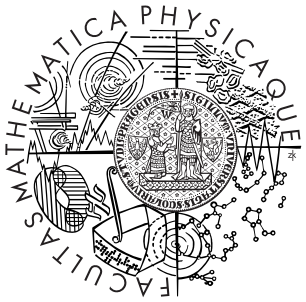
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bc. Lýdia Godušová

Statistické modely pro kapitálové modely pojišťoven – studium storen v životním pojištění

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Miroslav Šimurda, Ph.D

Studijní program: Matematika
Studijní obor: Finanční a pojistná matematika

Praha 2011

Rada by som sa poďakovala vedúcemu diplomovej práce, RNDr. Miroslavovi Šimurdovi, Ph.D, za čas strávený pri konzultáciách. Veľké poďakovanie patrí Michalovi Švagerkovi za pomoc pri používaní softvéru, podporu a motiváciu k dokončeniu práce. Ďakujem aj Veronike Janákovej, Kláre Karasovej, Michalovi Keselymu a Matejovi Vitáskovi za jazykovú korektúru.

Prehlasujem, že som túto diplomovú prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, hlavne skutočnosť, že Univerzita Karlova v Prahe má právo na uzatvorenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe dňa 12.4.2011

Obsah

Názov práce: Statistické modely pro kapitálové modely pojišťoven – studium storen v životním pojištění
Autor: Bc. Lýdia Godušová
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedúci diplomovej práce: RNDr. Miroslav Šimurda, Ph.D

Abstrakt: V tejto práci sa zaoberáme modelovaním miery storna v životnom poistení. K tomu predkladáme teoretický základ *lineárnych regresných modelov* a ich rozšírenia, *zovšeobecnených lineárnych modelov*. V teoretickej časti taktiež popisujeme výber modelu a spôsob jeho testovania. V druhej časti práce popisujeme závislosť miery storna na individuálnych a makroekonomických parametroch, tak ako boli skúmané vo svete. V poslednej časti aplikujeme teoretické znalosti zovšeobecnených lineárnych modelov. Dáta analyzujeme v štatistickom programe `R` a vysvetľujeme proces hľadania modelu, ktorý ich najlepšie popisuje. Interpretujeme výstupy z `R` a odhady získané z výsledného modelu porovnávame s pomerovou analýzou dát.

Kľúčové slová: zovšeobecnené lineárne modely, deviácie, reziduá, vierohodnosť, miera storna.

Title: Statistical models for insurance capital models - the study of lapse in the Life Insurance
Author: Bc. Lýdia Godušová
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Miroslav Šimurda, Ph.D

Abstract: This work deals with the topic of lapse rate modelling in the field of Life Insurance. First, the theoretical apparatus is established: the *linear models* and their extension, *generalized linear models*. Furthermore, we describe the process of model selection and evaluation. In the second part of this work we describe the influence of various individual as well as macroeconomical parameters on the lapse rate. We summarize the findings of previous works in this field. The last part introduces models in statistical software `R` based on generalized linear models and describes the process of their selection and evaluation. Outputs from these models are interpreted and compared to the ratio analysis results.

Keywords: generalized linear models, deviance, residuals, likelihood, lapse rate.

Úvod

Diplomová práca sa zaoberá modelovaním storna v životnom poistení. Poisťovne uzatvárajú poistné zmluvy väčšinou na dlhšiu dobu a počítajú s príjmom zjednaného poistného. Poistník však má možnosť poistnú zmluvu stornovať a poistiteľ, najčastejšie poisťovňa, tak stratí predpokladaný príjem poistného v budúcnosti. Miera storna je pre poisťovňu dôležitá z hľadiska likvidity a ziskovosti. Ovplyvňuje hodnotu portfólia celej poisťovne. Masívne predčasné ukončenie poistiek predstavuje pre poistiteľa hrozbu. Tá je potom vystavená riziku úrokovej sadzby. Názory na storno v Českej republike sú však rôzne, od snahy zachytiť storno pomocou štandardných štatistických modelov až po názor, že storno nie je štrukturalizované riziko. Základom predloženej diplomovej práce je preskúmanie a zmapovanie súčasnej situácie v oblasti stornovania životných poistných zmlúv vo svete, kde sa storná modelujú, a prenesenie poznatkov do Českej republiky, kde toto riziko nie je veľmi sledované a skúmané.

V prvej časti práce je popísaný teoretický základ zovšeobecnených lineárnych modelov *(generalized linear model)*, ktoré sa používajú pri modelovaní storien v životnom poistení vo svete. Výhoda týchto modelov je v tom, že berú do úvahy nielen jednotlivé parametre, ale aj ich korelácie a interakcie. Zovšeobecnené lineárne modely sú popísané s väčším dôrazom na binomické a Poissonovo rozdelenie, ktoré sa používajú pri aplikácii týchto modelov v oblasti životného poistenia.

Miera storna závisí na rôznych ukazovateľoch, či už individuálnych alebo makroekonomických. Druhá časť práce je z praktického hľadiska jedna z najdôležitejších a je zameraná na rešerš vedeckých článkov od autorov, ktorí tieto modely používajú na modelovanie storien v životnom poistení vo svete. V tejto časti práce porovnávame závislosť miery storna na individuálnych parametrov (vek poistníka pri uzatváraní poistnej zmluvy, pohlavie, rodinný stav poisteného, povolanie poisteného, trvanie poistnej zmluvy, výška poistného, frekvencia platenia poistného, typ poistnej zmluvy) a makroekonomických parametrov (miera nezamestnanosti, miera inflácie, miera hospodárskeho rastu, sezónny efekt, finančná kríza). Taktiež hľadáme príčiny závislosti miery storna na vybraných parametroch.

Cieľom práce je aj aplikácia získaných teoretických znalostí a hľadanie modelu, ktorý najlepšie vysvetľuje dáta. Zovšeobecnené lineárne modely sú aplikované na dáta v tretej časti práce. V poslednej kapitole porovnávame, ako sa mení pravdepodobnosť storna v závislosti na individuálnych parametroch (vstupný vek, pohlavie, rodinný stav poisteného, deti, mesačný príjem poistníka, typ poistnej zmluvy, poistná čiastka, distribúcia, rok zjednania poistnej zmluvy a veľkosť sídla, v ktorom poistenec býva). V Českej republike je pre poisťovňu najväčším rizikom stornovanie poistnej zmluvy v prvých dvoch rokoch poistenia[1]. Z toho dôvodu modelujeme stornovanie poistnej zmluvy v prvých dvoch rokoch trvania poistenia. Model je naprogramovaný v štatistickom programe `R`.

[1]Pokiaľ klient stornuje poistnú zmluvu v prvých dvoch rokoch od uzavretia poistnej zmluvy, poisťovňa klientovi nevyplatí žiadnu čiastku, pretože ešte nie sú splatené poplatky spojené s uzavrením zmluvy. Po dvoch rokoch zaplateného poistného sa poistenému pri výpovedi poistnej zmluvy vypláca tzv. odkupné.

Kapitola 1

Storno v životnom poistení

V životnom poistení pod stornom *(lapse)* rozumieme ukončenie poistnej zmluvy, kedy krytie poistnej zmluvy zaniká. Storno poistnej zmluvy nastáva v prípade, že poistník zanedbá povinnosť platiť poistné alebo podá výpoveď poistnej zmluvy. Pre poistiteľa je dôležité vedieť, aká je pravdepodobnosť takého ukončenia platenia poistného, pretože každý poistiteľ počíta s príjmom poistného počas celej dĺžky trvania poistnej zmluvy.

1.1 Dôvody stornovania poistnej zmluvy

Hlavné hypotézy vysvetľujúce racionálne poistníkove rozhodnutie predčasne ukončiť poistnú zmluvu sú napríklad tieto:

- Poistenci používajú hotovosť odkupného ako fond na mimoriadne udalosti, keď čelia finančnej katastrofe alebo v období straty zamestnania a nutnej potreby hotovosti.
- Poistenci používajú poistenie len ako formu sporenia, ktorí stornujú poistnú zmluvu v prípade získania lepších podmienok sporenia.
- Poistenci kontrolujú trh u konkurencie. Získanie vyššej poistnej čiastky za rovnakú výšku poistného, alebo naopak nižšieho poistného pri rovnakej poistnej čiastke u iného poistiteľa často vedie k stornovaniu poistnej zmluvy.

Je ťažké modelovať racionálne správanie poistníka a jeho rozhodnutie ukončiť poistnú zmluvu. Niektorí poistníci sú racionálnejší a pri príležitosti získania výhodnejšej poistnej zmluvy za lepšie podmienky túto možnosť využijú. Avšak nie každý poistenec sa správa úplne racionálne. Môžeme však modelovať vplyv makroekonomických ukazovateľov, ako je napríklad miera nezamestnanosti či úrokova miera, na stornovanie poistnej zmluvy.

Kapitola 2

Teoretický základ modelov

V tejto kapitole najskôr teoreticky vysvetlíme *lineárne modely (Linear Models)* a následne ich rozšírenie na *zovšeobecnené lineárne modely (Generalized Linear Models)*.

Než sa pustíme do teórie, dohodneme sa na terminológii hlavných premenných. Meranú veličinu y nazývame vysvetľovaná premenná, závislá premenná, regresand. Veličiny $x_1, \ldots, x_k$ nazývame vysvetľujúce premenné, nezávislé premenné, regresory a prediktory.

2.1 Lineárny regresný model

Lineárne regresné modely vyjadrujú vzťahy medzi meranou veličinou $\mathbf{y} = (y_1, \ldots, y_n)^T$ a vysvetľujúcimi veličinami $\mathbf{X}$, ktoré sa považujú za pevné, prostredníctvom lineárnej funkcie.

Definícia 2.1.1. Lineárny regresný model sa v maticovom zápise definuje predpisom

$$\mathbf{y} = \mathrm{E}\,(\mathbf{y}) + \varepsilon; \qquad \mathrm{E}\,(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\beta. \tag{2.1}$$

Vektor $\beta = (\beta_1, \ldots, \beta_k)^T \in \mathbb{R}^k$ v (2.1) je *vektor neznámych parametrov modelu*, inak nazývaných aj *regresné parametre*, ktorý budeme odhadovať z dát. Tento vektor vyjadruje vplyv vysvetľujúcich premenných na modelovanú veličinu. Parameter β_1 sa niekedy nazýva *intercept*.

Vektor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ je vektor nezávislých, vzájomne nekorelovaných náhodných veličín s normálnym rozdelením: $\varepsilon \sim N\,(0, \sigma^2 I_n)$. Tento vektor sa niekedy nazýva aj *vektor neznámych chýb (errors)* alebo aj *disturbancia*. Každá zložka vektoru chýb ε reprezentuje rozdiel medzi hodnotami, ktoré sme vysvetlili pomocou systematickej zložky μ_i, a hodnotami, ktoré sme pozorovali y_i.

X je konštrukčná matica s n riadkami a k stĺpcami, ktorej riadky zodpovedajú hodnotám jednotlivých meraní a stĺpce vysvetľujúcim premenným pri týchto meraniach. Pokiaľ chceme, aby model bol jednoznačne definovaný, tak matica

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

musí byť regulárna, teda musí mať lineárne nezávislé stĺpce.

Poznámka 2.1.2. *Prediktory X by mali byť nezávislé. Pokiaľ sú závislé, tak jeden parameter môžeme vyjadriť ako lineárnu kombináciu tých ostatných. Túto vlastnosť nazývame multikolinearita.*

Lineárny regresný model vníma pozorovania y_i ako realizáciu náhodnej premennej $\mathbf{y}$, ktorá má n zložiek, ktoré sú nezávislé, rovnako rozdelené, a každá realizácia y_i je kombináciou systematickej zložky $\mathrm{E}\left[\mathbf{y}_i\right] = \mu_i$ a náhodnej zložky ε_i.

Definícia 2.1.3. Model (2.1) môžeme prepísať pre i-tú zložku náhodnej veličiny $\mathbf{y}$ takto

$$\mu_i = \mathrm{E}\left(\mathbf{y}_i\right) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}, \qquad i = 1, \dots, n, \tag{2.2}$$

kde X_{i1} je rovný jedničke.

Predpoklady modelu

Pre zhrnutie klasického lineárneho modelu môžeme definovať tieto predpoklady:

(LM1) *Lineárny vzťah*: Medzi strednou hodnotou vektoru $\mathbf{y}$ a zložkami vysvetľujúcich premenných je lineárny vzťah

$$\mathrm{E}\left(\mathbf{y}\right) = \boldsymbol{\mu} = \sum_{j=1}^{k} X_j \beta_j.$$

(LM2) *Nezávislosť*: $\varepsilon_1, \dots, \varepsilon_n$ sú vzájomne nezávislé.

(LM3) *Rozdelenie*: Každá zložka vektoru ε má normálne rozdelenie so strednou hodnotou 0 a rozptylom σ^2.

2.1.1 Hľadanie parametrov modelu

Zložky parametrického vektoru β v modeli (2.1) chceme odhadovať tak, aby model čo najlepšie vysvetľoval dáta. To dosiahneme tým, že minimalizujeme sumu štvorcov jednotlivých zložiek vektorov reziduí $\mathbf{y} - \mathbf{X}\beta$. Túto metódu nazývame *metóda najmenších štvorcov.*

Hľadáme $\hat{\beta}$, ktorý je odhadom β:

$$\begin{aligned}\hat{\beta} &= \arg\min_\beta \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \\ &= \arg\min_\beta (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta),\end{aligned}$$

kde $\hat{\mathbf{y}}_i = \mathbf{X}_i^T \hat{\beta}$ je odhadom $\mathrm{E}(y_i)$.

To ľahko dostaneme tak, že parciálne derivácie podľa jednotlivých zložiek vektoru β položíme rovné nule

$$\begin{aligned}0 &= \frac{\partial}{\partial \beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &= -\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) + \left[(\mathbf{y} - \mathbf{X}\beta)^T (-\mathbf{X}) \right]^T = \\ &= -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta).\end{aligned}$$

Po úprave teda dostávame:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y}. \tag{2.3}$$

Poznámka 2.1.4. *Pretože* $\mathbf{X}$ *má lineárne nezávislé stĺpce, odhad* $\hat{\beta}$ *je určený jednoznačne.*

Veta 2.1.5. Odhad (2.3) je nestranným odhadom β a rozptyl odhadovaného parametru $\hat{\beta}$ vyjadríme ako $\mathrm{Var}\,\hat{\beta} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Dôkaz. Odhad (2.3) je nestranný, ak platí $\mathrm{E}\,\hat{\beta} = \beta$.

$$\mathrm{E}\,\hat{\beta} = \mathrm{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y}\right] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathrm{E}\,[\mathbf{y}] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{X}\beta = \beta.$$

Rozptyl $\hat{\beta}$ je

$$\begin{aligned}\mathrm{Var}\,\hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathrm{Var}\,[\mathbf{y}]\, \mathbf{X} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T(\sigma^2 I)\mathbf{X} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}.\end{aligned}$$

□

Na základe odhadu (2.3) môžeme pre odhady závislej premennej a reziduí písať nasledujúce poznámky:

- Vektor $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ je najlepšia aproximácia vektoru $\mathbf{y}$. Po dosadení parametru $\hat{\beta}$ dostávame
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}_{\mathbf{H}}\mathbf{y} = \mathbf{H}\mathbf{y}, \tag{2.4}$$
kde $\mathbf{H}$ je *projekčná matica.*

- Vektor $\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \underbrace{(I - \mathbf{H})}_{\mathbf{M}}\mathbf{y} = \mathbf{M}\mathbf{y}$ sa nazýva *vektor reziduí.*

- $\mathbf{M}$ je *projekčná matica do ortogonálneho podpriestoru.*

- Matice $\mathbf{H}$ a $\mathbf{M}$ sú symetrické
$$\mathbf{H} = \mathbf{H}^T, \quad \mathbf{M} = \mathbf{M}^T, \tag{2.5}$$
idempotentné
$$\mathbf{HH} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}, \mathbf{MM} = \mathbf{M} \tag{2.6}$$
a navzájom kolmé
$$\mathbf{HM} = 0, \quad \mathbf{H} + \mathbf{M} = I. \tag{2.7}$$

Poznámka 2.1.6. *Modely s neúplnou hodnosťou nebudeme uvažovať, pretože ich môžeme previesť na menší podmodel.*

Poznámka 2.1.7. *Za predpokladu normálneho rozdelenia vektoru*
$$\mathbf{y} \sim N\left(\mathbf{X}\beta, \sigma^2\right)$$
môžeme pre $\hat{\beta}$ písať
$$\hat{\beta} \sim N\left(\beta, \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2\right),$$
potom $\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2} \sim \chi^2_{n-k}$; $\sigma^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k}$, ako sa dočítame v knihe [22].

2.2 Zovšeobecnené lineárne modely

V praxi sa často stáva, že s danými údajmi nie sme schopní splniť všetky predpoklady klasického lineárneho modelu. V takom prípade môžeme použiť iný, rozšírený model, ktorého predpoklady už splniť dokážeme. Takýmito modelmi sú *zovšeobecnené lineárne modely*, ktoré tvoria rodinu štatistických modelov. Zahrňujú napríklad klasické lineárne modely, analýzu rozptylu, logistické modely.

Zovšeobecnené lineárne modely boli prvýkrát predstavené autormi Nelderom a Wedderburnom v článku [8] v roku 1972 a špecifikujú vzťah medzi strednou hodnotou náhodnej veličiny $\mathbf{y}$ a funkciou lineárnej kombinácie nezávislých parametrov modelu, kde stredná hodnota $\mathbf{y}$ je vyjadrená takto

$$\mathrm{E}\,(\mathbf{y}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T .$$

Definícia 2.2.1. Zovšeobecnený lineárny model môžeme definovať ako

$$\mathbf{y} = \mathrm{E}\,(\mathbf{y}) + \varepsilon; \quad \mathrm{E}\,(\mathbf{y}) = g^{-1}(\mathbf{X}\beta). \tag{2.8}$$

Matica $\mathbf{X}$ z (2.8) je známa matica s n riadkami a k stĺpcami, vektor $\beta = (\beta_1, \ldots, \beta_k)^T$ je vektor neznámych parametrov modelu. Vektor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ je vektor nezávislých náhodných veličín resp. vektor neznámych chýb. Funkciu $g(.)$ nazývame *transformačná funkcia*, ktorá je monotónna, diferencovateľná a existuje k nej inverzná funkcia g^{-1}.

Definícia 2.2.2. *Lineárny prediktor* $\boldsymbol{\eta}$ je lineárna funkcia parametrov modelu:

$$\eta_i = \sum_{j=1}^{k} X_{ij}\beta_j, \qquad i = 1, \ldots, n.$$

Predpoklady modelu

Predpoklady pre zovšeobecnený lineárny model analogicky k predpokladom klasického lineárneho modelu sú tieto:

(GLM1) *Transformačná funkcia*: Prepojenie medzi strednou hodnotou vektoru $\mathbf{y}$ a zložkami vysvetľujúcich premenných je prostredníctvom *transformačnej funkcie* s lineárnym prediktorom.

$$\eta_i = g(\mu_i) \quad \Longrightarrow \quad \mu_i = g^{-1}(\eta_i),$$

kde $g(\cdot)$ je *transformačná funkcia (link function).*

(GLM2) *Nezávislosť*: $y_1, \ldots, y_n$ sú vzájomne nezávislé.

(GLM3) *Rozdelenie*: Rozdelenie náhodnej veličiny $\mathbf{y}$ pochádza z *rodiny exponenciálnych rozdelení.*

Ak porovnáme predpoklady klasického a zovšeobecneného lineárneho modelu, mení sa nám prvý a tretí predpoklad. Namiesto normálneho rozdelenia veličiny $\mathbf{y}$ máme všeobecnejšiu triedu exponenciálnych rozdelení. V prvom predpoklade

použijeme namiesto identickej funkcie všeobecnejšiu tzv. *transformačnú funkciu*, ktoré vysvetlíme neskôr.

Príklady zovšeobecnených lineárnych modelov

1. Lineárna regresia:
   - $y_i \sim \mathrm{N}\ (\mu_i, \sigma^2)$
   - $\mathrm{E}\, y_i = \mu_i$
   - $g\,(\mu_i) = \mu_i$
2. Logistická regresia:
   - $y_i \sim \mathrm{Alt}\,(p_i)$
   - $\mathrm{E}\, y_i = \mu_i = p_i$
   - $\eta_i = g\,(\mu_i) = \log\frac{\mu_i}{1-\mu_i}$
   - $\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$
3. Loglineárny model:
   - $y_i \sim \mathrm{Po}\,(\lambda_i)$
   - $\mathrm{E}\, y_i = \mu_i = \lambda_i$
   - $\eta_i = g\,(\mu_i) = \log\mu_i$
   - $\mu_i = e^{\eta_i}$

2.2.1 Exponenciálna trieda rozdelení

V (GLM3) predpokladáme, že náhodná veličina y má rozdelenie pochádzajúce z *exponenciálnej triedy rozdelení (exponenciálneho typu rozdelení, the exponential family of distributions)*. Exponenciálna trieda rozdelení zahŕňa diskrétne i spojité rozdelenia, napr. binomické, exponenciálne, gamma, normálne, Poissonovo rozdelenie. Funkcia hustoty pravdepodobnosti pre túto triedu rozdelení zapísaná v kanonickom tvare vyzerá takto

$$f\,(y;\theta,\phi) = exp\left(\frac{y\theta - b\,(\theta)}{a\,(\phi)} + c\,(y,\phi)\right), \tag{2.9}$$

kde θ je *kanonický (canonical, natural) parameter* súvisiaci so strednou hodnotou, ϕ nazývame *disperzný (dispersion, scale) parameter, parameter škály* súvisiaci s rozptylom, $a\,(\phi)$ je kladná a spojitá funkcia, $b\,(\theta)$ je dvakrát diferencovateľná konvexná funkcia a $c\,(y,\phi)$ je tzv. normalizačný faktor, ktorý je nezávislý od parametra θ a zabezpečuje, aby integrál hustoty bol rovný 1.

Strednú hodnotu a rozptyl náhodnej veličiny y z exponenciálnej triedy rozdelení vypočítame pomocou momentovej vytvárajúcej funkcie s využitím nasledujúceho tvrdenia.

Tvrdenie 2.2.3. *Ak y má hustotu (2.9) a $b(\theta)$ je dvakrát spojito diferencovateľná, potom existuje momentová vytvárajúca funkcia funkcie y s nasledovným tvarom:*

$$M(t) = \mathrm{E}\, e^{ty} = exp\left(\frac{b\big(\theta + a(\phi)\, t\big) - b(\theta)}{a(\phi)}\right)$$

a $M(t)$ je dvakrát diferencovateľná v bode 0. Platí, že $\mathrm{E}(y) = b'(\theta)$ a $\mathrm{Var}(y) = a(\phi)\, b''(\theta)$.

Dôkaz. Označíme množinu A, ktorá je daná takými $\mathbf{y}$, pre ktoré je hustota (2.9) kladná:

$$\begin{aligned}
M(t) &= \int_A exp(ty) \cdot exp\left(\frac{y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) dy \\
&= \int_A exp\left(\frac{t\cdot y\cdot a(\phi) + y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) dy \\
&= \int_A exp\left(\frac{y\cdot\big(\theta + a(\phi)\cdot t\big) - b(\theta)}{a(\phi)} + c(y,\phi)\right) dy \\
&= \int_A exp\Bigg(\frac{y\cdot\big(\theta + a(\phi)\cdot t\big) - b(\theta) + b\big(\theta + a(\phi)\cdot t\big) - b\big(\theta + a(\phi)\cdot t\big)}{a(\phi)} \\
&\qquad + c(y,\phi)\Bigg) dy \\
&= exp\left(\frac{b\big(\theta + a(\phi)\cdot t\big) - b(\theta)}{a(\phi)}\right) \\
&\quad \underbrace{\int_A \left(\frac{y\cdot\big(\theta + a(\phi)\cdot t\big) - b\big(\theta + a(\phi)\cdot t\big)}{a(\phi)} + c(y,\phi)\right) dy}_{1} \\
&= exp\left(\frac{b\big(\theta + a(\phi)\cdot t\big) - b(\theta)}{a(\phi)}\right).
\end{aligned}$$

Ak v definícii hustoty exponenciálnej triedy rozdelení (2.9) zameníme parameter θ za parameter $\big(\theta + a(\phi)\big)$, dostaneme integrál

$$\int_A \left(\frac{y\cdot\big(\theta + a(\phi)\cdot t\big) - b\big(\theta + a(\phi)\cdot t\big)}{a(\phi)} + c(y,\phi)\right) dy,$$

ktorý je rovný jednej.

Aby sme získali strednú hodnotu a rozptyl $\mathbf{y}$, vypočítame prvú a druhú deriváciu momentovej vytvárajúcej funkcie:

$$M'(t) = exp\left(\frac{b\big(\theta + a(\phi)\cdot t\big) - b(\theta)}{a(\phi)}\right)\cdot b'\big(\theta + a(\phi)\cdot t\big),$$

$$M''(t) = exp\left(\frac{b\big(\theta + a(\phi)\cdot t\big) - b(\theta)}{a(\phi)}\right)\cdot \big[b'\big(\theta + a(\phi)\cdot t\big)\big]^2 +$$

$$exp\left(\frac{b\big(\theta + a(\phi)\cdot t\big) - b(\theta)}{a(\phi)}\right)\cdot b''\big(\theta + a(\phi)\cdot t\big)\cdot a(\phi).$$

Do prvej derivácie momentovej vytvárajúcej funkcie dosadíme za parameter $t = 0$ a dostaneme strednú hodnotu $\mathbf{y}$

$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{y}) = M'(0) = b'(\theta). \tag{2.10}$$

Druhý moment $\mathbf{y}$ dostaneme dosadením $t = 0$ do druhej derivácie momentovej vytvárajúcej funkcie a rozptyl teda vypočítame takto:

$$\begin{aligned}\mathrm{Var}[\mathbf{y}] &= \mathrm{E}\left(\mathbf{y}^2\right) - [\mathrm{E}(\mathbf{y})]^2 = M''(0) - [M'(0)]^2 \\ &= [b'(\theta)]^2 + a(\phi)\, b''(\theta) - [b'(\theta)]^2 = a(\phi)\, b''(\theta).\end{aligned} \tag{2.11}$$

□

Poznámka 2.2.4. *Z (2.10) vieme, že* $\boldsymbol{\mu} = b'(\theta)$. *Funkcia* $b'(\theta)$ *je monotónna, pretože platí* $\mathrm{Var}[\mathbf{y}] = a(\phi)\, b''(\theta) > 0$. *Rozptyl* $\mathbf{y}$ *potom môžeme napísať ako*

$$\mathrm{Var}[\mathbf{y}] = a(\phi)\, b''(\theta) = a(\phi)\, b''\left[(b')^{-1}(\mu)\right] = a(\phi)\, V(\mu),$$

kde funkcia $V(.)$ *sa nazýva* ***rozptylová funkcia****.*

Príklad 2.2.5. Pravdepodobnostná funkcia Poissonovho rozdelenia má takýto tvar

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = exp\{y\log\lambda - \lambda - \log(y!)\}.$$

Podľa vzorca (2.9) máme $\theta = \log\lambda$, $b(\theta) = e^\theta$ a $\phi = 1$.

Stredná hodnota náhodnej veličiny y, ktorá má Poissonovo rozdelenie, je rovná

$$\mathrm{E}\,(y) = b'(\theta) = \lambda$$

a jej rozptyl je

$$\mathrm{Var}\,(y) = b''(\theta) = b'(\theta) = \lambda.$$

Príklad 2.2.6. Pravdepodobnostná funkcia binomického rozdelenia má tvar

$$\begin{aligned} f(y; n, \mu) &= \binom{n}{y} \mu^y (1-\mu)^{n-y} \\ &= \exp\left(y\log\mu + (n-y)\log(1-\mu) + \log\binom{n}{y} \right) = \\ &= \exp\left(y\log\tfrac{\mu}{1-\mu} + n\log(1-\mu) + \log\binom{n}{y} \right), \end{aligned}$$

kde $y = 0, 1, \ldots, n;\ \mu = 0, 1.$

Stredná hodnota binomického rozdelenia je

$$\mathrm{E}\,(y) = n\mu$$

a rozptyl je

$$\mathrm{Var}\,(y) = n\mu(1-\mu).$$

Pre prehľad uvádzame parametre a značenie základných rozdelení patriacich do exponenciálnej triedy rozdelení v tabuľkách:

	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Normálne	$\frac{\phi}{\omega}$	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$
Poissonovo	$\frac{\phi}{\omega}$	e^{θ}	$-\ln(y!)$
Binomické/m	$\frac{\phi}{\omega}$	$\ln\left(1 + e^{\theta}\right)$	$\ln\left(\frac{m}{my}\right)$
Gamma	$\frac{\phi}{\omega}$	$-\ln(-\theta)$	$\frac{1}{\phi}\ln\left(\frac{y}{\phi}\right) - \ln(y) - \ln\left(\Gamma\left(\frac{1}{\phi}\right)\right)$

Tabuľka 2.1: Zhrnutie parametrov základných rozdelení patriacich do exponenciálnej triedy rozdelení.

	Značenie	θ
Normálne	$N(\mu, \sigma^2)$	μ
Poissonovo	$P(\mu)$	$\ln\mu$
Binomické/m	$B(\mu, \pi)/m$	$\ln\left(\frac{\mu}{1-\mu}\right)$
Gamma	$G(\mu, \nu)$	$\frac{1}{\mu}$

Tabuľka 2.2: Značenie rozdelení patriacich do exponenciálnej triedy rozdelení.

Rozptylová funkcia

Z poznámky 2.2.4 vieme, že pre rozptyl y platí

$$\text{Var}\ (y) = b''(\theta)\, a(\phi) = V(\mu)\, a(\phi),$$

kde $V(\mu)$ je rozptylová funkcia. Tá závisí na kanonickom parametri, a tým pádom aj na strednej hodnote. Pre Poissonovo a binomické rozdelenie je $a(\phi) = 1$, a teda pre tieto rozdelenia je $V(\mu)$ rovná rozptylu y.

	ϕ	$V(\mu)$
Normálne	σ^2	1
Poissonovo	1	μ
Binomické/m	$\frac{1}{m}$	$\mu(1-\mu)$
Gamma	ν^{-1}	μ^2

Tabuľka 2.3: Zhrnutie parametrov rozdelení patriacich do rodiny exponenciálnych rozdelení.

2.2.2 Transformačná funkcia

Transformačná funkcia (link function), nazývaná taktiež *spojovacia* či *linková funkcia*, je dôležitým pojmom v teórii zovšeobecnených lineárnych modelov. Voľba *transformačnej funkcie* $g(.)$ je vedľa voľby rozdelenia základným predpokladom zovšeobecneného lineárneho modelu. Transformačná funkcia je prostá a diferencovateľná. Popisuje vzťah strednej hodnoty $\boldsymbol{\mu} = \text{E}\ (y)$ a lineárneho prediktoru $\boldsymbol{\eta}$, pričom platí

$$g\big(\text{E}\ (\mathbf{y})\big) = g(\boldsymbol{\mu}) = \boldsymbol{\eta}.$$

Pretože transformačná funkcia je funkciou jednej premennej a je prostá, strednú hodnotu μ môžeme vyjadriť ako

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}).$$

Transformačná funkcia je definovaná ako lineárna kombinácia vysvetľujúcich premenných a neznámeho vektoru parametrov modelu β v tvare

$$\boldsymbol{\eta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k \mathbf{x}_k = \mathbf{x}^T \beta,$$

kde $\mathbf{x} = (x_1, \ldots, x_k)^T$ je známy vektor vysvetľujúcich premenných. Pri n pozorovaniach náhodných veličín $y_1, \ldots, y_n$ dostávame n-rozmerný vektor

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T = \begin{pmatrix} g\big(\mathrm{E}\ (y_1)\big) \\ g\big(\mathrm{E}\ (y_2)\big) \\ \vdots \\ g\big(\mathrm{E}\ (y_n)\big) \end{pmatrix} = \mathbf{X}\beta,$$

kde $\mathbf{X}$ je matica typu $n \times k$ a $x_i = (x_{i1}, \ldots, x_{ik})^T$ je i-ty riadok matice $\mathbf{X}$.

V klasickom lineárnom modeli bola transformačnou funkciou identická funkcia, tj. $\eta = \mu$. Transformačná funkcia musí spĺňať základné predpoklady rozdelení. Napríklad u binomického rozdelenia potrebujeme takú transformačnú funkciu, ktorá zobrazuje hodnoty z $\langle 0, 1\rangle$ na interval $\langle -\infty, \infty\rangle$.

Pre Poissonovo rozdelenie potrebujeme mať parameter μ kladný a parameter η môže byť záporný. Preto je vhodnejšie namiesto identickej transformačnej funkcie použiť logaritmickú: $\eta = \log \mu$, alebo inverzne $\mu = e^{\eta}$, čo zaručuje $\mu > 0$.

Každé rozdelenie z tabuľky 2.3 má tzv. *kanonickú transformačnú funkciu (canonical link function)*. Kanonická transformačná funkcia zjednodušuje tvar vierohodnostnej funkcie. Pre túto funkciu platí $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \theta$, kde θ je kanonický parameter z exponenciálnej triedy rozdelení. Parameter θ môžeme teda napísať ako $\theta = g(b'(\theta))$.

Príklad 2.2.7. Príklady najčastejších kanonických transformačných funkcií pre rôzne rozdelenia podľa [12]:

- Normálne rozdelenie:
  - identita: $g(\mu) = \mu$
  - logaritmus: $g(\mu) = \log\mu$
  - inverzia: $g(\mu) = \frac{1}{\mu}$
- Binomické rozdelenie:
  - logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
  - probit: $g(\mu) = \Phi^{-1}(\mu)$

  - doplnkový log-log: $g(\mu) = \log\big(-\log(\mu)\big)$
  - logaritmus: $g(\mu) = \log\mu$

- Poissonovo rozdelenie:
  - logaritmus: $g(\mu) = \log\mu$
  - identita: $g(\mu) = \mu$
  - druhá odmocnina: $g(\mu) = \sqrt{\mu}$

- Gamma rozdelenie:
  - inverzia: $g(\mu) = \frac{1}{\mu}$
  - logaritmus: $g(\mu) = \log\mu$
  - identita: $g(\mu) = \mu$

Transformačná funkcia pri modelovaní pravdepodobnosti je často volená ako logit funkcia s binomickým rozdelením náhodnej zložky modelu, čo spoločne vedie na logistický model. Transformačná funkcia *logit* zobrazuje hodnoty z $(0, 1)$ na interval $(-\infty, \infty)$, ako sa dočítame v článku [11].

	kanonická transformačná funkcia $\theta(\mu)$
Normálne	μ
Poissonovo	$\ln\mu$
Binomické/m	$\ln\left(\frac{\mu}{1-\mu}\right)$
Gamma	$\frac{1}{\mu}$

Tabuľka 2.4: Kanonická transformačná funkcia pre základné rozdelenia z rodiny exponenciálnych rozdelení.

2.2.3 Hľadanie riešenia zovšeobecnených lineárnych modelov

Vyjadrenie vierohodnostnej funkcie

V zovšeobecnenom lineárnom modeli hľadáme riešenie *metódou maximálnej vierohodnosti (maximum likelihood estimation)*. Pomocou tejto metódy získame hodnoty, ktoré maximalizujú pravdepodobnosť získania našej množiny dát. Pretože y_i sú nezávislé, vierohodnostnú funkciu náhodných veličín y_i píšeme takto

$$L(y;\theta,\phi)=\prod_{i=1}^{n} f(y_i;\theta_i,\phi)=\prod_{i=1}^{n} exp\left(\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right). \tag{2.12}$$

Pri odhadovaní parametrov zovšeobecneného lineárneho modelu sa používa logaritmická transformácia, ktorá polohu extrému neovplyvní, avšak sa jednoduchšie derivuje. Túto transformáciu si označíme ako

$$l_i=\log f(y_i;\theta_i,\phi).$$

Po zlogaritmovaní vierohodnostnej funkcie $L(y;\theta,\phi)$ teda dostávame *logaritmickú vierohodnostnú funkciu (log-likelihood function)* l

$$\begin{aligned} l=l(y;\theta,\phi) &= \log L(y;\theta,\phi) \\ &= \sum_{i=1}^{n}\log\left[f(y_i;\theta_i,\phi)\right] \\ &= \sum_{i=1}^{n}\left[\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right]=\sum_{i=1}^{n} l_i. \end{aligned} \tag{2.13}$$

Príklad 2.2.8. Poissonovo rozdelenie:

$$l(y;\lambda_i)=\sum_{i=1}^{n}\ln f(y_i;\lambda)=\sum_{i=1}^{n}\left(-\lambda_i+y_i\ln\lambda_i-\ln(y_i!)\right).$$

Predtým, než ukážeme odhad parametru θ, potrebujeme definovať skórový vektor a Fisherovu informačnú funkciu.

Skórový vektor

Definícia 2.2.9. *Skórový vektor* príslušný k hustote f je definovaný ako prvá derivácia l podľa parametru θ:

$$\mathbf{U}(\theta,y)=\frac{\partial l}{\partial\theta}=\frac{\partial\log f(y,\theta,\phi)}{\partial\theta}=\frac{1}{f(y,\theta,\phi)}\frac{\partial f(y,\theta,\phi)}{\partial\theta} \tag{2.14}$$

Ak zderivujeme funkciu $f(y_i, \theta, \phi)$, dostaneme

$$\begin{aligned}
\mathbf{U}(\theta, y_i) &= \frac{1}{f(y_i, \theta, \phi)} \frac{\partial f(y_i, \theta, \phi)}{\partial \theta} \\
&= \frac{exp\left(\frac{y_i\theta_i - b(\theta)}{a(\phi)} + c(y_i, \phi)\right)}{exp\left(\frac{y_i\theta_i - b(\theta)}{a(\phi)} + c(y_i, \phi)\right)} \frac{y_i - b'(\theta)}{a(\phi)} \\
&= \frac{1}{a(\phi)}\big(y_i - b'(\theta)\big).
\end{aligned}$$

Poznámka 2.2.10. *Z toho, že $c(y_i, \phi)$ je funkcia nezávislá na θ, plynie to, že nie je závislá ani na $\boldsymbol{\mu}$, a teda ani na β, a preto parameter $c(y_i, \phi)$ nie je významný pre riešenie maximálnej vierohodnostnej funkcie.*

Veta 2.2.11. Stredná hodnota derivácie l je nula

$$\mathrm{E}\,\mathbf{U}(\theta, y) = \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0. \tag{2.15}$$

A taktiež platí

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right)^2 = 0. \tag{2.16}$$

Dôkaz. Dôkazy nájdeme v knihe [6]. □

Z rovnosti (2.15) teda máme

$$0 = \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = \frac{\mathrm{E}(y) - b'(\theta)}{a(\phi)} = \frac{\mu - b'(\theta)}{a(\phi)}$$

a z toho dostávame strednú hodnotu náhodnej veličiny y

$$\mathrm{E}(y) = \mu = b'(\theta) = \left(\frac{\partial b(\theta)}{\partial \theta}\right). \tag{2.17}$$

Stredná hodnota (referr-e3) je určená prvou deriváciou b podľa θ. K rovnakému výsledku sme sa dopracovali aj pomocou momentovej vytvárajúcej funkcie, viď (2.10).

Definícia 2.2.12. Pre skórovú funkciu definujeme *skórovú štatistiku* vzorcom

$$\mathbb{U}(\theta) = \sum_{i=1}^{n} \mathbf{U}(\theta_i, y_i). \tag{2.18}$$

Definícia 2.2.13. Nech existuje matica druhých parciálnych derivácií f $\frac{\partial^2 f(y;\theta,\phi)}{\partial \theta^2}$.

Potom *Fisherovou informačnou funkciou* príslušnú k hustote f nazývame:

$$\mathbf{I}(\theta, y) = \frac{\partial^2 l}{\partial \theta^2} = \frac{\partial^2 \log f(y, \theta, \phi)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}.$$

Veta 2.2.14. Z rovnosti (2.16) plynie, že rozptyl náhodnej veličiny y je

$$\text{Var}[y] = b''(\theta)\, a(\phi) = \left(\frac{\partial^2 b(\theta)}{\partial \theta^2} a(\phi)\right). \tag{2.19}$$

Dôkaz.

$$\begin{aligned} 0 &= \text{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \text{E}\left(\frac{\partial l}{\partial \theta}\right)^2 \\ &= \frac{b''(\theta)}{a(\phi)} + \text{E}\left[\frac{(y - b'(\theta))^2}{a^2(\phi)}\right] \\ &= \frac{b''(\theta)}{a(\phi)} + \frac{\text{E}[y^2] - 2\text{E}[y]\, b'(\theta) + b'(\theta)^2}{a^2(\phi)} \\ &= \frac{1}{a^2(\phi)}\left(b''(\theta)\, a(\phi) + \underbrace{\text{E}\left(y^2\right) - \text{E}(y)^2}_{-\text{Var}[y]}\right) \end{aligned}$$

a z toho už dostávame rozptyl y: $\text{Var}[y] = b''(\theta)\, a(\phi)$. □

Poznámka 2.2.15. *Z [22] máme rozptyl skórového vektoru*

$$\text{Var}\,\mathbf{U}(\theta, y_i) = \frac{1}{a(\phi)} b''(\theta).$$

Navyše platí:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} \text{N}\left(0, \frac{a(\phi)}{b''(\theta)}\right).$$

Systém vierohodnostných funkcií

Definícia 2.2.16. *Systémom vierohodnostných rovníc* rozumieme

$$\mathbf{U}(\theta, y) = 0, \tag{2.20}$$

kde $\mathbf{U}$ je skórový vektor. Maximálne vierohodný odhad je riešením systému vierohodnostných rovníc. Odhadovaný parameter $\hat{\theta}$ nazývame maximálne vierohodným odhadom, ktorý je riešením (2.20).

Poznámka 2.2.17. *Z definície 2.14 plynie, že systém vierohodnostných funkcií*

môžeme zapisovať aj v tvare

$$\frac{\partial l}{\partial \theta} = 0.$$

Odhad $\hat{\theta}^{(0)}$ označme ako počiatočný odhad parametru θ. Potom postupnosť odhadov

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} + \left[\sum_{i=1}^{n} \mathbf{I}_i \left(\hat{\theta}^{(m)}\right)\right]^{-1} \mathbb{U}\left(\hat{\theta}^{(m)}\right); \qquad m = 0, 1, \ldots, \tag{2.21}$$

kde $\mathbf{I}_i(.)$ je Fisherova informačná funkcia, nazývame postupnosť odhadov získaných modifikovanou Newton-Rapsonovou iteračnou metódou, viď [6].

Maximálny vierohodný odhad

Aby sme získali maximálny vierohodný odhad parametru β_j, vyjadríme deriváciu l podľa β_j

$$U_j(\beta) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \beta_j}\right].$$

Z reťazového pravidla vieme, že platí

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{\partial l_i}{\partial \theta_i}\right)\left(\frac{\partial \theta_i}{\partial \mu_i}\right)\left(\frac{\partial \mu_i}{\partial \beta_j}\right), \tag{2.22}$$

kde $j = 1, \ldots, k$.

Teraz si to rozoberieme po častiach. Výraz $\left(\frac{\partial l}{\partial \theta_i}\right)$ sme už vyjadrili v (2.14). Ďalšie zložky vyjadríme takto:

- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial \theta_i}{\frac{\partial b(\theta_i)}{\partial \theta_i}} = \frac{\partial \theta^2}{\partial^2 b(\theta_i)} = \frac{1}{\frac{\partial^2 b(\theta_i)}{\partial \theta^2}} = \frac{1}{V(\mu_i)}$
- $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$
- a zo vzťahu (2.15) vieme, že $b'(\theta_i) = \mu_i$.

Rovnicu (2.22) upravíme podľa práve zavedených rovností

$$U_j(\beta) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) x_{ij}, \tag{2.23}$$

kde $j = 1, \ldots, k$ sú všeobecné odhadovacie rovnice pre triedu rozdelení na odhad $\hat{\beta}$ parametru β a $V(\mu_i)$ je variačná funkcia. $U_j(\beta)$ je skórová funkcia parametru β.

Derivácie l podľa β_j položíme rovné nule a nájdeme regresné koeficienty. Z poznámky 2.2.4 vieme, že $V(\mu_i) = \frac{\mathrm{Var}(y)}{a(\phi)}$. Vo vzorci (2.23) môžeme za $a(\phi)$ dosadiť $\frac{\phi}{w_i}$. Regresné koeficienty môžeme odhadnúť bez znalosti disperzného parametru ϕ, pretože ten sa skráti, ak sú parciálne derivácie položené nule. Parametre β_j odhadneme krokovou metódou najmenších štvorcov.

Kovariačná matica U_j má tvar

$$\mathfrak{I}_{jp} = \mathrm{E}\left[U_j U_p\right], \tag{2.24}$$

z ktorej získavame **informačnú maticu**

$$\begin{aligned} \mathfrak{I}_{jp} &= \mathrm{E}\left\{\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) x_{ij} \sum_{l=1}^{n} \frac{y_l - \mu_l}{V(\mu_l)} \left(\frac{\partial \mu_l}{\partial \eta_l}\right) x_{lp}\right\} \\ &= \sum_{i=1}^{n} \frac{\mathrm{E}\left[(y_i - \mu_i)^2\right] x_{ij} x_{ip}}{\left[V(\mu_i)\right]^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \end{aligned} \tag{2.25}$$

a pretože $\mathrm{E}\left[(y_i - \mu_i)(y_l - \mu_l)\right] = 0$ pre $i \neq l$ a to z dôvodu, že $\mathbf{y}$ sú nezávislé. Použitím $\mathrm{E}\left[(y_i - \mu_i)^2\right] = \mathrm{Var}(y)$ môžeme (2.25) zjednodušiť takto

$$\mathfrak{I}_{jp} = \sum_{i=1}^{n} \frac{x_{ij} x_{ip}}{\left[V(\mu_i)\right]} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{2.26}$$

Rovnicu (2.21) zovšeobecníme:

$$\mathbf{b}^{(m+1)} = \mathbf{b}^{(m)} + \left[\mathfrak{I}^{(m)}\right]^{-1} \mathbb{U}^{(m)}; \qquad m = 0, 1, \ldots ,$$

kde $\mathbf{b}^{(m+1)}$ je vektor odhadov parametrov $\beta_1, \ldots, \beta_k$ v $(m+1)$-ej iterácii. Matica $\left[\mathfrak{I}^{(m)}\right]^{-1}$ je inverzná k informačnej matici s elementami $\mathfrak{I}_{jk}$ daných v (2.25) a $\mathbb{U}^{(m)}$ je vektor elementov daných v (2.23), všetko vztiahnuté k $\mathbf{b}^{(m)}$.

Maticu $\mathfrak{I}$ môžeme prepísať aj do tvaru, viď [3],

$$\mathfrak{I} = \mathbf{X}^T \mathbf{W} \mathbf{X}, \tag{2.27}$$

kde $\mathbf{X}$ je regresná matica a $\mathbf{W}$ je diagonálna matica typu $n \times n$ so zložkami $\omega_{ii} = \frac{1}{\mathrm{Var}[y]} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$.

Podľa [3] dostaneme iteratívne rovnice. Najprv však definujeme závislú premennú z.

Definícia 2.2.18. Závislá premenná z je linearizovaná forma alebo linearizovaná odozva *(adjusted depend variable)* transformačnej funkcie aplikovanej

na y. Váhy sú funkciou fitovaných hodnôt $\hat{\mu}$. Proces je krokový, pretože upravená závislá premenná z a váhy ω závisia na fitovanej hodnote, ktorá je aktuálne dostupná. Závislá premenná z je rozvoj $g(y_i)$ v Taylorovom ráde okolo hodnoty $\hat{\mu}$

$$g(y) \simeq g(\mu) + (y - \mu) g'(\mu).$$

Upravená závislá premenná má teda tvar

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i) g^{-1}(\hat{\mu}_i),$$

kde derivácia transformačnej funkcie je vyčíslená ako μ_i.

Definícia 2.2.19. Iteratívna rovnica má tvar

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m+1)} = \mathbf{X}^T\mathbf{W}\mathbf{z}.$$

Poznámka 2.2.20. *Maximálne vierohodný odhad $\hat{\beta}$ v zovšeobecnenom lineárnom modeli spĺňa rovnicu*

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}\mathbf{z}\right). \tag{2.28}$$

Podrobný popis riešenia iteračne vážených najmenších štvorcov je popísaný v knihe [7].

Na konci iteračného procesu sa dopracujeme k odhadom parametru β, konečnej variačnej matici $\mathbf{W}$ a asymptotickej normalite

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \phi\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\right). \tag{2.29}$$

Odhad parametru ϕ

Disperzný parameter odhadujeme metódou momentov, ako môžeme vidieť v knihe [10]

$$\mathrm{E}\left[\frac{(y_i - \mu_i)^2}{V(\mu_i)}\right] = \frac{\phi}{V(\mu_i)}\mathrm{Var}\,(\mu_i) = \phi.$$

Pearsonova štatistika je

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

kde $\frac{X^2}{\phi} \sim \chi^2$.

Odhad ϕ je

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

2.2.4 Kvázi-vierohodnosť

Pri riešení zovšeobecnených lineárnych rovníc vierohodnostných funkcií predpokladáme, že poznáme konkrétne rozdelenie zodpovedajúce pozorovaným parametrom y. V mnohých prípadoch však toto rozdelenie nepoznáme. Aby sme vypočítali deviácie, potrebujeme poznať vierohodnostnú funkciu. K vierohodnostnej funkcii potrebujeme poznať rozdelenie. Pre tieto prípady zavádzame tzv. *funkciu kvázi-vierohodnosti (quasi-likelihood function)*, ktorá vyžaduje len predpoklad vzájomnej nezávislosti pozorovaní a znalosť rozptylovej funkcie a transformačnej funkcie.

Pokiaľ je teda μ_i strednou hodnotou náhodnej premennej y_i a $V(\mu_i)$ je známa rozptylová funkcia, tak z poznámky 2.2.4 máme $\text{Var}(y_i) = a(\phi) V(\mu_i)$. Navyše predpokladáme, že y_i sú nezávislé. Ďalej predpokladáme, že

$$X^T\beta = \boldsymbol{\eta} = (g(\mu_1), \ldots, g(\mu_n))^T,$$

kde $g(.)$ je známa transformačná funkcia. Potom pre každé pozorovanie definujeme funkciu kvázi-vierohodnosti $Q_i(y_i, \mu_i)$ vzťahom

$$\frac{\partial Q_i(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{a(\phi) V(\mu_i)},$$

respektíve, viď [6],

$$Q_i(y_i, \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - \tilde{\mu}}{a(\phi) V(\tilde{\mu})} d\tilde{\mu}.$$

	Kvázi-vierohodnosť
Normálne	$-\frac{(y-\mu)^2}{2}$
Poissonovo	$y\log\mu - \mu$
Binomické/m	$y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$
Gamma	$-\frac{y}{\mu} - \log\mu$

Tabuľka 2.5: Kvázi-vierohodnostné funkcie základných rozdelení patriacich do rodiny exponenciálnych rozdelení.

Združená funkcia kvázi-vierohodnosti je daná vzťahom

$$Q = \sum_{i=1}^{n} Q_i(y_i, \mu_i).$$

Môžeme vidieť, že pre túto rovnicu požadujeme znalosť parametra $V(\mu_i)$ skôr než znalosť celého rozdelenia y_i. Ak chceme odhad maximálneho kvázivierohodného parametru β zovšeobecnených lineárnych modelov, zderivujeme túto rovnicu q podľa jednotlivých zložiek β_j a položíme rovnú nule. Týmto dostaneme

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} Q_i(y_i, \mu_i) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi) V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0.$$

Tento vzorec sme už popisovali vo vzorci (2.23).

2.3 Výber modelu

Pri modelovaní dát pomocou zovšeobecnených lineárnych modelov sa snažíme vybrať vhodný model pomocou vhodného rozdelenia y a voľby vhodnej transformačnej funkcie. Transformačnú funkciu volíme kanonickú, ktorú môžeme dostať na základe reziduí. Model vyberáme podľa toho, aby čo najlepšie vysvetľoval dáta. Dobrý štatistický model sa od pozorovaných dát veľmi nelíši a je úsporný, teda obsahuje čo najmenej parametrov. Všeobecne platí, že jednoduchší model s menej parametrami, ktorý dobre popisuje dáta, je lepší než model takmer dokonale popisujúci dáta s veľa parametrami. Najmenší a zároveň najjednoduchší model je tzv. *null model*. Najkomplexnejší je *maximálny, úplný (full)* alebo *saturovaný (saturated) model*.

Null model má 1 parameter reprezentujúci spoločné μ pre všetky y, reprezentuje situáciu, kedy nie je žiadny vzťah medzi prediktorom a parametrom y, teda máme len jeden parameter θ.

V *saturovanom modeli* sú dáta vysvetlené presne. Vektor $\beta_{\max}$ má n zložiek.

Problém počtu parametrov je ten, že pokiaľ máme veľa parametrov modelu, môžeme síce lepšie vysvetľovať, ale na úkor stupňov voľnosti a teda nulovej sily testu.

V praxi je null model príliš jednoduchý a full model nenesie žiadnu novú informáciu, pretože nesumarizuje dáta, ale iba ich opakuje v plnom rozsahu. Úplný model nám však dáva základ pre mieru rozporu pre model o n parametroch.

Proces voľby modelu je niekedy považovaný za cestu nahradzovania množiny dát y množinou hodnôt μ odvodených z modelu, ktorý má obvykle menej parametrov. Všeobecne nemá množina dát y rovnaké hodnoty ako μ a nás zaujíma, aké veľké sú tieto rozdiely, pretože malé sa môžu zanedbať, veľké však nie.

Miera toho, ako dobre bude model vysvetľovať dáta indikuje adekvátnosť modelu pre popis štruktúry dát. Mier je niekoľko, a všeobecne se nedá povedať, ktorá je najoptimálnejšia. Preto sa táto miera používa v kontexte hypotéz. Väčšina mier vychádza z maximálnej hodnoty vierohodnostnej funkcie pre daný model, teda podľa čoho meriame *odchýlku*.

2.3.1 Testovanie hypotéz o parametroch zovšeobecneného lineárneho modelu

Pri aplikácii zovšeobecnených lineárnych modelov je cieľom vybrať vhodný model pomocou voľby rozdelenia pravdepodobností z exponenciálnej triedy rozdelení pre závislú premennú $\mathbf{y}$ a voľbu vhodnej transformačnej funkcie. Transformačnú funkciu môžeme voliť kanonickú, pretože zjednodušuje tvar vierohodnostnej funkcie. Avšak dôležitejšie je vybrať taký model, ktorý je praktický a ľahko interpretovateľný.

Pre testovanie nulovej hypotézy $H_0 : g(\mu) = X_0\beta_0$, že testovaný model dobre vysvetľuje dáta, sa používa rozdiel saturovaného modelu a testovaného modelu. Parameter μ je predpoklad strednej hodnoty y, ktorý má nezávislé zložky, rovnako rozdelené z rodiny exponenciálnych rozdelení. Nech $l\left(\hat{\beta}_0\right)$ a $l\left(\hat{\beta}_{\max}\right)$ sú maximálne vierohodnostné funkcie modelu.

Pokiaľ platí hypotéza H_0, potom

$$2\left[l\left(\hat{\beta}_{\max}\right) - l\left(\hat{\beta}_0\right)\right] \sim \chi^2_{k_0-k_1} \tag{2.30}$$

a teda testovaný model vysvetľuje dáta rovnako dobre ako saturovaný model. Ak nulová hypotéza neplatí, potom maximálny model bude mať podstatne vyššiu vierohodnosť než testovaný model.

Vzorec (2.30) sa dá použiť len pre modely, kde je disperzný parameter ϕ známy a to napríklad pre Poissonovo alebo binomické rozdelenie, viď [6]. Prípad, že disperzný parameter nepoznáme, preberieme neskôr.

2.3.2 Deviácia

Ak máme vybraný model, mali by sme odhadnúť parametre a ohodnotiť presnosť odhadu. *Deviácia* je meradlom toho, ako je náš model vhodný *(goodness of fit)*, ako zovšeobecnený lineárny model zodpovedá dátam. Ak so zovšeobecnenými lineárnymi modelmi pracujeme v praxi, pre modelovanie je najlepšie mať čo najväčšie množstvo dát, ktoré môžeme interpretovať. Práve táto kvantita dát

je *deviácia* modelu a za predpokladu $a_i = \frac{\phi}{\omega_i}$ je definovaná ako rozdiel logaritmu vierohodnosti pre maximálny model a logaritmu vierohodnostnej funkcie. Tvar je nasledujúci

$$\begin{aligned} D = D(y, \hat{\mu}) &= 2\phi \left[l\left(\hat{\beta}_{\max}\right) - l\left(\hat{\beta}\right) \right] \\ &= \sum_{i=1}^{n} 2 \left[y_i \left(\theta_i^* - \hat{\theta}_i\right) - b(\theta_i^*) + b\left(\hat{\theta}_i\right) \right] \\ &= \sum_{i=1}^{n} d_i, \end{aligned} \tag{2.31}$$

kde $l\left(\hat{\beta}_{\max}\right)$ je maximálna vierohodnostná funkcia saturovaného modelu s n parametrami. Je to najväčšia hodnota vo vierohodnosti, akú je vôbec možné dosiahnuť, $l(\hat{\mu}, \phi; y)$ je logaritmická vierohodnostná maximalizácia cez β po opravené hodnoty disperzného parametru ϕ. Potom θ^* a $\hat{\theta}$ znázorňujú maximum vierohodnostných odhadov kanonického parametru pre saturovaný model a model, ktorý študujeme. Parameter $c(y_i, \phi)$ je pre obidva logaritmy vierohodnosti rovnaký, takže sa vo výsledku vyruší.

	Deviácia
Normálne	$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$
Poissonovo	$\sum_{i=1}^{n} 2y_i \log(y_i/\hat{\mu}_i) - 2(y_i - \hat{\mu}_i)$
Binomické	$\sum_{i=1}^{n} 2\left[y_i \log(y_i/\hat{\mu}_i) + (m - y_i) \log\left[(m - y_i)/(m - \hat{\mu}_i)\right]\right]$
Gamma	$\sum_{i=1}^{n} 2\left[-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\right]$

Tabuľka 2.6: Deviácia pre rozdelenia.

S deviáciou je úzko spojená aj *škálovaná deviácia* definovanou vzťahom

$$D^* = D^*(y, \hat{\mu}) = \frac{D(y, \hat{\mu})}{\phi}, \tag{2.32}$$

ktorá závisí na disperznom parametri. V Poissonovon a binomickom rozdelení, u ktorých je disperzný parameter rovný jednotke, je deviácia a škálovaná deviácia rovnaká.

Poznámka 2.3.1. *Ako môžeme vidieť v [6], pre veľký počet dát pre vierohodnostný pomer z rovnice (2.30) asymptoticky platí*

$$D^*(y, \hat{\mu}) \sim \chi^2_{n-k}, \tag{2.33}$$

$$\hat{\phi} = \frac{D}{n-k}.$$

Z definície deviácie môžeme odvodiť *celkový pomer pravdepodobnosti (generalized likelihood ratio)*, ktorý je vyjadrený ako

$$D_0^* - D_1^* \sim \chi^2_{k_1-k_0}, \tag{2.34}$$

kde D_i^* je škálovaná deviácia modelu i, ktorá má k_i parametrov. Toto je však užitočné len v prípade, že poznáme hodnotu disperzného parametru.

V prípade, že hodnotu disperzného parametru nepoznáme, stále máme aproximované výsledky (2.33) a (2.34). Ak $D_0^* - D_1^*$ a D_1^* sú považované za asymptoticky nezávislé, tak sa v [6] dočítame, že to pri veľkom množstve dát implikuje

$$F = \frac{\frac{D_0^* - D_1^*}{k_1 - k_0}}{\frac{D_1^*}{n-k_1}} \sim F_{k_1-k_0, n-k_1}. \tag{2.35}$$

2.3.3 Reziduá

Jednou z najdôležitejších úloh v štatistickom modelovaní je kontrola správnosti modelu. V prípade lineárnych modelov je taká kontrola založená na overovaní reziduí modelu, ktoré obsahujú všetky informácie o údajoch, ktoré nie sú vysvetlené v modelu jeho systematickej časti.

Hlavný dôvod, prečo u zovšeobecnených lineárnych modelov nezjednodušujeme overovanie reziduí, $r_i = y_i - \hat{\mu}_i$, je náročnosť kontroly predpokladaného vzťahu stredných hodnôt a rozptylov reziduí. Preto overovanie správnosti modelov robíme metódou normalizácie reziduí.

Závislú náhodnú veličinu, ktorá je normálne rozdelená, môžeme vyjadriť vo forme $y = \hat{\mu} + (y - \hat{\mu})$, kde $\hat{\mu}$ je fitovaná hodnota a $y - \hat{\mu} = r$ je reziduum.

Existuje viacero typov reziduí, ale uvedieme len tie najznámejšie:

- Pearsonove reziduá,
- Deviačné reziduum.

Pearsonove reziduá

Asi najjednoduchšou cestou k normalizácii reziduí je rozdeliť ich podľa pomeru kvantity k ich štandardnej deviácii podľa fitovaného modelu. Tým získame

Pearsonovo reziduum, ktoré je definované ako

$$\hat{r}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},$$

ktoré by malo mať približne nulovú strednú hodnotu a rozptyl ϕ a kde rozptylová funkcia $V(\mu) \equiv b''(\theta)$ je očakávaný rozptyl. Pearsonova štatistika má tvar

$$\sum_{i=1}^{n} (\hat{r}_i^p)^2 = \chi^2$$

a je analogická k reziduálnej sume štvorcov.

Ak máme všetky regresory diskrétne, tak platí

$$\hat{\phi} = \frac{1}{n-k}\chi^2.$$

Pearsonovu štatistiku však nemôžeme použiť k otestovaniu podmodelu.

Deviačné reziduá

Deviácie u týchto reziduí hrajú podobnú úlohu ako reziduálny súčet štvorcov u klasických lineárnych modelov a majú tvar

$$\hat{r}_i^d = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

kde d_i je i-tý prvok deviácie, viď (2.31).

Suma štvorcov deviačného rezidua je rovná deviácii

$$\sum_{i=1}^{n} \left(\hat{r}_i^d\right)^2 = deviácia = \sum_{i=1}^{n} d_i.$$

Pokiaľ v modeli, v ktorom poznáme všetky parametre, vieme vyčísliť deviácie, potom $D^* \sim \chi_n^2$, z čoho môžeme pre jednoduché dáta napísať $d_i \sim \chi_1^2$ a to implikuje $r_i^d \sim N(0,1)$.

Príklad 2.3.2. Poissonovo rozdelenie má deviačné reziduum v tvare

$$\hat{r}_i^d = \text{sign}(y - \hat{\mu}) \left[2\left(y\log\frac{y}{\mu} - y + \hat{\mu}\right)\right]^{\frac{1}{2}}.$$

Štandardizované deviačné reziduá majú v porovnaní s $\hat{r}_i^d$ navyše jednotkový

rozptyl a sú definované vzťahom

$$\hat{r}_i^{ds} = \frac{\hat{r}_i^d}{\sqrt{\phi\,(1-h_i)}} = \frac{\text{sign}\,(y_i - \hat{\mu}_i)}{\sqrt{\phi\,(1-h_i)}}\sqrt{d_i},$$

kde h_i sú tzv. *páky (leverage)*, ktoré popisujú vplyv i-tého merania na model (1 - veľký vplyv, 0 - malý vplyv). Jedná sa o diagonálne prvky projekčnej matice, ktorú sme uviedli v prípade lineárneho regresného modelu vo vzťahu (2.4).

2.3.4 Binomický model

Najjednoduchší pravdepodobnostný model je binomický logit a probit model, ktorý má závislé premenné len v 2 kategóriach: udalosť A nastala (Y=1), udalosť A nenastala (Y=0). Udalosť A nastáva s pravdepodobnosťou π a nenastáva s pravdepodobnosťou $1 - \pi$.

Hustota binomického rozdelenia je

$$f_i\,(y_i) = \binom{n_i}{y_i}\pi_i^{y_i}\,(1-\pi_i)^{n_i - y_i}\,.$$

Z tabuľky 2.1 v kapitole 2.2.1 máme

$$b\,(\theta_i) = n_i \log\left(1 + e^{\theta_i}\right).$$

Logit model

Kim Changki vo svojom článku [18] modeloval vplyv makroekonomických ukazovateľov logistickou regresiou *(Logit Model)*, kde logistická funkcia, viď [5], má nasledujúci tvar

$$ln\left(\frac{q_s}{1-q_s}\right) = \beta_0 + \beta_1 V_1 + \ldots \beta_n V_n,$$

kde q_s je miera storna, β_i je odhadovaný koeficient a V_i je vysvetľujúca premenná.

Binomické rozdelenie patrí do exponenciálnej triedy rozdelení, viď príklad 2.2.6. Kanonický parameter θ je logit funkcia parametru π_i

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

Logistická regresia je používaná pre binomické rozdelenie a považovaná za model výberu. Tento model sa používa, ak chceme modelovať *dichotomické uda-*

losti (áno, nie). Logistická funkcia je užitočná, pretože zo vstupných parametrov, ktoré majú rozsah od mínus nekonečna do nekonečna, dostaneme výstupné hodnoty v rozsahu od 0 do 1. Logistická regresia analyzuje binomicky rozdelené dáta $y_i = B(n_i, \mu_i)$, kde n_i je počet pokusov a p_i je pravdepodobnosť úspechu. Ak označíme y vysvetľujúcu premennú (rozhodnutie poistníka o ukončení poistnej zmluvy), máme

$$y = \begin{cases} 1, & \text{ak poistník zmluvu predčasne ukončí} \\ 0, & \text{inak} \end{cases}$$

Predpoklady:

- $\eta = \mathbf{X}\beta$
- $\mu_i = g^{-1}(\eta_i)$

Logistická funkcia má tvar

$$g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \eta.$$

Inverzná logistická funkcia má tvar

$$\mu = g^{-1}(\mathbf{X}\beta) = g^{-1}(\eta) = \frac{exp(\eta)}{1 + exp(\eta)}.$$

Fisherov skórovací algoritmus v logistickej regresii

Máme logistický model vyjadrený v tvare

$$\eta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\mu_i}{n_i - \mu_i}\right),$$

čo môžeme prepísať ako

$$\eta_i = \log(\mu_i) - \log(n_i - \mu_i).$$

Derivácia podľa μ_i je

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i} = \frac{n_i}{\mu_i(n_i - \mu_i)} = \frac{1}{n_i \pi_i (1 - \pi_i)}.$$

Závislá premenná, s ktorou pracujeme, je

$$z_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i},$$

čo upravíme na tvar

$$z_i = \eta_i + \frac{y_i - n_i\pi_i}{n_i\pi_i\left(1-\pi_i\right)}.$$

Postupné váhy sú

$$\omega_i = \frac{1}{b''\left(\theta_i\right)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2} = \frac{\left(\pi_i\left(1-\pi_i\right)\right)^2}{n_i\pi_i\left(1-\pi_i\right)} = n_i\pi_i\left(1-\pi_i\right).$$

Binomická deviácia

Predpokladáme, že $\hat{\mu}_i$ je maximálny vierohodnostný odhad parametru μ_i, kde $\hat{\mu}_i = y_i$ v saturovanom modeli.

Z (2.31) vieme, že platí

$$\begin{aligned} D &= 2\sum\left[y_i\log\left(\frac{y_i}{n_i}\right) + \left(n_i - y_i\right)\log\left(\frac{n_i - y_i}{n_i}\right)\right. \\ &\quad \left. -y_i\log\left(\frac{\hat{\mu}_i}{n_i}\right) - \left(n_i - y_i\right)\log\left(\frac{n_i - \hat{\mu}_i}{n_i}\right)\right] \\ &= 2\sum\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \left(n_i - y_i\right)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]. \end{aligned}$$

Probit model

Probit model je nazvaný tiež *štandardné normálne inverzné rozdelenie*. Probitová funkcia má tvar

$$g^{-1}\left(\eta\right) = \phi\left(\eta\right).$$

Kapitola 3

Rešerš

3.1 História

Odstúpenie poistníka od životnej poistnej zmluvy bolo do prvej polovice 20. storočia takmer neznámou témou. Jednými z prvých riešiteľov problému častého stornovania poistky boli Charles F. B. Richardson a John M. Hartwell, ktorí sa tejto problematike začali venovať v polovici 20. storočia. Získali údaje o počte stornovaných poistiek a vydali článok *Lapse Rate* [20], ktorý popisuje, aká je pravdepodobnosť predčasného vypovedania poistnej zmluvy v závislosti na **individuálnych ukazovateľoch** ako je napríklad vek či príjem poistníka. V tomto článku ešte nenájdeme štatistické modely popisujúce mieru storna, ale skôr len pomerové analýzy, preto tento článok nie je významný z hľadiska modelovania, dáva nám však pravdepodobnosť stornovania poistnej zmluvy na základe jednotlivých ukazovateľov. Dôsledkom toho sa budeme na tento článok v rešerši častejšie odkazovať.

Úvodné štúdie boli založené na relatívnej početnosti storien v závislosti na individuálnych ukazovateľov rozdelených do niekoľkých skupín. V 80. rokoch 20. storočia sa postupne začalo modelovať odstúpenie od životnej poistnej zmluvy a tejto problematike sa venovalo stále viac a viac ľudí. V 21. storočí sa začalo modelovať stornovanie poistiek v životnom poistení aj v závislosti na **makroekonomických ukazovateľoch** ako je napríklad miera nezamestnanosti či miera inflácie.

V tejto rešerši je uvedená miera stornovania poistiek v závislosti na jednotlivých individuálnych a makroekonomických ukazovateľov.

3.2 Individuálne ukazovatele

Mieru stornovania poistky môže ovplyvniť mnoho individuálnych ukazovateľov. V minulosti boli skúmané aj tieto ukazovatele: *príjem* poistníka, *povolanie* poistníka, *výška poistného*, *vstupný vek* poisteného pri uzatvorení poistnej zmluvy, *frekvencia platenia poistnej zmluvy*, *pohlavie* poisteného, *trvanie* poistky, *typ* poistky a *agent poisťovne*. Postupne si rozoberieme všetky tieto faktory a zameriame sa na to, aký vplyv má daný ukazovateľ na mieru storna.

3.2.1 Príjem poistníka

Príjem poistníka je podľa Richardsona a Hartwella [20] najdôležitejším faktorom, ktorý ovplyvňuje pravdepodobnosť storna poistiek. Aj v [21] sa dočítame, že príjem poistníka významne vplýva na mieru storna[1]. Väčšina článkov sa však tomuto parametru toľko nevenuje.

Pri modelovaní storna je vhodné faktorizovať premenné príjmu, teda rozdeliť ich do niekoľkých skupín. Jednou možnosťou je rozdelenie ročného príjmu do týchto skupín: menej ako $3000, $3001 až $5000, $5001 až $7500, $7501 až $15000 a nad $15001, ako to použili autori článku [20]. Poistník s príjmom nižším ako $3000 ročne má vysokú pravdepodobnosť storna poistnej zmluvy[2]. Môžeme teda zhrnúť, že poistník s vyšším príjmom je menej náchylný poistku ukončiť predčasne.

V dnešnej dobe by však bolo vhodnejšie použiť iné rozdelenie na faktory, pretože od prvej polovice 20. storočia sa hodnota doláru v dôsledku nielen inflácie posunula a $3000 ako ročný príjem nie je v dnešnej dobe reálny.

3.2.2 Povolanie poistníka

Parameter *povolanie* sa ťažko delí do menších skupín. Musíme brať do úvahy takmer každé povolanie ako samostatnú skupinu. Do faktorov môžeme spojiť len také povolania, ktoré majú rovnaký charakter práce a približne rovnaký plat, napríklad skupina poľnohospodárov a robotníkov. Rozdiely v stornovaní medzi povolaniami s rovnakými príjmami môžeme vidieť napríklad na povolaniach poľnohospodárov a študentov. Relatívna početnosť ukončenia poistky poľnohospodárov je o dosť vyššia než napríklad u študentov[3]. Najväčšie rozdiely

[1] Autori modelovali storná pomocou redukovanej formy logistickej regresie. Parameter je významný na hladine 1 %.

[2] 24,6 % z 20483 poistných zmlúv v období od 1. januára do 30. júna 1946.

[3] V článku [20] študenti stornujú poistnú zmluvu s pravdepodobnosťou 10,0 %, predavači s pravdepodobnosťou 31,7 %. Tieto percentá sú všeobecné pre všetky príjmové skupiny.

v skupinách zamestnaní sú v príjmovej triede $3001-$5000[4]. V triedach príjmu 1-3 platí, že čím vyšší plat platca poistného poberá, tým sa rozdiely v stornovaní medzi triedami zamestnania zmenšujú. V triede vysokého príjmu sú hodnoty nevyspytateľné. Hodnoty sú odzrkadlením nedostatočných údajov, a teda výsledné hodnoty sú nepresvedčivé.

Povolanie a príjem sú v istej korelácii. Avšak existujú povolania, ktoré majú rôzne postavenia a teda poistníci s tým istým povolaním majú úplne rôzne príjmy, napríklad sa jedná o povolanie junior a senior. To nám trochu problematizuje modelovanie storna v závislosti na parametri *povolanie*, pretože poistník s povolaním junior bude stornovať poistnú zmluvu s väčšou pravdepodobnosťou, než poistník s povolaním senior, a to z dôvodu, že sa správajú podľa *príjmu* poistníka. Z toho dostávame záver, že parameter *povolanie* nemá tak veľký vplyv na stornovanie ako parameter *príjmu* poistníka.

3.2.3 Výška poistného

Výška poistného je relevantný faktor hlavne v neživotnom poistní, ktorý sa môže v priebehu poistenia meniť. Z [1] vieme a ako by sme aj očakávali, zdražením poistného sa zvyšuje pravdepodobnosť storna. V životnom poistení sa poistné pre už uzavreté poistné zmluvy nezvyšuje, ale v tomto odvetví môžeme pozorovať poistníkovo správanie v prípade, že sa na trhu objaví poistenie, ktoré je s rovnakými parametrami lacnejšie než má klient poisťovne v danej chvíli zjednané. V takom prípade sa pravdepodobnosť storna zvýši.

3.2.4 Vstupný vek poistníka pri uzatvorení poistky

Parameter *vstupný vek* poistníka je jeden z najvýznamnejších pri modelovaní storna, viď [19][5]. Aj v [21] sa dočítame, že vstupný vek poistníka významne vplýva na mieru storna[6]. Vo všeobecnosti môžeme tvrdiť, že čím nižší *vek* poistníka, tým vyššia pravdepodobnosť vypovedania poistnej zmluvy. Medzi *vekom* poistníka a jeho *príjmom* je istá korelácia. Pri vyššom vstupnom veku má poistník pravdepodobne väčší príjem a teda nižšiu pravdepodobnosť stornovania zmlúv. Vplyv týchto faktorov na pravdepodobnosť storna je korelovaný. Vo vyšších príjmových skupinách vek prestáva ovplyvňovať rozhodnutie poistníka o ukončení poistky.

[4]Študenti stornujú poistnú zmluvu s pravdepodobnosťou 7,2 %, predavači s pravdepodobnosťou 30,7 %.

[5]Autori použili binomickú odozvu s logit transformačnou funkciou zovšeobecnených lineárnych modelov.

[6]Redukovaná forma logistickej regresie. Parameter je významný na hladine 1 %.

Pravdepodobnosť storna s rastúcim vekom poisteného klesá, ako sa zhodujú články [18], [13][7] a [9][8]. Aj v článku [1] sa dozvedáme, že mladší poistníci medzi vekmi 20-29 stornujú poistku podstatne častejšie než starší poistníci. Je to pravdepodobne z toho dôvodu, že mladší poistníci majú viac času a nadšenia pri hľadaní lepšej ponuky, a možno tiež ako výsledok všeobecne horšej finančnej situácie mladých ľudí, a preto majú väčší záujem o nájdenie konkurenčnej ceny, ako sa môžeme dočítať v článku [11].

Rozdiel relatívnej početnosti ukončenia poistky medzi vekovou skupinou 10-19 rokov a skupinou 20-29 rokov je zaujímavý. Stornovanie poistky poisteného vo veku 20-29 je vyššie, ako vo veku 10-19, a to z dôvodu, že osoba vo veku 10-19 rokov takmer nikdy nie je platcom poistného, ale platia ju rodičia, ktorí sú väčšinou vo vyššej triede príjmu.

Renshaw a Haberman v [19] rozdelili *Vstupný vek poisteného* na 3 kategórie: 15-29, 30-39 a 40-64 rokov. Miera storna má tendenciu mierne klesať s rastúcim vekom pre poistenia pre prípad dožitia a dočasných poistení. V kategórii 40-64 rokov je tento pokles najvýraznejší.

3.2.5 Frekvencia platenia poistky

Vplyv *frekvencie platenia poistky* je najviac rozdielny pre poistníkov s nižším príjmom, ako sa môžme dočítať v [20]. Poistka s mesačnou frekvenciou platenia má najvyššiu pravdepodobnosť stornovania a pravdepodobnosť je menšia pre štvrťročnú, polročnú frekvenciu a najnižšia pre ročnú frekvenciu platenia[9].

3.2.6 Pohlavie poisteného

Životná poistná zmluva má pravdepodobnosť dlhšieho trvania u žien než u mužov [20], ale rozdiely sú takmer zanedbateľné[10]. To, že parameter pohlavia na relatívnu početnosť ukončenia poistnej zmluvy nemá žiaden vplyv, sa môžeme dočítať v článkoch [11], [13].

[7] Pravdepodobnosť storna v prvom roku poistenia pre vekovú skupinu 25-34 rokov je 11,8 %, pre vekovú skupinu 35-44 rokov je to 9,6 %, pre vekovú skupinu 45-54 rokov je to 7,8 % a pre 55-64 rokov je to 5,6 %.

[8] Vek je v článku rozdelený na 3 skupiny: 15-29 rokov, 30-39 rokov, 40-64 rokov.

[9] Pre mesačnú frekvenciu platenia je pravdepodobnosť storna 25,1 %, pre štvrťročnú je 24,6 %, pre polročnú je 21,3 % a pre ročnú frekvenciu platenia je pravdepodobnosť storna 14,6 %.

[10] Celková pravdepodobnosť storna pre ženy je 17,3 % a pre mužov 18,6 %.

3.2.7 Fajčiar / nefajčiar

Lebel vo svojich štúdiách [13] zisťoval rozdiely v ukončovaní poistky medzi fajčiarom a nefajčiarom. Došiel k záverom, že fajčiari častejšie stornujú životnú poistnú zmluvu. V prvom roku poistenia je pravdepodobnosť stornovania viac než dvojnásobná[11]. V neskorších rokoch poistenia sa rozdiel zmenšuje[12].

3.2.8 Trvanie poistky

Lebel v článku [13] skúmal pravdepodobnosť odstúpenia od poistnej zmluvy podľa doby trvania v závislosti na kalendárnom roku[13]. Síce sú isté rozdiely v stornovaní poistky v závislosti na kalendárnom roku, ale miera storna konverguje v neskorších rokoch poistenia.

Podľa [18] a [4] však miera storna v prvých piatich rokoch trvania poistky mierne stúpa. V šiestom a siedmom roku je pravdepodobnosť stornovania vyššia než v predchádzajúcich rokoch a v prvých troch mesiacoch ôsmeho roku je pravdepodobnosť stornovania najvyššia, takmer 16 percent. Potom miera storna klesá. Takmer 30 percent z ukončených poistných zmlúv je stornovaných práve v ôsmom roku jej trvania, viď [18]. Dôvody takéhoto rozhodovania popísali v článku [4] S. H. Cox a Y. Lin, ktorí toto poistníkovo rozhodovanie vysvetľujú ako dôsledok nulovosti poplatkov za stornovanie poistky v ôsmom roku poistenia, viď podkapitolu 3.3.6. Poistník si môže uvedomovať, že ak odloží rozhodovanie o predčasnom ukončení poistnej zmluvy o rok, poplatky za stornovanie sa znížia. Poistníkovi sa teda vyplatí počkať. Ale v ôsmom roku poistenia, kedy sa poplatky za odstúpenie od zmluvy dostanú na nulu, poistník už nemá dôvod kvôli poplatkom so stornom poistky vyčkávať. Je však potrebné upozorniť, že tento záver dostávame z amerických dát, kde je poisťovníctvo na inej úrovni ako v Českej republike, viď podkapitolu 3.3.6.

Autori článku [11] tvrdia, že pravdepodobnosť storna je výrazne vyššia do 10 rokov trvania poistnej zmluvy než v neskoršom období trvania poistky[14]. Tí poistní klienti, ktorí poistnú zmluvu nevypovedali v prvých desiatich rokoch trvania poistky pravdepodobne už poistnú zmluvu nevypovedajú. Pokiaľ však modelujeme trvanie poistky podľa rôznych typov poistnej zmluvy, ako to môžeme vidieť v článku, miera storna sa takmer nemení.

[11] V prvom roku poistenia je pravdepodobnosť stornovania poistnej zmluvy pre fajčiara 15,8 % a pre nefajčiara 7,6 %.

[12] Napríklad v šiestom roku poistenia fajčiar stornuje s pravdepodobnosťou 4,8 % a nefajčiar 3,2 %.

[13] Doba trvania poistky po rokoch 1 až 13 rokov v kalendárnych rokoch 1994-1998.

[14] Skúmané obdobie: 1991 - 2007. Počet zozbieraných údajov: 279 000 stornovaných poistiek z celkového počtu 6 129 000 poistiek.

Autori článku [4] modelovali mieru storna v závislosti na trvaní poistky *Tobit modelom*. Pokiaľ modelujeme stredné hodnoty miery storna, Tobit model dobre popisuje hodnoty u všetkých dôb trvania zmlúv. Avšak pri použití modelu je potrebné byť opatrný, pokiaľ chceme predpovedať medián miery storna podľa dĺžky trvania poistky. Tobit model totiž medián nadhodnocuje[15].

Renshaw a Haberman v [19] a [9] rozdelili *Trvanie poistky* do 3 kategórií: krátka (1-3 roky trvania), stredná (4-8 rokov trvania) a dlhá kategória (9 a viac rokov trvania). V tomto článku je tento parameter jeden z najdôležitejších, ktorý ovplyvňuje mieru storna[16]. Podľa článku [19], v ktorom autori porovnávali vzájomnú kovarianciu parametrov na zovšeobecnené lineáre modely, je najvýznamnejšia práve interakcia medzi typom poistky a dobou trvania[17]. Poistky krátkeho trvania (2-3 roky) sú najviac náchylné k zániku. Sklon k zániku sa znižuje s rastúcou dobou trvania poistnej zmluvy.

3.2.9 Typ poistnej zmluvy

Zo sekcie trvania poistky vieme, že čím je dlhšia doba trvania poistky, tým viac klesá pravdepodobnosť storna. Renshaw a Haberman v článku [19] a [9] porovnávali storno u týchto typov poistenia:

- poistenie v prípade dožitia s podielom na zisku (*with-profit endowment policy*),
- poistenie v prípade dožitia bez podielu na zisku (*non-profit endowment policy*),
- poistenie v prípade smrti s podielom na zisku (*with-profit whole-life policy*),
- poistenie v prípade smrti bez podielu na zisku (*non-profit whole-life policy*),
- dočasné poistenie (*temporary policy*),
- investičné poistenie (*unit-linked policy*).

Interakciu medzi *typom poistky* a dobou trvania poistky modelovali binomickou odozvou GLM[18]. Pre prvý typ poistky pravdepodobnosť storna bola vyššia

[15]Medián skúmaných dát je pre ôsmy rok poistenia 0,291 a pre odhad parametru modelom Tobit je 0,317. Stredná hodnota je rovnaká pre dáta a pre model a to 0,319 pre ôsmy rok poistenia.

[16]V porovnaní s typom poistky, vstupným vekom a typom agenta má doba trvania najväčší vplyv na stornovanie poistnej zmluvy. Vstupný vek a typ poistky vplývajú na mieru storna približne rovnako.

[17]Autori použili test významnosti v zovšeobecnených lineárnych modeloch pri normálnom rozdelení pozorovaní a binomickom rozdelení s logit transformačnou funkciou.

[18]Zovšeobecnené lineárne modely (Generalized Linears Models).

počas trvania poistky 4-8 rokov, než u poistky pre prípad dožitia bez výplaty plnenia pre dobu trvania 0-3 roky, čo je proti trendu u ostatných typov poistiek. Možné vysvetlenie je také, že dáta sú zaznamenané v dobe, kedy bola priemerná dĺžka vykazovania približne 7 rokov.

Renshaw a Haberman v [19] došli k záveru, že poistenia bez podielu na zisku majú mieru storna vyššiu než poistenia s podielom na zisku, a to či pre poistenie pre prípad dožitia alebo pre prípad smrti. Dočasné poistky vykazujú podobnú mieru storna ako poistky bez podielu na zisku. Investičné životné poistenia pre mladú vekovú skupinu majú vyššiu mieru storna než iné poistenia, a vo veku 40-54 rokov je miera storna podobná ako u poistenia s podielom na zisku.

3.2.10 Agent poisťovne

Zo súhrnu môžeme vidieť, že niektoré faktory sú na sebe viac alebo menej závislé. Je tu však ešte jeden dôležitý faktor, a tým je charakteristika agenta predávajúceho poistnú zmluvu, viď [19]. Agent zastáva dôležitú úlohu v záujmoch poisťovne a jeho motivácia predania zmluvy je kľúčová. Nový agent má štatisticky uzavrených viac zmlúv, ktoré se neskôr stornujú. Pokiaľ je motivácia agenta len osobný peňažný zisk pre seba, bude sa snažiť predať aj nevýhodné poistky, ktoré klienti poisťovne možno nebudú schopní splácať, a tým sa zvýši pravdepodobnosť predčasného vypovedania poistnej zmluvy. Ďalší spôsob zisku agenta je v pretáčaní zmlúv, čo znamená, že agent presvedčí poistníka, aby poistku stornoval a uzavrel inú zmluvu.

V [19] sa autori zaoberali modelovaním *veku poisteného*, *dĺžkou trvania* poistnej zmluvy, *typom* poistenia a *agentom*. *Agent* poisťovne bol z týchto štyroch faktorov najmenej významný. To však neznamená, že pravdepodobnosť stornovania poistnej zmluvy poistníkov od rôznych agentov nie je významným parametrom a nie je potrebné tento parameter modelovať.

3.3 Makroekonomické ukazovatele

Nielen individuálne, ale aj makroekonomické ukazovatele majú vplyv na rozhodovanie poistníka o predčasnom ukončení poistnej zmluvy. Existujú poistníci, ktorí životné poistenie nepovažujú len za pomoc pri úraze či zabezpečení rodiny v prípade smrti, ale životné poistenie používajú aj ako sporenie. Pokiaľ však poistenec má zjednané investičné poistenie a výnosnosť zmluvy je menšia než očakával, poistník je náchylný poistnú zmluvu ukončiť a hotovosť uložiť do výnosnejšej alternatívy. Ďalším dôvodom pre stornovanie poistnej zmluvy je

náhla stráta práce a potreba získať hotovosť, ktorú môže klient dostať práve predčasným ukončením poistnej zmluvy a vyplatením odbytného. Hlavné príčiny zapríčiňujúce mieru storna, ktorými sa v práci zaoberáme, sú:

- *miera nezamestnanosti*,
- *úroková miera*,
- *miera hospodárskeho rastu*,
- *sezónny efekt*,
- výskyt *finančného šoku*
- a *poplatky* za zrušenie poistky.

3.3.1 Miera nezamestnanosti

Odhad parametru miery nezamestnanosti logistickou regresiou podľa [18] má vysokú hodnotu[19]. To znamená, že miera storna se mení pozitívne v závislosti na pohybe miery nezamestnanosti - pri rastúcej miere nezamestnanosti rastie aj stornovanie poistných zmlúv. Dobré ekonomické podmienky majú pre poistiteľa pozitívny vplyv na rozhodovanie poistníka, a teda pravdepodobnosť stornovania poistky je menšia.

V článku [4], kde autori modelovali mieru storna pomocou *Tobit modelu*, sa dozvedáme, že miera nezamestnanosti je významne negatívne korelovaná s úrokovou mierou. Vysoko korelované premenné sú redundantné a spôsobujú *problém kolinearity*[20], preto tento parameter pri modelovaní autori [4] vynechali.

Vplyv miery nezamestnanosti na odstúpenie od poistnej zmluvy je významný ako v krátkom aj dlhom období trvania poistnej zmluvy. Tento poznatok je popísaný v článku [2].

3.3.2 Úroková miera

O vplyve úrokovej miery na stornovanie poistnej zmluvy sa dočítame napríklad v [2]. V tomto článku nachádzame, že vplyv úrokovej miery na stornovanie poistky nie je významný pri krátkom trvaní poistnej zmluvy. Avšak pri dlhom trvaní poistnej zmluvy je tento vplyv už o dosť významnejší.

[19]Kórejské dáta: 50.6348, americké dáta: 24.3694 pri logistickej regresii, 3 roky trvania poistnej zmluvy.

[20]Sundberg R.: *Collinearity*, Volume 1, Encyclopedia of Environmetrics, John Wiley & Sons, Ltd, 2002, strany 365-366.

V porovnaní s mierou nezamestnanosti, má táto väčší dopad na storno zmluvy pri krátkom trvaní poistky. Úroková sadzba má však významnejší ekonomický dopad na mieru storna než miera nezamestnanosti. Poistník ďaleko viac reaguje na náhodnú zmenu (šok) v úrokových sadzbách, než na náhodnú zmenu v miere nezamestnanosti.

Ak berieme do úvahy rozdiel medzi referenčnou úrokovou sadzbou na trhu a technickou úrokovou mierou, tak miera storna začne rásť, keď tento rozdiel začne byť pre trh významný, viď [18].

Pri uzatváraní životnej poistnej zmluvy nám poisťovňa garantuje tzv. technickú úrokovú mieru. Čím je táto nižšia, tým je pravdepodobnosť uzatvárania zmlúv menšia než pri vyššej technickej úrokovej miere. Ak už poistník má uzavretú poistnú zmluvu, tak pokiaľ poisťovne začnú garantovať pre novouzavreté poistné zmluvy vyššiu technickú úrokovú mieru, poistník s vyššou pravdepodobnosťou ukončí súčasnú poistnú zmluvu a uzatvorí novú, pre neho výhodnejšiu poistnú zmluvu. Je teda väčšia pravdepodobnosť, že novozískané zmluvy budú poskytovať rovnaké pokrytie za nižšie poistné. Poistenci majú tendenciu stornovať svoju aktuálnu poistnú zmluvu, aby využili vyššie výnosy alebo nižšie poistné, ktoré je na trhu k dispozícii.

3.3.3 Miera hospodárskeho rastu

Miera storna závisí negatívne na miere hospodárskeho rastu, teda pravdepodobnosť stornovania poistky klesá pri dobrých ekonomických podmienkach, ako sa píše v [18][21].

3.3.4 Sezónny efekt

Podľa článku [18] sú niektoré odhadované parametre logistickej regresie pre sezónny efekt kladné a niektoré záporné. A preto síce dokážeme povedať, v ktorom mesiaci sa trochu častejšie stornujú zmluvy, avšak všetky hodnoty sú malé. Z toho získavame záver, že sezónny efekt má len malý vplyv na predčasné ukončenie poistných zmlúv.

3.3.5 Finančný šok

Kim Changki sa vo svojom článku [18] zaoberal modelovaním ukazovateľov v *extrémnych* podmienkach, ktoré definoval takto:

[21] Kórejské dáta: -5.3360, americké dáta: -2.6450 pri logistickej regresii, 3 roky trvania poistnej zmluvy.

Definícia 3.3.1. *Extrémne podmienky* znamenajú viac ako dve štandardné odchýlky od očakávanej úrovne za rôznych ekonomických podmienok a v kombinácii s rôznymi charakteristikami poistky.

V období extrémnych podmienok (alebo náhlych zmien na finančnom trhu) sa pravdepodobnosť storna mení viac, než by sme očakávali, a stornovosť rastie, pokiaľ na trhu nastane finančný alebo ekonomický šok. Napríklad počas finančnej krízy v Južnej Kórei v decembri 1998 miera storna vykazovala náhle stúpanie.

Veľký vplyv na sféru poisťovníctva má aj dnešná finančná kríza. Poistení klienti poisťovne, ovplyvňovaní informáciami o stave finančných trhov sa na základe nie príliš optimistických prognóz ďalšieho vývoja často rozhodnú poistnú zmluvu vypovedať.

3.3.6 Poplatky

Parameter *poplatky* nepatrí úplne medzi makroekonomické ukazovatele, ale skôr medzi racionálne správanie poistníkov. Pre jednoduchosť však ponecháme tento popis parametru poplatky v kapitole makroekonomických ukazovateľov.

Poisťovňa poistníkovi účtuje rôzne poplatky za vedenie životného poistenia. V článku [14] sa dozvedáme, že motivácia stornovať poistku je vysoká, pokiaľ:

$$PV(\text{benefitov}) < PV(\text{poplatkov za vedenie poistnej zmluvy}),$$

kde PV je *súčasná hodnota (present value).*

Ak sa navýši poplatok, potom poistník okamžite stornuje poistku. V prípade účtovania poplatku za zrušenie zmluvy sa táto motivácia znižuje, teda motivácia stornovať poistku je vysoká, pokiaľ

$$\begin{aligned} PV(\text{benefitov}) \quad < \quad & PV(\text{poplatkov za vedenie poistnej zmluvy}) \\ & + \text{Poplatok za stornovanie poistky} \end{aligned}$$

Stornovací poplatok

Väčšina kontraktov pripisuje celé poistné na účet kapitálovej hodnoty a posudzuje stornovací poplatok až v okamihu, kedy poistník od poistnej zmluvy odstúpi. Poplatok za stornovanie poistky *(surrender charge)* sa mení v závislosti na trvaní poistnej zmluvy. Väčšinou v prvom roku je okolo 7%, a každým ďalším rokom trvania poistky klesá až na 0%, viď [4]. Podľa [18] sa veľkosť poplatkov

pohybuje väčšinou medzi 7-10 percentami z hodnoty účtu a klesá k nule u 6- až 10- ročných poistných zmlúv.

Na tomto mieste je potrebné zdôrazniť, že všetky uvedené články o stornovacích poplatkoch popisujú americké dáta, kde forma poistenia je iná než forma v Českej republike. V USA sa pri stornovaní poistnej zmluvy vypláca poistníkovi odkupné očistené o poplatky. V Českej republike sa odkupné vypláca až po 2 rokoch zaplateného poistného, kedy sa splatia poisťovni všetky poplatky náklady na zjednanie zmluvy. To znamená, že v USA je možné pre klienta poisťovne získať nejakú výplatu z poistenia i pri stornovaní poistnej zmluvy v prvom či druhom roku poistenia. V Českej republike sa klientovi vyplatí odbytné väčšinou až od tretieho roku trvania poistnej zmluvy podľa svojich nárokov. Pre českú poisťovňu, ktorá v prvých dvoch rokoch poistenia ešte nemá od klienta splatené všetky náklady, je najväčšie riziko stornovania poistnej zmluvy do dvoch rokov poistenia, pretože je stratová. Z toho dôvodu sa my v praktickom modeli zameriame na stornovanie poistných zmlúv v prvých dvoch rokoch poistenia.

Kapitola 4

Aplikácia na dátach

V tejto kapitole ukážeme použitie zovšeobecnených lineárnych modelov na dátach. Nájdeme model, ktorý najlepšie popisuje dáta, a odhady parametrov porovnáme s pomerovou analýzou.

Spočiatku sme chceli získať dáta od českých poisťovní, a teda modelovať storno na reálnych dátach. Všetky poisťovne sú však citlivé na svoje dáta a odmietajú ich poskytnúť verejnosti. Z toho dôvodu sme sa rozhodli použiť generované dáta a na nich vysvetliť a model a prácu v programe `R`. Je preto potrebné zdôrazniť, že výsledky, ktoré dostaneme, nemusia zodpovedať realite.

Rozhodli sme sa pre generovanie dát reprezentujúcich ukončenie životného poistenia počas prvých dvoch rokoch od uzavretia poistnej zmluvy. Ako sme už uvádzali v rešerši, pre poisťovňu je najväčšia strata, pokiaľ klient stornuje poistnú zmluvu práve v prvých dvoch rokoch poistenia, pokiaľ ešte poisťovni nesplatil všetky náklady, ktoré súvisia s uzavretím poistenia.

Dáta, ktoré máme k dispozícii, obsahujú skúsenosti v rozmedzí 10 rokov a 851 955 zmlúv.

V nasledujúcom zozname nájdeme skúmané faktory tak, ako by boli uvedené klientom alebo boli inak známe pri uzatváraní zmlúv:

- pohlavie,
- vek,
- manželský stav,
- deti,
- mesačný zárobok v tisícoch korún,

- typ zjednaného poistenia,
- poistná čiastka v tisícoch korún,
- sprostredkovateľ poistnej zmluvy,
- rok zjednania,
- sídlo
- a informácia o stornovaní poistnej zmluvy v prvých dvoch rokoch.

4.1 Klasifikácia dát

V prvom rade je potrebné dáta istým spôsobom rozdeliť do tried, tzv. faktorizovať. Dôležité je rozdelenie do tried najmä takých veličín, ktoré majú veľkú doménu, ako je vek či poistná čiastka. V dátach máme ľudí vo veku 18-69. Ak by sme mali ľudí rozdelených po vekoch, mali by sme 52 tried, čo sa ťažko modeluje. Jednoduchšie je rozdeliť si ľudí do vekových kategórii po 5 alebo 10 rokov. Ľudí by sme mali rozdeliť do tried podľa toho, ako sa líšia pravdepodobnosti stornovania ľudí v rôznych vekových kategóriách. Taktiež záleží na počte ľudí v triede. Pokiaľ máme málo mladých ľudí, málo seniorov, ale veľký počet ľudí stredného veku, rozdelíme ľudí tak, aby v každej triede bol približne rovnaký počet ľudí.

Naše dáta sú však vytvorené umelo. Už proces generovania dát bol ovplyvnený tým, aké rozdelenie do tried sme použili. Z toho dôvodu nebudeme popisovať hľadanie rozdelenia do tried, ale použijeme triedy z generovaných dát. Rozdelenie do tried je nasledujúce:

- `S`: pohlavie (Muž, Žena),
- `A`: vek (A1: 18-29, A2: 30-39, A3: 40-49, A4: 50-59, A5: 60+),
- `M`: manželský stav (M0: slobodný/rozvedený, M1: ženatý/vydatá),
- `C`: deti (C0: žiadne, C1: 1 a viac),
- `E`: mesačný zárobok v tisícoch korún (E1: < 10; E2: $\langle 10, 20)$; E3: $\langle 20, 30)$; E4: 30+),
- `T`: typ zjednaného poistenia, (T1: smrť bez podielu na zisku, T2: smrť s podielom na zisku, T3: dožitie bez podielu na zisku, T4: dožitie s podielom na zisku, T5: investičné poistenie),

- `I`: poistná čiastka v tisícoch korún (I1: $(0, 500)$, I2: $\langle 500, 1000)$, I3: 1000+),
- `O`: sprostredkovateľ poistnej zmluvy / distribúcia (O1, O2, O3, O4, O5),
- `Y`: rok zjednania (kalendárny rok zjednania: Y1: 1996-1997, Y2: 1998-1999, Y3: 2000-2001, Y4: 2002-2003, Y5: 2004-2005),
- `R`: sídla (počet obyvateľov žijúcich v rovnakom sídle: R1: $< 10\ 000$, R2: $\langle 10\ 000, 50\ 000)$, R3: $\langle 50\ 000, 100\ 000)$, R4: $> 100\ 000$)
- a `Lapse`: príznak o stornovaní poistnej zmluvy v prvých dvoch rokoch (0 - nie, 1 - áno).

Pokiaľ máme v dátach parameter ako je manželský stav či počet detí a nie je potrebné daný parameter rozdeliť do tried, pretože má malý počet skupín. Také parametry môžu mať hodnoty 0, 1. Problém však môže nastať pri modelovaní v programe `R`, ktorý ich môže považovať za hodnoty a nie triedy. Z toho dôvodu použijeme v programe `R` funkciu `factor` a teda prevod na triedy[1], viď kód 4.1.

```
> tabulkaStoren <- read.table("tabulkaStorien.csv",
header=TRUE,sep=";")
> tabulkaStoren$S<-factor(tabulkaStoren$S)
> tabulkaStoren$A<-factor(tabulkaStoren$A)
> tabulkaStoren$M<-factor(tabulkaStoren$M)
> tabulkaStoren$E<-factor(tabulkaStoren$E)
> tabulkaStoren$T<-factor(tabulkaStoren$T)
> tabulkaStoren$I<-factor(tabulkaStoren$I)
> tabulkaStoren$O<-factor(tabulkaStoren$O)
> tabulkaStoren$C<-factor(tabulkaStoren$C)
> tabulkaStoren$Y<-factor(tabulkaStoren$Y)
> tabulkaStoren$R<-factor(tabulkaStoren$R)
> attach(tabulkaStoren)
```

Kód 4.1: Rozdelenie do skupín v programe `R`.

Ak už máme dáta načítane a rozdelené v skupinách, pozrieme sa na celkovú štatistiku dát pomocou funkcie `Summary` v kóde 4.2. Vo výstupe vidíme počty ľudí rozdelených do skupín.

Rovnomernému rozdelenie dát len potvrdzuje, že dáta sú vytvorené umelo. Máme polovicu mužov a žien (`M`: 425977 a `F`: 425978) a ľudí rozdelených rovnomerne do veľkosti sídla (parameter `R`). Taktiež si môžeme všimnúť, že program `R`

[1]Po použití funkcie `factor` už nebude program `R` považovať číslice 0 a 1 manželského stavu za hodnoty, ale ako triedy.

```
> summary(tabulkaStoren)
       ID          S            A               M
 Min.   :     1   F:425978   A1:170442   M0:114072
 1st Qu.:212990   M:425977   A2:170809   M1:737883
 Median :425978              A3:170319
 Mean   :425978              A4:169965
 3rd Qu.:638967              A5:170420
 Max.   :851955

  E           T           I           O           Y
 E1:213392   T1:170094   I1:284044   O1:170399   Y1:170460
 E2:212193   T2:170795   I2:284576   O2:170479   Y2:171140
 E3:213084   T3:170203   I3:283335   O3:169635   Y3:169583
 E4:213286   T4:170357               O4:170866   Y4:170359
             T5:170506               O5:170576   Y5:170413

    C         R              Lapse
C0:114467   R1:212988   Min.   :0.0000
C1:737488   R2:212989   1st Qu.:0.0000
            R3:212989   Median :0.0000
            R4:212989   Mean   :0.2029
                        3rd Qu.:0.0000
                        Max.   :1.0000
```

Kód 4.2: Funkcia `Summary` pre všetky generované dáta.

nepovažuje hodnoty parametru `Lapse` za skupiny, ale počíta napríklad maximálnu hodnotu, medián či strednú hodnotu.

4.2 Zoskupenie dát a práca v programe R

Ak chceme modelovať dáta, potrebujeme ich zoskupiť do skupín. To znamená nájsť všetky skupiny ľudí s rovnakými parametrami (rovnaké pohlavie, rovnaká kategória veku, manželského stavu, ...) okrem parametru `Lapse`. Do tabuľky pridáme počet stornovaných zmlúv `LapseNum`, sumu parametru `Lapse`, a celkový počet položiek v skupine `PolNum` a načítame data do programu `R`. Prvých 5 riadkov tabuľky je uvedených v kóde 4.3.

Ako **Skúmané kategórie** označme tie kategórie, ktoré máme v modelovaných datách.

Označme ako **skupinu ľudí s rovnakými kategóriami** takú skupinu ľudí, ktorí patria do rovnakých *skúmaných kategórií* (napríklad majú rovnaký vek, sú v rovnakej príjmovej kategórii,..).

```
> tabulkaStorenBezInterakci[1:5,]
  S  A  M  E  T  I  O  Y  C  R LapseNum PolNum
1 F A1 M0 E1 T1 I1 O1 Y1 C0 R3        5      9
2 F A1 M0 E1 T1 I1 O1 Y1 C0 R4        2      9
3 F A1 M0 E1 T1 I1 O1 Y1 C1 R3        2      8
4 F A1 M0 E1 T1 I1 O1 Y1 C1 R4        1      7
5 F A1 M0 E1 T1 I1 O1 Y2 C0 R3        0      4
```

Kód 4.3: Pohľad na prvých 5 riadkov tabuľky.

Poznámka 4.2.1. *V kóde 4.3 sú skúmané kategórie pohlavie, vek, manželský stav, zárobok, zjednané poistenie, poistná čiastka, distribúcia, rok zjednania, počet detí a sídlo (počet obyvateľov mesta, v ktorom dotyčný býva). Skupina ľudí s rovnakými kategóriami je v kóde 4.3 napríklad prvý riadok tabuľky. Prvá skupina ľudí s rovnakými kategóriami sú ženy, ktoré patria do prvej vekovej kategórie, sú bezdetné, slobodné alebo rozvedené, ich mesačný príjem je menší než 10 tisíc Kč, majú zjednané poistenie pre prípad smrti bez podielov na zisku s poistnou čiastkou do 500 tisíc Kč, zjednali si poistenie v distribúcii O1 v roku 1996-1997 a bývajú v bydlisku s počtom obyvateľov menším než 10 tisíc ľudí. Takýchto ľudí sme v súbore dát mali 9 (viď v prvom riadku kódu 4.3 stĺpec* `PolNum`*), z ktorých 5 ľudí poistenie stornovalo (viď* `LapseNum`*).*

Ak chceme vidieť, ktoré parametre majú aký vplyv na stornovanie, skúsime model súčtu jednotlivých parametrov, viď kód 4.4, a pozrieme sa na zhrnutie, kroré nám ponúka `R` v kóde 4.5.

```
> glm1 <- glm(cbind(LapseNum,PolNum - LapseNum) ~ S + A + M +
    E + T + I + O + Y + C + R,
    family = binomial(link = "logit"),
    data = tabulkaStorenBezInterakci)
```

Kód 4.4: Jednoduchý model súčtu parametrov

Funkcia `Summary` poskytuje dostatok informácii o modeli, viď kód 4.5. Položka `Call` nám len pripomína volanie premennej, ktorú sumarizujeme. `(Deviance) Residuals` nám udávajú zhrnutie minima `Min`, prvého kvantilu `1Q`, mediánu `Median`, tretieho kvantilu `3Q` a maxima `Max` reziduí.

Koeficienty `Coefficients` sa vzťahujú k odhadovaným parametrom. Prvé dva stĺpce `Estimate` a `Std. Error` sa vzťahujú k odhadom $\hat{\beta}$. Stĺpec `Std. Error` je v teórii vyjadrený vo vzorci (2.28). Tretí stĺpec `z value` je *z-hodnota*, teda odhad vydelený štandardnou chybou

$$T_j = \frac{\texttt{Estimate}}{\texttt{Std. Error}}.$$

```
> summary(glm1)
Call:
glm(formula = cbind(LapseNum, PolNum - LapseNum) ~ S + A + M +
 E + T + I + O + Y + C + R, family = binomial(link = "logit"),
 data = tabulkaStorenBezInterakci)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0587  -0.8562  -0.2993   0.6446   4.7274

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.9991239  0.0148728 -67.178  < 2e-16 ***
SM          -0.0014545  0.0077212  -0.188    0.851
AA2         -0.0808570  0.0096601  -8.370  < 2e-16 ***
AA3         -0.3318490  0.0103275 -32.133  < 2e-16 ***
AA4         -0.2814784  0.0105178 -26.762  < 2e-16 ***
AA5         -0.3467333  0.0107622 -32.218  < 2e-16 ***
MM1         -0.0654888  0.0090165  -7.263 3.78e-13 ***
EE2         -0.0926543  0.0074820 -12.384  < 2e-16 ***
EE3         -0.2256437  0.0076262 -29.588  < 2e-16 ***
EE4         -0.3187816  0.0077399 -41.187  < 2e-16 ***
TT2          0.1974978  0.0087454  22.583  < 2e-16 ***
TT3          0.3590047  0.0085794  41.845  < 2e-16 ***
TT4          0.4410034  0.0085010  51.876  < 2e-16 ***
TT5         -0.3088128  0.0095053 -32.489  < 2e-16 ***
II2         -0.1352584  0.0065507 -20.648  < 2e-16 ***
II3         -0.2793295  0.0067090 -41.635  < 2e-16 ***
OO2          0.0136894  0.0087410   1.566    0.117
OO3          0.2551620  0.0084800  30.090  < 2e-16 ***
OO4         -0.0001694  0.0087572  -0.019    0.985
OO5          0.0034975  0.0087531   0.400    0.689
YY2          0.1682463  0.0087485  19.232  < 2e-16 ***
YY3          0.2024196  0.0087294  23.188  < 2e-16 ***
YY4          0.1282176  0.0088081  14.557  < 2e-16 ***
YY5          0.1745576  0.0087521  19.945  < 2e-16 ***
CC1         -0.2107379  0.0089113 -23.648  < 2e-16 ***
RR2          0.0069130  0.0077161   0.896    0.370
RR3          0.0115619  0.0077105   1.499    0.134
RR4                 NA         NA      NA       NA

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
...
```

Kód 4.5: Zhrnutie modelovaných dát.

Pre parameter `AA2` z `glm1` v kóde 4.5 je z-hodnota vyjadrená takto

$$T_j = \frac{-0.0808570}{0.0096601} = -8.370 \quad .$$

Z-hodnota reprezentuje informáciu, ako ďaleko je tento parameter od nuly. Pokiaľ testujeme hypotézu $H_0 : \beta_j = 0$, tak $T_j \sim t_{n-p}$, kde n je počet pozorovaní a p je počet odhadovaných parametrov, viď [6].

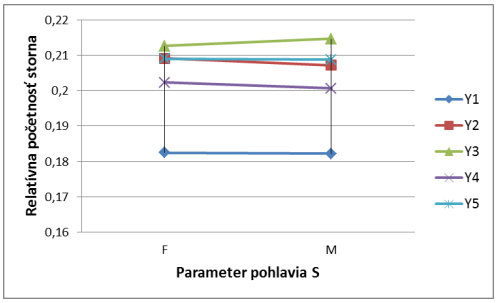
Štvrtým stĺpcom je p-hodnota (`Pr(>|z|)`), ktorá vyjadruje pravdepodobnosť, že za platnosti H_0 pozorujeme údaje, ktoré svedčia proti H_0 viac, než skutočne pozorované dáta. P-hodnota teda vyjadruje pravdepodobnosť, s akou pozorujeme za platnosti H_0 dáta viac odporujúce H_0. Hypotézu H_0 zamietame na hladine α práve vtedy, keď táto hodnota je menšia než α. Z toho dôvodu predpokladáme, že parametre, ktoré majú vysokú p-hodnotu, nebudú v správnom modeli zahrnuté. Pri p-hodnote môžeme pozorovať hviezdičky. Ich význam je popísaný v `Signif. codes`. Pokiaľ je p-hodnota v rozmedzí 0.1 a 1, pri parametri sa nezobrazí žiaden znak. Ak je medzi 0.05 až 0.1, zobrazí sa bodka. Ak je p-hodnota medzi 0 a 0.001, zobrazia sa tri hviezdičky. Čím viac hviezdičiek máme, tým významnejší je daný parameter.

`Null deviance` je deviácia pre nulový model popísaný v kapitole 2.3. Pojem `Residual deviance` je deviácia pre skúmaný model.

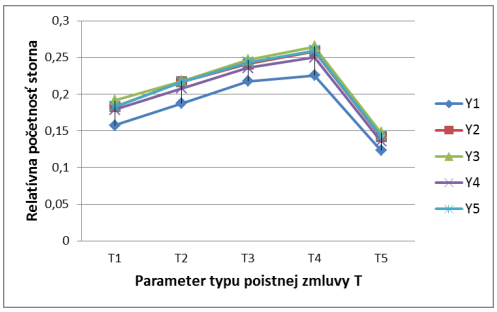
4.3 Závislosť parametrov

V podkapitole 2.3 sme uviedli, že pri modelovaní je lepší jednoduchší model s čo najmenej parametrami. Pri hľadaní toho správneho modelu je dôležité odhadnúť parametre, na ktorých storno nezávisí a prípadne ich z modelu vylúčiť. Najprv zistíme závislosť storna na jednotlivých parametroch podľa rokov zjednania poistenia. Cieľom pre zistenie závislosti je predikcia do budúcich rokov, z toho dôvodu hľadáme závislosti práve podľa parametru rok uzavretia poistnej zmluvy. Na nevýznamné parametre nás môže naviesť aj výstup z jednoduchého modelu súčtu všetkých parametrov, kde podľa p-hodnoty (a počtu hviezdičiek) môžeme odhadnúť, na ktorých parametroch storno nezávisí. Podľa kódu 4.5 sú parametre pohlavie `S` a sídlo `R` kandidátmi na vyradenie z modelu. Ešte sa pozrieme na pomerovú analýzu parametrov, z ktorej dostávame relatívne početnosti stornovaných zmlúv. Tieto závislosti ukážeme na grafoch 4.1 a 4.2.

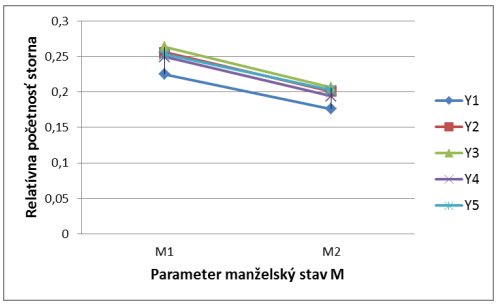
V grafe 4.1b je vidieť závislosť relatívnej početnosti storna na parametri typ poistky `T`. Krivky pre všetky roky `Y1` až `Y5` vykazujú rovnaký trend. Storno



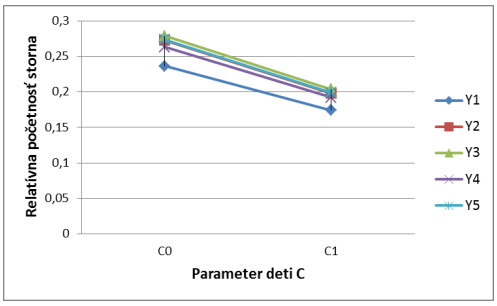
(a) Závislosť relatívnej početnosti storna na pohlaví S podľa roku Y.



(b) Závislosť relatívnej početnosti storna na typu zjednaného poistenia T podľa roku Y.



(c) Závislosť relatívnej početnosti storna na rodinnom stave M podľa roku Y.



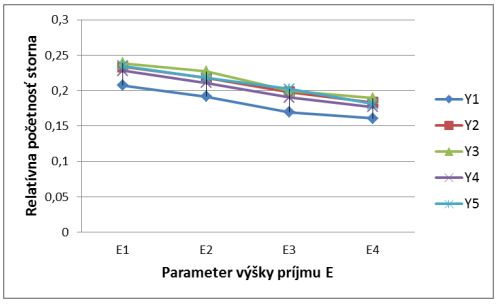
(d) Závislosť relatívnej početnosti storna na počtu detí C podľa roku Y.

Obr. 4.1: Závislosť relatívnej početnosti storna na parametroch podľa rokov Y.

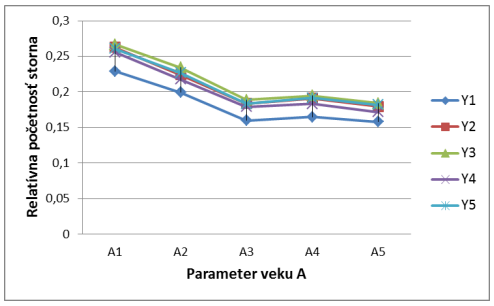
na tomto parametri teda pravdepodobne bude závisieť v čase konzistentne. Naopak graf 4.1a pre pohlavie vykazuje rôzne trendy, mierny rast pre `Y3` a mierny pokles pre `Y2` a `Y4`. Všeobecne sú však priamky takmer konštantné, takže storno na tomto parametri nezávisí.

Aj na grafoch 4.2a a 4.2b je vidieť rovnaký trend pre rôzne roky uzavretia poistnej zmluvy. Oba parametre sú kandidátmi do modelu, pretože vykazujú istú závislosť.

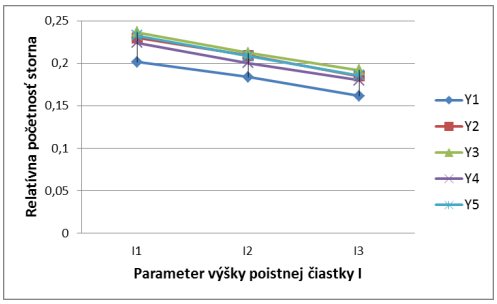
Závislosť storna na výške poistnej čiastky má tiež rovnaký trend v rôznych rokoch, viď 4.2c. Naopak závislosť storna na veľkosti sídla, ktoré máme znázornenú na grafe 4.2d, nevykazuje úplne rovnaký trend pre všetky roky a navyše všetky priamky tvoria takmer konštantnú krivku. Z toho usudzujeme, že storno na sídle vôbec nezávisí. Zaujímavý je graf závislosti storna na distribúcii 4.2e, pretože v dvoch rokoch je práve na jednej distribúcii `O3` pravdepodobnosť storna výrazne vyššia než u ostatných distribúcii. Pravdepodobne to bude súvisieť s interakciou medzi sprostredkovateľom poistných zmlúv a rokom zjednania poistnej zmluvy. Je možné, že stornovanie u distribúcie `O3` bolo v rokoch `Y2` a `Y3` výrazne vyššie zo strany sprostredkovateľa poistných zmlúv z dôvodu podvodov. Takže tento graf 4.2e nám indikuje interakciu medzi parametrami `O` a `Y`.



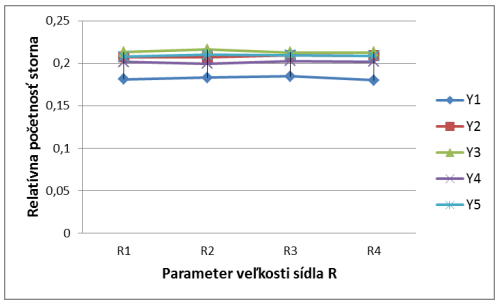
(a) Závislosť relatívnej početnosti storna na príjmu E podľa roku Y.



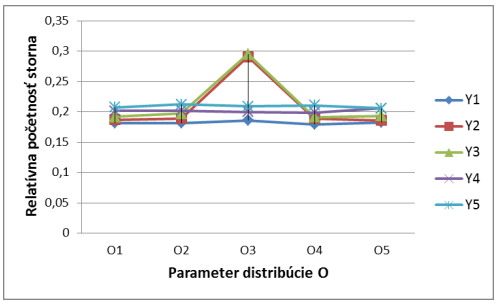
(b) Závislosť relatívnej početnosti storna na veku A podľa roku Y.



(c) Závislosť relatívnej početnosti storna na výške poistnej čiastky I podľa roku Y



(d) Závislosť relatívnej početnosti storna na veľkosti sídla R podľa rokov Y



(e) Závislosť relatívnej početnosti storna na distrubúcii O podľa roku Y

Obr. 4.2: Závislosť relatívnej početnosti storna parametroch podľa roku Y.

Podľa kódu 4.5[2] a podľa grafov 4.1a a 4.2d môžeme tvrdiť, že *pohlavie* a *sídlo* nemajú na stornovanie poistných zmlúv takmer žiaden vplyv, takže týmito parametrami sa nebudeme v ďalšom texte zaoberať. Z distribúcie je významný len parameter `O3`, preto parameter budeme skúmať, hlavne v interakcii s inými parametrami.

4.4 Interakcie medzi parametrami

V predošlej úvahe sme z modelu odstránili parametre *pohlavie* a *sídlo*. Teraz zistíme, či interakcie medzi jednotlivými parametrami majú vplyv na stornova-

[2] `Signif. codes` vysvetľuje, že parametre označené troma hviezdičkami majú na model veľký vplyv. Parametre bez hviezdičiek naopak na model nemajú žiaden vplyv.

```
> glm2 <- glm(cbind(LapseNum,PolNum - LapseNum) ~ A * M,
    family = binomial(link = "logit"), data = tabulkaStoren)
> summary(glm2)

Call:
glm(formula = cbind(LapseNum, PolNum - LapseNum) ~ A * M,
    family = binomial(link = "logit"), data = tabulkaStoren)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8927  -0.9050  -0.3301   0.7104   5.8251

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error  z value Pr(>|z|)
(Intercept)     -1.034983   0.007796 -132.761  < 2e-16 ***
AA2             -0.180937   0.019825   -9.127  < 2e-16 ***
AA3             -0.433863   0.028871  -15.028  < 2e-16 ***
AA4             -0.475169   0.045055  -10.546  < 2e-16 ***
AA5             -0.439444   0.010174  -43.194  < 2e-16 ***
M1              -0.076069   0.011119   -6.841 7.86e-12 ***
AA2:M1           0.023224   0.022223    1.045   0.2960
AA3:M1           0.018361   0.030635    0.599   0.5489
AA4:M1           0.104522   0.046180    2.263   0.0236 *
AA5:M1                 NA         NA       NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 108674  on 75006  degrees of freedom
Residual deviance: 103796  on 74998  degrees of freedom
AIC: 231494

Number of Fisher Scoring iterations: 4
```

Kód 4.6: Interakcia medzi parametrom A a M

nie poistných zmlúv. Najprv vyskúšame interakciu medzi parametrami vekom `A` a manželským stavom `M`. Z kódu 4.6 vidíme, že interakcia medzi A a M nemá vplyv na stornovanie[3]. Taktiež podľa grafu 4.3a môžeme vidieť, že síce storno závisí na veku `A`, krivky však vykazujú rovnaký trend pre oba manželské stavy. Z toho vidíme, že medzi vekom a manželským stavom nie je žiadna interakcia.

Ak v programe R modelujeme jednotlivé závislosti, je potrebné v R vyhodnocovať dáta, ktoré sú zoskupené podľa jednotlivých parametrov, v našom prípade

[3]Interakcia `A4` a `M` má nevýznamný vplyv, čo je štatisticky bezpredmetné.

podľa veku `A` a manželského stavu `M`. Kód `Summary` nám síce môže pomôcť určiť, ktoré parametre medzi sebou interakciu majú, avšak tieto modely neberú do úvahy ostatné parametre, takže informácia, ktorú model ponúka, je len pomocná. Grafy, ktoré uvádzame v 4.3, majú lepšiu výpovednú hodnotu a v porovnaní s tvorbou v `R` sú jednoduchšie. V programe R načítame všetky dáta a pomocou funkcie `xtabs` dostaneme kontingenčné tabuľky, pomocou ktorých napríklad v programe Microsoft Excel vytvoríme grafy 4.3.

Takto vyhľadáme interakcie medzi všetkými parametrami. V diplomovej práci vykreslíme len relevantné interakcie, viď grafy 4.3. Z 4.3 a 4.2e môžeme tvrdiť, že na storno majú vplyv len interakcie A*T, M*T, C*T a O*Y.

4.5 Model

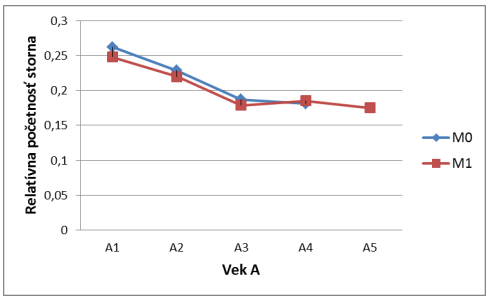
4.5.1 Binomický model s interakciami

V tejto časti vysvetlíme princíp hľadania modelu, ktorý najlepšie vysvetľuje dáta. Dáta máme rozdelené do skupín a identifikáciu, či táto zmluva s danými parametrami bola alebo nebola stornovaná. Takáto štruktúra dát indikuje použitie binomického modelu. Najčastejšie sa binomický model používa s transformačnou funkciou logit. Tento model nazveme `glmBi`, v ktorom zahrnieme interakcie `A*T`, `M*T`, `C*T` a `O*Y`, o ktorých máme indikácie z podkapitoly 4.4. Pozrieme sa na funkciu `Summary` v dodatku B a na grafy 4.4, ktoré nám vykreslí program `R`, a vďaka ktorým je možné posúdiť adekvátnosť modelu.

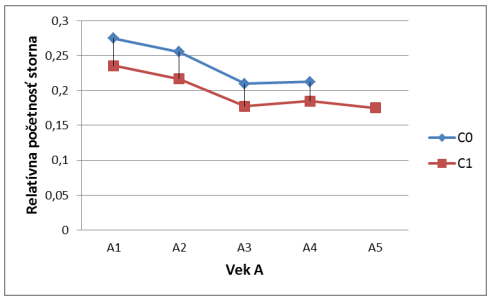
Funkcia `plot()` znázorňuje štyri diagnostické diagramy, ako sa dočítame v [6]. Deviačné reziduá zobrazujú rôznymi spôsobmi. Ľavý horný graf predstavuje závislosť deviačných reziduí $\hat{\varepsilon}_i$ oproti hodnotám lineárneho prediktoru modelu η_i. Program `R` nesprávne označuje tento graf na ose x ako `Predicted values`. V tom prípade by nemohli byť hodnoty na ose y pod nulou. V skutočnosti osa x je lineárny prediktor η_i.

Ľavý horný diagram nám pomáha vyhodnotiť správnosť predpokladov o rozdelení, transformačnej funkcie a iných predpokladov. Reziduá by mali byť rovnomerne rozptýlené nad a pod nulou a nezávislé na lineárnom prediktore η, pretože stredná hodnota reziduí by mala byť nula ($\mathrm{N} \sim (0, 1)$). V opačnom prípade by nám graf indikoval buď chybnú štruktúru modelu alebo vynechanie dôležitej premennej.

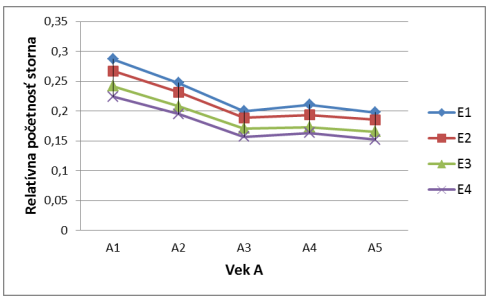
Pravý horný graf je tzv. *Q-Q (quantile-quantile) graf, kvantil – kvantilový diagram.* Graficky porovnáva experimentálne kvantily meraní y a teoretické kvantily



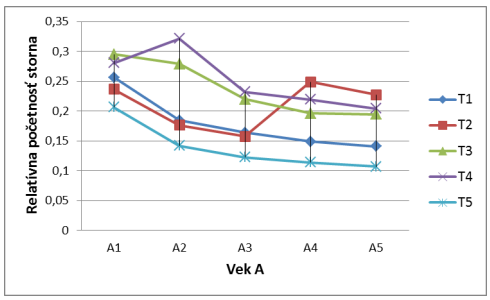
(a) Interakcia medzi A a M



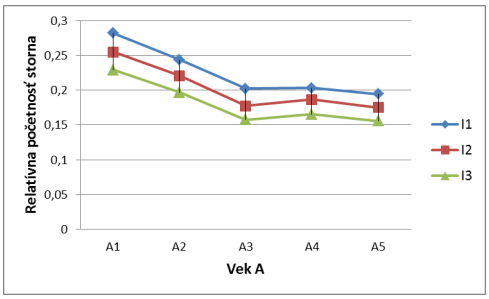
(b) Interakcia medzi A a C



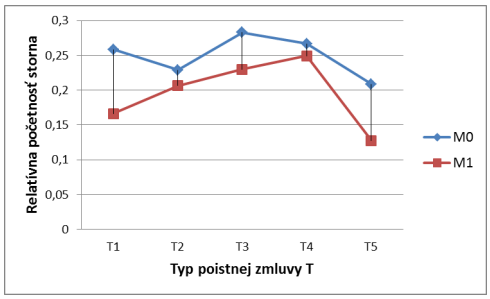
(c) Interakcia medzi A a E



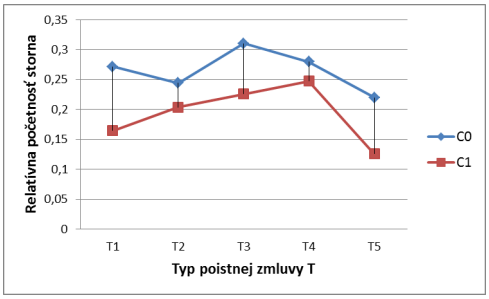
(d) Interakcia medzi A a T



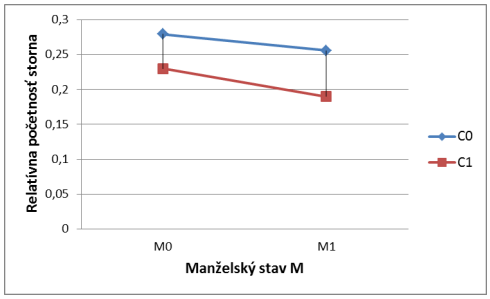
(e) Interakcia medzi A a I



(f) Interakcia medzi M a T

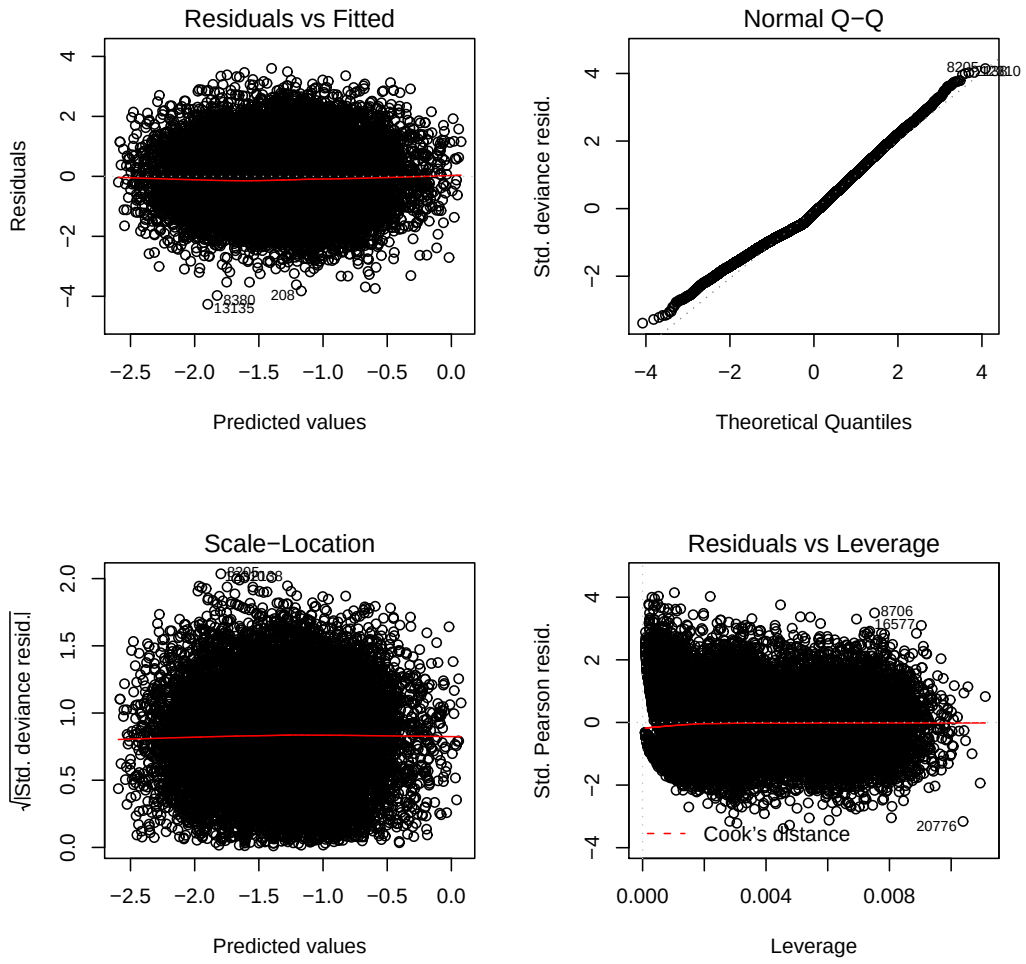


(g) Interakcia medzi C a T



(h) Interakcia medzi C a M

Obr. 4.3: Interakcia medzi parametrami



Obr. 4.4: Grafy znázorňujúce binomický model pre všetky dáta.

rozdelenia, prípadne porovnáva kvantily dvoch experimentálnych meraní. Graf je používaný práve k overeniu normality deviačných reziduí. Pokiaľ deviačné reziduá splňujú predpoklad normality, mali by body (vyjadrujúce hodnoty) ležať na priamke $y = x$.

Ľavý dolný graf `Scale-Location` znázorňuje odmocninu z absolútnej hodnoty štandardizovaných deviačných reziduí (viď podkapitolu 2.3.3) a môže upozorniť na zmenu variability reziduí s predikovanou hodnotou.

Posledný z týchto štyroch grafov je v pravom dolnom rohu, nazýva sa `Residuals vs Leverage`. Tento graf znázorňuje štandarizované rezidua oproti tzv. pákam *(leverage)*, viď podkapitolu 2.3.3. Slúži k identifikácii hodnôt s príliš veľkým vplyvom na odhad parametrov modelu. Kombinácia veľkých zostatkových reziduí a vysokého *leverage* znamená, že zodpovedajúci údaj má podstatný vplyv na celkový odhad. Jedným z príkladov miery vplyvu konkrétnych dát je tzv. *Cookova vzdialenosť (Cook's distance)*. Je to miera, ktorá udáva, aký veľký vplyv

má každé pozorovanie na odhadovaný model.

Z grafov 4.4 môžeme pozorovať, že model popisuje dáta uspokojivo. Ľavý horný obrázok je rovnomerne rozvrstvený pod a nad nulou. Správny Q-Q diagram zodpovedá funkcii $y = x$, pričom graf na obrázku má podobnú krivku. Celkovo môžeme zhodnotiť, že binomický model s transformačnou funkciou logit a zvolenými premennými je vhodný k aplikácií na uvedené dáta, čo potvrdzujú aj grafy 4.4.

4.5.2 Poissonov model s interakciami

Hoci binomický model vhodne popisuje dáta, pre porovnanie ukážeme sa aj na Poissonov model, ktorý je prirodzenou alternetívou k binomickému modelu. Model pomenujeme `glmPo` a necháme si vykresliť graf 4.5. Syntax zovšeobecnených lineárnych modelov je pre Poissonov model v programe `R` iná než pre binomický model, viď kód 4.7. Pozrime sa teda na Poissonov model pomocou grafu 4.5.

```
> glmPo <- glm(LapseNum ~ A * T + M * T + C * T + E +
  I + O * Y, family = poisson(link = "log"),
  offset = log(PolNum), data = tabulkaStoren)
```

Kód 4.7: Poissonov model.

Ako môžeme vidieť v grafe 4.5, Q-Q diagram má tvar správneho modelu. Ľavý horný graf má síce niekoľko výbežkov v tvare oblúku, ktoré sú charakteristické pre menšiu skupinu dát. Tieto výbežky však postupne prechádzajú do súvisle rozvrstvenej oblasti bodov, ktoré sú rovnomerne rozvrstvené nad a pod nulou. Výbežky sú spôsobené malým počtom storien v niektorých kategóriách. Tento graf teda nezamieta správnosť modelu, práve ho naopak podporuje. Taktiež ostatné grafy podporujú hypotézu, že Poissonov model správne vysvetľuje napozorované hodnoty. Pozrieme sa ešte na funkciu `Summary` pre Poissonov model, viď kód 4.8.

4.5.3 Ďalšie rozdelenia

Hoci Poissonov model vysvetľuje dáta dobre, vyskúšali sme v programe `R` aj ostatné rozdelenia. Žiaden lepší model sme nenašli. Napríklad Gamma rozdelenie vždy skončilo chybou, pretože v našich dátach sú aj nulové hodnoty, pri ktorých sa Gamma rozdelenie nedá použiť. Skúšali sme Poissonovo rozdelenie s transformačnou funkciou odmocnina `sqrt` a identickou transformačnou funkciou `identity` a obe nebolo možné modelovať, pretože dáta nie su vhodné na použitie

```
glm(formula = LapseNum ~ A * T + M * T + C * T + E + I + O *
    Y, family = poisson(link = "log"), data = tabulkaStoren,
    offset = log(PolNum))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1527  -0.7684  -0.1205   0.5469   2.9935

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.0515432  0.0204355 -51.457  < 2e-16 ***
AA2         -0.1567210  0.0196992  -7.956 1.78e-15 ***
AA3         -0.2459707  0.0210582 -11.681  < 2e-16 ***
AA4         -0.3320803  0.0218941 -15.168  < 2e-16 ***
AA5         -0.3780298  0.0226038 -16.724  < 2e-16 ***
TT2         -0.1732735  0.0224685  -7.712 1.24e-14 ***
TT3          0.0082666  0.0213612   0.387 0.698763
TT4         -0.1578221  0.0218896  -7.210 5.60e-13 ***
TT5         -0.1943057  0.0230509  -8.429  < 2e-16 ***
M           -0.1654211  0.0176165  -9.390  < 2e-16 ***
C           -0.2560559  0.0174306 -14.690  < 2e-16 ***
EE2         -0.0706039  0.0065292 -10.814  < 2e-16 ***
EE3         -0.1736046  0.0067072 -25.883  < 2e-16 ***
EE4         -0.2486553  0.0068432 -36.336  < 2e-16 ***
II2         -0.1043818  0.0057432 -18.175  < 2e-16 ***
II3         -0.2180044  0.0059303 -36.761  < 2e-16 ***
OO2          0.0008431  0.0179613   0.047 0.962562
OO3          0.0247803  0.0178555   1.388 0.165190
OO4         -0.0098342  0.0179534  -0.548 0.583857
OO5          0.0064656  0.0179262   0.361 0.718340
YY2          0.0285977  0.0178128   1.605 0.108393
YY3          0.0579589  0.0177405   3.267 0.001087 **
YY4          0.1065581  0.0174880   6.093 1.11e-09 ***
YY5          0.1338201  0.0173611   7.708 1.28e-14 ***
AA2:TT2     -0.0448636  0.0280925  -1.597 0.110267
AA3:TT2     -0.0587728  0.0299589  -1.962 0.049788 *
AA4:TT2      0.4927322  0.0298001  16.535  < 2e-16 ***
AA5:TT2      0.4524553  0.0307105  14.733  < 2e-16 ***
AA2:TT3      0.1620403  0.0260655   6.217 5.08e-10 ***
AA3:TT3      0.0179716  0.0280752   0.640 0.522093
AA4:TT3     -0.0042374  0.0292546  -0.145 0.884832
AA5:TT3      0.0380551  0.0299768   1.269 0.204267
```

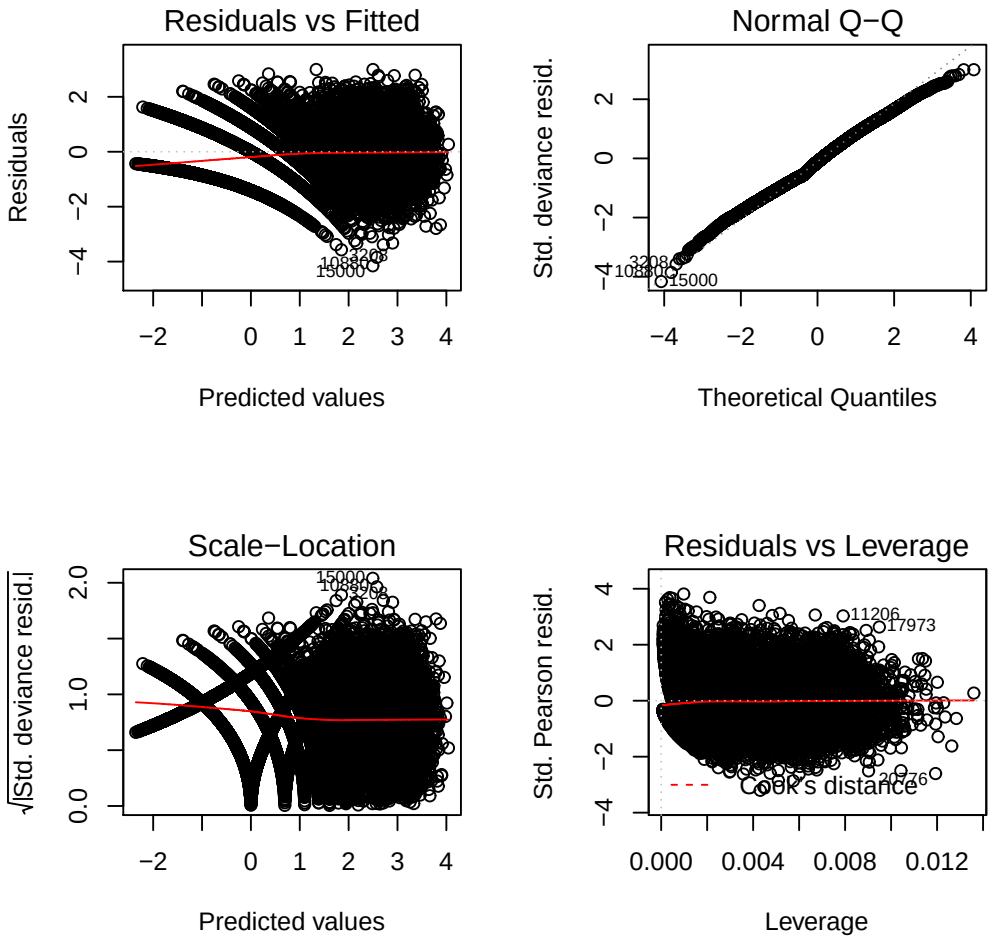
```
AA2:TT4      0.2549118  0.0257480   9.900  < 2e-16 ***
AA3:TT4      0.0105910  0.0278348   0.380 0.703578
AA4:TT4      0.0368244  0.0288416   1.277 0.201680
AA5:TT4      0.0109942  0.0297025   0.370 0.711276
AA2:TT5     -0.0195988  0.0297530  -0.659 0.510078
AA3:TT5     -0.0504197  0.0318645  -1.582 0.113577
AA4:TT5     -0.0241478  0.0331972  -0.727 0.466979
AA5:TT5     -0.0278462  0.0342482  -0.813 0.416179
TT2:M        0.1109593  0.0252664   4.392 1.13e-05 ***
TT3:M        0.1722473  0.0238704   7.216 5.36e-13 ***
TT4:M        0.2503469  0.0239692  10.445  < 2e-16 ***
TT5:M       -0.0199615  0.0264764  -0.754 0.450889
TT2:C        0.0894464  0.0249523   3.585 0.000337 ***
TT3:C        0.1057922  0.0235205   4.498 6.86e-06 ***
TT4:C        0.2691273  0.0236739  11.368  < 2e-16 ***
TT5:C       -0.0323589  0.0261959  -1.235 0.216732
OO2:YY2      0.0104185  0.0251564   0.414 0.678764
OO3:YY2      0.4179739  0.0240231  17.399  < 2e-16 ***
OO4:YY2      0.0223481  0.0251764   0.888 0.374723
OO5:YY2     -0.0132816  0.0252108  -0.527 0.598319
OO2:YY3      0.0258633  0.0250355   1.033 0.301574
OO3:YY3      0.4006743  0.0239479  16.731  < 2e-16 ***
OO4:YY3      0.0037044  0.0251261   0.147 0.882792
OO5:YY3     -0.0012633  0.0250463  -0.050 0.959773
OO2:YY4     -0.0044491  0.0247730  -0.180 0.857472
OO3:YY4     -0.0340823  0.0247388  -1.378 0.168301
OO4:YY4     -0.0023755  0.0247966  -0.096 0.923680
OO5:YY4      0.0129554  0.0246992   0.525 0.599912
OO2:YY5      0.0198020  0.0245479   0.807 0.419858
OO3:YY5     -0.0170937  0.0245145  -0.697 0.485621
OO4:YY5      0.0252564  0.0245563   1.029 0.303710
OO5:YY5     -0.0143154  0.0245891  -0.582 0.560442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 38662  on 22113  degrees of freedom
Residual deviance: 18041  on 22050  degrees of freedom
AIC: 84155

Number of Fisher Scoring iterations: 4
```

Kód 4.8: Funkcia Summary pre Poissonov model.



Obr. 4.5: Poissonov model pre všetky dáta.

týchto transformačných funkcií, ako sme popisovali v podkapitole 2.2.2 výber transformačnej funkcie.

4.5.4 Správne rozdelenie

Pri hľadaní takého modelu, ktorý najlepšie vysvetľuje dáta, sme našli dve rozdelenia, ktoré vysvetľujú dáta dobre a to binomické rozdelenie a Poissonovo rozdelenie. Oba modely sú dobré. Avšak pre jednoduchšiu intepretáciu, hlavne vďaka jednoduchšej transformačnej funkcie, si zvolíme Poissonov model za správny. U Poissonovho modelu používame logaritmickú transformačnú funkciu a teda odhady parametru pri interpretácii stačí pretransformovať cez exponenciálu. V binomickom modeli je použitá transformačná funkcia logit, takže odhady by sme museli najprv pretransformulovať z logit funkcie a až potom použiť exponenciálu.

4.5.5 Testovanie oprávnenosti zahrnutia premenných

Chceme otestovať, či máme najvhodnejší model. Na to použijeme test `Anova`, kde otestujeme aktuálny správny model s inými podobnými modelmi, ktoré majú viac parametrov, alebo naopak niektoré uberieme. Aktuálne správny model 4.7 načítame do `R` pod názvom `glmHlavny`. Najprv otestujeme tento model s modelom, ktorý má navyše parameter pohlavia `S`, ktorý sme v podkapitole 4.3 vylúčili. Pripravíme si dáta v Microsoft Access s parametrami ako v kóde 4.7 a pridáme k tomu ešte parameter pohlavie a dáta zoskupíme a uložíme do `HlavnyS`. V programe `R` vytvoríme model podobný 4.7 s tým rozdielom, že pridáme súčet pohlavia `S` a používame novú tabuľku dát. Tým dostaneme druhú premennú `glmHlavnyS` a tieto modely otestujeme pomocou funkcie `Anova`, viď kód 4.9.

```
> glmHlavnyS <- glm(LapseNum ~ S + A * T + M * T + T * C + E +
  I + O * Y, family = poisson(link = "log"),
  offset = log(PolNum),data = HlavnyS)
> glmHlavny2 <- glm(LapseNum ~ A * T + M * T + T * C + E + I +
  O * Y, family = poisson(link = "log"),
  offset = log(PolNum),data = HlavnyS)
> anova(glmHlavny2, glmHlavnyS, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
Model 2: LapseNum ~ S + A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     41272      33777
2     41271      33777  1  0.27675    0.5988
```

Kód 4.9: Funkcia Anova aplikovaná na správny model a model s parametrom S.

Podľa výsledkov testu Anova v kóde 4.9, p-hodnota je dosť veľká, takže nezamietame model `glmHlavny` v prospech `glmHlavnyS`. Z toho dôvodu považujeme model bez parametru `S` lepší, pretože má menej parametrov. Skúsime ešte porovnať model `glmHlavny` s rovnakým modelom okrem parametru `A`, viď kód 4.10.

Vo funkcii `Anova` sa zameriame na veľkosť p-hodnoty v stĺpci `p-value`. Pokiaľ tá je veľmi malá, znamená to, že zamietame model `glmHlavnyBezA` v prospech `glmHlavny`. V našom prípade je `p-value` malá, takže správny model je `glmHlavny`. V dodatku C je výpis ďalších výstupov iných podmodelov modelu `glmHlavny` a všetky majú malú p-hodnotu, teda vždy zamietame podmodely v prospech modelu `glmHlavny`. Z toho usudzujeme, že model, popísaný v 4.7, najlepšie vysvetľuje dáta.

```
> glmHlavny <- glm(LapseNum ~ A * T + M * T + T * C + E + I +
  O * Y, family = poisson(link = "log"), offset = log(PolNum),
  data = tabulkaHlavny)
> glmHlavnyBezA <- glm(LapseNum ~ M * T + T * C + E + I +
  O * Y, family = poisson(link = "log"), offset = log(PolNum),
  data = tabulkaHlavny)
> anova(glmHlavnyBezA, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ M * T + T * C + E + I + O * Y
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22070      21640
2     22050      18041 20   3598.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Kód 4.10: Funkcia Anova aplikovaná na správny model a model bez parametru A.

4.6 Porovnanie výsledkov

V tejto sekcii porovnáme závery obdržané na základe jednoduchých pomerových ukazovateľov (pomer stornovaných zmlúv daného typu ku všetkým zmluvám daného typu) a pomocou Poissonovho zovšeobecneného lineárneho modelu s logaritmickou transformačnou funkciou. Výsledky sú vyjadrené v grafe 4.6, kde jedna krivka znázorňuje modelovanú hodnotu a druhá jednoduchú pomerovú analýzu.

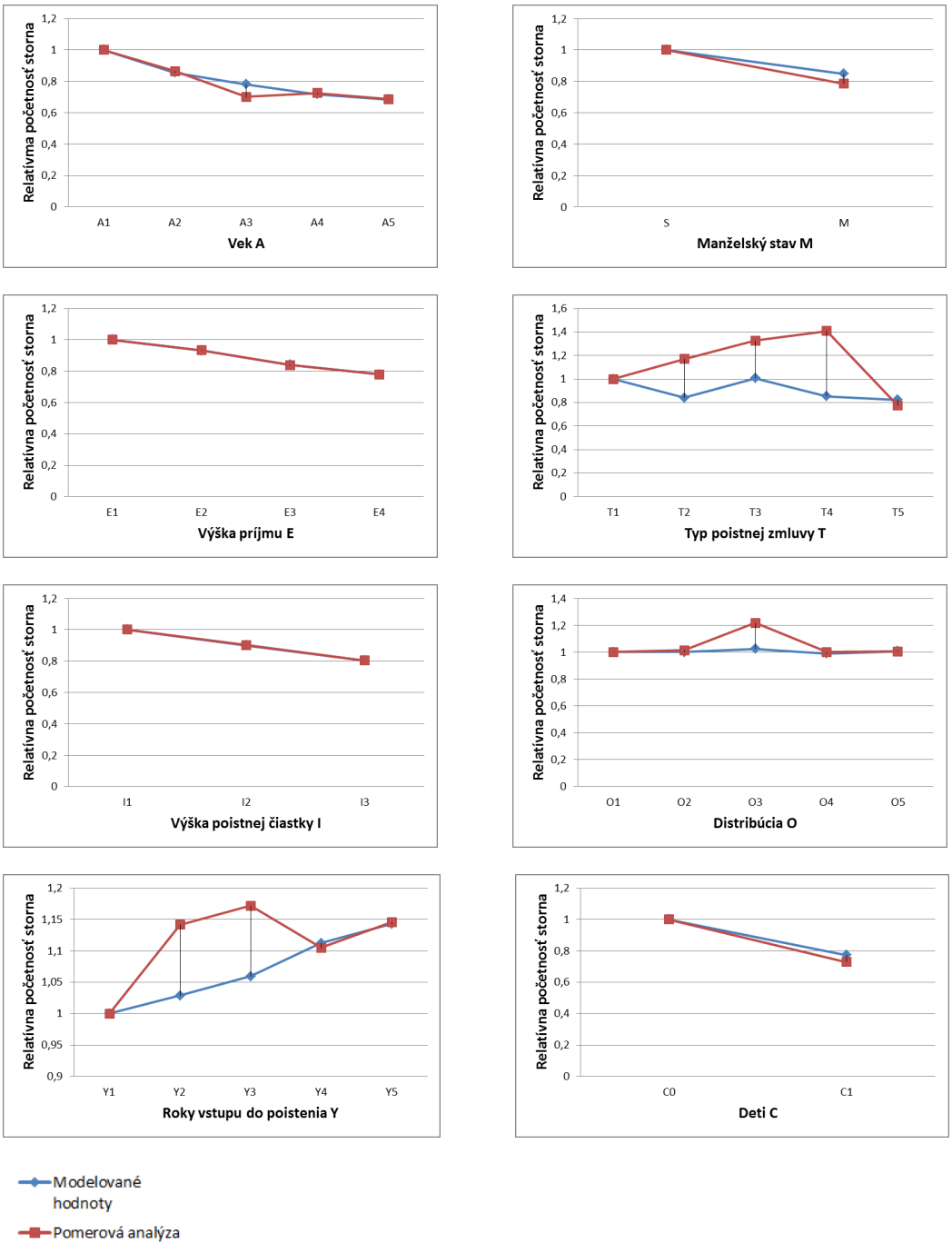
V grafe 4.6 sú znázornené modelované hodnoty a pomerová analýza. K zisťovaniu pomerovej analýzy sme v prvom rade vypočítali relatívnu početnosť storna pre každý parameter zvlášť. Tým sme získali pravdepodobnosti stornovania p pre daný parameter, z ktorého vypočítame multiplikátor, ktorý zanesieme do grafu. Za základ pre počítanie multiplikátora stanovíme vždy prvú skupinu, takže pre vek `A` to bude skupina `A1`. Potom z každej hodnoty p vezmeme jej podiel so základnou hodnotou. Pokiaľ by sme v modeli používali transformačnú funkciu logit, získané pravdepodobnosti p pomerovou analýzou by sme pretransformovali pomocou vzorca do hodnoty q,

$$q = \ln\left(\frac{p}{1-p}\right).$$

Modelované hodnoty dostaneme z modelu, ktorý považujeme za správny, v našom prípade z Poissonovho modelu. Použijeme hodnoty `Estimate` z výstupu

R, viď kód 4.8. Z týchto hodnôt urobíme exponenciálu a tie taktiež zanesieme do grafu.

Keď sa pozrieme na grafy, pomerová analýza sa často líši od modelovaných dát. Vidíme to najmä na grafoch veku, typu poistnej zmluvy, distribúcie a roku zjednania poistnej zmluvy. Na grafu Typu poistnej zmluvy pomerová analýza vykazuje rastúci trend pre typy T2, T3, T4. Naopak namodelované dáta majú pre typ poistnej zmluvy iný trend. Závislosť storna na type poistného produktu je podľa modelu malá. Najväčší rozdiel je u produktu T4, kde podľa pomerovej analýzy má najväčšiu pravdepodobnosť stornovania poistnej zmluvy, ale podľa modelovaných dát je menej stornovaný než referenčný produkt T1. Tieto rozdiely vznikajú z dôvodu interakcií a korelácií. Pomerová analýza nedokáže zohľadniť interakcie ani korelácie a zovšeobecnené lineárne modely zohľadňujú korelácie automaticky a interakcie pomocou modelov, ktoré používame.



Obr. 4.6: Porovnanie výsledkov

Záver

V diplomovej práci sme vysvetlili, čo je storno v životnom poistení, v akých prípadoch nastáva a čo spôsobuje, že sa poistník rozhodne poistnú zmluvu stornovať.

V teoretickej časti sme v krátkosti predstavili lineárne regresné modely, aby sme mohli na nich vysvetliť zovšeobecnené lineárne modely. Definovali sme exponenciálnu triedu rozdelení a zvolili sme transformačnú funkciú, oboje charakteristické pre zovšeobecnené lineárne modely. Poslednou časťou teórie je popis výberu modelu, jeho hodnotenie deviáciou a kontrola správnosti pomocou reziduí.

Tretiu kapitolu sme venovali zhrnutiu záverov doterajších článkov venovaných tejto problematike vo svete. Porovnávali sme závislosť storna na individuálnych a makroekonomických parametroch.

V poslednej kapitole sme aplikovali získané teoretické znalosti na generované dáta. Vysvetlili sme na nich proces hľadania správneho modelu pomocou analýzy interakcií a odstraňovania nepotrebných parametrov. Našli sme dva modely, ktoré vysvetľujú dáta uspokojivo, a to binomický model s transformačnou funkciou logit a Poissonov model s logaritmickou transformačnou funkciou. Pretože Poissonov model je jednoduchší, najmä vďaka jednoduchšej transformačnej funkcii, zvolili sme si tento model a porovnali odhadované parametre s relatívnou početnosťou dát. Odhadované parametre sa od nej v niektorých prípadoch líšili výrazne; a to hlavne z dôvodu, že zovšeobecnený lineárny model berie do úvahy závislosti medzi parametrami a ich interakcie, zatiaľ čo relatívna početnosť nie. V práci vysvetľujeme a interpretujeme výsledky zo štatistického programu `R`.

Literatúra

[1] Anderson D., Feldblum S., et al., 2007: *A Practitioner's Guide to Generalized Linear Models*, Casualty Actuarial Socity.

[2] Kuo W., Tsai Ch., Chen W. K., 2003: *An Empirical Study on the Lapse Rate: The Cointegration Approach*, The Journal of Risk and Insurance, Vol. 70, No. 3, 489–508.

[3] Dobson A. J., 2002: *An introduction to generalized linear models*, Second edition, Chapman and Hall/CRC, Boca Raton.

[4] Cox S.H., Lin Y., 2006: *Annuity Lapse Rate Modeling: Tobit or nor Tobit.*, Society of Actuaries.

[5] Faraway J. J., 2006: *Extending the linear model with R: Generalized linear, Mixed Effects and Nonparametric Regression Models*, Chapman and Hall/CRC, Boca Raton.

[6] Wood S. N.: *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC, Boca Raton.

[7] McCullagh P., Nelder J. A., 1999: *Generalized Linear Models*, Chapman and Hall, Boca Raton.

[8] Nelder J. A., Wedderburn R. W. M., 1972: *Generalized Linear Models*, Journal of the Royal Statistical Society, 135, 370–384.

[9] Renshaw A. E., Haberman S. J., 1996: *Generalized Linear Models and Actuarial Science*, The Statistician, Vol. 45, No. 4, 407–436.

[10] Geyer Ch. J., 2003: *Generalized Linear Models in R*, University of Minnesota, študijný materiál.

[11] Cerchiara R.R., Edwards M., Gambini A., 2008: *Generalized linear models in life insurance: decrements and risk factor analysis under Solvency II*, Interntional AFIR Colloquium.

[12] Irwin M. E., 2005: *Generalized Linear Models Introduction*, Harvard University Statistics Department, študijný materiál.

[13] Lebel D., 1999: *Lapse Experience Under Lapse - Supported Policies*, Canadian Institute of Actuaries.

[14] Marcsik J., Wion M., 2008: *Managing and Modelig Policyholder Behavior Risks*, Society of Actuaries, Equity-based Insurance Guarntees Conferece.

[15] Reiskytl J., Siegel S., 2005: *Policyholder Behavior in the Tail Risk Management Working Group, Variable Annuity Guaranteed Benefits Survey Results*, Society of Actuaries.

[16] Giroux G. B., et al., 2010: *Predictive Modeling for Life Insurers*, Towers Watson.

[17] Zvára K., 2008: *Regrese*, Matfyzpress, Praha.

[18] Changki K., 2006: *Report to the Policyholder Behavior in the Tail Subgroups Project*, Technical report, Society of Actiaries.

[19] Renshaw A. E., Haberman S. J., 1986: *Statistical analysis of life assurance lapses*, Journal of the Institute of Actuaries **113**, 459–497.

[20] Richardson Ch. F. B., Hartwell J. M., 2006: *Transactions of socety of actuaries, Vol III*, University of Chicago Press, 1951, 338–396

[21] Fang H., Kung E., 2010: *Why Do Life Insurance Policyholders Lapse? The Roles of Income, Health and Bequest Motive Shocks*, Department of Economics, Duke University.

[22] Anděl, J., 2002: *Základy matematické statistiky*, Matfyzpress

Dodatok A

Numerické riešenie vierohodnostných rovníc

Pri hľadaní numerických riešení použijeme Newton-Raphsonovu integračnú metódu

$$\beta^{m+1} = \beta^m - H_m^{-1} U_m,$$

kde $H_m = H(\beta^m)$ a $H = H(\beta) = \left(\frac{\partial^2 L(\beta)}{\partial\beta_i \partial\beta_j}\right)$, $i, j = 1, \ldots, n$.

Vyjadríme U_m a q_m ako $U_m = U(\beta^m) = X q_m$, kde U je skórový vekto a $q_m = q(\beta^m) = X q_m$.

Parameter β^m je m-tá aproximácia maximálne vierohodnostného odhadu $\hat{\beta}$, $m = 1, 2, \ldots$ udáva index integračného kroku.

V ďalších úvahách pre zjednodušenie výpočtov algoritmu, viď [3], nahradíme maticu druhých derivácií $H(\beta)$ maticou stredných hodnôt

$$J(\beta) = -\mathrm{E}\, H(\beta) = -\mathrm{E}\left(\frac{\partial^2 l(\beta)}{\partial\beta_i \partial\beta_j}\right), \qquad i, j = 1, \ldots, n$$

a maticu J nazveme Fisherovou informačnou maticou parametru β.

Dostávame integračný proces

$$\beta^{m+1} = \beta^m - J_m^{-1} U_m,$$

kde $J_m = J(\beta)$.

Prvky matice J označíme J_{ij} a dostaneme

$$J_{ij} = \frac{\partial l}{\partial\beta_i}\frac{\partial l}{\partial\beta_j} = \frac{x_{il} x_{lj}}{var(y_l)}\left(\frac{\partial\mu_l}{\partial\eta_l}\right)^2.$$

Vidíme, že Fisherovu informačnú maticu môžeme zapísať v tvare

$$J(\beta) = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

kde $\mathbf{W} = \mathbf{W}(\beta)$ je diagonálna matica s prvkami $\omega_i = \frac{1}{\mathrm{Var}(y_i)} \left(\frac{\partial \mu_l}{\partial \eta_l}\right)^2$ a označíme $\mathbf{W}_m = \mathbf{W}(\beta^m)$.

Integračný proces prepíšeme na tvar

$$\beta^{m+1} = \beta^m - \left(\mathbf{X}^T \mathbf{W}_m \mathbf{X}\right)_m^{-1} U_m.$$

Ak upravíme túto rovnicu a položíme

$$\mathbf{z}_m = \mathbf{X}\beta^m + \mathbf{W}_m^{-1} q_m,$$

dostávame integračný proces v tvare

$$\mathbf{X}^T \mathbf{W}_m \mathbf{X} \beta^{m+1} = \mathbf{X}^T \mathbf{W}_m \mathbf{z}_m.$$

Tento tvar je rovnaký ako u normálnych rovníc pre lineárny model, ktorý sme získali metódou vážených najmenších štvorcov, kde miesto vektoru $\mathbf{y}$ na pravej strane rovnice je vektor $\mathbf{z}$. Hľadaný parameter β je nutné hľadať prepočítavaním iteratívne z rovnice. Takže $\hat{\beta} = \lim_{m \to \infty} \beta^m$. Táto metóda sa nazýva *iteratívna metóda vážených najmenších štvorcov*. Je ešte nutné zdôrazniť, že váhová matica $\mathbf{W}$ sa v každom kroku mení a je ju treba neustále prepočítavať.

Dodatok B

Funkcia Summary pre binomický model

V dodatku je výstup z programu R pre binomický model tak, ako je popísaný v kapitole 4.5.1.

```
> summary(glmBi)

Call:
glm(formula = cbind(LapseNum, PolNum - LapseNum) ~ A * T +
    M * T + C * T + E + I + O * Y,
    family = binomial(link = "logit"), data=tabulkaStoren)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2657  -0.8398  -0.1362   0.6298   3.6013

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.643788   0.023739 -27.120  < 2e-16 ***
AA2           -0.201963   0.022285  -9.063  < 2e-16 ***
AA3           -0.308931   0.023644 -13.066  < 2e-16 ***
AA4           -0.410861   0.024465 -16.794  < 2e-16 ***
AA5           -0.463520   0.025189 -18.402  < 2e-16 ***
TT2           -0.237061   0.026341  -9.000  < 2e-16 ***
TT3            0.024316   0.025551   0.952 0.341277
TT4           -0.212374   0.025912  -8.196 2.49e-16 ***
TT5           -0.273212   0.026872 -10.167  < 2e-16 ***
M             -0.218974   0.020285 -10.795  < 2e-16 ***
C             -0.338386   0.020078 -16.853  < 2e-16 ***
```

```
EE2          -0.093473   0.007511 -12.445  < 2e-16 ***
EE3          -0.226349   0.007654 -29.571  < 2e-16 ***
EE4          -0.320987   0.007768 -41.321  < 2e-16 ***
II2          -0.136839   0.006575 -20.811  < 2e-16 ***
II3          -0.281379   0.006733 -41.789  < 2e-16 ***
OO2           0.001030   0.020097   0.051 0.959122
OO3           0.031083   0.020008   1.553 0.120304
OO4          -0.012297   0.020076  -0.613 0.540178
OO5           0.008051   0.020067   0.401 0.688273
YY2           0.035917   0.019966   1.799 0.072027 .
YY3           0.073220   0.019918   3.676 0.000237 ***
YY4           0.135248   0.019695   6.867 6.55e-12 ***
YY5           0.170448   0.019586   8.703  < 2e-16 ***
AA2:TT2      -0.052779   0.031695  -1.665 0.095865 .
AA3:TT2      -0.069999   0.033576  -2.085 0.037092 *
AA4:TT2       0.620065   0.033650  18.427  < 2e-16 ***
AA5:TT2       0.557919   0.034541  16.152  < 2e-16 ***
AA2:TT3       0.208789   0.030127   6.930 4.20e-12 ***
AA3:TT3       0.001931   0.032075   0.060 0.951994
AA4:TT3      -0.033607   0.033237  -1.011 0.311958
AA5:TT3       0.014905   0.033994   0.438 0.661055
AA2:TT4       0.345204   0.029878  11.554  < 2e-16 ***
AA3:TT4      -0.012947   0.031879  -0.406 0.684643
AA4:TT4       0.010466   0.032898   0.318 0.750385
AA5:TT4      -0.029050   0.033768  -0.860 0.389642
AA2:TT5      -0.012155   0.033127  -0.367 0.713675
AA3:TT5      -0.042992   0.035221  -1.221 0.222228
AA4:TT5      -0.008454   0.036552  -0.231 0.817099
AA5:TT5      -0.010714   0.037617  -0.285 0.775794
TT2:M         0.148260   0.028942   5.123 3.01e-07 ***
TT3:M         0.228116   0.027873   8.184 2.74e-16 ***
TT4:M         0.338171   0.027942  12.102  < 2e-16 ***
TT5:M        -0.011524   0.029977  -0.384 0.700654
TT2:C         0.121789   0.028621   4.255 2.09e-05 ***
TT3:C         0.125016   0.027557   4.537 5.72e-06 ***
TT4:C         0.356475   0.027667  12.884  < 2e-16 ***
TT5:C        -0.019884   0.029658  -0.670 0.502585
```

```
OO2:YY2   0.013102   0.028206   0.465 0.642276
OO3:YY2   0.565523   0.027312  20.706  < 2e-16 ***
OO4:YY2   0.028281   0.028221   1.002 0.316282
OO5:YY2  -0.016803   0.028259  -0.595 0.552118
OO2:YY3   0.032838   0.028130   1.167 0.243069
OO3:YY3   0.546470   0.027267  20.042  < 2e-16 ***
OO4:YY3   0.004379   0.028197   0.155 0.876597
OO5:YY3  -0.001504   0.028134  -0.053 0.957373
OO2:YY4  -0.005731   0.027899  -0.205 0.837231
OO3:YY4  -0.043111   0.027875  -1.547 0.121959
OO4:YY4  -0.003364   0.027904  -0.121 0.904043
OO5:YY4   0.016998   0.027838   0.611 0.541468
OO2:YY5   0.025887   0.027713   0.934 0.350232
OO3:YY5  -0.021195   0.027684  -0.766 0.443909
OO4:YY5   0.032313   0.027705   1.166 0.243483
OO5:YY5  -0.018351   0.027738  -0.662 0.508248
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49498  on 22113  degrees of freedom
Residual deviance: 23330  on 22050  degrees of freedom
AIC: 83288

Number of Fisher Scoring iterations: 4
```

Kód B.1: Pokračovanie binomického modelu dát s interakciami AT, MT a OY

Dodatok C

Submodely Poissonovho modelu

V tomto dodatku ukážeme niektoré podmodely Poissonovho rozdelenia, ako sme to rozoberali v podkapitole 4.5.5. Všetky podmodely sme skúmali s hlavným modelom 4.7. Všetky podmodely zamietame v prospech modelu 4.7.

```
> anova(glmHlavnyBezE, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A * T + M * T + T * C + I + O * Y
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22053      19601
2     22050      18041  3   1559.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-----------------------------------------------
> anova(glmHlavnyBezI, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A * T + M * T + T * C + E + O * Y
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22052      19401
2     22050      18041  2   1359.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(glmHlavnyBezO, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A * T + M * T + T * C + E + I + Y
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22070      20527
2     22050      18041 20   2485.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-----------------------------------------------
> anova(glmHlavnyBezY, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A * T + M * T + T * C + E + I + O
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22070      19940
2     22050      18041 20   1898.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-----------------------------------------------
> anova(glmHlavnyBezC, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A * T + M * T + E + I + O * Y
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22055      18643
2     22050      18041  5   601.42 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-----------------------------------------------
> anova(glmHlavnyBezT, glmHlavny, test="Chisq")
Analysis of Deviance Table

Model 1: LapseNum ~ A + M + C + E + I + O * Y
Model 2: LapseNum ~ A * T + M * T + T * C + E + I + O * Y
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     22078      28459
2     22050      18041 28    10418 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```