato imata po mojem mnenju mnogo večji vpliv na današnje razmere kot ostali produkti drugačnih oblik. Prav tako je v tem obdobju močno naraslo število zavarovanj v tej zavarovalnici.

Za pojasnjevalne spremenljivke sem vzel podatke o zavarovalni polici, zavarovalcu in zavarovalnem zastopniku, ki je to polico prodal zavarovalcu. Izbral sem zavarovalca in ne zavarovanca, ker se mi zdi, da je prekinitev pogodbe odvisno od osebe, ki plačuje zavarovanje in ne od osebe, ki je zavarovana. Podatki o polici so: *zavarovalni produkt, status police, datum začetka zavarovanja, datum konca zavarovanja, datum prekinitve zavarovanja* (če je bilo sploh prekinjeno) in *frekvenca plačevanja premije*. Želel bi si tudi podatek o višini premije in zavarovalne vsote, vendar bi bila ta dva podatka težje pridobljena iz baze podatkov zavarovalnice, posebno pri neaktivnih policah (kapitalizirane in odkupljene), zato sem ju izpustil. Zavarovalec je opisan s podatki o *datumu rojstva, spolu in poštni številki*. Zastopnika pa opisujejo *datum rojstva, spol, poštna številka, poslovna enota* in *prodajna pot*.

Iz zgornjih spremenljivk sem generiral nekatere nove. Iz *datuma začetka* in *datuma konca* sem naredil spremenljivko *doba*. Ta spremenljivka pove, za kolikšno dobo

je stranka sklenila zavarovanje. Iz spremenljivke *zavarovalni produkt* sem ustvaril novo spremenljivko *tip zavarovanja*, ki lahko zavzame vrednosti ‚riziko', ‚klasika'in ‚naložbeno'zavarovanje. Ker dotična zavarovalnica že vrsto let nima v svoji ponudbi klasičnega življenjskega zavarovanja, sem ta zavarovanja izpustil. Obravnaval sem torej le naložbena zavarovanja in riziko zavarovanja.

V opisani tabeli je vsaka vrstica predstavljala eno zavarovanje. Taka struktura ne prikaže aktivnosti police pred prekinitvijo ali iztekom. To je bil razlog, da sem generiral novo pojasnjevalno spremenljivko *trajanje*. Ta spremenljivka pove, koliko let je preteklo med začetkom zavarovanja in prekinitvijo ali iztekom zavarovanja. Nato sem za vsako polico generiral nove vrstice. Vsaka polica je imela toliko zapisov, kot je bila vrednost spremenljivke *trajanje*. Vrstice iste police pa so se razlikovale v tem, da je vsaka imela drugo številko od 1 do vrednosti *trajanja*. Ta podatek prikazuje *leto zavarovanja*. Zapisom, ki so imeli *leto zavarovanja* manjše od *trajanja*, sem dodelil oznako ‚aktivna', ostalim pa pustil oznako statusa police. S tem sem ustvaril novo spremenljivko *aktivnost*. Dodal pa sem tudi spremenljivko *koledarsko leto*, ki pove, v katerem koledarskem letu je posamezna zavarovalna polica bila aktivna oz. prekinjena.

Iz *datuma rojstva* zavarovalca in zastopnika sem ustvaril spremenljivki *starost zavarovalca* in *starost zastopnika*. Starost obeh oseb se je spreminjala z vsakim *letom zavarovanja*, zato se tudi pri vsaki polici starost oseb povečuje. Iz spremenljivke *poštna številka* zavarovalca sem ustvaril dve novi: Prva prikazuje *okraj* zavarovalca, iz poštne številke sem izluščil samo prvo številko. Temu koraku je botrovalo veliko različnih poštnih številk v Sloveniji, zato sem število kategorij lokacije zmanjšal na okraje. Iz *poštne številke* sem naredil tudi spremenljivko *mestna oseba*, ki pokaže ‚Da', če zavarovalec prihaja iz kraja s poštno številko, ki je večkratnik števila 1000, v ostalih primerih pa ‚Ne'. Zdi se mi, da je to dober približek dejanskemu stanju.

V portfelju zavarovalnice so tudi stranke, ki imajo stalno prebivališče izven Slovenije. Za te stranke sem v spremenljivko *okraj* zapisal ‚Tujina'. Poseben status imajo tudi pravne osebe. Pri teh strankah ne moremo opazovati spremenljivk spol in starost, zato pravne osebe v teh dveh spremenljivkah zavzameta vrednost ‚pravna'.

Med bolj podrobnim pregledom podatkov sem zaznal nekaj anomalij. Nekatere stranke v bazo niso bile vnešene kot stranke s tujim prebivališčem, zato sem moral sam prečistiti in odstraniti zapise brez slovenskih naslovov. Največ napak se je pojavljalo pri datumu rojstva zavarovalnega zastopnika. V preveč primerih se je kot datum rojstva pojavil zapis 01.01.1900 ali 01.01.1950, v nekaterih primerih pa je bila letnica rojstva že iz tretjega tisočletja, kar je za zastopnika nemogoče. Ocenil sem, da ima spremenljivka *starost zastopnika* preveč napak in jo zato odstranil. Nekaj napak je bilo tudi pri starosti zavarovalcev, zato sem očitno napačne primere izločil, obstaja pa možnost, da so nekatere še ostale. Odstranil sem vse police z zavarovalci nad starostjo 84 let, saj jih je bilo premalo in so izkrivljali modele. Upam pa si trditi, da na model kot celoto nimajo velikega vpliva, saj sem zaobjel ogromno množico podatkov.

Na začetku sem tipično številskim pojasnjevalnim spremenljivkam določil kategorije in jih spremenil v kategorične spremenljivke. Ta korak se storil, da bi lažje in hitreje zasledil, v kakšnem razmerju (ne nujno linearnem) sta posamezna pojasnjevalna spremenljivka in proučevana spremenljivka. S tem sem zagotovil večjo natančnost modela.

Končne pojasnjevalne spremenljivke (v oklepaju so navedene vrednosti spremenljivke), ki sem jih vključil v model, so:

- spremenljivke zavarovalne police:
  - *tip zavarovanja* (riziko, naložbeno),
  - *zavarovalna doba* (od 1 leta do 40 let ali vseživljenjsko),
  - *frekvenca plačevanja* (enkratno, letno, polletno, četrtletno, mesečno),
  - *možnost odkupa police* (Da, Ne),
  - *leto zavarovanja* - v katerem letu trajanja je polica (1-17),
  - *koledarsko leto* - iz katerega koledarskega leta je podatek (2002 - 2018);
- spremenljivke zavarovalca:
  - *spol zavarovalca* (moški, ženska, pravna oseba),
  - *starost zavarovalca* (0 - 84 let, pravna oseba),
  - *okraj zavarovalca* (1 - 6 in 8,9 in tujina),
  - *mestni* - ali je zavarovalec iz mesta (Da, Ne);
- spremenljivke zavarovalnega zastopnika:
  - *spol zastopnika* (moški, ženska),
  - *poslovna enota* (0-9),
  - *prodajna pot* (specialne agencije, specialni zastopniki, univerzalne agencije, univerzalni zastopniki, direktni marketing, nekaj posameznih agencij).

Proučevana slučajna spremenljivka je *aktivnost*, ki lahko zavzame kategorije (aktivna, prekinjena, kapitalizirana, odkup).

Modele sem gradil v s programskim jezikom R, v okolju RStudio. Ta dobro podpira statistično analizo podatkov, v posebnem grajenje posplošenih linearnih modelov. Za potrebe teh modelov sem uporabljal knjižnico *stat*. Zaradi različne narave in vsebine možnih prekinitev sem zgradil 3 modele; posebej za prekinitve, kapitalizacije in odkupe.

8.4. **Gradnja modelov.**

8.4.1. *Prekinitve.* Prekinitve življenjskega zavarovanja so možne tako pri riziko zavarovanju kot tudi pri naložbenem zavarovanju, vendar pri naložbenem le v prvih dveh ali treh letih, odvisno od trajanja zavarovanja in posameznega produkta. Če stranka preneha plačevati premijo po preteku tega obdobja, se polica kapitalizira.

Iz podatkovne množice sem za gradnjo modela prekinitev vzel le police, ki so pri spremenljivki *aktivnost* zavzeli vrednost 'aktivna' ali vrednost 'prekinjena'. Kapitaliziranih in odkupljenih polic torej nisem obravnaval. Odstranil sem tudi vsa naložbena zavarovanja, ki imajo vrednost *leto zavarovanja* večjo od 3, ker pri teh prekinitev ni možna. Za ta model tudi ni pomembna spremenljivka *možnost odkupa police*, zato sem jo odstranil.

Na začetku sem gradil model brez interakcij: Linearna cenilka η je bila torej sestavljena iz vseh zgoraj naštetih pojasnjevalnih spremenljivk, razen spremenljivke *možnost odkupa police*. V R-u sem poklical naslednji ukaz za gradnjo modela:

```
glm(AKTIVNOST~., family = binomial, data = prekinitve)
```

Funkcija `glm()` zgradi posplošen linearni model, podamo ji formulo, torej v kakšni zvezi želimo proučevano slučajno spremenljivko z ostalimi pojasnjevalnimi spremenljivkami. V zgornjem primeru je `AKTIVNOST` proučevana spremenljivka, vse ostale spremenljivke (stolpci) v tabeli pa so pojasnjevalne spremenljivke. Izbrana porazdelitev je binomska, v tem primeru bo povezovalna funkcija logit. Podatki, s katerimi fukcija model zgradi, pa so se v mojem primeru nahajali pod imenom `prekinitve`.

Za preučitev deviance modela prekinitev sem uporabil funkcijo `anova()`. Funkcija najprej izračuna devianco za ničti model in poda prostostne stopnje. Potem dodaja posamično pojasnjevalne spremenljivke v enakem vrstem redu, kot so podane v tabeli. Za vsako spremenljivko poda število prostostnih stopenj (df), devianco modela, ki vsebuje vse prejšnje spremenljivke, vključno z dotično spremenljivko. Podane so tudi prostostne stopnje tega modela in razlika devianc. Na podlagi vseh teh podatkov je izračunana tudi p-vrednost testa razlike devianc s porazdelitvijo χ^2. Rezultati so podani v tabeli 2.

Tabela 2. Tabela devianc modela prekinitev.

Spremenljivka	df	D	Δ df	Δ D	$p - vrednost$
-	433656	297795,9	-	-	-
leto zavarovanja	433643	287832,2	13	9963,65741	0,000000
frekvenca plačevanja	433639	279986,9	4	7845,29225	0,000000
zavarovalna doba	433607	277159,9	32	2827,06209	0,000000
tip zavarovanja	433606	275460,1	1	1699,77577	0,000000
spol zavarovalca	433604	275265,3	2	194,78485	$5,046701 \cdot 10^{-43}$
starost zavarovalca	433588	273258,1	16	2007,15626	0,000000
okraj zavarovalca	433580	271428,1	8	1830,08751	0,000000
mestni	433579	271396,6	1	31,46184	$2,033981 \cdot 10^{-08}$
poslovna enota	433570	269198,2	9	2198,43912	0,000000e+00
spol zastopnika	433569	269109,9	1	88,25988	$5,739581 \cdot 10^{-21}$
prodajna pot	433561	267233,9	8	1875,94743	0,000000
koledarsko leto	433545	263920,9	16	3313,00297	0,000000

Opazil sem, da ima tudi po vseh vključenih pojasnjevalnih spremenljivkah model še vedno velik delež deviance ničtega modela, natančneje, devianca se je zmanjšala le za 11,5 %. Menim, da je razlog v veliki variabilnosti podatkov in da ni na voljo še več pojasnjevalnih spremenljivk, ki bi morda bolje pojasnile dogajanje. Iz tabele je razvidno tudi, da so vse pojasnjevalne spremenljivke v modelu statistično značilne in izboljšujejo prileganje modela podatkom. Največja razlika devianc je pri spremenljivki *leto zavarovanja*, sledijo *frekvenca plačevanja*, *koledarsko leto* in *zavarovalna doba*.

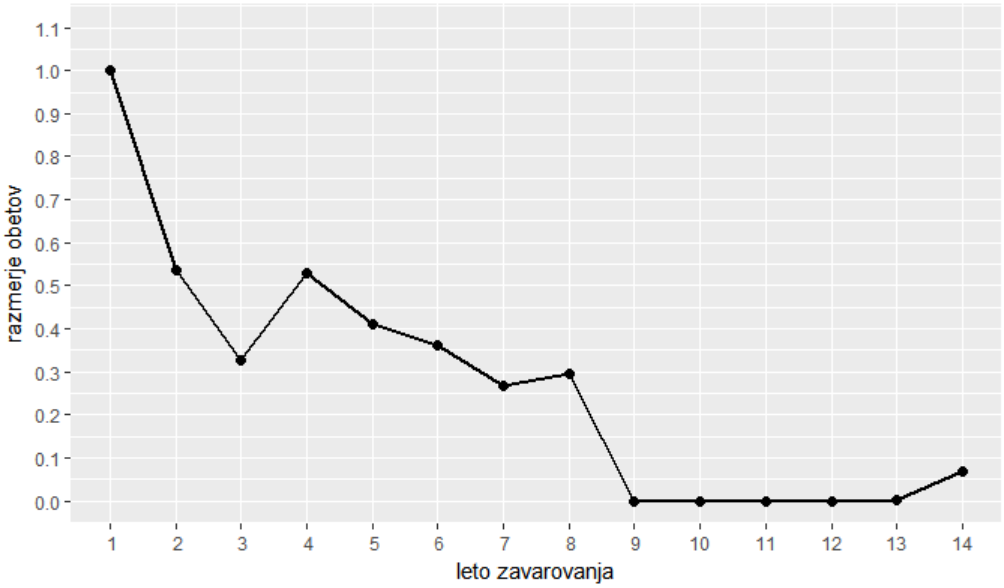
V nadaljevanju je z grafi ponazorjen vpliv posameznih spremenljivk na prekinitev. Grafi predstavljajo razmerje obetov v odvisnosti od kategorij pojasnjevalne spremenljivke. Obet je razmerje verjetnosti, razmerje obetov pa je mera med izpostavljenostjo določenemu faktorju in izidom. Za lažje razumevanje bom navedel primer: Če vzamemo *ženski spol* za bazo in je razmerje obetov pri kategoriji *moški spol* 2, potem je obet prekinitve za moškega dvakrat večji v primerjavi z žensko. V zakup pa so upoštevane tudi vse ostale spremenljivke v modelu. Razmerje obetov sem iz ocen za koeficiente dobil z eksponentno funkcijo: $exp(\text{logit}(p)) = \frac{p}{1-p}$

Opazimo, da je največja verjetnost prekinitve pogodbe življenjskega zavarovanja v prvem letu trajanja, razmerje obetov pa se potem z leti manjša. Lokalni vrh ima graf še v četrtem in osmem letu.

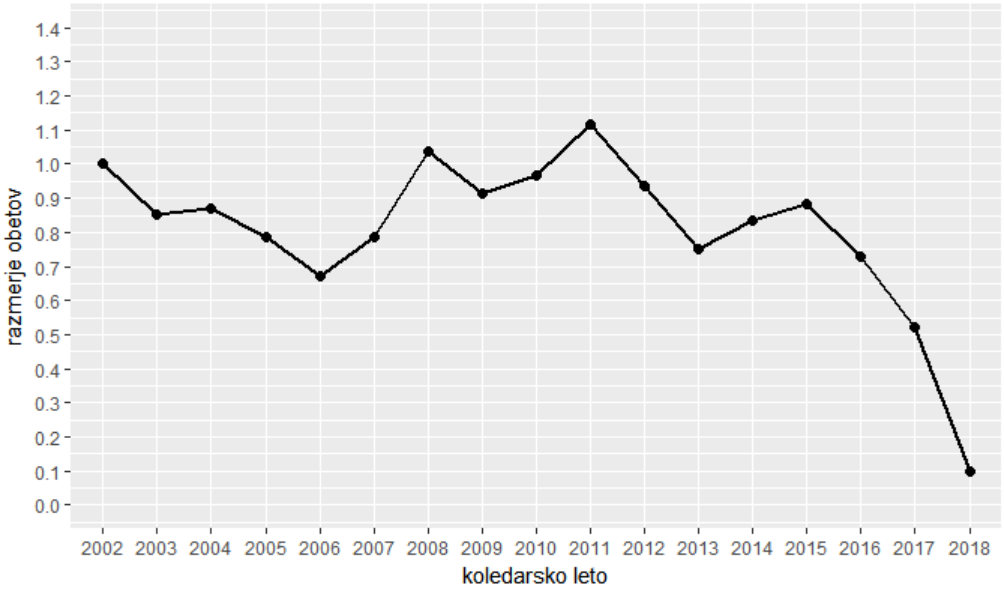
Iz grafa razmerja obetov v odvisnosti od koledarskega leta lahko razberemo, da so obeti do leta 2006 padali, nato vzpenjali in dosegli vrhova v letih 2008 in 2011. Možno je, da sta ta vrhova posledica gospodarske krize. Graf potem z leti spet pada.

Pri zavarovalni dobi se opazi, da ima največji obet prekinitve pogodba s petletnim trajanjem. Pojasnilo lahko iščemo v tem, da so življenjska zavarovanja s tako kratko

Slika 1. Graf razmerja obetov v odvisnosti od leta zavarovanja



Slika 2. Graf razmerja obetov v odvisnosti od koledarskega leta

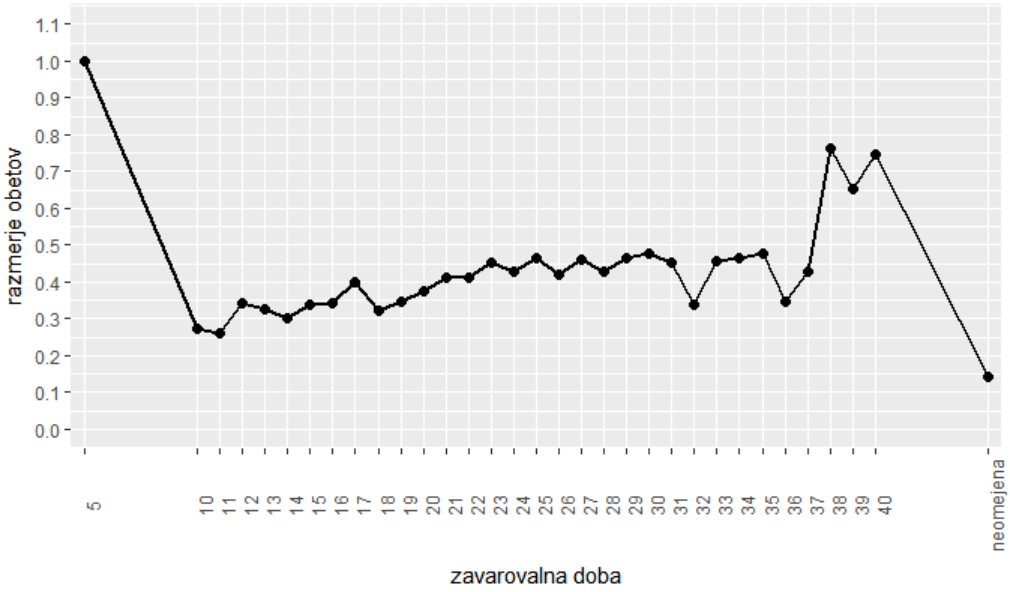


zavarovalno dobo prodana redkeje in niso ponujena stranki, saj so vsa zavarovanja z dobo manj kot 10 let obdavčena. Portfelj 5-letnih življenjskih zavarovanj je lahko zato popačen. Naprej pa lahko vidimo, da obeti naraščajo od 10 let pa do 40 let. Najmanjše obete za prekinitev pa imajo zavarovanja z neomejenim trajanjem, torej vseživljenjska zavarovanja.

Pri grafu starosti opazimo visok vrh pri kategoriji od 15 - 19 let. Čeprav so kot zavarovalci navedene mladoletne osebe, najverjetneje premijo plačujejo starši. Po vrhu obeti padajo in se ustalijo, manjši spust je samo pri letih od 60 do 69 let.

Zdi se mi, da pri nobeni od pojasnjevalnih spremenljivk, ki so zgoraj prikazane z razmerjem obetov, ne moremo zaznati direktnega trenda, zato sem jih pustil v obliki kategoričnih spremenljivk.

Slika 3. Graf razmerja obetov v odvisnosti od zavarovalne dobe



Slika 4. Graf razmerja obetov v odvisnosti od starosti zavarovalca

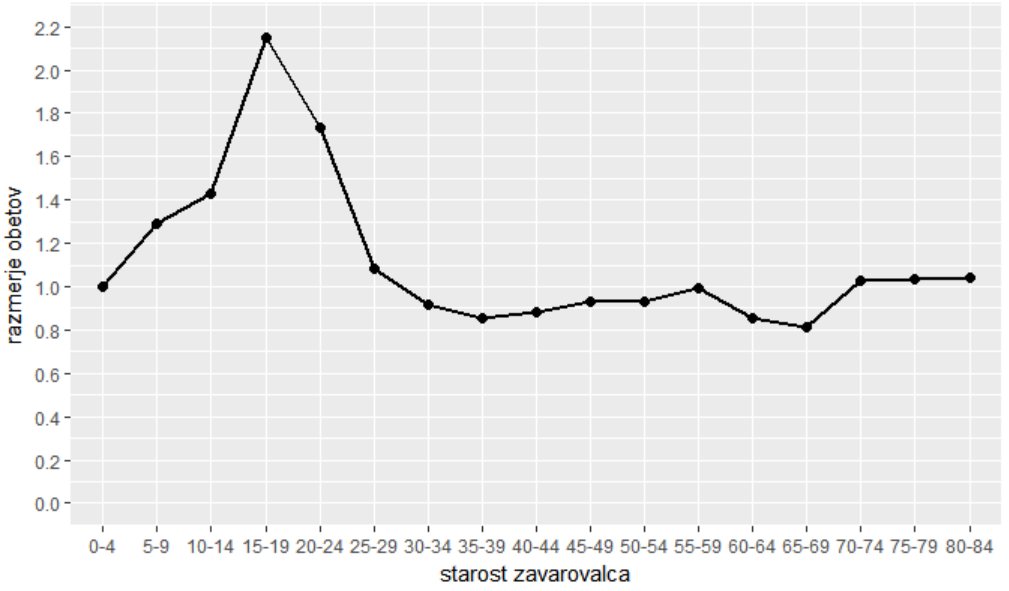


Tabela 3. Razmerja obetov frekvenc plačevanja.

Kategorija	mesečna	četrtletna	polletna	letna	enkratna
Razmerje obetov	1	1,339	1,082	1,547	0,0004

Pri spremenljivki *frekvenca plačevanja* se vidi, da imajo daleč najmanjši obet prekinitve police z enkratnim plačilom premije. Enkratna plačila se po navadi pojavljajo pri naložbenih zavarovanjih, zato bi bilo nesmiselno prekiniti pogodbo, ker bi stranka izgubila veliko denarja. Presenetili pa so me obeti pri četrtletnem, polletnem in letnem plačevanju, saj sem pričakoval nižje obete prekinitve kot pri mesečni frekvenci plačevanja.

Obet prekinitve zavarovanja pri riziko *tipu zavarovanja* je 2,29-krat višji kot pri naložbenem zavarovanju. To se zdi smiselno, saj so prekinitve pri naložbenem zavarovanju možne le v prvih nekaj letih, tudi kazen za prekinitev zavarovanja je nedvomno večja pri naložbenih zavarovanjih, saj poleg zavarovalne komponente propade tudi varčevalna komponenta zavarovanja.

Pri pojasnjevalni spremenljivki *spol zavarovalca* je obet prekinitve pri moških višji za 6,8 % od obeta pri ženskah. Pri pravnih osebah pa je obet kar 54,5 % višji.

Tabela 4. Razmerja obetov okrajev zavarovalca.

Okraj	1	2	3	4	5	6	8	9	Tujina
Razmerje obetov	1,0	1,609	1,577	1,001	0,96	1,146	1,372	1,709	1,866

V tabeli razmerja obetov okrajev zavarovalca lahko pridemo do zanimivega zaključka, da se obeti prekinitev zelo razlikujejo v vzhodni in zahodni Sloveniji. Poštne številke od s prvo številko 1, 4, 5, 6 (območja okrog Ljubljane, Kranja, Nove Gorice, Kopra) imajo podobne, nižje obete, poštne številke s prvo številko 2, 3, 8, 9 (območja okrog Maribora, Celja, Novega mesta, Murske Sobote) pa imajo veliko večje obete prekinitve. Največji obet pa imajo zavarovalci s prebivališčem izven Slovenije.

Iz spremenljivke *mestni* je razvidno, da je obet prekinitve pri zavarovalcih iz mestnih okolij višji za 7,4 % kot pri zavarovalcih iz ostalih delov Slovenije.

Tabela 5. Razmerja obetov poslovnih enot.

Poslovna enota	0	1	2	3	4	5	6	7	8	9
Razmerje obetov	1,95	1,0	0,84	1,18	0,95	1,24	0,96	1,02	0,82	1,12

Iz tabele 5 je razvidno, da imajo veliko večje obete prekinitve tiste police, ki so jih sklenili zastopniki iz agencij (poslovna enota 0), mreža lastnih zastopnikov pa ima manjše obete, se pa razlikujejo od poslovne enote do poslovne enote. Agencije so morda zdaj ali v preteklosti spodbujale zavarovalce k prekinitvi zavarovanja in jim ponudili novega. Lahko produkt druge zavarovalnice (neekskluzivne agencije) ali pa nov produkt iste. S tem bi morda dobili novo provizijo za sklenitev življenjskega zavarovanja.

Pri *spolu zastopnika* imajo višji obet ženske zastopnice. Ta je za 4 % višji od obeta moških zastopnikov.

Tabela 6. Razmerja obetov prodajnih poti.

Prodajna pot	A1	A2	A3	dir. trženje	ost.
Razmerje obetov	0,23	1,00	0,40	0,38	0,93
Prodajna pot	spec. zastop.	spec. agenc.	uni. zastop.	uni. agenc.	
Razmerje obetov	1,02	0,84	1,00	1,06	

V razmerjih obetov pri prodajnih poteh opazimo, da imata agenciji A1 in A2 in direktno trženje veliko manjši obet prekinitve kot preostale prodajne poti. Največje obete pa ima jo specialni zstopniki in univerzalne agencije. Hkrati je spremenljivka *prodajna pot* korelirana s spremenljivko *poslovna enota*, saj so vse agencije vezane na poslovno enoto 0. Zato so obeti posameznih agencij pri tej spremenljivki toliko nižji.

8.4.2. *Kapitalizacije.* Kapitalizacija življenjskega zavarovanja je možna pri naložbenih življenjskih zavarovanjih in ne pri riziko zavarovanjih. Prav tako se ne more kapitalizirati zavarovanje, ki je imelo le enkratno premijo.

Iz podatkov sem zato izločil vsa riziko zavarovanja in vsa zavarovanja z enkratnim plačilom zavarovalne premije. Izločil sem tudi prekinjene police in odkupljene police. Za ta model tudi ni pomembna spremenljivka *možnost odkupa police*, zato sem jo odstranil. Prav tako je nepomembna spremenljivka *tip zavarovanja*, ker so v tem primeru samo naložbena zavarovanja.

Tako kot pri modelu prekinitev sem tudi tukaj zgradil model brez interakcij, model sem zgradil s funkcijo `glm()`, za povezovalno funkcijo sem izbral logit, proučevana spremenljivka pa je bila *aktivnost*.

Tabela 7. Tabela devianc modela kapitalizacij.

Spremenljivka	df	D	Δ df	Δ D	$p - vrednost$
-	543007	94193,32	-	-	-
leto zavarovanja	542991	85652,98	16	8540,337	0,000000
frekvenca placevanja	542988	85649,01	3	3,977270	$2,639290 \cdot 10^{-1}$
zavarovalna doba	542957	84572,48	31	1076,524	$2,099440 \cdot 10^{-206}$
spol zavarovalca	542955	84547,39	2	25,09220	$3,558754 \cdot 10^{-06}$
starost zavarovalca	542939	84460,12	16	87,26465	$7,957328 \cdot 10^{-12}$
okraj zavarovalca	542931	84330,02	8	130,1077	$2,687364 \cdot 10^{-24}$
mestni	542930	84330,01	1	0,01049812	$0,9183913$
poslovna enota	542921	84155,95	9	174,0550	$8,814593 \cdot 10^{-33}$
spol zastopnika	542920	84127,89	1	28,06575	$1,172627 \cdot 10^{-07}$
prodajna pot	542913	83946,92	7	180,9672	$1,217158 \cdot 10^{-35}$
koledarsko leto	542897	80582,97	16	3363,946	0,000000

V modelu za analizo in napoved kapitalizacij zavarovanj je delež preostale devianc malo manjši kot pri modelu prekinitev. Znaša 85,5 % v primerjavi z 88,5 % pri prekinitvah. Tudi tu se mi zdi, da je velika devianca modela posledica velike variabilnosti podatkov in mogoče manjkajočih pojasnjevalnih spremenljivk. Iz tabele devianc lahko razberemo tudi, da imata spremenljivki *frekvenca plačevanja* in *mestni* p-vrednosti 0,26 in 0,92. Če vzamemo stopnjo zaupanja 5 % potem zavrnemo hipotezo, da ti dve spremenljivki nastopata v modelu kapitalizacij. Zato sem ju izločil. Največ deviance pojasni spremenljivka *leto zavarovanja*, sledi je *koledarsko leto* in še *zavarovalna doba*.

Tudi tu so vplivi pojasnjevalnih spremenljivk podani z razmerji obetov.

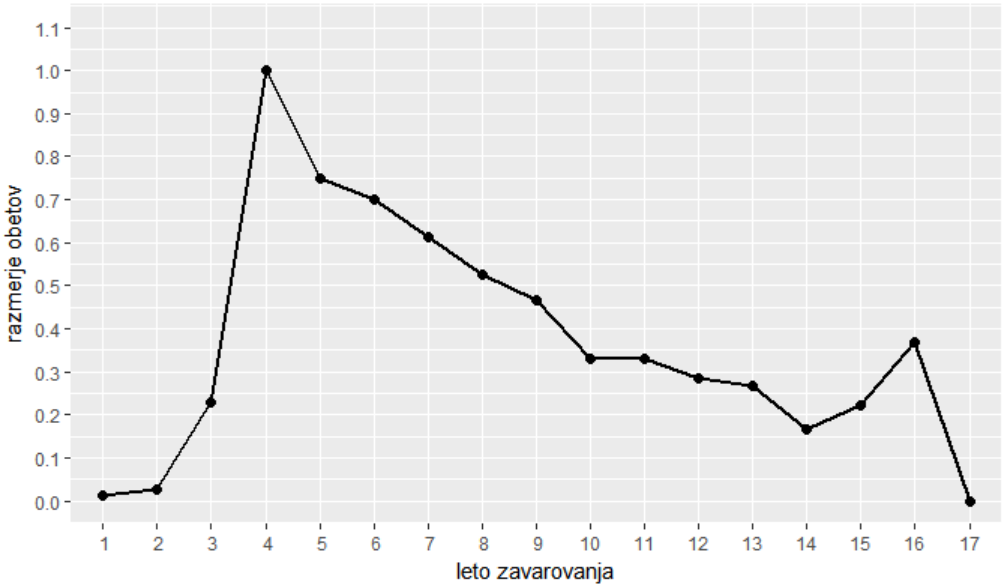
Iz grafa razmerja obetov v odvisnosti od leta zavarovanja lahko opazimo visok skok v četrtem letu zavarovanja, potem sledi počasen spust obetov. Četrto leto je najbolj izrazito, ker je v veliko primerih kapitalizacija možna šele po treh letih, zavarovalci pa to možnost hitro izkoristijo.

Obet kapitalizacije zavarovanja je bil do leta 2009 relativno nizek, potem pa se je močno povečal in dosegel vrh v letu 2013. Zdi se mi, da sta opažen efekt povzročila gospodarska kriza in večja brezposelnost v Sloveniji.

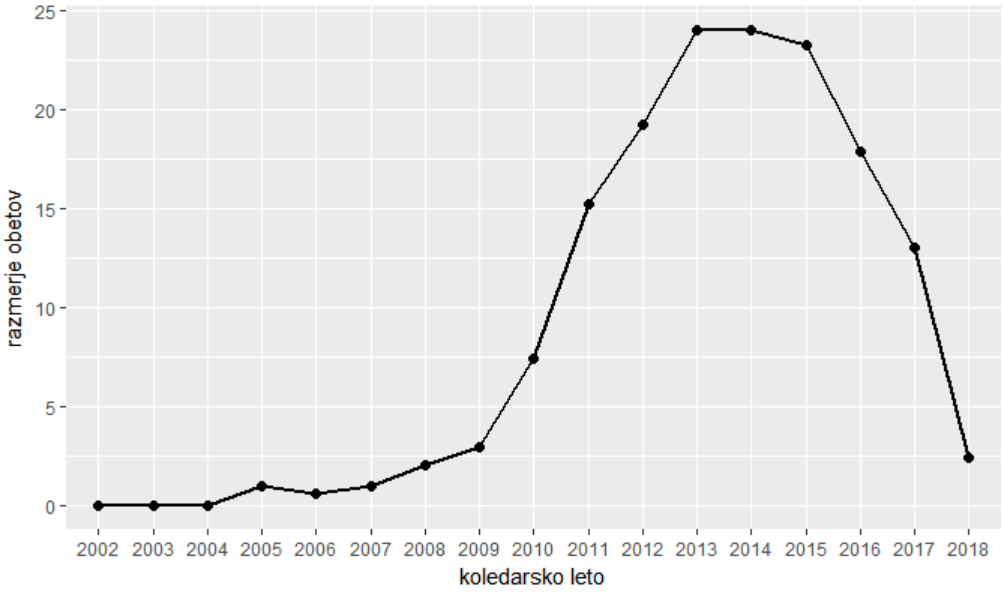
Razmerje obetov pri zavarovalni dobi raste z leti. V grafu lahko opazimo nekaj nihanja, sploh pri dobi od 31 do 40 let. Razlog je majhno število podatkov za te zavarovalne dobe, saj se večina zavarovanj sklene z zavarovalno dobo do 30 let.

V grafu spremenljivke starost opazimo rahlo vzpenjanje obetov kapitalizacije zavarovanja do 59 leta, potem pa se graf strmo obrne navzdol. Močno izstopa kategorija

Slika 5. Graf razmerja obetov v odvisnosti od leta zavarovanja



Slika 6. Graf razmerja obetov v odvisnosti od koledarskega leta



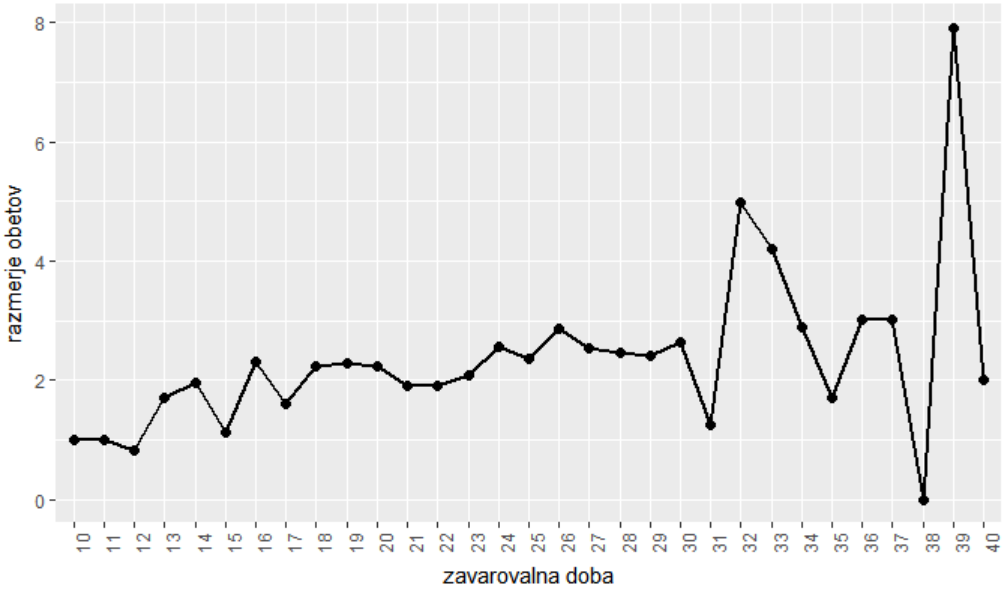
od 75 do 79 let, vendar se mi zdi, da je to posledica manjše baze podatkov zavarovanj z zavarovalci te starostne kategorije.

Pri kapitalizacijah ima ženska zavarovalka spet višji obet od moškega zavarovalca, obet je višji za 8,7 %. Še višji je pri pravnih osebah, pri katerih je obet kapitalizacije višji za 51 % od moških zavarovalcev.

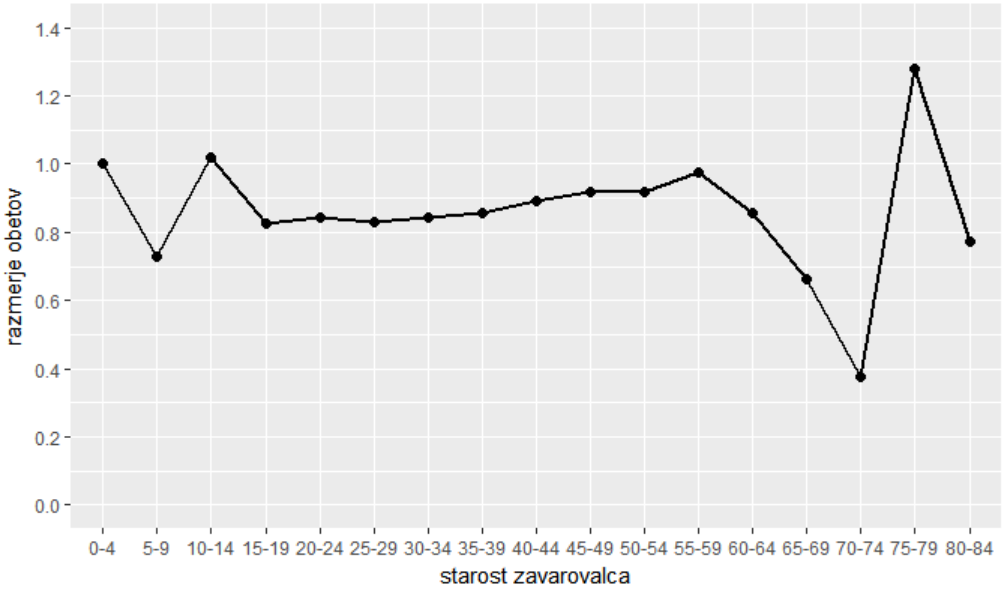
Tabela 8. Razmerja obetov okrajev zavarovalca.

Okraj	1	2	3	4	5	6	8	9	Tujina
Razmerje obetov	1,0	1,13	1,19	1,07	1,62	1,13	1,13	0,99	1,24

Slika 7. Graf razmerja obetov v odvisnosti od zavarovalne dobe



Slika 8. Graf razmerja obetov v odvisnosti od starosti zavarovalca



Pri kapitalizacijah ni enake očitne razmejitve zavarovalcev na vhod in zahod. Najvišji obet imajo v okolici Nove Gorice, najmanjšega pa v Prekmurju in v Ljubljanski regiji.

Tabela 9. Razmerja obetov poslovnih enot.

Poslovna enota	0	1	2	3	4	5	6	7	8	9
Razmerje obetov	2,29	1,0	1,57	2,12	1,75	2,24	1,87	2,12	1,54	2,35

Iz razmerij obetov poslovnih enot lahko opazimo, da ima najmanjše obete za kapitalizacijo zavarovanja poslovna enota 1, največje pa poslovna enota 9 in agencije

(poslovna enota 0). Vidno je veliko nihanje obetov med posameznimi poslovnimi enotami.

Tudi pri kapitalizacijah imajo zastopnice višji obet od svojih moških kolegov, ampak je razlika zelo majhna. Ženske imajo za 0,26 % višje obete od moških.

Tabela 10. Razmerja obetov prodajnih poti.

Prodajna pot	A1	A2	A3	ostalo
Razmerje obetov	0,94	1,12	5,16	1,18
Prodajna pot	spec. zastop.	spec. agenc.	uni. zastop.	uni. agenc.
Razmerje obetov	1,06	1,23	1,0	1,11

Pri pregledu prodajnih poti najbolj izstopa agencija A3, ki ima kar 5,16-krat višje obete za kapitalizacijo zavarovanja kot univerzalni zastopniki. Agencija A1 pa ima najnižje obete.

8.4.3. *Odkupi.* Odkup police življenjskega zavarovanja je prekinitev zavarovanja na prošnjo stranke, zavarovalnica pa zavarovalcu izplača varčevalni del premije. Odkup je možen samo pri zavarovanjih z varčevalno komponento, zato sem pri gradnji modela iz podatkovne množice izbral samo naložbena življenjska zavarovanja. Odstranil sem prekinjene in kapitalizirane police. V množico sem zajel samo tiste police, ki so imele v tistem letu zavarovanja možen odkup. S funkcijo `glm()` sem naredil posplošen linearni model in s funkcijo `anova()` dobil tabelo devianc.

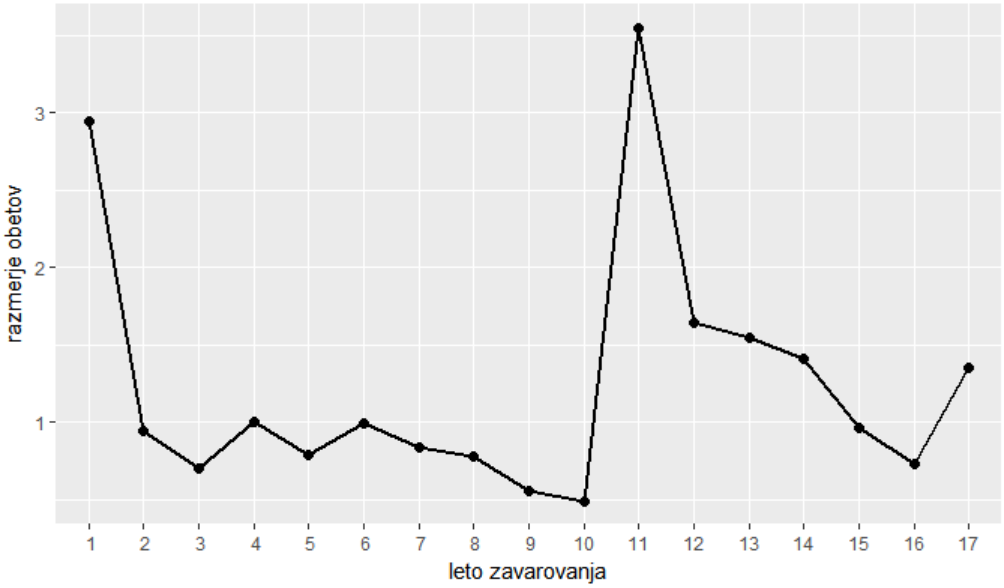
Tabela 11. Tabela devianc modela odkupov.

Spremenljivka	df	D	Δ df	Δ D	$p - vrednost$
-	651088	301620,0	-	-	-
leto zavarovanja	651072	267313,4	16	34306,60846	0,00000
frekvenca placevanja	651068	265937,5	4	1375,83407	$1,20114 \cdot 10^{-296}$
zavarovalna doba	651036	263146,8	32	2790,72097	0,00000
spol zavarovalca	651033	260447,5	2	92,61563	$7,74054 \cdot 10^{-21}$
starost zavarovalca	651017	259689,8	16	757,73677	$6,52554 \cdot 10^{-151}$
okraj zavarovalca	651009	258193,2	8	1496,58673	$7,34729 \cdot 10^{-318}$
mestni	651008	258191,8	1	1,37615	0,2407581
poslovna enota	650999	258046,5	9	145,37109	$8,00972 \cdot 10^{-27}$
spol zastopnika	650998	257944,4	1	102,09496	$5,29231 \cdot 10^{-24}$
prodajna pot	650991	257513,7	7	430,70841	$6,15476 \cdot 10^{-89}$
koledarsko leto	650975	250842,1	16	6671,57138	0,00000

S pojasnjevalnimi spremenljivkami, zaobjetimi v model, je bilo pojasnjeno 17 % deviance ničtega modela. To je več tako od modela prekinitev kot tudi od modela kapitalizacij. Iz tabele je razvidno, da spremenljivka *mestni* nima dovolj majhne p-vrednosti, zato sem jo iz modela odstranil. Vse ostale spremenljivke zadoščajo χ^2 testu s stopnjo zaupanja 5 %. Daleč največja razlika devianc je pri spremenljivki *leto zavarovanja*, sledita ji *koledarsko leto* in *zavarovalna doba*. V nadaljevanju so predstavljeni vplivi pojasnjevalnih spremenljivk z razmerji obetov.

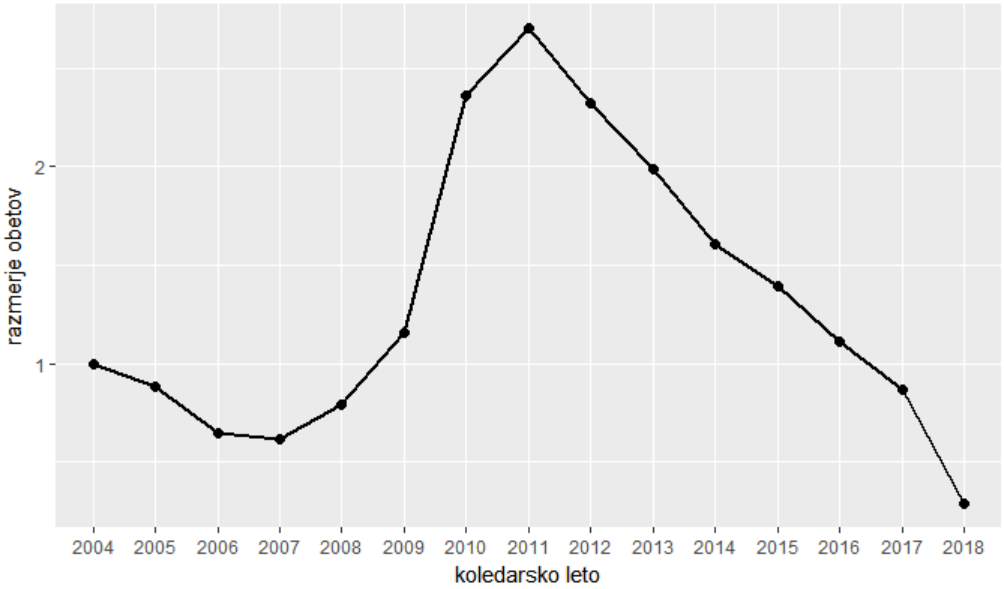
Iz grafa razmerja obetov v odvisnosti od leta zavarovanja se opazita dva vrhova. Pojavita se pri prvem letu zavarovanja in enajstem letu zavarovanja. Slednjega je lažje pojasniti - po desetih letih je življenjsko zavarovanje neobdavčeno, prav tako

Slika 9. Graf razmerja obetov v odvisnosti od leta zavarovanja



kateri koli prihodek iz tega naslova, zato se veliko število zavarovancev odloči za odkup police po preteku te dobe. Večje obete v prvem letu zavarovanja pa bi lahko pojasnili s specifiko zavarovalnice. Ta je za zavarovanja v prejšnjem informacijskem sistemu ponujala odkupe in takojšnje plačilo enkratne premije v nov zavarovalni produkt. Več teh zavarovalcev pa se je potem tudi hitro odločilo za odkup novega zavarovanja.

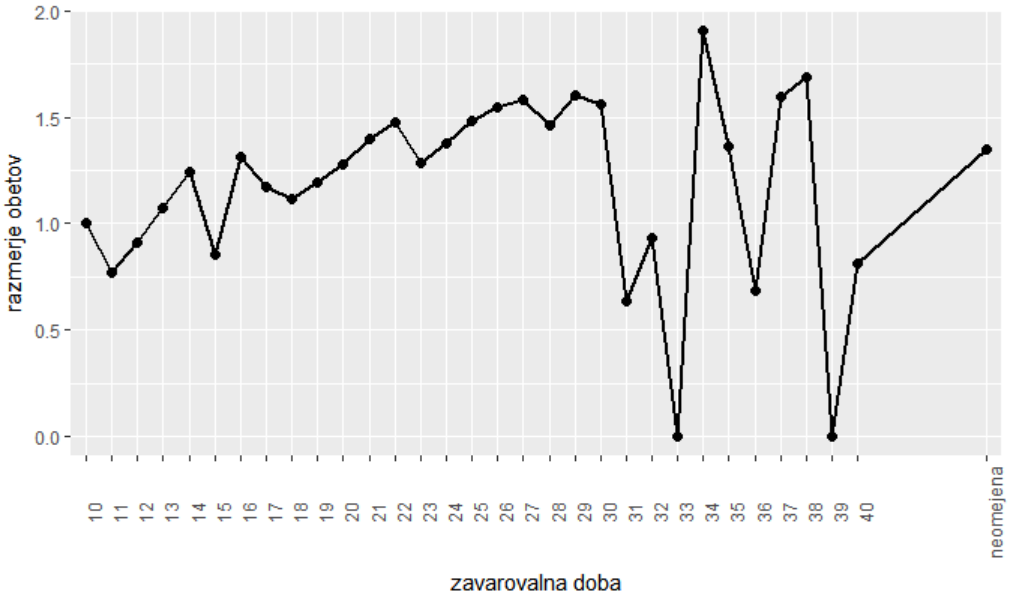
Slika 10. Graf razmerja obetov v odvisnosti od koledarskega leta



Iz razmerja verjetij pojasnjevalne spremenljivke *koledarko leto* se vidi visok vrh v letu 2011. Vse od leta 2010 do 2013 so obeti veliko višji kot v ostalih letih. Opazi se velik vpliv ekonomske krize in pomanjkanje finančnih sredstev zavarovancev, posledica odkupov je lahko strah še pred večjo izgubo sredstev v varčevalnem delu police

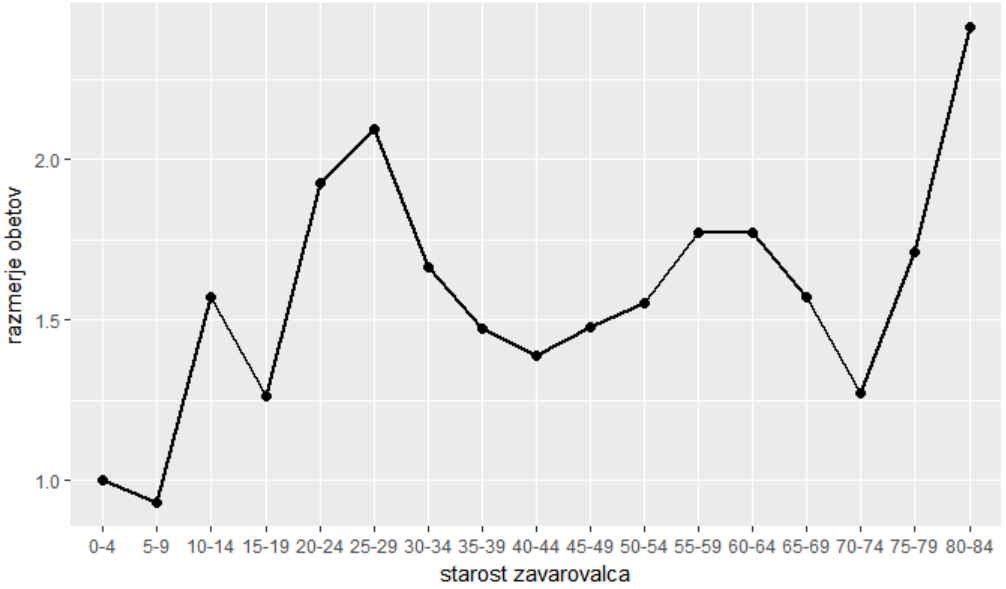
življenjskega zavarovanja, mogoče je v odkupe privedlo tudi nezaupanje prebivalstva v finančne institucije.

Slika 11. Graf razmerja obetov v odvisnosti od zavarovalne dobe



Obeti odkupov se povečujejo z zavarovalno dobo do zavarovanj s 30-letnim trajanjem - to je razvidno iz grafa obetov v odvisnosti od zavarovalne dobe. Po tej dobi začne graf močno nihati, spet razlog tiči v premajhnem vzorcu zavarovanj s tako dobo.

Slika 12. Graf razmerja obetov v odvisnosti od starosti zavarovalca



Obeti odkupov zavarovalcev v odvisnosti od starosti zavarovalca imajo vrh v kategoriji od 25 do 29 let, potem se spuščajo do kategorije 40 - 44 let, nato pa spet naraščajo do kategorije 55 - 59 let. Vse skupaj se še ponovi s padcem do 70 - 74 let in spet dvig do kategorije 80 - 84 let.

Ženske zavarovalke imajo za 1,7 % večje obete za odkup police od moških, pravne osebe pa za kar 146 % večje obete.

Tabela 12. Razmerja obetov okrajev zavarovalca.

Okraj	1	2	3	4	5	6	8	9	Tujina
Razmerje obetov	1,0	1,48	1,37	1,05	0,99	0,86	1,39	1,65	1,72

Pri obetih za odkupe spet opazimo veliko razdeljenost med vzhodno in zahodno Slovenijo. Z višjimi obeti izstopajo regije okrog Maribora, Celja, Novega mesta in Murske Sobote. Najbolj pa izstopajo zavarovalci iz tujine, tako kot pri prekinitvah in kapitalizacijah.

Tabela 13. Razmerja obetov poslovnih enot.

Poslovna enota	0	1	2	3	4	5	6	7	8	9
Razmerje obetov	1,0	1,33	0,82	0,84	1,0	1,31	1,05	0,88	0,99	0,99

Največje obete odkupa imajo zavarovanja, sklenjena v poslovnih enotah 1 in 5, ki močno izstopata iz povprečja. Najmanjši obeti pa so v poslovnih enotah 2 in 3.

Tako kot pri kapitalizacijah in prekinitvah imajo večji obet odkupov zavarovanja, sklenjena pri ženskih zastopnicah; obeti so za 7,6 % višji.

Tabela 14. Razmerja obetov prodajnih poti.

Prodajna pot	A1	A2	A3	ostalo
Razmerje obetov	0,72	1,13	1,15	0,93
Prodajna pot	spec. zastop.	spec. agenc.	uni. zastop.	uni. agenc.
Razmerje obetov	1,07	1,04	1,0	1,04

Iz podatkov o razmerjih obetov prodajnih poti izstopa agencija A1 z najmanjšimi obeti za odkup, največje obete pa imata agenciji A2 in A3. Zastopniki iz mreže zavarovalnice imajo podobne obete.

9. Zaključek

V delu sem obravnaval posplošene linearne modele in njihovo uporabo pri analiziranju prekinitev, kapitalizacij in odkupov polic življenjskega zavarovanja. Posplošeni linearni modeli imajo večjo robustnost kot linearna regresija, omogočajo modeliranje spremenljivk z različnimi porazdelitvami iz eksponentne družine. V primeru, ki sem ga obravnaval, se je izkazalo, da se modeli niso najbolje prilegali podatkom, bolj kot za napovedovanje prihodnjih vrednosti so uporabni za analiziranje vpliva posameznih pojasnjevalnih spremenljivk. Ti modeli bi bili lahko nadgrajeni z različnimi interakcijami posameznih spremenljivk in dodatnimi, novimi pojasnjevalnimi spremenljivkami. Vseeno pa omogočajo globlji vpogled v obnašanje zavarovalcev in zavarovalnih zastopnikov dotične zavarovalnice. Posplošene linearne modele v mojem primeru bi lahko nadgradil tudi z drugimi metodami strojnega učenja in drugimi modeli, ki se vedno bolj uporabljajo v praksi, vseeno večina izmed njih nima takega vpogleda v vpliv posamezne spremenljivke, kar nam izvrstno omogočajo posplošeni linearni modeli.

Slovar strokovnih izrazov

ANCOVA analiza kovariance
ANOVA analiza variance
covariate številska spremenljivka
deviance devianca
exponential family eksponentna družina
factor kategorična spremenljivka
generalized linear models posplošeni linearni modeli
goodnes-of-fit prileganje
interaction interakcija
lapsed insurance prekinjeno zavarovanje
life insurance življenjsko zavarovanje
link function povezovalna funkcija
logistic regression logistična regresija
minimally sufficient statistics minimalna zadostna statistika
null model ničti model
paid up insurance kapitalizirano zavarovanje
parsimony skopost, enostavnost
saturated model poln model
surrendered insurance odkupljeno zavarovanje

Literatura

[1] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher in N. Thandi, *A Practitioner's Guide to Generalized Linear Models*, verzija 1. 2. 2007, [10.7.2018]; dostopno na `https://www.towerswatson.com/DownloadMedia.aspx?media={E7F1DAFE-D085-4169-81CE-C22ED018FBA3}`.

[2] P. J. Boland, *Statistical and Probabilistic Methods in Actuarial Science*, Interdisciplinary Statistics , CRC Press, Boca Raton, 2007, 221–264.

[3] R. R. Cerchiara, M. Edwards in A. Gambini, *Generalized Linear Models in Life Insurance: Decrements and Risk Factor Analysis under Solvency II*, Working paper presented at the AFIR Colloquium in Rome, 2008; dostopno na `http://www.actuaries.org/AFIR/Colloquia/Rome2/Cerchiara_Edwards_Gambini.pdf`

[4] FSA, *Survey of the persistency of life and pension policies*, Financial Services Authority, London, 2007.

[5] T. D. Little, *The Oxford Handbook of Quantitative Methods*, Volume 2: Statistical Analysis , Oxford University Press, New York, 2013, 26–51.

[6] G. Rodriguez, *Lecture Notes on Generalized Models*, verzija 12. 9. 2007, [10.7.2018], dostopno na `http://data.princeton.edu/wws509/notes/`.

[7] Slovensko zavarovalno združenje, *Življenjsko zavarovanje*, verzija 1. 6. 2010, [ogled 18. 7. 2018], dostopno na `https://www.zav-zdruzenje.si/wp-content/uploads/2017/11/%C5%BDivljenjsko-zavarovanje-bru%C5%A1ura.pdf`.