MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Vojtěch Neumann

Škodní inflace v pojištění aut

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Ing. Pavel Kříž, Ph.D.
Studijní program: Finanční matematika
Studijní obor: MFMP

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V .... dne ....

Podpis autora

Chtěl bych poděkovat za cenné rady, konstruktivní kritiku a podporu během psaní svému vedoucímu práce Mgr. Ing. Pavlovi Křížovi, Ph.D. .

Speciální poděkování také patří mým milým kolegům z pojišťovny, za poskytnutí dat k analýze.

Rád bych také poděkoval své rodině a přátelům, kteří mi poskytli morální podporu a byli trpěliví, když jsem byl zaneprázdněn psaním práce.

Název práce: Školní inflace v pojištění aut

Autor: Vojtěch Neumann

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Ing. Pavel Kříž, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá praktickým využitím zobecněných lineárních modelů s cílem analyzovat škodní inflaci na povinném ručení v autopojištění. K tomu jsou poskytnuta aktuální data z české pojišťovny. V práci je detailně zkonstruován zobecněný lineární model pomocí stanovených kritérií. Z modelu je identifikován vliv inflace a určena její výše za dané období.

Klíčová slova: škodní inflace; neživotní pojištění; zobecněné lineární modely; gama rozdělení

Title: Claims inflation in car insurance

Author: Vojtěch Neumann

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Ing. Pavel Kříž, Ph.D., department

Abstract: This thesis explores the practical use of generalized linear models. The aim of the thesis is to analyze the claims inflation for Motor Third Party Liability Insurance. For this purpose, current data from a Czech insurance company are provided. In the thesis, a generalized linear model is constructed in detail based on specified criteria. From the model, the effect of inflation is identified and its value for the given period is determined.

Keywords: claims inflation; non-life insurance; generalized linear models; gamma distribution

Obsah

Úvod

Inflace je důležitý ekonomický pojem, který označuje míru růstu cenové hladiny zboží a služeb v určitém období. Měření inflace přidáváme vysokou hodnotu, jelikož ovlivňuje kupní sílu jednotlivců a podniků i celkové zdraví ekonomiky. Proto má klíčový význam v pojišťovnictví, které je citlivější na změny a je důležité ji zohlednit v predikci výdajů pojišťovny a stanovení sazebníku. V posledních letech je toto téma eskalováno od skončení pandemické krize a míry inflace zveřejňovány Českou národní bankou, ale i světovými bankami, dosahují výjimečně nepředvídatelných hodnot.

Trh pojišťovnictví obsahuje velkou míru rivality a společnost je v tomto směru ovlivňována nejlepší nabídkou pojistky. Špatná identifikace trendu cen oprav a nákladů může přivést pojišťovnu do ztráty a při nepřiměřené reakci může hrozit případný bankrot. Identifikace míry inflace je zásadní součástí zajištění toho, aby pojistné přesně odráželo rizika spojená s pojištěním vozidla a aby pojišťovny i zákazníci byli chráněni před finančními ztrátami. Pojišťovny léta operují s bohatým počtem modelů na tvoření sazebníků, netto pojistného a rezervování, ale v minulosti se tento trh nepotkal s nestandardními výkyvy. Vzhledem k této skutečnosti se inflace dostala na popředí probíraných témat.

Cíl této bakalářské práce je identifikace škodní inflace v autopojištění v povinném ručení na českém trhu. Kvůli nejasné závislosti lineárních a nelineárních vztahů různých ukazatelů jsem se rozhodl pokusit se identifikovat míru inflace pomocí metodiky zobecněných lineárních modelů. Pomocí této metodiky můžeme zohlednit obsáhlou škálu faktorů a určit jak každý faktor ovlivňuje vysvětlovanou proměnnou.

Tato práce se rozděluje na část teoretickou, kde jsou vysvětleny základní principy zobecněných lineárních modelů se zaměřením na jeden konkrétní model a na část praktickou, kde na aktuálních datech české pojišťovny zkonstruujeme model a za daných předpokladů identifikujeme výši inflace.

1. Lineární regrese

Tato kapitola vychází z (Anděl, 2007), kapitola 12 a Zvára, 2008, kapitola 2.

Lineární regrese se zabývá zkoumáním vztahu mezi spojitou veličinou Y a vektorem X, který obsahuje alespoň jednu veličinu. Složky vektoru X jsou spojité či diskrétní veličiny. Chceme zjistit, jakým způsobem složky vektoru X ovlivňují střední hodnotu EY, ale nikoliv rozptyl. Na základě toho můžeme ku příkladu předpovídat hodnotu Y pro dané X. Předpokládáme data z n nezávislých pozorování ve tvaru $(Y_i,X_i), i = 1, \dots n$, kde X_i je $p < n$ složkové.

Definice 1 (Lineární model)**.** *Lineárním modelem* nazýváme statistický model, který popisuje vztah mezi vysvětlujícími proměnnými $X_0, X_1, X_2, \dots, X_p$ a závislou proměnnou Y pomocí lineární funkce:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i,$$

kde $\beta_0, \beta_1, \dots, \beta_p$ jsou neznámé *regresní koeficienty* a $\varepsilon_0, \dots, \varepsilon_n$ jsou nezávislé náhodné veličiny, nazývané *chybové členy*. Mají nulovou střední hodnotu a konstantní rozptyl.

Většinou volíme $X_{i1} = 1$ pro všechna i. Pak získáme výhodný tvar

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \tag{1.1}$$

a β_0 nazýváme *absolutní člen.*

Poznámka Zde si zavedeme názvy proměnných regresních modelů, které budeme používat v této práci.

- Náhodná veličina Y_i se nazývá *odezva*. Alternativní názvy: *vysvětlovaná, závislá*
- Složky X_i nazýváme *regresory*. Alternativní název: *vysvětlující*

Rovnici z (1.1) můžeme přepsat jako

$$\begin{aligned} E(Y_i|X_i) &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \\ \operatorname{var}(Y_i|X_i) &= \sigma^2. \end{aligned}$$

V tomto zápise je zdůrazněno, že lineární regresní model popisuje vztah mezi podmíněnou střední hodnotou Y pro dané X pomocí lineárního vztahu a předpokládá, že rozptyl Y je pro všechny hodnoty X stejný a nezávislý na nich.

Značení. Lineární regresní model (1.1) můžeme maticově zapsat jako:

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

kde Y_i je vysvětlovaná proměnná pro i-té pozorování, X_{ij} je hodnota j-té vysvětlující proměnné pro i-té pozorování, β_j je regresní koeficient pro j-tou vysvětlující proměnnou a ϵ_i je chybový člen pro i-té pozorování. Matici X nazýváme *regresní matice* a předpokládáme, že sloupce jsou lineárně nezávislé. Matice X má tedy plnou hodnost.

Jinými slovy model tvrdí, že závislou proměnnou Y lze modelovat jako lineární funkci regresorů X, kde koeficienty β určují sílu a směr tohoto lineárního vztahu. Chyby ϵ_i představují náhodnou odchylku Y od podmíněné střední hodnoty.

1.1 Odhad parametru

V této sekci se budeme zabývat metodou, kterou využíváme na získání odhadu regresních koeficientů. Kromě metody maximální věrohodnosti se v lineární regresi používá i metoda nejmenších čtverců, která má stejné hodnoty jako metoda maximální věrohodnosti, za předpokladu, že chybové členy mají normální rozdělení. Důkaz nalezneme v (Seber and Lee, 2003), kapitola 3, sekce 3.5.

1.1.1 Metoda maximální věrohodnosti (MLE)

Pro lineární model tvaru (1.1) můžeme použít odhad **MLE** na regresní koeficienty $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, což v tomto případě budou hodnoty, které maximalizují věrohodnostní funkci, což je sdružené pravděpodobnostní rozdělení pozorovaných dat vzhledem k parametrům modelu.

Za předpokladu, že chybový člen ϵ_i má normální rozdělení se střední hodnotou 0 a rozptylem σ^2, pak Y má normální rozdělení se střední hodnotou $X\beta$ a rozptylem σ^2. Věrohodnostní funkci můžeme odvodit z hustoty:

Hustota pro každé pozorování i

$$f(Y_i|\beta_0, \beta_1, ..., \beta_p, \sigma^2, \mathbf{X}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2}{2\sigma^2}\right).$$

Pro získání *věrohodnostní funkce* musíme vynásobit hustoty pro všechna n pozorování:

$$L(\beta_0, \beta_1, ..., \beta_p, \sigma^2|\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} f(Y_i|\beta_0, \beta_1, ..., \beta_p, \sigma^2, \mathbf{X}).$$

Logaritmická věrohodnostní funkce má tvar

$$\begin{aligned} \ln L(\beta, \sigma^2|\mathbf{X}, \mathbf{Y}) &= \ln \left[\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2\right)\right] \\ &= \ln \frac{1}{(2\pi\sigma^2)^{n/2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 \end{aligned}$$

$$\ln L(\beta, \sigma^2|\mathbf{X}, \mathbf{Y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2\right), \quad (1.2)$$

kde $\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{np})$ je $n \times p$ matice nezávislých proměnných a $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ je $n \times 1$ vektor závislých proměnných.

Odhady MLE regresních koeficientů a rozptylu chyb lze získat maximalizací funkce věrohodnosti vzhledem k parametrům:

$$\hat{\beta} = \arg \max_{\beta_0, \beta_1, \ldots, \beta_p, \sigma^2} L(\beta_0, \beta_1, \beta_2, \ldots, \beta_p, \sigma^2 | \mathbf{X}, \mathbf{Y}).$$

2. Zobecněný lineární model

Celá kapitola (kromě jednotlivých odvození pro gama rozdělení) vychází z Ohlsson and Johansson (2010), kapitola 2.

Námi dříve popsaný základní lineární model není zcela vhodný pro modelování cen v neživotním pojištění, jelikož vyžaduje normální rozdělení náhodné chyby a funkční závislost v lineárním tvaru. Počet pojistných událostí sleduje diskrétní pravděpodobnostní rozdělení na nezáporných celých čísel a výše pojistných událostí jsou také nezáporné a mívají těžký chvost směrem doprava. To nastává z důvodu existence škod, které mají diametrálně vyšší pojistné plnění na rozdíl od ostatních, ale v počtu je jich velmi málo. Příkladem v pojištění nemovitostí většina škod bude ve spojení s vodovodními problémy, poškozením z důsledku přírodních vlivů či přepětí. Tato množina škod bude sledovat podobnou výši, ale nastanou výjimečně i totální škody v důsledku velkých požárů či katastrof. Právě tyto škody způsobují nesymetričnost rozdělení škod.

2.1 Rozdělení exponenciální třídy

Zobecněným lineárním modelem (Generalized linear model - **GLM**) můžeme modelovat vysvětlovanou proměnnou s jiným rozdělením než normálním. Při použití těchto modelů předpokládáme, že vysvětlovaná proměnná pochází z určité *třídy exponenciálních rozdělení.*

Definice 2 (Rozdělení exponenciálního třídy - **EDM**)**.** Řekněme, že rozdělení je exponenciální třídy, pokud má jeho hustota (v případě spojitého rozdělení) nebo pravděpodobnostní funkce (pro diskrétní rozdělení) následující tvar:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi/w} + c(y,\phi,w)\right). \tag{2.1}$$

V 2.1 představuje y konkrétní realizaci pozorovaných dat z Y, která má rozdělení exponenciální třídy. Konstanta $w > 0$ je známá a používá se k úpravě (tzv. scaling) *parametru rozptýlení* ϕ, pro který platí $\phi \in (0,\infty)$. Parametr rozptýlení udává, jak moc je rozdělení prvků daného rozdělení vzdálené od střední hodnoty. θ je neznámý parametr nabývající reálných hodnot z nějaké otevřené podmnožiny $\mathbb{R}$, který odhadujeme. Dále se předpokládá, že *kumulantová funkce* $b(\theta)$ je dvakrát spojitě diferencovatelná a existuje $(b')^{-1}$. Funkce c je normalizační logaritmická konstanta, ale jelikož nezávisí na parametru θ, nebudeme se jí zabývat. Lze se o ní dočíst v Jorgensen, 1997, kapitola 5. V této knize nalezeneme i odvození, že do třídy EDM patří známá rozdělení jako normální, Poissonovo, gama, binomické a samozřejmě mnoho dalších splňujících definici 2.1. Pro účely této práce se budeme zabývat pouze gama rozdělením. Lineární regrese 1.1 je speciální případ GLM pro vysvětlovanou proměnnou pocházející z normálního rozdělení.

Speciálně pro GLM má m-té nezávislé pozorování funkci hustoty ve tvaru:

$$f(y_m,\theta_m,\phi,w_m) = \exp\left(\frac{y_m\theta_m - b(\theta_m)}{\phi/w_m} + c(y_m,\phi,w_m)\right). \tag{2.2}$$

Jednotlivé y_i jsou nezávislé. Parametr rozptýlení ϕ je stejný pro všechna m. Konstanta w_m a parametr θ_m se můžou lišit mezi pozorováními.

Základní momentové charakteristiky rozdělení z EDM jsou:

$$\mu_i \equiv E(Y_i) = b'(\theta_i) \tag{2.3}$$

$$Var(Y_i) = \phi v(\mu_i)/w_i, \tag{2.4}$$

kde $v(\mu_i) = b''(b'^{-1}(\mu_i)$ nazýváme *funkce rozptylu*. Tato odvození jsme převzali z (Ohlsson and Johansson, 2010), Lemma 2.1.

2.1.1 Rozdělení výše škod

V následující části se přesuneme na fundamentální model této práce. Námi hledané rozdělení musí mít kladné hodnoty a být nesymetricky vychýlené doprava. Proto neuvažujeme normální rozdělení (a lineární model 1.1), a i když bychom nalezli více kandidátů, gama rozdělení se rozšířilo pro modelování výše škod, viz (Murphy et al., 2000).

Funkce hustoty gama rozdělení je

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}; x > 0, \alpha > 0, \beta > 0, \tag{2.5}$$

kde $\Gamma(\alpha)$ je Gama funkce, $\Gamma(\alpha,\beta)$ označujeme gama rozdělení s parametry α, β. Parametr α je často nazývaný *tvarový parametr*, jelikož větší hodnoty α vedou k výraznějšímu zkosení a vyšší poloze vrcholu. Parametr *rozsahu* β ovlivňuje rozptyl rozdělení tak, že větší hodnota β vede k většímu rozptylu.

Vlastnosti gama rozdělení:

- Rozdělení má známou střední hodnotu α/β a rozptyl α/β^2.
- Jestliže $X_1, X_2, \ldots, X_n \sim \Gamma(\alpha_i,\beta)$ jsou nezávislé stejně rozdělené náhodné veličiny s tvarovými parametry $\alpha_1, \alpha_2, \ldots, \alpha_n$ rozsahu β, pak jejich součet $Y = X_1 + X_2 + \ldots + X_n$ je gama rozdělená náhodná veličina s parametrem $\alpha = \sum_{i=1}^n \alpha_i$ a stejným parametrem rozsahu β.

Druhý bod lehce ukážeme za pomocí momentové vytvořující funkce gama rozdělení, která pro X_i má tvar $M_X(r) = \left(\frac{\beta}{r-\beta}\right)^{\alpha_i}$,$r < \beta$. Pak součet Y n nezávislých náhodných veličin získáme
$M_Y(r) = M_X(r)^n = \prod_{i=1}^n M_{X_i}(r) = \left(\frac{\beta}{r-\beta}\right)^{n\alpha}$.

Předpokládáme tedy, že jednotlivé výše škod jsou vzájemně nezávislé a mají gama rozdělení se stejným parametrem rozsahu. Označíme-li i-tou škodu jako S_i a počet škod n, pak $S = \sum_{i=1}^n S_i$ má díky druhé vlastnosti gama rozdělení také gama rozdělení $\Gamma(n\alpha,\beta)$. Vlastnosti průměrné výše škody $Y = S/n$ si pak odvodíme z S:

$$P(Y \leq y) = P(S/n \leq y) = P(S \leq ny)$$

a hustotu

$$f_Y(y) = nf_S(ny) = \frac{(n\beta)^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-n\beta y}; y > 0, \alpha > 0, \beta > 0. \tag{2.6}$$

Tedy $Y \sim \Gamma(\alpha n,\beta n)$ a střední hodnota se upraví na EY=α/β. Nyní ukážeme, že gama rozdělení patří do třídy EDM a tedy je typem 2.1.

Přepíšeme si hustotu gama rozdělení jako

$$f(y) = \exp(\log(\frac{\beta^\alpha}{\Gamma(\alpha)}) + (\alpha - 1)\log(y) - \beta y)$$
$$f(y) = \exp\big(- \beta y + \alpha \log \beta + (\alpha - 1)\log(y) - \log \Gamma(\alpha)\big).$$

Vynásobíme první dva členy $\frac{1/\alpha}{1/\alpha}$, z čehož získáme tvar

$$f(y) = \exp\big(\frac{-\frac{\beta}{\alpha}y + \log \beta}{1/\alpha} + (\alpha - 1)\log(y) - \log \Gamma(\alpha)\big).$$

Dále substitujeme $\theta = -\beta/\alpha$, $\phi = 1/\alpha$. Z této substituce si vyjádříme $\beta = -\theta\alpha = -\theta/\phi$, tedy $\log \beta = \log(-\theta) - \log \phi$. Toto dosadíme do rovnice:

$$f(y) = \exp\big(\frac{\theta y + log(-\theta)}{\phi} - \frac{\log \phi}{\phi} + (1/\phi - 1)\log(y) - \log \Gamma(1/\phi)\big) \tag{2.7}$$

Z čehož nám vznikne tvar

$$f(y) = \exp\big(\frac{\theta y - (-log(-\theta))}{\phi} + c(y,\phi,w)\big),$$

kde pro $b(\theta) = -\log(-\theta)$ a $w = 1$, získáme tvar 2.1. Tímto jsme dokázali, že gama patří do třídy exponenciálních rozdělení (EDM).
Z vyjádření 2.3 a získaných znalostí o gama rozdělení získáváme vlastnosti:

$$b(\theta_i) = -log(-\theta_i) \tag{2.8}$$
$$b'(\theta_i) = -1/\theta_i = \mu_i \tag{2.9}$$
$$b''(\theta_i) = 1/\theta_i^2 = \mu_i^2 \tag{2.10}$$
$$Var(Y_i) = \phi\mu_i^2/w_i. \tag{2.11}$$

2.2 Linková funkce

Mějme odezvu $Y = (Y_1, \ldots Y_n)$ ze třídy exponenciálních rozdělení 2.1 se střední hodnotou $\mu = (\mu_1, \ldots, \mu_n)$, rozptylem σ^2 a X je regresní matice. Modelujeme vztah mezi Y a X pomocí *lineárního prediktoru*, který je lineární kombinací regresorů a koeficientů $\eta = X\beta$. Doposud jsme se již setkali s lineárním prediktorem při modelování střední hodnoty v 1.1 (E$Y = \mu$), kde jsme měli $\eta_i = \mu_i$. Pro zobecněné lineární modely již nemusíme modelovat pouze střední hodnotu, ale můžeme ji transformovat pomocí *linkové funkce*, která může být lineární či nelineární.

Definice 3 (Linková funkce)**.** Striktně monotónní a dvakrát diferencovatelnou funkci g, transformující střední hodnotu na lineární prediktor $g(\mu_i) = \eta_i = x_i^\mathsf{T}\beta$, nazýváme *linková funkce*.

2.2.1 Kanonická linková funkce

Výběr linkové funkce závisí na typu a chování naší vysvětlované proměnné a typu problému, který se snažíme řešit. Na jeden model dat můžeme aplikovat více linkových funkcí v závislosti na tom, co chceme získat pomocí našeho modelu. První volba linkové funkce může být *kanonická funkce*. Pro GLM vždy existuje dobře definovaná kanonická linková funkce, která je odvozena z exponenciální funkce hustoty odezvy. Taková linková funkce však nemusí vždy vyhovovat konkrétnímu modelu.

Již známe transformaci:

$$\begin{array}{ccccc} \text{parametr} & & \text{střední hodnota} & & \text{lineární prediktor} \\ \theta & \xrightarrow{\quad b'(\theta) \quad} & \mu & \xrightarrow{\quad g(\mu) \quad} & \eta \end{array}$$

Z obrázku vyčteme vztahy: $\eta = g(b'(\theta))$ a $\theta = (b')^{-1}(g^{-1}(\eta))$.

Speciální volbou linkové funkce $g(\mu) = (b')^{-1}(\mu)$ dostaneme přímý vztah mezi prediktorem a parametrem $\theta : \theta = \eta$. Tato volba zjednodušuje řadu výpočtů a je v jistém smyslu přirozená, proto ji nazýváme *linková funkce*.

Například v 1.1 jsme použili jako linkovou funkci identitu. Pro gama rozdělení máme z 2.8 $b'(\theta) = -\frac{1}{\theta}$, tedy inverzní link $g(\mu) = -\frac{1}{\mu}$. Tento výraz může nabývat záporných hodnot a inverz této funkce, který určuje střední hodnotu může být záporný, proto ho na modelování výše škod nepoužijeme.

K modelování výše škod pojistných událostí použijeme logaritmický link $g(\mu) = log(\mu)$. Při použití této linkové funkce si zaručíme, že předpovídané hodnoty jsou vždy kladné, jelikož výše škod musí být vždy kladná. Kromě toho umožňuje aditivní vlastnost predikčních proměnných na logaritmicky transformovanou proměnnou odezvy, což je užitečné pro interpretaci výsledků modelu ve smyslu procentuální změny závažnosti pojistných událostí spojené s jednotkovou změnou predikční proměnné. To vychází z následujícího vztahu:

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

Linkovou funkci volíme $g(x) = \log(x)$, z čehož máme $g^{-1}(x) = \exp(x)$ a získáváme vztah

$$\begin{aligned} \mu &= \exp(\eta) = \exp(X\beta) \\ &= \exp(\beta_0 + \beta_1 X_1 +, \dots, +\beta_p X_p), \end{aligned}$$

kde β_i je vektor parametrů faktoru X_i. Inverz linkové funkce, tedy transformaci $g^{-1}(\eta) = \mu$ nazýváme *funkce střední hodnoty*.

2.3 Odhad parametru

Odhad parametrů v zobecněném lineárním modelu je prováděn pomocí maximálně věrohodných odhadů (viz. metoda MLE pro lineární model 1.2). Tato metoda v tomto případě předpokládá, že napozorovaná data vysvětlované proměnné pocházejí z exponenciální třídy rozdělení 2.1. Hledáme pro parametry hodnoty, které maximalizují věrohodnostní funkci, stejně jako v 1.2.

Logaritmická věrohodnostní funkce pro model nadý vztahem 2.2 θ je

$$\ell(\theta;\phi,y) = \frac{1}{\phi}\sum_{i}^{n} w_i(y_i\theta_i - b(\theta_i)) + \sum_{i}^{n} c(y_i,\phi,w_i). \tag{2.12}$$

Nyní lze vidět, že ϕ se nepodílí na maximalizaci logaritmické věrohodnostní funkce při odhadu vektoru parametru θ. Odhady parametru $\beta = (\beta_1,\ldots,\beta_p)$ se získávají pomocí skórové funkce

$$\frac{1}{\phi}\sum_{i} w_i \frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)} x_{ij}, \tag{2.13}$$

kterou položíme pro $j = 1,2\ldots,p$ nule a vztahu β s lineárním prediktorem. Odvození nalezneme v Ohlsson and Johansson (2010), sekce 2.3.2.

Řešení musí splňovat

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\sum_{j} x_{ij}\beta_j).$$

Řešení $\mu_i = y_i$ je pouze v modelu, který má stejný počet parametrů, jako pozorování. Takový model se nazývá *saturovaný model*. Má nejpřesnější odhad, ze všech modelů, ovšem nemůžeme ho použít na žádnou predikci.

Parametry se hledají numericky pomocí iteračních a optimalizačních algoritmů, jelikož nemají explicitní řešení. O používaných iteračních metodách se dočteme v Ohlsson and Johansson, 2010, sekce 3.2.3.

2.3.1 Tvar použitého modelu

V kapitole 2.1.1 jsme rozebírali podobnost rozdělení výše a škod gama rozdělení a proto předpokládáme, že použití gama rozdělení na modelování výše škod je vhodný model. V sekci 2.2.1 zdůvodňujeme použití logaritmu jako linkové funkce pro modelování výše pojistných událostí.

Model s logaritmickou linkovou funkcí a vysvětlovanou proměnnou z gama rozdělení lze napsat jako

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}. \tag{2.14}$$

Odhad parametrů β je prováděn metodou MLE a věrohodnostní logaritmická funkce nabývá tvaru

$$\ell(\theta;\phi,y) = \frac{1}{\phi}(\sum_{i}^{n} y_i\theta_i + log(\theta_i)) + \sum_{i}^{n}(1/\phi - 1)\log(y_i) - \log\Gamma(1/\phi) - \frac{log\phi}{\phi}.$$

Věrohodnostní funkce pro model má tvar

$$\ell(\beta) = \frac{1}{\phi}\sum_{i=1}^{n}\Big(y_i\exp(x_i\beta_i) - log(-1/\exp(x_i\beta_i)) + c(y_i,\phi)\Big),$$

kde funkci c lze numericky vyhodnotit.

3. Testování modelů

Pro praktické použití a k samotné konstrukci modelu potřebujeme mezi sebou porovnávat submodely, či porovnávat celé různorodé modely. Budeme muset rozhodovat, které regresory chceme použít, jaká je jejich významnost a správně interpretovat výsledek modelu. V této části probereme teoretické aspekty kalkulace odchylek a pomocných parametrů a v praktické části popíši své rozhodování a dosažené závěry pomocí získaných teoretických znalostí.

3.1 Testování shody modelu s daty

V této kapitole se budeme zabývat, jak odpovídají data vytvořenému modelu. Fit je celosvětově používaný pojem, převzaný z angličtiny, pro tvorbu modelu na základě dat a používáme ho v kontextu odhadu kvality fitu a fitování modelu, tedy posuzování, zda model odpovídá datům.

3.1.1 Škálovaná odchylka

Informace o škálované odchylce vychází (Ohlsson and Johansson, 2010), kapitola 3.1. Odvození konkrétního tvaru pro gama model je rozšířením literatury. Jako první měřítko kvality přizpůsobení modelu k datům použijeme *škálovanou odchylku* D^*, statistiku definovanou následujícím výrazem

$$D^* = 2\big(\ell(y;\hat{\theta}^*,\phi) - \ell(y;\hat{\theta},\phi)\big). \tag{3.1}$$

- $\ell(y;\hat{\theta}^*,\phi)$ je maximálně dosažitelná hodnota logaritmické věrohodnostní funkce, která by byla získána, kdyby model dokonale odpovídal datům (v případě saturovaného modelu $\mu_i = y_i$)
- $\ell(y;\hat{\theta},\phi)$ je maximální logaritmická věrohodnost, kterou získává fitovaný model, když je aplikován na skutečná data

Předpokládáme, že logaritmické věrohodnostní funkce mají stejné ϕ.

Označme h funkci inverzní k b'. Funkce h popisuje vztah $\theta_i = h(\mu_i)$. Pak si vyjádříme 3.1 pomocí 2.12 jako

$$D^* = \frac{2}{\phi}\sum_i w_i\Big[y_i h(y_i) - b(h(y_i)) - y_i h(\hat{\mu_i}) + b(h(\hat{\mu_i}))\Big] \tag{3.2}$$

Pro gama rozdělení si tvar nejdříve upravíme a poté odvodíme přesný:

$$D^* = \frac{2}{\phi}\sum_i w_i\Big[y_i\big(h(y_i) - h(\hat{\mu_i})\big) - (b(h(y_i)) - b(h(\hat{\mu_i}))\Big].$$

Připomeneme známá vyjádření parametrů $b(\theta) = -\log(-\theta)$, $b'(\theta) = -1/\theta$, $\mu = b'(\theta)$, pak $h(\mu) = -1/\mu$.

Pro kontrolu vidíme $b'(h(\mu)) = b'(-1/\mu) = \frac{-1}{-1/\mu} = \mu$. Dosadíme tyto vztahy do rovnice a získáme

$$\begin{aligned}
D^* &= \frac{2}{\phi}\sum_i w_i\left[y_i\left(\frac{-1}{y_i}+\frac{1}{\hat{\mu}_i}\right)-\Big(b(-1/y_i)-b(-1/\hat{\mu}_i)\Big)\right]\\
&= \frac{2}{\phi}\sum_i w_i\left[y_i\frac{-\hat{\mu}+y_i}{y_i\hat{\mu}_i}-(-\log(1/y_i)+\log(1/\hat{\mu}_i))\right]\\
&= \frac{2}{\phi}\sum_i w_i\left[\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}-(\log y_i - log\hat{\mu}_i)\right]\\
&= \frac{2}{\phi}\sum_i w_i\left[\frac{y_i}{\hat{\mu}_i}-1-\log\frac{y_i}{\hat{\mu}_i}\right].
\end{aligned}$$

Škálovaná odchylka pro gama model je

$$D^*(y,\hat{\mu}) = \frac{2}{\phi}\sum_i w_i(y_i/\hat{\mu}_i - 1 - \log(y_i/\hat{\mu}_i)). \tag{3.3}$$

Test věrohodnostním poměrem

Zde si popíšeme, jak se provádí test věrohodnostním poměrem, známý pod zkratkou **LRT** (Likehood ratio test).
Mějme *model* s p parametry a *submodel* s $p-a \geq 1$ ($a \in \mathbb{N}$) parametry. *Model* obsahuje stejné parametry jako *submodel* a má o a parametrů více. O submodelu říkáme, že je *vnořený* vůči modelu. Nulovou hypotézu stanovíme jako

H_0 : *submodel platí* oproti alternativě H_1 : *model platí, ale submodel ne*

LR statistiku definujeme jako:

$$LR = 2\log\left(\frac{L(model)}{L(submodel)}\right),$$

kde L je maximalizovaná věrohodnost modelu či submodelu. LR statistika má za platnosti H_0 přibližně χ^2 rozdělení o a stupních volnosti, viz Mittelhammer et al., 2000, strana 66. Při počítání LR statistiky se odhaduje i parametr ϕ. Pokud je parametr ϕ stejný v obou věrohodnostních funkcí, získáváme rozdíl škálovaných odchylek 3.1.

Nyní jsme získali nástroj na posouzení dvou velice podobných modelů, ve smyslu použitých regresorů. Použijeme toto především při determinaci důležitosti jednoho konkrétního parametru. Například si přidáme do modelu parametr a budeme testovat věrohodnostním poměrem, zda se nám pomocí testovaného regresoru model statisticky významně vylepší. Můžeme takto otestovat více regresorů a na základě porovnání výsledků testů si vybrat regresor s nejlepším zlepšením a zahrnout ho do modelu. Nebude-li žádný z dostupných parametrů vylepšovat stávající model, můžeme již mít optimální model. Tato metoda se nazývá *forward stepwise selection.*

3.2 AIC - Akaikovo informační kritérium

Tento text vychází z Awad, 1996, kapitola 1.

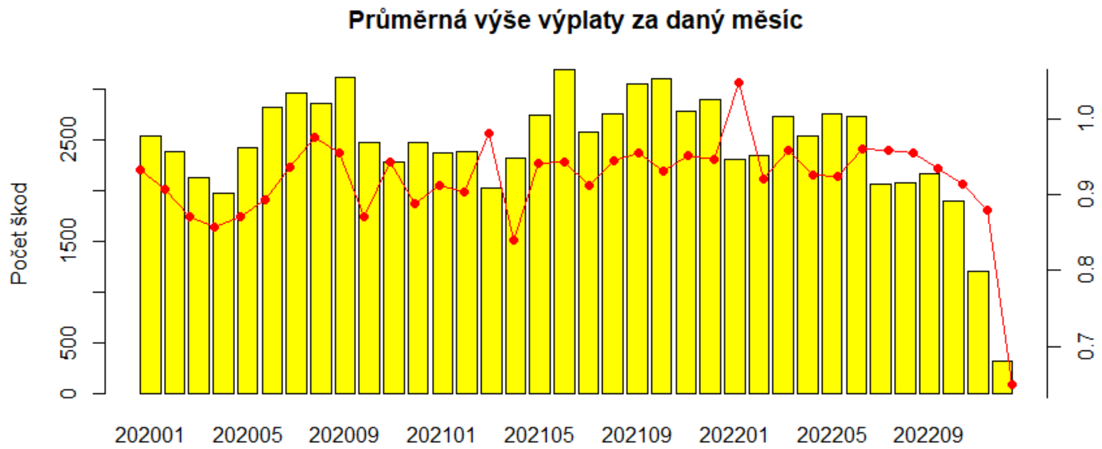
V této sekci si definujeme nástroj, který nám umožní posuzovat modely, které nejsou vnořené. Akaikovo informační kritérium se používá pro výběr nejlepšího modelu z celé množiny modelů. Pomocí AIC se snažíme najít model, který nejlépe vysvětluje data s pomocí nejmenšího počtu parametrů - skládá se tedy z kombinace dvou faktorů: přizpůsobení modelu datům a počtu vysvětlujících v modelu.

Definice 4 (Akaikovo informační kritérium)**.** Nechť p je počet parametrů v daném modelu. Pak Akaikovo informační kritérium definujeme jako:
AIC $=$ - 2 * $\log(maxL(y;\hat{\mu},\hat{\phi})) + 2p =$ - 2 * $\ell(y;\hat{\mu},\hat{\phi}) + 2p$

Nižší hodnota AIC indikuje, že model má lepší rovnováhu mezi přizpůsobením se datům a jednoduchostí modelu. Čím vyšší je odhad logaritmické věrohodnostní funkce, tím blíže je model k pravému rozdělení.

4. Praktická část

Cíl praktické části této Bakalářské práce je identifikace škodní inflace v autopojištění na povinném ručení na českém trhu. Tyto komplexní hodnoty nelze měřit pouze jednodimenzionálním pohledem, jelikož do výše vyplacených škod vstupuje více vlivů. Pohlédněme na tyto reálné hodnoty průměrné výplaty z dat pojišťovny na škodách uzavřených v období 2020-2022, rozdělené do měsíců v obrázku 4.1. Zobrazujeme si pouze poslední dva uběhlé roky. Sloupce představují počty uzavřených škod dle měsíce vzniku a červené body průměrnou výplatu na škodě, která je zde kvůli konfidentalitě podělena celkovou průměrnou výplatou přes všechny škody. Zaměříme-li se na porovnání meziroční průměrné výplaty, získali bychom odhad navýšení v jednotkách procent. To neodpovídá hodnotám ČNB a Českého statistického úřadu a ani realitě. V posledních měřených měsících si můžeme všimnout méně uzavřených škod a dokonce se nám průměrná výše škody snižuje.



Obrázek 4.1: Průměrná výše výplaty v měsících

Přidáme-li do úvahy hodnoty z tabulky 4.1, kde vyčteme průměrnou výplatu škody dle trvání času vypořádání škody, vidíme, že škody rychleji uzavřené (v prvních měsících) mají menší výši, zatímco déle trvající uzavření pojistné události vede až k dvojnásobným výplatám. Dokážeme tímto vysvětlit pokles průměrné výše škod za poslední kvartál a ptát se, jak tedy vypadá opravdu reálná výše škody za poslední kvartál po vyloučení tohoto vlivu? Jaké další faktory mají dopad na průměrnou výši škod a jak je každý z nich vlivný?

Tabulka 4.1: Průměrná výplata dle prodlení uzavření škody od vzniku

Prodleni	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Vyplata	0.72	0.90	0.97	1.10	1.18	1.38	1.54	1.72	1.69	1.62	1.79	2.14	1.56	1.82	2.05	1.46	1.60	1.56	2.10

Proto jsem se rozhodl na problematiku nahlédnout pomocí zobecněných lineárních modelů a pokusit se identifikovat míru inflace s jejich pomocí.

4.1 Popis dat

Data jsou poskytnuty ze známé české pojišťovny a množina je vybraná na základě data prvního uzavření mezi 2020-2022. Byla-li škoda znovuotevřena a v čase do stáhnutí dat změněna výše škody, výplata je aktuální. Výše škod je bez regresu a bez externích nákladů (honoráře, využití databází a dalších zdrojů informací apod.). Množina je vybrána na základě následující úvahy: Škodní inflaci chceme měřit jako navýšení průměrné výplaty v čase. Bereme pouze uzavřené škody, abychom znali celkovou výši škody (rezervy nejsou spolehlivé) a zahrnuli všechny výplaty na škodě. Životní (zdravotní) škody jsou vyřazeny, jelikož se chovají odlišně. U životních škod je běžné, že uzavření trvá roky a především nahlášení škody má také velkou prodlevu, jelikož nastávají situace, kde zpětně poškozený zjistí, že vzniklé problémy se vážou k dané škodě. Nulové škody jsou také odebrány.

Škody máme ještě rozdělené dle počtu výplat, kvůli většímu detailu, na co byla výplata použita. V datech máme faktory:

- **MES_1_UZ** - Měsíc prvního uzavření škody ve formátu YYYYMM.
- **HLASENI_UZAVRENI_M** - Doba do uzavření škody v měsících od vzniku škody. Nabývá hodnot 0, 1, . . . , 18, kde v hodnotě 18 jsou zahrnuty i všechny hodnoty větší.
- **VYPLATA** - Velikost výplaty podělená průměrnou výplatou.
- **ZARAZENI** - Zařazení škody dle typu výplaty. Kategorie jsou podrobně rozepsány níže.
- **TYP_SUBCLAIMU** - Rozřazení, zda je škoda vozidlová či majetková.
- **VUZ_ROKY** - Stáří vozu poškozeného, máme-li informaci. Hodnoty N/A, 0, . . . , 18, 18+.
- **POPIS_KODU_VOZIDLA** - Rozřazení do kategorií: vozidlo osobní, motocykl, návěs, autobus, nákladní, přívěs, tahač, ostatní, N/A.
- **STAT_SKODY** - Země škody dle mezinárodních zkratek: AT - Austria, BE - Belgium, CH - Switzerland, CZ - Czech Republic, DE - Germany, ES - Spain, FR - France, GB - United Kingdom, HR - Croatia, HU - Hungary, IT - Italy, NL - Netherlands, PL - Poland, SK - Slovakia, xxx - ostatní.
- **POCET_AUT** - Počet poškozených vozidel v nehodě. Hodnoty 1, 2, 3 a N/A.
- **FIL** - Informace o smlouvě - flotilová, leasingová a retailová.
- **MESIC_VZNIKU** - Datum vzniku ve formátu YYYYMM. Použito na vypočítání prodlení uzavření škody, v regresním modelu proměnnou nepoužíváme.

Kategorie zařazení škody:

- TŠ - Totální škoda vozidla. Vůz je neopravitelný, či cena opravy je větší než cena vozidla.
- zahraniční - Škoda se stala v zahraničí.
- odtah - Cena pouze za odtah vozidla.
- ostatní majetek - Majetková škoda.
- náhradní vozidlo - Cena za náhradní vozidlo.
- ROZPOČET - Určí se částka dle vyhodnocení stávající ceny vozidla a závažnosti škody. Ta je vyplacena přímo klientovi. Klient si opravu řeší sám.
- hasiči - Částka vyplacená za zásah hasičů.
- ZNAČKA - Vozidlo opraveno ve značkovém servisu. Pojišťovna má se servisem smlouvu.
- NEZNAČKA - Vozidlo opraveno ve neznačkovém servisu. Pojišťovna má se servisem smlouvu.
- NESMLUVNÍ-NEZNAČKA - Vozidlo bylo opraveno v neznačkovém a nesmluvním servisu. Pojišťovna nemá se servisem smlouvu.
- skla - Cena pouze za opravu či výměnu skla.
- other - Ostatní.

4.2 Stavba modelu

Výběr modelu

V GLM závisí volba rozdělení pro proměnnou odezvy na povaze dat a zkoumané problematice. Již jsme zmiňovali v sekci 2.1.1 Rozdělení výše škod, jaký je důvod použití gama rozdělení k modelování výše škod. V případě analýzy výše pojistných událostí se předpokládá, že proměnná odezvy má gama rozdělení a proto používáme gama GLM k modelování vztahu mezi odezvou a vysvětlujícími proměnnými a k odhadu vlivu těchto proměnných na průměr hodnoty odezvy.

V praxi nejčastěji používanou linkovou funkcí pro modelování výše pojistných událostí s gama rozdělením je logaritmická linková funkce, viz Murphy et al., 2000. Již jsme zmiňovali v sekci 2.2.1, že výhodou logaritmické linkové funkce je, že dokáže pracovat s kladnými spojitými proměnnými odezvy, což je běžná vlastnost výše pojistných událostí. Kromě toho umožňuje aditivní vlastnost predikčních proměnných na logaritmicky transformovanou proměnnou odezvy, což je užitečné pro interpretaci výsledků modelu ve smyslu procentuální změny závažnosti pojistných událostí spojené s jednotkovou změnou predikční proměnné.

Navíc logaritmická spojovací funkce má vlastnost, že predikované hodnoty proměnné odezvy zůstávají kladné, což je důležité, protože gama rozdělení má nosič pouze na kladných reálných číslech, oproti od kanonické linkové funkce.

Metody pro vytváření modelů

V této sekci si představíme některé z metod používaných při stavění modelu. Všechny metody jsou iterativní s nadefinovaným koncovým stavem. První představovaná metoda se nazývá *Forward selection*. Funguje na metodě přidávání predikčních proměnných do modelu, dokud není splněno kritérium zastavení. V každé iteraci je do modelu přidána predikční proměnná, která dle zvoleného kritéria přispívá k modelu nejvíce. Proces pokračuje, dokud žádná další proměnná nesplňuje kritérium přidání do modelu. Výhodou Forward selection je, že je oproti ostatním metodám výpočetně nenáročná, zejména pokud je počet predikčních proměnných relativně malý. Nemusí však identifikovat nejlepší podmnožinu prediktorů, zejména pokud existují kolineární proměnné. Může se také stát, že postupným přidáváním proměnných se některé již zařazené regresory, které se na začátku zdáli jako velmi významné, stanou redundantními.

Další metodou používanou při sestavování GLM je *Backward elimination*, která začíná s plným modelem (model se zařazenými všemi faktory). Spočívá v iterativním odstraňování predikčních proměnných z modelu, dokud není splněno kritérium zastavení. V každé iteraci je z modelu odstraněna jedna predikční proměnná, která dle zvoleného kritéria přispívá k modelu nejméně. Proces pokračuje, dokud žádná další proměnná nesplňuje kritérium odebrání. Výhodou zpětné eliminace je, že dokáže identifikovat významnou podmnožinu prediktorů, zejména pokud existují kolineární proměnné. Po eliminaci nám v modelu zůstanou pouze proměnné, které nejlépe vysvětlují odezvu. Existuje zároveň risk, že některé z dříve odebraných faktorů se pro model stanou zásadními poté, co odebereme jiné proměnné. Důvodem je fakt, že významnost proměnné může záviset na přítomnosti či nepřítomnosti jiných proměnných v modelu.

Kombinovaná metoda *Forward-Backward* výběru využívá výhod obou metod. Zahrnuje iterativní přidávání a odstraňování predikčních proměnných z modelu, dokud není splněno kritérium zastavení. V každé iteraci, za splnění kritérií pro přidání a vyřazení, je do modelu přidána predikční proměnná, která podle zvoleného kritéria přispívá k modelu nejvíce a odstraněna proměnná, která přispívá modelu nejméně. Proces pokračuje, dokud žádná další proměnná nesplňuje kritérium zastavení, tedy nelze již žádnou proměnnou přidat a odebrat. Výhodou kombinovaného postupu Forward-Backward je, že dokáže určit nejlepší podmnožinu prediktorů z hlediska zvoleného kritéria, ovšem může být výpočetně náročná. V širším slova smyslu je důležité poznamenat, že metoda stále může trpět na problémy souvisejícími s kolinearitou a nemusí vždy identifikovat nejlepší podmnožinu prediktorů.

4.3 Vytváření modelu

K vytvoření modelu použijeme statistický program R kvůli transparentnosti kódu. V příloze nalezneme celý mnou vytvořený mechanický postup vytvoření modelu a zde se budeme věnovat pouze výsledkům a verifikaci se zabudovanými funkcemi R.

4.3.1 Průběh jedné iterace v kombinované metodě

Kritérium AIC (4) slouží k rozhodování, zda přidání určitého faktoru zlepší stávající model. Nicméně je důležité si uvědomit, že konkrétní hodnota AIC sama o sobě neobsahuje žádnou informaci o modelu. Hodnotu AIC nelze přímo interpretovat jako přínosnou nebo nepřínosnou. Místo toho lze hodnoty AIC pouze porovnávat mezi různými modely a na základě toho určit, který model nejlépe odpovídá datům. Porovnávání hodnot AIC modelů mezi sebou může pomoci v rozhodování, který model nejlépe odhaduje výši škod.

Na základě hodnoty AIC budeme rozhodovat o výběru faktorů. Uvedeme si zde první iteraci algoritmu. Počáteční model, tzv. null_model nadefinujeme jako:

```
model <- glm(data$Vyplata ~ 1, data = data,
family = Gamma(link = "log"))
predictor_vars <- c( "HLASENI_UZAVRENI_M", "ZARAZENI",
"TYP_SUBCLAIMU", "STAT_SKODY", "VUZ_ROKY",
"POPIS_KODU_VOZIDLA","POCET_AUT","FIL","MES_1_UZ")
#faktory k dispozici
model
```

```
Call:glm(formula = data$Vyplata ~ 1, family = Gamma(link = "log"),
    data = data)

Coefficients:
(Intercept)
  6.837e-14

Degrees of Freedom: 100658 Total (i.e. Null);  100658 Residual
Null Deviance:          138900
Residual Deviance: 138900        AIC: 20130
```

KROK 1: Vygenerujeme jednotlivé modely se závislostí na každém dostupném faktoru a porovnáme hodnoty AIC vygenerovaných modelů.

```
AIC(model) #AIC našeho null modelu
selected_models <- create_glm_models(data, predictor_vars) #volání
funkce na vytvoření jednotlivých modelů, viz příloha
model_selection_summary(selected_models)
```

```
[1] 201257.9

"Model 1: HLASENI_UZAVRENI_M | AIC: 191842.38 | Scaled Deviance:
128429.45"
"Model 2: ZARAZENI | AIC: 147398.35 | Scaled Deviance: 87852.50"
"Model 3: TYP_SUBCLAIMU | AIC: 201097.28 | Scaled Deviance:
138727.52"
"Model 4: STAT_SKODY | AIC: 189481.96 | Scaled Deviance:125929.07"
```

```
"Model 5: VUZ_ROKY | AIC: 196431.40 | Scaled Deviance: 133426.77"
"Model 6: POPIS_KODU_VOZIDLA | AIC: 193411.98 | Scaled Deviance:
130143.32"
"Model 7: POCET_AUT | AIC: 199658.57 | Scaled Deviance: 137082.75"
"Model 8: FIL | AIC: 197908.67 | Scaled Deviance: 135112.63"
"Model 9: MES_1_UZ | AIC: 200325.85 | Scaled Deviance: 137768.46"
```

Zde máme vytvořené jednotlivé modely, kde je do našeho modelu (v kódu proměnná model) přidán zvlášť každý z faktorů. Největší zlepšení (v odhadu parametrů modelu), dle kritéria AIC, nastane při přidání faktoru ZARAZENI. Přidáme ho do modelu:

```
model <- glm(Vyplata ~ ZARAZENI, data = data, family =
Gamma(link = "log"))
predictor_vars <- c( "HLASENI_UZAVRENI_M", "TYP_SUBCLAIMU",
"STAT_SKODY", "VUZ_ROKY","POPIS_KODU_VOZIDLA","POCET_AUT",
"FIL","MES_1_UZ")
pridane = c("ZARAZENI") #faktor jiz v modelu
elimination_list = c("ZARAZENI")
#faktor v modelu pro testování odebrání
```

Nyní otestujeme, zda se v modelu některá ze zařazených faktorů nestala redundantní.

```
selected_models <-backwards_elimination(model,elimination_list)
AIC(model)
model_selection_summary(selected_models)
```

```
## [1] 147398.4
## [1] "Model 1: ZARAZENI | AIC: 201257.94 |
Scaled Deviance: 138914.01"
```

Máme zde vygenerovaný pouze jeden submodel, jelikož v modelu se zatím nachází pouze jeden faktor. Jeho vyřazením bychom se vrátili zpět na začátek kroku 1 a tedy je intuitivní, že ho zde nevyřazujeme.

V kroku 2 opakujeme tento proces, tedy nejprve určíme, zda lze některý z faktorů zařadit do modelu a případně vybrat ten, který přinese největší zlepšení. Poté otestujeme, zda by se model nezlepšil odebráním některého z již zahrnutých faktorů. Tento proces opakujeme do té doby, kdy na základě porovnání AIC nebude možné v jednom kroku přidat či odebrat faktor.

4.3.2 Výsledek modelu

Aplikováním iterační metody sestavíme model. V příloze se nachází jednotlivé kroky vytvořené metody, které se zcela shodují s výstupem zabudované funkce *stepAIC* z knihovny MASS. Metoda proběhne úspěšně do konce s následujícími kroky: Počáteční AIC null modelu je 201257.9. Iterační metoda postupně přidává faktory v pořadí : zarazeni, hlaseni_uzavreni_m, vuz_roky, stat_skody, pocet_aut, mes_1_uz, popis_kodu_vozidla, fil, typ_subclaimu. Všechny faktory se podílí na zlepšení modelu. Metoda s vybranými daty v žádném kroku neodebírá faktory a na základě hodnoty AIC neprokáže po přidání každého z faktorů redundantnost jiného faktoru. Finální model má hodnotu AIC=134423.6 a nabývá tvaru

```
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
STAT_SKODY + POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA +
FIL + TYP_SUBCLAIMU, data = data, family = Gamma(link = "log")).
```

Rozepíšeme jednotlivé rozdíly ve výsledné hodnotě AIC:

AIC	Rozdíl	Přidaný faktor
201257.9		
147398.4	53859.5	ZARAZENI
143471.6	3926.8	HLASENI_UZAVRENI_M
140456.5	3015.1	VUZ_ROKY
138018.6	2437.9	STAT_SKODY
136373.3	1645.3	POCET_AUT
135635.6	737.7	MES_1_UZ
134762.4	873.2	POPIS_KODU_VOZIDLA
134483.2	279.2	FIL
134423.6	59.6	TYP_SUBCLAIMU

Tabulka 4.2: AIC hodnoty a rozdíly mezi jednotlivými modely

Postupné snižování hodnoty AIC nám umožňuje nalézt model, který poskytuje nejlepší kompromis mezi přesností modelu a jednoduchostí (počtem zahrnutých regresorů). Nicméně, je důležité býti obezřetný a ptát se, zda přidání dalšího faktoru do modelu přináší dostatečné zlepšení výkonu modelu vzhledem k jeho složitosti. Pro ověření statistické významnosti nově přidaného faktoru použijeme věrohodnostní test 3.1.1. Pokud bychom nebyli opatrní a přidávali bychom nové proměnné do modelu bez ověření jejich významnosti, mohlo by to vést k jeho přizpůsobení příliš specifickým datům, což by vedlo k neschopnosti modelu vyloučit náhodné vlivy a nemuseli bychom identifikovat systematické trendy. V důsledku bychom dostali špatnou predikci pro nová data. Přidáním faktoru FIL se zmenšila hodnota AIC o 279.2. Provedeme věrohodnostní test pro model s faktorem a jemu vnořený model 3.1.1 na hladině $\alpha = 0.05$ s H_0: platí submodel_t a H_1:platí model_t.

```
library(lmtest)
submodel_t <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
+ STAT_SKODY + POCET_AUT + POPIS_KODU_VOZIDLA + MES_1_UZ, family =
=Gamma(link = "log"), data = data)

model_t    <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
+ STAT_SKODY + POCET_AUT + POPIS_KODU_VOZIDLA + MES_1_UZ+ FIL, family =
=Gamma(link = "log"), data = data)
lrtest <- lrtest(model_t,submodel_t)
```

Výsledek testu je:

```
Likelihood ratio test

Model 1: Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
+ STAT_SKODY + POCET_AUT + POPIS_KODU_VOZIDLA + MES_1_UZ + FIL

Model 2: Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
+ STAT_SKODY + POCET_AUT + POPIS_KODU_VOZIDLA + MES_1_UZ
  #Df LogLik Df  Chisq Pr(>Chisq)
1  79 -67163
2  77 -67304 -2 283.28  < 2.2e-16
```

Výsledek testu ukazuje, že Model 1 (model_t) je při vysvětlování odezvy lepší než Model 2 (submodel_t), což naznačuje testovací statistika Chisq = 283.28, s 2 stupni volnosti a velmi malá p-hodnota ($< 2.2 * 10^{-16}$), která je mnohem menší než hladina významnosti 0.05. Na základě tohoto výsledku zamítáme nulovou hypotézu s velkou rezervou ve prospěch H_1. To naznačuje, že proměnná FIL je důležitým prediktorem *Výplaty* a měla by být zahrnuta do modelu.

Provedeme stejný test pro poslední přidanou proměnnou TYP_SUBCLAIMU.

```
model_t    <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
+ STAT_SKODY + POCET_AUT + MES_1_UZ+ POPIS_KODU_VOZIDLA  + FIL +
+ TYP_SUBCLAIMU, data = data, family = Gamma(link = "log"))
submodel_t <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +
+ STAT_SKODY + POCET_AUT + MES_1_UZ+ POPIS_KODU_VOZIDLA  + FIL,
family = Gamma(link = "log"), data = data)
lrtest <- lrtest(model_t,submodel_t)
```

S výsledkem testu

```
Likelihood ratio test

Model1:Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY+
    POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA + FIL + TYP_SUBCLAIMU
Model2:Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY+
    POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA + FIL
  #Df LogLik Df  Chisq Pr(>Chisq)
1  80 -67132
2  79 -67163 -1 61.522  4.378e-15
```

Výsledek testu vypovídá, že Model 1 (model_t) je při vysvětlování odezvy lepší než Model 2 (submodel_t) s testovací statistikou Chisq = 65.522, s 1 stupněm volnosti a velmi malou p-hodnotou. Zamítáme nulovou hypotézu ve prospěch alternativy.

4.4 Škodní inflace

S dokončeným modelem se přesuneme na téma škodní inflace. Rozhodli jsme se měřit inflaci pomocí zahrnutí faktoru času do GLM jako další vysvětlující proměnnou pro výši škody. Celkovou výši škody stanovujeme k datu uzavření škody, z čehož získáváme zařazení pro časový údaj o škodě. Proto v modelu máme zahrnut faktor měsíc uzavření škody. Zaměříme se nyní pouze na tento faktor a otázku, zda po zohlednění ostatních vlivů má tento faktor vliv na výši výplaty a případně jaký?

Nejprve zodpovíme na otázku, zda lze prokázat statisticky významnou hodnotu faktoru MES_1_UZ v regresním modelu. Provedeme věrohodnostní test stejným způsobem a výsledek testu je

```
Likelihood ratio test

Model 1: Vyplata~ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY+
   + POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA + FIL + TYP_SUBCLAIMU
Model 2: Vyplata~ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY+
  +  POCET_AUT + POPIS_KODU_VOZIDLA + FIL + TYP_SUBCLAIMU
  #Df LogLik  Df  Chisq Pr(>Chisq)
1 114 -66987
2  79 -67564 -35 1154.2  < 2.2e-16
```

Testová statistika je vysoká a p-hodnota $< 2.2 * 10^{-16}$, tedy nulovou hypotézu zamítáme s velkou rezervou. Pomocí výsledku věrohodnostní test víme, že faktor je v modelu důležitý. V postaveném modelu můžeme zjistit, jak se konkrétní výše výplaty chová v závislosti na čase. Vytvoříme si tu modelový příklad, kterému stanovíme pevné faktory a budeme hýbat pouze tím, kdy byla škoda uzavřena.

```
for (date in c(202001, 202112, 202212)) {
  zarazeni <- "TŠ"
  hlaseni_uzavreni_m <- 3
  vuz_roky <- 5
  stat_skody <- "CZ"
  pocet_aut <- 2
  popis_kodu_vozidla <- "OSOBNÍ"
  fil <- "flotily"
  typ_subclaimu <- "V"
  MES_1_UZ <- date

  newdata <- data.frame(ZARAZENI = zarazeni,
                        HLASENI_UZAVRENI_M = factor(hlaseni_uzavreni_m),
                        VUZ_ROKY = factor(vuz_roky),
                        STAT_SKODY = stat_skody,
                        POCET_AUT = factor(pocet_aut),
                        POPIS_KODU_VOZIDLA = popis_kodu_vozidla,
                        FIL = fil,
                        TYP_SUBCLAIMU = typ_subclaimu,
                        MES_1_UZ=factor(MES_1_UZ)
  )

  pred <- predict(model, newdata = newdata, type = "response")
```

```
  print(pred)
}
```

Jako výstup získáme tři hodnoty 6.5209, 7.485288, 7.865677, které vypovídají o výši výplaty vzhledem k celkové průměrné výši výplaty dat. Hodnoty se nám opravdu liší a o dost rostou. Hodnoty jsou obecně takto velké, jelikož máme totální škodu a dvě poškozená auta. Ve identickém scénáři s jediným rozdílem zařazení "skla"místo TŠ bychom měli hodnoty 1.102781, 1.265873, 1.330203, které také reflektují růst v čase.

Z modelu se zaměříme na odhadnuté koeficienty regresoru MES_1_UZ. Model si stanovil hodnotu u 202001 jako *základ* a ostatní koeficienty jsou vyjádřeny ve vztahu vůči tomuto základu. V následující tabulky máme vyjádřeny hodnoty parametrů z postaveného modelu.

Tabulka 4.3: Regresní výsledky parametru MES_1_UZ

Proměnná	**Koeficient**	**Std. Error**	**t-value**
MES_1_UZ202002	-0.0719	0.0343	-2.098
MES_1_UZ202003	-0.0328	0.0347	-0.943
MES_1_UZ202004	0.0071	0.0371	0.192
MES_1_UZ202005	-0.0145	0.0381	-0.382
MES_1_UZ202006	0.0092	0.0356	0.257
MES_1_UZ202007	-0.0004	0.0365	-0.011
MES_1_UZ202008	0.0288	0.0367	0.783
MES_1_UZ202009	-0.0198	0.0358	-0.553
MES_1_UZ202010	-0.0219	0.0358	-0.611
MES_1_UZ202011	-0.0125	0.0370	-0.339
MES_1_UZ202012	-0.0076	0.0368	-0.206
MES_1_UZ202101	-0.0211	0.0360	-0.586
MES_1_UZ202102	0.1826	0.0372	4.913
MES_1_UZ202103	0.0499	0.0357	1.397
MES_1_UZ202104	0.0378	0.0381	0.994
MES_1_UZ202105	0.0526	0.0371	1.419
MES_1_UZ202106	0.0459	0.0361	1.271
MES_1_UZ202107	0.1156	0.0394	2.938
MES_1_UZ202108	0.0299	0.0374	0.800
MES_1_UZ202109	0.0484	0.0365	1.325
MES_1_UZ202110	0.0580	0.0361	1.607
MES_1_UZ202111	0.0789	0.0347	2.274
MES_1_UZ202112	0.1379	0.0355	3.886
MES_1_UZ202201	0.0950	0.0358	2.649
MES_1_UZ202202	0.1427	0.0361	3.953
MES_1_UZ202203	0.1177	0.0339	3.484
MES_1_UZ202204	0.1987	0.0371	5.191
MES_1_UZ202205	0.1677	0.0356	4.709
MES_1_UZ202206	0.1340	0.0353	3.791
MES_1_UZ202207	0.3049	0.0379	8.034
MES_1_UZ202208	0.2519	0.0362	6.951
MES_1_UZ202209	0.1854	0.0368	5.043
MES_1_UZ202210	0.1930	0.0359	5.377
MES_1_UZ202211	0.2002	0.0356	5.631
MES_1_UZ202212	0.1875	0.0365	5.140

Z kapitoly 2.3.1 Tvar pro gama model si dokážeme vyjádřit vztah

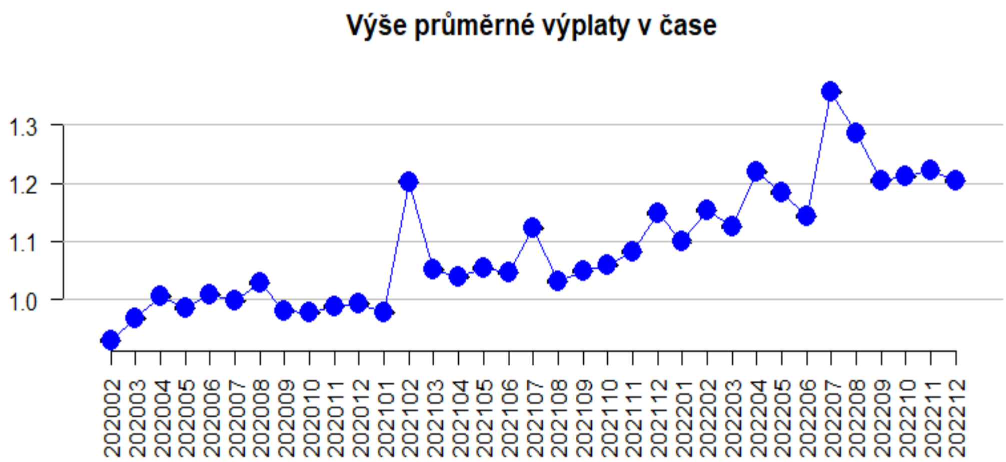
$$\hat{y} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p + x_p).$$

Dopočítáme jednotlivé hodnoty koeficientů.

Tabulka 4.4: Dopočítané hodnoty koeficientů

Proměnná	**Koeficienty**	exp(**Koeficienty**)
MES_1_UZ202002	-0.0719	0.9303
MES_1_UZ202003	-0.0328	0.9677
MES_1_UZ202004	0.0071	1.0071
MES_1_UZ202005	-0.0145	0.9856
MES_1_UZ202006	0.0092	1.0092
MES_1_UZ202007	-0.0004	0.9996
MES_1_UZ202008	0.0288	1.0292
MES_1_UZ202009	-0.0198	0.9804
MES_1_UZ202010	-0.0219	0.9783
MES_1_UZ202011	-0.0125	0.9876
MES_1_UZ202012	-0.0076	0.9924
MES_1_UZ202101	-0.0211	0.9792
MES_1_UZ202102	0.1826	1.2007
MES_1_UZ202103	0.0499	1.0515
MES_1_UZ202104	0.0378	1.0380
MES_1_UZ202105	0.0526	1.0539
MES_1_UZ202106	0.0459	1.0466
MES_1_UZ202107	0.1156	1.1228
MES_1_UZ202108	0.0299	1.0303
MES_1_UZ202109	0.0484	1.0493
MES_1_UZ202110	0.0580	1.0590
MES_1_UZ202111	0.0789	1.0824
MES_1_UZ202112	0.1379	1.1477
MES_1_UZ202201	0.0950	1.0992
MES_1_UZ202202	0.1427	1.1534
MES_1_UZ202203	0.1177	1.1242
MES_1_UZ202204	0.1987	1.2205
MES_1_UZ202205	0.1677	1.1826
MES_1_UZ202206	0.1340	1.1433
MES_1_UZ202207	0.3049	1.3565
MES_1_UZ202208	0.2519	1.2856
MES_1_UZ202209	0.1854	1.2037
MES_1_UZ202210	0.1930	1.2127
MES_1_UZ202211	0.2002	1.2212
MES_1_UZ202212	0.1875	1.2051

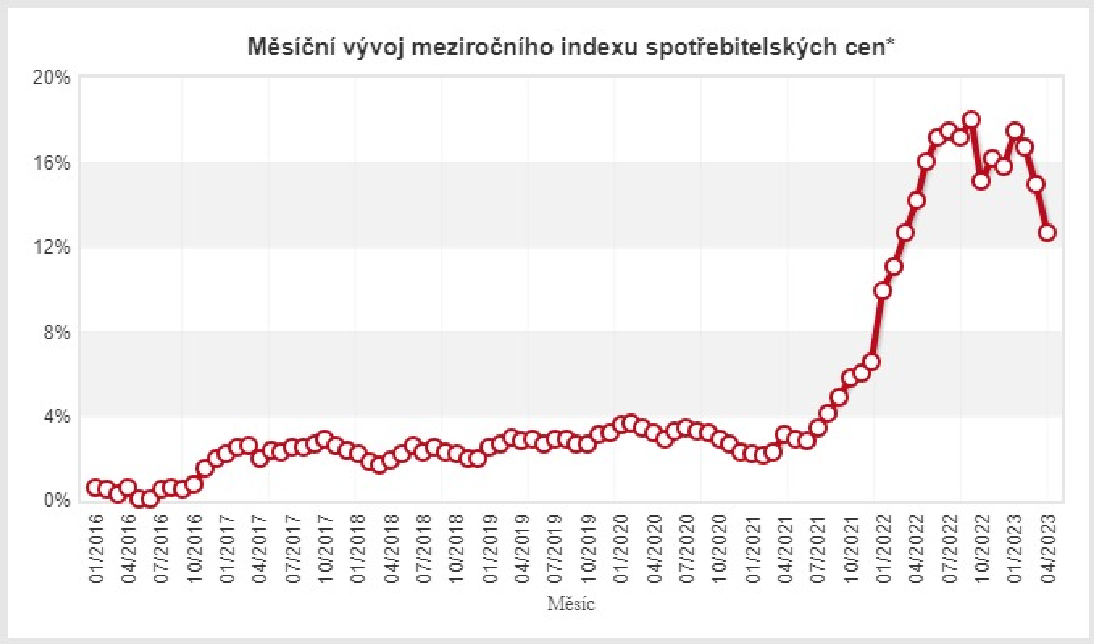
Dopočítané hodnoty vznikly izolováním faktoru času v regresním modelu na daných datech. V úvodu do praktické části zmiňujeme problematiku měření inflace a ve skutečnosti nevyřazujeme, spíše upozorňujeme na fakt provázanosti inflace na socio-ekonomických vlivech celkového hospodaření států a aktuálního stavu trhu. Nemůžeme také zcela vyřadit potenciální existenci jiných komplexnějších nelineárních vztahů. Pokud se v této práci omezíme na předpoklad, že všechny další vlivy jsou zanedbatelné a důležité vlivy máme v této práci a modelu zahrnuté, tato metoda by nám měla přiblížit charakter inflace. V následujícím grafu lze pozorovat dopočítané hodnoty, které zde prohlašujeme za škodní inflaci. Lze stále pozorovat lehký šum, tedy je možné, že může existovat ještě jiný statisticky významný faktor, nebo se může jednat o chybu dat (například v měsíci 202102 mohli být zaznamenány velmi nespecifické škody a stálo by za zvážení data blíže prozkoumat a případně škodu či sadu škod vyřadit z modelu, jelikož můžou zkreslovat výsledek). Je také vidět jistý trend.



Obrázek 4.2: Výše průměrné výplaty dle měsíce

Budeme-li tedy chtít odpovědět na otázku, jakou inflaci měříme za poslední dva uběhlé roky v autopojištění na povinném ručení, koeficient váhy v posledních měsících roku 2022 ukazuje navýšení přibližně o 20 % od ledna roku 2020. Meziroční inflaci za období 2020/12 - 2021/12 spočítáme jako $\frac{1.1477}{0.9924}$, jelikož obě hodnoty jsou vyjádřeny vzhledem k 2020/01. Získáme cca 15.6% meziroční nárůst cen a za období 2021/12 - 2022/12 stejnou metodou získáváme 5 %. Porovnáme tyto hodnoty s daty z Českého statistického úřadu. Následující graf představuje v každém bodě vývoj indexu spotřebitelských cen ke stejnému měsíci předchozího roku, který měří změny nákupní síly ve spotřebním koši. Je zobrazena procentní změna cenové hladiny ve vykazovaném měsíci daného roku proti stejnému měsíci předchozího roku. Vyjadřuje se k celému stavu české ekonomiky, ale stále můžeme vidět stejný trend růstu - největší růst inflace jsme zaznamenali v prvním kvartálu roku 2022. Třetí kvartál v růstu stagnoval a vysoké hodnoty lehce klesají. Z výsledků našeho modelu je vidět velmi podobný trend náhlého růstu a stagnace. Je důležité myslet na to, že námi zvýrazněný graf zobrazuje navýšení daného měsíce vůči základu a graf ČSÚ je oproti měsíci minulého roku.

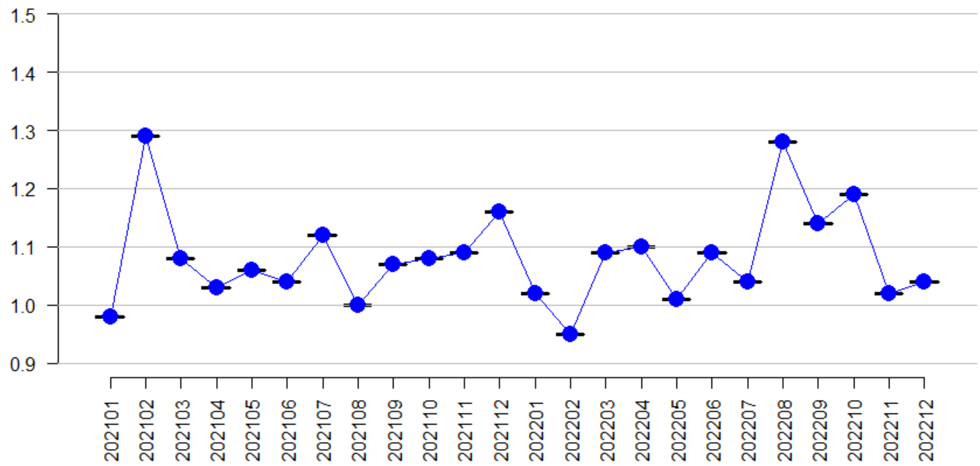
Zdroj obrázku:Český statistický úřad (2023).



Obrázek 4.3: Měsíční vývoj meziročního indexu spotřebitelských cen

Date	01	02	03	04	05	06	07	08	09	10	11	12
2021	0.98	1.29	1.08	1.03	1.06	1.04	1.12	1.00	1.07	1.08	1.09	1.16
2022	1.02	0.95	1.09	1.10	1.01	1.09	1.04	1.28	1.14	1.19	1.02	1.05

Tabulka 4.5: Meziroční nárůst v každém měsíci oproti stejnému měsíci minulého roku



Obrázek 4.4: Meziroční inflace každého měsíce oproti stejnému měsíci minulého roku

5. Závěr

V bakalářské práci jsme detailně vysvětlili základní pojmy a principy teorie zobecněných lineárních modelů. Rozebrali jsme fundamentální části pro tvorbu modelu, mezi které patří výběr rozdělení exponenciální třídy a výběr linkové funkce, který musí zohledňovat předpokládaný vztah mezi lineárním prediktorem a střední hodnotou vysvětlované proměnné. Nedílnou součástí jsou základní metody používané pro posuzování kvality modelů. Pro gama rozdělení jsme dokázali tvrzení, že patří do exponenciální třídy, ukázali kanonický link a odvodili přesný tvar škálované odchylky.

V praktické části jsme znalosti aplikovali a rozebrali základní metodiky stavby konkrétního zobecněného lineárního modelu pro aktuální data z české pojišťovny. Interpretovali jsme výsledky použitých algoritmů, které jsou získány pomocí iteračních algoritmů. Detailní postup metodiky se nachází v příloze. Jeden krok iterace je znázorněn na metodě vytvořené autorem a shoduje se s již zabudovanými metodami statistického programu. Výsledky algoritmu a vzniklé submodely jsme mezi sebou porovnali.

V práci je dokázáno, že časové období má zásadní vliv na výši škody. Vliv časového období je zanalyzován a konkrétně určen. Ukázali jsme jednotlivé meziroční míry inflace a naměřili jejich hodnotu za časové období 2020/2022. Zjištěním je, že inflace za dvouleté období dosahuje 20 % a jednotlivé meziroční míry navýšení dosahují 15 % a 5 %. Výsledek je validován za pomoci celkových statistik z Českého statistického úřadu, kde je pozorován podobný trend.

Literatura

J. Anděl. *Statistické metody*. Čtvrté opravené vydání. Matfyzpress, Praha, 2007. ISBN 80-7378-003-8.

Adnan M Awad. Properties of the akaike information criterion. *Microelectronics Reliability*, 36(4):457–464, 1996.

B. Jorgensen. *The theory of exponential dispersion models and analysis of deviance*. Chapman and Hall, London, 1st edition, 1997. ISBN 978-0412055514.

Ron C. Mittelhammer, George G. Judge, and Douglas J. Miller. *Econometric Foundations*. Cambridge University Press, New York, 2000. ISBN 0-521-62394-4.

K. P. Murphy, M. J. Brockman, and P. K. W. Lee. Using generalized linear models to build dynamic pricing systems for personal lines insurance. In *CAS Winter 2000 Forum*, Washington, D.C., 2000.

Esbjorn Ohlsson and Bjorn Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. EAA Lecture Notes. Springer-Verlag Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-10791-7.

G.A.F. Seber and A.J. Lee. *Linear Regression Analysis*. Wiley, New York, druhé vydání edition, 2003. ISBN 0-471-41540-5.

K. Zvára. *Regrese*. Vydání první. Matfyzpress, Praha, 2008. ISBN 978-80-7378-041-8.

Český statistický úřad. Měsíční vývoj meziročního indexu spotřebitelskych cen, 2023. URL `https://www.czso.cz/csu/czso/inflace_spotrebitelske_ceny`.

Seznam tabulek

6. Přílohy

První příloha - kód R. Obsahuje mnou vytvořenou proceduru na stavbu modelu, ve které jsou mechanickým postupem detailně znázorněny jednotlivé kroky v budování modelu.

Druhá příloha - použití implementované metody stepAIC z library(MASS).

Příloha č.1

```r
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data <- read_excel("C:/Users/vojta/Desktop/Škola/Bakalářská práce - Škodní inflace v autopojištění/Data
_prakticka_cast/Data_R.xlsx")
data <- data[, !(colnames(data) %in% c("IND_SML", "MESIC_VZNIKU"))]

data <- within(data, {
  FIL = factor(FIL)
  TYP_SUBCLAIMU = factor(TYP_SUBCLAIMU)
  STAT_SKODY = factor(STAT_SKODY)
  ZARAZENI = factor(ZARAZENI)
  POPIS_KODU_VOZIDLA = factor(POPIS_KODU_VOZIDLA)
  POCET_AUT = factor(POCET_AUT)
  HLASENI_UZAVRENI_M =factor(HLASENI_UZAVRENI_M)
  VUZ_ROKY = factor(VUZ_ROKY)
  Vyplata = as.numeric(Vyplata)
  MES_1_UZ = factor(MES_1_UZ)

})

# funkce vytvarejici modely pro vsechny predictor_vars
create_glm_models <- function(data, predictor_vars) {
  selected_models <- list()
  for (var in predictor_vars) {
    formula <- as.formula(paste0("data$Vyplata ~",pridane,"+", var))
    model <- glm(formula, data = data, family = Gamma(link = "log"))
    selected_models[[var]] <- model
  }
  return(selected_models)
}

#funkce na vypocet AIC a scaled deviance
model_selection_summary <- function(selected_models) {
  for (i in seq_along(selected_models)) {
    model <- selected_models[[i]]
    Aic_model <- AIC(model)
    Scaled_deviance <- deviance(model, scale = TRUE)
    print(paste(
      sprintf("Model %d: %s", i, names(selected_models)[i]),
      sprintf("AIC: %.2f", Aic_model),
      sprintf("Scaled Deviance: %.2f", Scaled_deviance),
      sep = " | "
    ))
  }
}

#funkce na vytvareni submodelu při backwards eliminaci
backwards_elimination <- function(model, elimination_list) {
  selected_models <- list()
  for (var in elimination_list) {
    formula <- formula(update(model, as.formula(paste(".~.-", paste(c(promenne, c(var)), collapse="-
"))))) #odebiram prommene (odebrane) a act var
    new_model <- glm(formula, data = data, family = Gamma(link = "log"))
    selected_models[[var]] <- new_model
  }
  return(selected_models)
}

#Volání výše naprogramovaných funkcí:
```

```
#eliminační
#selected_models <-backwards_elimination(model,elimination_list)
#přidávací
#selected_models <- create_glm_models(data, predictor_vars)
#Vyhodnoceni AIC a SCALED DEVIANCE
#model_selection_summary(selected_models)


####################################################################################################
##############################################
summary(data)
```

```
##       FIL        HLASENI_UZAVRENI_M    MES_1_UZ     TYP_SUBCLAIMU
##  flotily:24878   0      :36116      202203 : 3551   M:11325
##  leasing:11813   1      :25223      202001 : 3479   V:89334
##  retail :63968   2      :13525      202002 : 3383
##                  3      : 6586      202111 : 3256
##                  18     : 4532      202003 : 3224
##                  4      : 3808      202206 : 3054
##                  (Other):10869      (Other):80712
##     Vyplata                       ZARAZENI       STAT_SKODY       VUZ_ROKY
##  Min.   :   0.0001   ZNAČKA            :20503   CZ     :93255   N/A    :44558
##  1st Qu.:   0.2262   náhradní vozidlo  :17960   DE     : 3160   18+    : 5220
##  Median :   0.4984   ROZPOČET          :16741   AT     :  896   3      : 3808
##  Mean   :   1.0000   NESMLUVNÍ-NEZNAČKA:12066   IT     :  885   2      : 3716
##  3rd Qu.:   1.0902   zahraniční        : 8525   SK     :  648   4      : 3628
##  Max.   :1259.7922   TŠ                : 8152   PL     :  305   1      : 3506
##                      (Other)           :16712   (Other): 1510   (Other):36223
##  POPIS_KODU_VOZIDLA POCET_AUT
##  OSOBNÍ  :56234     1  :88851
##  N/A     :37468     2  :  885
##  NÁKLADNÍ: 2881     3  :   21
##  OSTATNÍ : 1395     N/A:10902
##  TAHAČ   :  769
##  MOTOCYKL:  742
##  (Other) : 1170
```

```
#POČÁTEČNÍ STAV
model <- glm(data$Vyplata ~ 1, data = data, family = Gamma(link = "log"))
selected_models <- list() #prazdny list na vytvareni modelu v kazdem kroku
predictor_vars <- c( "HLASENI_UZAVRENI_M", "ZARAZENI", "TYP_SUBCLAIMU", "STAT_SKODY", "VUZ_ROKY","POPIS
_KODU_VOZIDLA","POCET_AUT","FIL","MES_1_UZ")
pridane = ("1") #promenne jiz pridane v modelu
elimination_list =c()  #promenne na testovani submodelu bez zarazenych promennych
promenne=c() #zde bude null, jelikoz budeme testovat pouze modely, kde odebirame po jedne promenne, ne
jako v backwards, kde jsme se odvíjeli od plneho modelu



####################################################################################################
##################################################



# KROK 1: Vygenerujeme jednotlivé modely se závislostí na každém parametru
AIC(model)  #AIC našeho null modelu
```

```
## [1] 201257.9
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: HLASENI_UZAVRENI_M | AIC: 191842.38 | Scaled Deviance: 128429.45"
## [1] "Model 2: ZARAZENI | AIC: 147398.35 | Scaled Deviance: 87852.50"
## [1] "Model 3: TYP_SUBCLAIMU | AIC: 201097.28 | Scaled Deviance: 138727.52"
## [1] "Model 4: STAT_SKODY | AIC: 189481.96 | Scaled Deviance: 125929.07"
## [1] "Model 5: VUZ_ROKY | AIC: 196431.40 | Scaled Deviance: 133426.77"
## [1] "Model 6: POPIS_KODU_VOZIDLA | AIC: 193411.98 | Scaled Deviance: 130143.32"
## [1] "Model 7: POCET_AUT | AIC: 199658.57 | Scaled Deviance: 137082.75"
## [1] "Model 8: FIL | AIC: 197908.67 | Scaled Deviance: 135112.63"
## [1] "Model 9: MES_1_UZ | AIC: 200325.85 | Scaled Deviance: 137768.46"
```

```
#nejvetsi zlepseni - ZARAZENI
model <- glm(Vyplata ~ ZARAZENI, data = data, family = Gamma(link = "log"))
predictor_vars <- c( "HLASENI_UZAVRENI_M", "TYP_SUBCLAIMU", "STAT_SKODY", "VUZ_ROKY","POPIS_KODU_VOZIDL
A","POCET_AUT","FIL","MES_1_UZ")
pridane = c("ZARAZENI")
elimination_list = c("ZARAZENI")

AIC(model)
```

```
## [1] 147398.4
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 201257.94 | Scaled Deviance: 138914.01"
```

```
#nevyřazujeme





#KROK 2
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: HLASENI_UZAVRENI_M | AIC: 143471.56 | Scaled Deviance: 84856.83"
## [1] "Model 2: TYP_SUBCLAIMU | AIC: 146898.82 | Scaled Deviance: 87468.22"
## [1] "Model 3: STAT_SKODY | AIC: 144013.75 | Scaled Deviance: 85267.21"
## [1] "Model 4: VUZ_ROKY | AIC: 143910.88 | Scaled Deviance: 85181.39"
## [1] "Model 5: POPIS_KODU_VOZIDLA | AIC: 146814.67 | Scaled Deviance: 87393.19"
## [1] "Model 6: POCET_AUT | AIC: 144170.31 | Scaled Deviance: 85400.78"
## [1] "Model 7: FIL | AIC: 146489.30 | Scaled Deviance: 87154.03"
## [1] "Model 8: MES_1_UZ | AIC: 146335.09 | Scaled Deviance: 86986.32"
```

```
#pridame HLASENI_UZAVRENI_M
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M, data = data, family = Gamma(link = "log"))
predictor_vars <- c( "TYP_SUBCLAIMU", "STAT_SKODY", "VUZ_ROKY","POPIS_KODU_VOZIDLA","POCET_AUT","FI
L","MES_1_UZ")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M")

#vyrazeni
AIC(model)
```

```
## [1] 143471.6
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 191842.38 | Scaled Deviance: 128429.45"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 147398.35 | Scaled Deviance: 87852.50"
```

```
#nevyřazujeme

#KROK 3
AIC(model)
```

```
## [1] 143471.6
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 143204.04 | Scaled Deviance: 84656.44"
## [1] "Model 2: STAT_SKODY | AIC: 141005.56 | Scaled Deviance: 83018.64"
## [1] "Model 3: VUZ_ROKY | AIC: 140456.48 | Scaled Deviance: 82610.04"
## [1] "Model 4: POPIS_KODU_VOZIDLA | AIC: 143033.74 | Scaled Deviance: 84519.66"
## [1] "Model 5: POCET_AUT | AIC: 141236.57 | Scaled Deviance: 83203.48"
## [1] "Model 6: FIL | AIC: 142795.22 | Scaled Deviance: 84351.81"
## [1] "Model 7: MES_1_UZ | AIC: 142563.45 | Scaled Deviance: 84131.62"
```

```
#pridame VUZ_ROKY
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY, data = data, family = Gamma(link = "lo
g"))
predictor_vars <- c( "TYP_SUBCLAIMU", "STAT_SKODY","POPIS_KODU_VOZIDLA","POCET_AUT","FIL","MES_1_UZ")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY")

#vyrazeni
AIC(model)
```

```
## [1] 140456.5
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 183784.97 | Scaled Deviance: 119999.71"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 143910.88 | Scaled Deviance: 85181.39"
## [1] "Model 3: VUZ_ROKY | AIC: 143471.56 | Scaled Deviance: 84856.83"
```

```
#nevyřazujeme

#KROK 4
AIC(model)
```

```
## [1] 140456.5
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 140189.08 | Scaled Deviance: 82414.50"
## [1] "Model 2: STAT_SKODY | AIC: 138018.56 | Scaled Deviance: 80835.57"
## [1] "Model 3: POPIS_KODU_VOZIDLA | AIC: 140185.56 | Scaled Deviance: 82401.80"
## [1] "Model 4: POCET_AUT | AIC: 138330.43 | Scaled Deviance: 81073.81"
## [1] "Model 5: FIL | AIC: 139827.26 | Scaled Deviance: 82151.08"
## [1] "Model 6: MES_1_UZ | AIC: 139503.87 | Scaled Deviance: 81869.98"
```

```
#zařadíme stát škody
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY, data = data, family = Gamma
(link = "log"))
predictor_vars <- c( "TYP_SUBCLAIMU","POPIS_KODU_VOZIDLA","POCET_AUT","FIL")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 138018.6
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 172597.70 | Scaled Deviance: 109122.52"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 140658.75 | Scaled Deviance: 82762.96"
## [1] "Model 3: VUZ_ROKY | AIC: 141005.56 | Scaled Deviance: 83018.64"
## [1] "Model 4: STAT_SKODY | AIC: 140456.48 | Scaled Deviance: 82610.04"
```

```
#nevyřazujeme

#KROK 5
AIC(model)
```

```
## [1] 138018.6
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 137924.30 | Scaled Deviance: 80767.00"
## [1] "Model 2: POPIS_KODU_VOZIDLA | AIC: 137728.27 | Scaled Deviance: 80617.57"
## [1] "Model 3: POCET_AUT | AIC: 136373.31 | Scaled Deviance: 79666.54"
## [1] "Model 4: FIL | AIC: 137654.69 | Scaled Deviance: 80573.80"
```

```
#zařadíme POCET_AUT
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY+POCET_AUT, data = data, famil
y = Gamma(link = "log"))
predictor_vars <- c( "TYP_SUBCLAIMU","POPIS_KODU_VOZIDLA","FIL","MES_1_UZ")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY+POCET_AUT")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY","POCET_AUT")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 136373.3
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 168522.63 | Scaled Deviance: 105384.55"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 138561.02 | Scaled Deviance: 81244.44"
## [1] "Model 3: VUZ_ROKY | AIC: 139367.26 | Scaled Deviance: 81826.31"
## [1] "Model 4: STAT_SKODY | AIC: 138330.43 | Scaled Deviance: 81073.81"
## [1] "Model 5: POCET_AUT | AIC: 138018.56 | Scaled Deviance: 80835.57"
```

```
#nevyřazujeme



#KROK 6
AIC(model)
```

```
## [1] 136373.3
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 136182.61 | Scaled Deviance: 79531.11"
## [1] "Model 2: POPIS_KODU_VOZIDLA | AIC: 135610.96 | Scaled Deviance: 79120.77"
## [1] "Model 3: FIL | AIC: 136080.37 | Scaled Deviance: 79457.93"
## [1] "Model 4: MES_1_UZ | AIC: 135407.75 | Scaled Deviance: 78941.16"
```

```
#zařadíme MES_1_UZ
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ, data = da
ta, family = Gamma(link = "log"))
predictor_vars <- c( "TYP_SUBCLAIMU","POPIS_KODU_VOZIDLA","FIL")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY","POCET_AUT","MES_1_UZ")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 135407.8
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 166994.65 | Scaled Deviance: 103951.88"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 137541.82 | Scaled Deviance: 80468.09"
## [1] "Model 3: VUZ_ROKY | AIC: 138532.10 | Scaled Deviance: 81176.51"
## [1] "Model 4: STAT_SKODY | AIC: 137436.43 | Scaled Deviance: 80387.65"
## [1] "Model 5: POCET_AUT | AIC: 137044.26 | Scaled Deviance: 80094.46"
## [1] "Model 6: MES_1_UZ | AIC: 136373.31 | Scaled Deviance: 79666.54"
```

```
#nevyřazujeme





#KROK 7
AIC(model)
```

```
## [1] 135407.8
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 135207.22 | Scaled Deviance: 78800.00"
## [1] "Model 2: POPIS_KODU_VOZIDLA | AIC: 134538.06 | Scaled Deviance: 78325.48"
## [1] "Model 3: FIL | AIC: 135099.78 | Scaled Deviance: 78723.80"
```

```
#zařadíme POPIS_KODU_VOZIDLA
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU
_VOZIDLA, data = data, family = Gamma(link = "log"))
predictor_vars <- c( "TYP_SUBCLAIMU","FIL")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU_VOZIDL
A")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY","POCET_AUT","MES_1_UZ","PO
PIS_KODU_VOZIDLA")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 134538.1
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 161353.38 | Scaled Deviance: 99016.80"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 136664.41 | Scaled Deviance: 79836.27"
## [1] "Model 3: VUZ_ROKY | AIC: 137770.65 | Scaled Deviance: 80622.11"
## [1] "Model 4: STAT_SKODY | AIC: 136351.54 | Scaled Deviance: 79610.44"
## [1] "Model 5: POCET_AUT | AIC: 136736.25 | Scaled Deviance: 79865.76"
## [1] "Model 6: MES_1_UZ | AIC: 135610.96 | Scaled Deviance: 79120.77"
## [1] "Model 7: POPIS_KODU_VOZIDLA | AIC: 135407.75 | Scaled Deviance: 78941.16"
```

```
#nevyřazujeme






#KROK 8
AIC(model)
```

```
## [1] 134538.1
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 134478.46 | Scaled Deviance: 78282.82"
## [1] "Model 2: FIL | AIC: 134263.13 | Scaled Deviance: 78132.49"
```

```
#zařadíme FIL
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU
_VOZIDLA+FIL, data = data, family = Gamma(link = "log"))
predictor_vars <- c( "TYP_SUBCLAIMU")
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU_VOZIDLA
+FIL")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY","POCET_AUT","MES_1_UZ","PO
PIS_KODU_VOZIDLA","FIL")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 134263.1
```

```
deviance(model,scaled=TRUE)
```

```
## [1] 78132.49
```

```r
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 161025.20 | Scaled Deviance: 98733.56"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 136359.88 | Scaled Deviance: 79619.12"
## [1] "Model 3: VUZ_ROKY | AIC: 137487.10 | Scaled Deviance: 80417.91"
## [1] "Model 4: STAT_SKODY | AIC: 135918.96 | Scaled Deviance: 79304.11"
## [1] "Model 5: POCET_AUT | AIC: 136384.86 | Scaled Deviance: 79615.59"
## [1] "Model 6: MES_1_UZ | AIC: 135352.20 | Scaled Deviance: 78937.29"
## [1] "Model 7: POPIS_KODU_VOZIDLA | AIC: 135099.78 | Scaled Deviance: 78723.80"
## [1] "Model 8: FIL | AIC: 134538.06 | Scaled Deviance: 78325.48"
```

```r
# FIL je vetsi pouze o stovky




#KROK 9
AIC(model)
```

```
## [1] 134263.1
```

```r
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: TYP_SUBCLAIMU | AIC: 134202.09 | Scaled Deviance: 78088.92"
```

```r
#zařadíme TYP_SUBCLAIMU
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU
_VOZIDLA+FIL+TYP_SUBCLAIMU, data = data, family = Gamma(link = "log"))
predictor_vars <- c( )
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU_VOZIDLA
+FIL+TYP_SUBCLAIMU")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY","POCET_AUT","MES_1_UZ","PO
PIS_KODU_VOZIDLA","FIL","TYP_SUBCLAIMU")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 134202.1
```

```r
deviance(model,scaled=TRUE)
```

```
## [1] 78088.92
```

```r
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 161016.12 | Scaled Deviance: 98724.13"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 136347.29 | Scaled Deviance: 79608.87"
## [1] "Model 3: VUZ_ROKY | AIC: 137372.48 | Scaled Deviance: 80335.24"
## [1] "Model 4: STAT_SKODY | AIC: 135901.35 | Scaled Deviance: 79290.37"
## [1] "Model 5: POCET_AUT | AIC: 136203.46 | Scaled Deviance: 79486.77"
## [1] "Model 6: MES_1_UZ | AIC: 135286.26 | Scaled Deviance: 78889.91"
## [1] "Model 7: POPIS_KODU_VOZIDLA | AIC: 134895.68 | Scaled Deviance: 78580.50"
## [1] "Model 8: FIL | AIC: 134478.46 | Scaled Deviance: 78282.82"
## [1] "Model 9: TYP_SUBCLAIMU | AIC: 134263.13 | Scaled Deviance: 78132.49"
```

```
#KROK 10
AIC(model)
```

```
## [1] 134202.1
```

```
selected_models <- create_glm_models(data, predictor_vars)
model_selection_summary(selected_models)
#již máme zahrnuty všechny parametry v modelu, žádný další přidat nemůžeme
model <- glm(Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M+VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU
_VOZIDLA+FIL+TYP_SUBCLAIMU, data = data, family = Gamma(link = "log"))
predictor_vars <- c( )
pridane = c("ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY +STAT_SKODY+POCET_AUT+MES_1_UZ+POPIS_KODU_VOZIDLA
+FIL+TYP_SUBCLAIMU")
elimination_list = c("ZARAZENI","HLASENI_UZAVRENI_M","VUZ_ROKY","STAT_SKODY","POCET_AUT","MES_1_UZ","PO
PIS_KODU_VOZIDLA","FIL","TYP_SUBCLAIMU")

#kontrola redundantnosti parametru
AIC(model)
```

```
## [1] 134202.1
```

```
deviance(model,scaled=TRUE)
```

```
## [1] 78088.92
```

```
selected_models <-backwards_elimination(model,elimination_list)
model_selection_summary(selected_models)
```

```
## [1] "Model 1: ZARAZENI | AIC: 161016.12 | Scaled Deviance: 98724.13"
## [1] "Model 2: HLASENI_UZAVRENI_M | AIC: 136347.29 | Scaled Deviance: 79608.87"
## [1] "Model 3: VUZ_ROKY | AIC: 137372.48 | Scaled Deviance: 80335.24"
## [1] "Model 4: STAT_SKODY | AIC: 135901.35 | Scaled Deviance: 79290.37"
## [1] "Model 5: POCET_AUT | AIC: 136203.46 | Scaled Deviance: 79486.77"
## [1] "Model 6: MES_1_UZ | AIC: 135286.26 | Scaled Deviance: 78889.91"
## [1] "Model 7: POPIS_KODU_VOZIDLA | AIC: 134895.68 | Scaled Deviance: 78580.50"
## [1] "Model 8: FIL | AIC: 134478.46 | Scaled Deviance: 78282.82"
## [1] "Model 9: TYP_SUBCLAIMU | AIC: 134263.13 | Scaled Deviance: 78132.49"
```

```
#Nepřidali jsme ani neodebrali žádný parametr - končíme proces
```

Příloha č.2

```r
library(readxl)
library(dplyr) #upozornění, že názvy fci se překrývají, budou se pouzivat ty z dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
data <- read_excel("C:/Users/vojta/Desktop/Škola/Bakalářská práce - Škodní inflace v autopojištění/Data
_prakticka_cast/Data_R.xlsx")
data <- data[, !(colnames(data) %in% c("IND_SML", "MESIC_VZNIKU"))]



data <- within(data, {
  FIL = factor(FIL)
  TYP_SUBCLAIMU = factor(TYP_SUBCLAIMU)
  STAT_SKODY = factor(STAT_SKODY)
  ZARAZENI = factor(ZARAZENI)
  POPIS_KODU_VOZIDLA = factor(POPIS_KODU_VOZIDLA)
  POCET_AUT = factor(POCET_AUT)
  HLASENI_UZAVRENI_M =factor(HLASENI_UZAVRENI_M)
  VUZ_ROKY = factor(VUZ_ROKY)
  Vyplata = as.numeric(Vyplata)
  MES_1_UZ = factor(MES_1_UZ)

})


plny_model <- glm(data$Vyplata ~ZARAZENI+HLASENI_UZAVRENI_M+VUZ_ROKY+STAT_SKODY+POCET_AUT+POPIS_KODU_VO
ZIDLA+FIL+TYP_SUBCLAIMU+MES_1_UZ, data = data, family = Gamma(link = "log"))
null_model <- glm(data$Vyplata ~ 1, data = data, family = Gamma(link = "log"))

step_both <- stepAIC(null_model, scope = list(lower = null_model, upper = plny_model), direction = "bot
h")
```

```
## Start:  AIC=201257.9
## data$Vyplata ~ 1
##
##                     Df Deviance    AIC
## + ZARAZENI          11    87852 198804
## + STAT_SKODY        14   125929 200656
## + HLASENI_UZAVRENI_M 18   128429 200786
## + POPIS_KODU_VOZIDLA  8   130143 200849
## + VUZ_ROKY          20   133427 201032
## + FIL                2   135113 201078
## + POCET_AUT          3   137083 201175
## + TYP_SUBCLAIMU      1   138728 201251
## <none>                   138914 201258
## + MES_1_UZ          35   137768 201272
##
## Step:  AIC=147398.4
## data$Vyplata ~ ZARAZENI
##
##                     Df Deviance    AIC
## + HLASENI_UZAVRENI_M 18    84857 146879
## + VUZ_ROKY          20    85181 146943
## + STAT_SKODY        14    85267 146947
## + POCET_AUT          3    85401 146950
## + FIL                2    87154 147273
## + MES_1_UZ          35    86986 147308
## + TYP_SUBCLAIMU      1    87468 147329
## + POPIS_KODU_VOZIDLA  8    87393 147329
## <none>                    87852 147398
## - ZARAZENI          11   138914 156839
##
## Step:  AIC=143471.6
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M
##
##                     Df Deviance    AIC
## + VUZ_ROKY          20    82610 142830
## + STAT_SKODY        14    83019 142942
## + POCET_AUT          3    83203 142976
## + MES_1_UZ          35    84132 143322
## + FIL                2    84352 143322
## + POPIS_KODU_VOZIDLA  8    84520 143385
## + TYP_SUBCLAIMU      1    84656 143413
## <none>                    84857 143472
## - HLASENI_UZAVRENI_M 18    87852 144344
## - ZARAZENI          11   128429 156661
##
## Step:  AIC=140456.5
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY
##
##                     Df Deviance    AIC
## + STAT_SKODY        14    80836 139938
## + POCET_AUT          3    81074 139989
## + MES_1_UZ          35    81870 140298
## + FIL                2    82151 140319
## + TYP_SUBCLAIMU      1    82414 140398
## + POPIS_KODU_VOZIDLA  8    82402 140408
## <none>                    82610 140456
## - VUZ_ROKY          20    84857 141109
```

```
## - HLASENI_UZAVRENI_M 18    85181 141213
## - ZARAZENI           11   120000 151959
##
## Step:  AIC=138018.6
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY
##
##                      Df Deviance    AIC
## + POCET_AUT           3    79667 137571
## + MES_1_UZ           35    80094 137801
## + FIL                 2    80574 137921
## + POPIS_KODU_VOZIDLA  8    80618 137950
## + TYP_SUBCLAIMU       1    80767 137994
## <none>                     80836 138019
## - STAT_SKODY         14    82610 138679
## - HLASENI_UZAVRENI_M 18    82763 138730
## - VUZ_ROKY           20    83019 138825
## - ZARAZENI           11   109123 148969
##
## Step:  AIC=136373.3
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY +
##     POCET_AUT
##
##                      Df Deviance    AIC
## + MES_1_UZ           35    78941 136100
## + POPIS_KODU_VOZIDLA  8    79121 136131
## + FIL                 2    79458 136279
## + TYP_SUBCLAIMU       1    79531 136311
## <none>                     79667 136373
## - POCET_AUT           3    80836 136920
## - STAT_SKODY         14    81074 137011
## - HLASENI_UZAVRENI_M 18    81244 137084
## - VUZ_ROKY           20    81826 137355
## - ZARAZENI           11   105385 148515
##
## Step:  AIC=135407.8
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY +
##     POCET_AUT + MES_1_UZ
##
##                      Df Deviance    AIC
## + POPIS_KODU_VOZIDLA  8    78325 135103
## + FIL                 2    78724 135299
## + TYP_SUBCLAIMU       1    78800 135336
## <none>                     78941 135408
## - MES_1_UZ           35    79667 135716
## - POCET_AUT           3    80094 136003
## - STAT_SKODY         14    80388 136133
## - HLASENI_UZAVRENI_M 18    80468 136167
## - VUZ_ROKY           20    81177 136532
## - ZARAZENI           11   103952 148414
##
## Step:  AIC=134538.1
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY +
##     POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA
##
##                      Df Deviance    AIC
## + FIL                 2    78132 134445
## + TYP_SUBCLAIMU       1    78283 134519
## <none>                     78325 134538
```

```
## - POPIS_KODU_VOZIDLA  8    78941 134832
## - MES_1_UZ          35    79121 134869
## - STAT_SKODY        14    79610 135158
## - HLASENI_UZAVRENI_M 18    79836 135264
## - POCET_AUT          3    79866 135309
## - VUZ_ROKY          20    80622 135656
## - ZARAZENI          11    99017 144949
##
## Step:  AIC=134263.1
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY +
##     POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA + FIL
##
##                      Df Deviance    AIC
## + TYP_SUBCLAIMU       1    78089 134243
## <none>                     78132 134263
## - FIL                 2    78325 134358
## - POPIS_KODU_VOZIDLA  8    78724 134551
## - MES_1_UZ           35    78937 134606
## - STAT_SKODY         14    79304 134837
## - HLASENI_UZAVRENI_M 18    79619 134990
## - POCET_AUT           3    79616 135019
## - VUZ_ROKY           20    80418 135396
## - ZARAZENI           11    98734 144817
##
## Step:  AIC=134202.1
## data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M + VUZ_ROKY + STAT_SKODY +
##     POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA + FIL + TYP_SUBCLAIMU
##
##                      Df Deviance    AIC
## <none>                     78089 134202
## - TYP_SUBCLAIMU       1    78132 134222
## - FIL                 2    78283 134295
## - POPIS_KODU_VOZIDLA  8    78581 134431
## - MES_1_UZ           35    78890 134531
## - STAT_SKODY         14    79290 134772
## - POCET_AUT           3    79487 134892
## - HLASENI_UZAVRENI_M 18    79609 134922
## - VUZ_ROKY           20    80335 135280
## - ZARAZENI           11    98724 144447
```

```
summary(step_both)
```

```
##
## Call:
## glm(formula = data$Vyplata ~ ZARAZENI + HLASENI_UZAVRENI_M +
##     VUZ_ROKY + STAT_SKODY + POCET_AUT + MES_1_UZ + POPIS_KODU_VOZIDLA +
##     FIL + TYP_SUBCLAIMU, family = Gamma(link = "log"), data = data)
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.2271660  0.1159427 -10.584  < 2e-16 ***
## ZARAZENInáhradní vozidlo    0.1720198  0.0551856   3.117 0.001827 **
## ZARAZENINESMLUVNÍ-NEZNAČKA  1.4053013  0.0552957  25.414  < 2e-16 ***
## ZARAZENINEZNAČKA            1.2910055  0.0580511  22.239  < 2e-16 ***
## ZARAZENIodtah              -0.9298731  0.0648654 -14.335  < 2e-16 ***
## ZARAZENIostatní majetek     0.8240833  0.0393130  20.962  < 2e-16 ***
## ZARAZENIother               0.4271468  0.0543765   7.855 4.03e-15 ***
## ZARAZENIROZPOČET            0.7947803  0.0549816  14.455  < 2e-16 ***
## ZARAZENIskla                0.2370887  0.0946197   2.506 0.012222 *
## ZARAZENITŠ                  2.0142659  0.0562131  35.833  < 2e-16 ***
## ZARAZENIzahraniční          1.4178292  0.0585299  24.224  < 2e-16 ***
## ZARAZENIZNAČKA              1.5913869  0.0551249  28.869  < 2e-16 ***
## HLASENI_UZAVRENI_M1         0.1334335  0.0120696  11.055  < 2e-16 ***
## HLASENI_UZAVRENI_M2         0.1474040  0.0147818   9.972  < 2e-16 ***
## HLASENI_UZAVRENI_M3         0.1857989  0.0195368   9.510  < 2e-16 ***
## HLASENI_UZAVRENI_M4         0.1684054  0.0248554   6.775 1.25e-11 ***
## HLASENI_UZAVRENI_M5         0.2032129  0.0299869   6.777 1.24e-11 ***
## HLASENI_UZAVRENI_M6         0.2434774  0.0352889   6.900 5.25e-12 ***
## HLASENI_UZAVRENI_M7         0.3263895  0.0408198   7.996 1.30e-15 ***
## HLASENI_UZAVRENI_M8         0.2687668  0.0462423   5.812 6.19e-09 ***
## HLASENI_UZAVRENI_M9         0.3183700  0.0517592   6.151 7.73e-10 ***
## HLASENI_UZAVRENI_M10        0.3082467  0.0583576   5.282 1.28e-07 ***
## HLASENI_UZAVRENI_M11        0.4596980  0.0600446   7.656 1.94e-14 ***
## HLASENI_UZAVRENI_M12        0.2769432  0.0539915   5.129 2.91e-07 ***
## HLASENI_UZAVRENI_M13        0.4153193  0.0644141   6.448 1.14e-10 ***
## HLASENI_UZAVRENI_M14        0.6821500  0.0767242   8.891  < 2e-16 ***
## HLASENI_UZAVRENI_M15        0.3077461  0.0853363   3.606 0.000311 ***
## HLASENI_UZAVRENI_M16        0.2667120  0.0940827   2.835 0.004585 **
## HLASENI_UZAVRENI_M17        0.2135059  0.0965526   2.211 0.027018 *
## HLASENI_UZAVRENI_M18        0.5562252  0.0236108  23.558  < 2e-16 ***
## VUZ_ROKY1                   0.0360088  0.0413621   0.871 0.383988
## VUZ_ROKY10                 -0.0791762  0.0443336  -1.786 0.074116 .
## VUZ_ROKY11                 -0.1649639  0.0444520  -3.711 0.000207 ***
## VUZ_ROKY12                 -0.2019735  0.0446054  -4.528 5.96e-06 ***
## VUZ_ROKY13                 -0.2105442  0.0446779  -4.712 2.45e-06 ***
## VUZ_ROKY14                 -0.3213428  0.0446957  -7.190 6.54e-13 ***
## VUZ_ROKY15                 -0.3545045  0.0458302  -7.735 1.04e-14 ***
## VUZ_ROKY16                 -0.3924467  0.0477650  -8.216  < 2e-16 ***
## VUZ_ROKY17                 -0.4413447  0.0490048  -9.006  < 2e-16 ***
## VUZ_ROKY18                 -0.5762839  0.0515572 -11.178  < 2e-16 ***
## VUZ_ROKY18+                -0.6749325  0.0409317 -16.489  < 2e-16 ***
## VUZ_ROKY2                   0.0371810  0.0409774   0.907 0.364223
## VUZ_ROKY3                   0.0477934  0.0408329   1.170 0.241818
## VUZ_ROKY4                   0.0369738  0.0411862   0.898 0.369335
## VUZ_ROKY5                   0.0276494  0.0416101   0.664 0.506379
## VUZ_ROKY6                  -0.0108030  0.0424174  -0.255 0.798969
## VUZ_ROKY7                  -0.0017195  0.0438230  -0.039 0.968701
## VUZ_ROKY8                  -0.0376828  0.0443784  -0.849 0.395813
## VUZ_ROKY9                  -0.0566006  0.0447618  -1.264 0.206059
```

```
## VUZ_ROKYN/A                   0.0605569  0.0427676   1.416 0.156792
## STAT_SKODYBE                 -0.3135831  0.1206327  -2.599 0.009338 **
## STAT_SKODYCZ                 -0.5561203  0.0577562  -9.629  < 2e-16 ***
## STAT_SKODYDE                  0.2089485  0.0538263   3.882 0.000104 ***
## STAT_SKODYES                 -0.7359583  0.1629350  -4.517 6.28e-06 ***
## STAT_SKODYFR                 -0.6051726  0.0962587  -6.287 3.25e-10 ***
## STAT_SKODYGB                  0.1215455  0.1102733   1.102 0.270368
## STAT_SKODYHR                 -0.7817803  0.1229452  -6.359 2.04e-10 ***
## STAT_SKODYHU                 -0.8179669  0.1407504  -5.811 6.21e-09 ***
## STAT_SKODYCH                 -0.1080754  0.1531311  -0.706 0.480333
## STAT_SKODYIT                 -0.5436967  0.0677061  -8.030 9.83e-16 ***
## STAT_SKODYNL                 -0.3893007  0.1341557  -2.902 0.003710 **
## STAT_SKODYPL                 -0.5891407  0.0947441  -6.218 5.05e-10 ***
## STAT_SKODYSK                 -0.5428742  0.0742047  -7.316 2.58e-13 ***
## STAT_SKODYxxx                -0.1531563  0.0985893  -1.553 0.120312
## POCET_AUT2                    1.0300005  0.0498555  20.660  < 2e-16 ***
## POCET_AUT3                    1.3927542  0.3099459   4.494 7.01e-06 ***
## POCET_AUTN/A                  0.8568267  0.0583856  14.675  < 2e-16 ***
## MES_1_UZ202002               -0.0718814  0.0342635  -2.098 0.035917 *
## MES_1_UZ202003               -0.0327619  0.0347302  -0.943 0.345518
## MES_1_UZ202004                0.0071344  0.0370915   0.192 0.847471
## MES_1_UZ202005               -0.0145361  0.0380778  -0.382 0.702649
## MES_1_UZ202006                0.0091675  0.0356437   0.257 0.797026
## MES_1_UZ202007               -0.0004186  0.0365484  -0.011 0.990861
## MES_1_UZ202008                0.0287506  0.0367250   0.783 0.433710
## MES_1_UZ202009               -0.0198014  0.0358239  -0.553 0.580442
## MES_1_UZ202010               -0.0218700  0.0357986  -0.611 0.541256
## MES_1_UZ202011               -0.0125477  0.0370315  -0.339 0.734731
## MES_1_UZ202012               -0.0075753  0.0368017  -0.206 0.836916
## MES_1_UZ202101               -0.0210983  0.0360198  -0.586 0.558050
## MES_1_UZ202102                0.1825629  0.0371567   4.913 8.97e-07 ***
## MES_1_UZ202103                0.0499096  0.0357173   1.397 0.162312
## MES_1_UZ202104                0.0378298  0.0380614   0.994 0.320267
## MES_1_UZ202105                0.0525963  0.0370733   1.419 0.155986
## MES_1_UZ202106                0.0458808  0.0360932   1.271 0.203670
## MES_1_UZ202107                0.1156401  0.0393619   2.938 0.003306 **
## MES_1_UZ202108                0.0299019  0.0373971   0.800 0.423958
## MES_1_UZ202109                0.0484001  0.0365211   1.325 0.185086
## MES_1_UZ202110                0.0580072  0.0360984   1.607 0.108075
## MES_1_UZ202111                0.0789244  0.0347002   2.274 0.022940 *
## MES_1_UZ202112                0.1379272  0.0354900   3.886 0.000102 ***
## MES_1_UZ202201                0.0949541  0.0358417   2.649 0.008068 **
## MES_1_UZ202202                0.1426848  0.0360927   3.953 7.71e-05 ***
## MES_1_UZ202203                0.1182292  0.0339384   3.484 0.000495 ***
## MES_1_UZ202204                0.1928050  0.0371410   5.191 2.09e-07 ***
## MES_1_UZ202205                0.1677369  0.0356186   4.709 2.49e-06 ***
## MES_1_UZ202206                0.1340035  0.0353472   3.791 0.000150 ***
## MES_1_UZ202207                0.3049266  0.0379552   8.034 9.55e-16 ***
## MES_1_UZ202208                0.2518972  0.0362387   6.951 3.65e-12 ***
## MES_1_UZ202209                0.1854268  0.0367720   5.043 4.60e-07 ***
## MES_1_UZ202210                0.1929636  0.0358884   5.377 7.60e-08 ***
## MES_1_UZ202211                0.2002111  0.0355533   5.631 1.79e-08 ***
## MES_1_UZ202212                0.1874962  0.0364774   5.140 2.75e-07 ***
## POPIS_KODU_VOZIDLAMOTOCYKL -0.0341849  0.0763618  -0.448 0.654392
## POPIS_KODU_VOZIDLAN/A        -0.2781816  0.0617321  -4.506 6.61e-06 ***
## POPIS_KODU_VOZIDLANÁKLADNÍ  0.2331118  0.0611727   3.811 0.000139 ***
## POPIS_KODU_VOZIDLANÁVĚS      0.0277700  0.0969214   0.287 0.774480
## POPIS_KODU_VOZIDLAOSOBNÍ     0.0839920  0.0558122   1.505 0.132352
```

```
## POPIS_KODU_VOZIDLAOSTATNÍ   0.1337463  0.0694927   1.925 0.054281 .
## POPIS_KODU_VOZIDLAPŘÍVĚS   -0.1638630  0.1196290  -1.370 0.170765
## POPIS_KODU_VOZIDLATAHAČ     0.3982930  0.0754787   5.277 1.32e-07 ***
## FILleasing                 -0.0987293  0.0161933  -6.097 1.09e-09 ***
## FILretail                  -0.1089977  0.0112639  -9.677  < 2e-16 ***
## TYP_SUBCLAIMUV              0.3165918  0.0690276   4.586 4.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 2.009822)
##
##     Null deviance: 138914  on 100658  degrees of freedom
## Residual deviance:  78089  on 100546  degrees of freedom
## AIC: 134202
##
## Number of Fisher Scoring iterations: 10
```