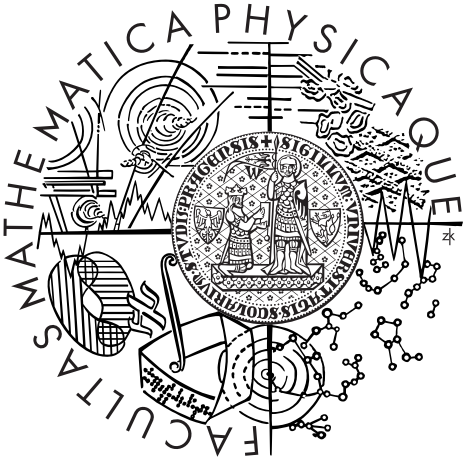
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Diplomová práce



Lenka Smolková

Regresní metody výpočtu rezerv na pojistná plnění a jejich praktická aplikace v pojištění motorových vozidel

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Jakub Strnad, PhD.
Studijní program: Matematika
Studijní obor: Finanční a pojistná matematika

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce Jakubu Strnadovi za jeho aktivitu a pomoc při vedení této práce, Janu Švábovi za dodání dat, Petru Jedličkovi za konzultace, rodičům a rodině, kteří mě podporovali během mých studií.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze, dne 13.12.2006 Lenka Smolková

Obsah

Abstrakt

Název práce: Regresní metody výpočtu rezerv na pojistná plnění a jejich praktická aplikace v pojištění motorových vozidel
Autor: Lenka Smolková
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Jakub Strnad, PhD.
e-mail vedoucího: jakub.strnad@cpp.cz
Abstrakt: Tato práce pojednává o modelování škodních trojúhelníků. V první části jsou popsány metody a modely užívané pro analýzu škodních trojúhelníků, důvody pro užití statistických metod v pojišťovnictví. Ve druhé části se aplikuje nejlepší možná metoda na mírně modifikovaná data. Cílem je popsat a aplikovat danou regresní metodu na vzorek dat.
Klíčová slova: vývojový trojúhelník, trend, regrese, odhad, předpověď

Abstract

Title: Regression Methods of Calculations of Reserves and their Practical Application to Automobile Insurance
Author: Lenka Smolková
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Jakub Strnad, PhD.
Supervisor's e-mail address: jakub.strnad@cpp.cz
Abstract: This paper deals with modelling of loss development array. The methods and models used for analysis of the array and the reason of using statistical methods in the insurance are described in the first part of the thesis. The second part is concerned with the best model applied on the slightly modified data. The aim of this work is presentation and application of the regression method on the data sample.
Keywords: loss development array, trend, regression, estimate, assumption to the future

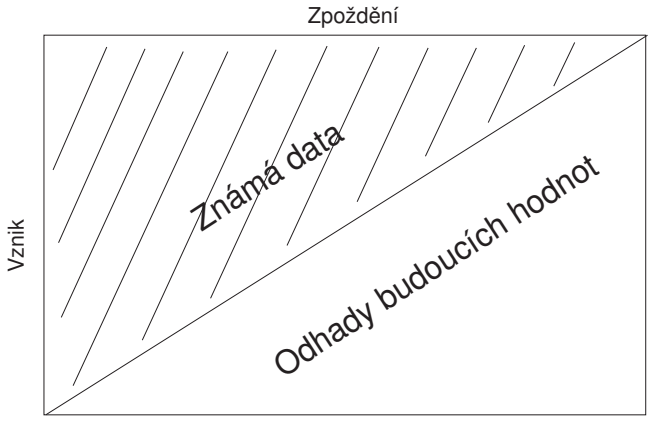
Úvod

Tato práce se zabývá modelováním a analýzou škodních trojúhelníků, které se používají především pro výpočet dvou složek rezerv na pojistná plnění: RBNS (rezervy na škody pojišťovně nahlášené, ale dosud nezlikvidované - Reported But Not Settled) a IBNR (rezervy na škody jež v minulosti vznikly, ale dosud nebyly pojišťovně hlášeny - Incurred But Not Reported) v neživotním pojištění.
Ve druhé kapitole jsou zmíněny modely deterministické, pomocí kterých získáme přímo hodnoty předpovědí budoucích rezerv na pojistná plnění. Přes modely popsané v kapitole třetí, kde se již počítá s náhodnou složkou (neboli chybou předpovědi), se dostáváme k hlavní části a tou je popis modelů PTF - Probabilistic Trend Family. Tato skupina pravděpodobnostních modelů připouští určité výkyvy v předpovídaných hodnotách.

Kapitola 1

Trojúhelníková schémata a definice základních pojmů

Pro výpočet rezervy na pojistná plnění na základě minulého škodního období se velmi často používají takzvané vývojové trojúhelníky. Data v takových trojúhelnících odpovídají nahlášeným nebo uhrazeným škodám, nebo počtu vzniklých pojistných událostí. Hodnoty v jednom řádku trojúhelníku jsou vztaženy ke stejnému referenčnímu období (například rok/čtvrtletí/měsíc, ve kterém škody nastaly) a data v jednom sloupci odpovídají časovému posunu (zpoždění) v úhradě nebo nahlášení příslušných pojistných událostí. Cílem výpočtu IBNR (nebo RBNS) je odhad hodnot pod diagonálou - tzn. hodnot do budoucna.



Užívání schématu vývojového trojúhelníku poskytuje jednoduchý přehled o vývoji škod a jejich rezerv. Existují dva typy vývojových trojúhelníků: kumulativní a nekumulativní neboli přírůstkový. Jestliže bude u nekumulativního trojúhelníku splněna podmínka stejného časového posunu (zpoždění), potom na diagonálách najdeme souhrnná data kalendářního období.

Definice 1. *Nechť písmenem $X_{i,j}$ značíme hodnoty kumulativního a písmenem $Y_{i,j}$ nekumulativního trojúhelníku. Potom:*

$$X_{i,j} = \sum_{s=1}^{j} Y_{i,s} \quad ,$$

kde $Y_{i,s}$ značí výši nebo počet škod vzniklých v roce i a zaplacených v roce $i+s$ a $X_{i,j}$ je součet nekumulativních hodnot počínaje obdobím i až do konce období $i+j$.

Schéma trojúhelníků vypadá následovně:

	0	1	2	...	j	...	$(t-2)$	$(t-1)$
1	$Y_{1,0}$	$Y_{1,1}$	$Y_{1,2}$	...	$Y_{1,j}$	...	$Y_{1,t-2}$	$Y_{1,t-1}$
2	$Y_{2,0}$	$Y_{2,1}$	$Y_{2,2}$	...	$Y_{2,j}$	...	$Y_{2,t-2}$	
⋮	...	...	...	...	...	...		
i	$Y_{i,0}$	...	...	...	$Y_{i,j}$			
⋮	...	...	...	...				
t	$Y_{t,0}$							

Tento trojúhelník předpokládá, že vývoj škod bude po $(t-1)$ období ukončen a odhad celkové výše/počtu škod vzniklých v roce i bude

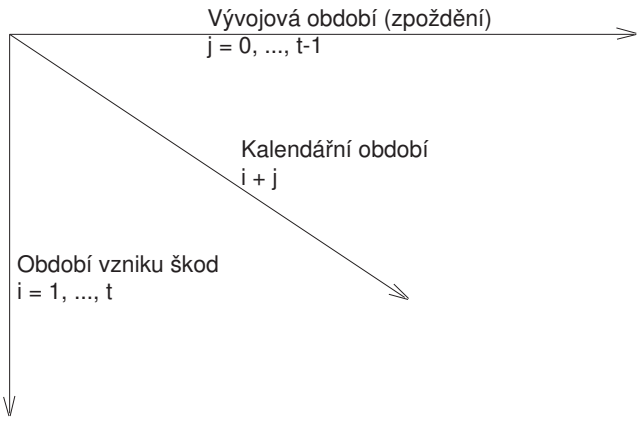
$$\hat{X}_{i,\infty} = X_{i,t-1}.$$

Pokud by vývoj škod dále pokračoval, je nutné opravit celkový počet škod $\hat{X}_{i,\infty}$ o část rezervy na pojistná plnění připadající na škody vzniklé/nahlášené v období vzniku i a vyplacené až po $i+t-1$. Pro odhady inkrementálních dat $Y_{i,j}$ pro $i+j \geq t$ (nebo kumulativních hodnot $X_{i,j}$, kdy $i+j \geq t$) jsou užívány různé statistické metody. Nejznámější z nich je metoda Chain-Ladder.
V diplomové práci se však budu zabývat především metodou PTF - Probabilistic Trend Family založenou na výpočtu vývojových faktorů pomocí regrese.

Definice 2. *Vývojovým faktorem rozumíme poměr dvou sousedních hodnot v řádku kumulativního trojúhelníku*

$$\zeta = \frac{X_{i,j}}{X_{i,j-1}}.$$

Následující obrázek znázorňuje geometrické směry ve vývojovém trojúhelníku:



Pro úplnost je nutno definovat pojem lineární regrese.

Definice 3. *Nechť* $\mathbb{Y}$ *je vektor náhodných veličin rozměru* $(n\times 1)$, $\mathbb{X}$ *je matice daných čísel* $(n \times k)$ *s lineárně nezávislými sloupci, kde* $k < n$, β *je neznámý vektor parametrů* $(k \times 1)$ *a* ε *náhodný vektor splňující podmínky* $E\varepsilon = 0$ *a* $Var\varepsilon = \sigma^2\mathbb{I}$, *kde* $\mathbb{I}$ *je jednotková matice a* σ^2 *neznámé.Potom*

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

je lineární regresní model.
Vektor β *se určuje pomocí metody nejmenších čtverců.*

Kapitola 2

Modely užívající deterministicky určené vývojové faktory

2.1 Deterministické vývojové faktory

Pro jednoduchost předpokládejme pouze jedno období vzniku škod. Označme $x(j)$ celkovou hodnotu výše/počtu pojistných plnění až do vývojového období j (kumulativní hodnota) a nechť $y(j) = ln[x(j)]$. Definujme :

$$\alpha = ln[x(0)] \tag{2.1}$$

$$\text{a} \qquad \gamma_j = y(j) - y(j-1). \tag{2.2}$$

Parametr α značí počáteční hodnotu/hladinu a parametr γ_j značí trend v logaritmickém měřítku mezi vývojovými období $j-1$ a j.
$y(j)$ tedy může být vyjádřeno pomocí rovnice:

$$\begin{aligned} y(j) &= y(0) + y(1) - y(0) + \ldots + y(j) - y(j-1) \\ &= \alpha + \sum_{s=1}^{j} \gamma_s. \end{aligned} \tag{2.3}$$

To znamená, že $y(j)$ se může vyjádřit jako počáteční hodnota plus součet diferencí až do vývojového období j. Tyto diference jsou považovány za trendy.

$$\begin{aligned} \gamma_j &= y(j) - y(j-1) \\ &= ln[x(j)] - ln[x(j-1)] \\ &= ln\left[\frac{x(j)}{x(j-1)}\right] \end{aligned} \tag{2.4}$$

Poměr $x(j)/x(j-1)$ je z definice 2 vývojovým faktorem. To znamená, že trendový parametr γ_j je logaritmus tohoto vývojového faktoru.

2.2 Modely

V této části budou popsány základní modely založené na deterministických vývojových faktorech. Nechť $x(i,j)$ je kumulativní hodnota škod odpovídající období vzniku i a vývojovému období j a $y(i,j) = ln[x(i,j)]$.

2.2.1 Cape Cod (CC)

Nechť α je počáteční hladina a nechť je stejná pro všechna období vzniku škod i. Nechť vývojové faktory $\gamma_1, \gamma_2, \ldots, \gamma_{t-1}$ jsou také stejné pro všechna období vzniku škod i. Za těchto předpokladů může být model zapsán rovnicí:

$$y(i,j) = \alpha + \sum_{s=1}^{j} \gamma_s. \tag{2.5}$$

Rovnice (2.5) popisuje deterministický CC model.

2.2.2 Chain Ladder (CL)

Model Chain Ladder je rozšíření techniky vývojových faktorů. V matematickém zápisu:

$$y(i,j) = \alpha_i + \sum_{s=1}^{j} \gamma_s. \tag{2.6}$$

Parametr α_i zde odpovídá období vzniku i. V modelu CC je α a γ_j konstantní pro všechna období vzniku škod. V modelu CL se tedy předpokládají parametry γ_j stejné pro všechna období vzniku škod – homogenita γ_j, ale různé hodnoty parametru α_i pro různá období vzniku škod – heterogenita α_i.

2.2.3 Cape Cod s konstantní inflací (CCI)

Označme trend během kalendářních období (zkráceně také inflaci) v logaritmickém měřítku jako ι_t. Hodnota $y(i,j)$ škod nastalých v roce i a zaplacených až do roku $i+j$ zahrnuje také inflační faktor ι_{i+j-1}. Hodnota faktoru v kalendářním roce $i+j$ za předpokladu konstantní inflace je $\iota \cdot (i+j-1)$. Tedy rovnici pro model, za tohoto předpokladu, lze zapsat:

$$y(i,j) = \alpha + \sum_{s=1}^{j} \gamma_s + \iota(i+j-1) \tag{2.7}$$

nebo

$$y(i,j) = \alpha + \iota(i-1) + \sum_{s=1}^{j} (\gamma_s + \iota). \tag{2.8}$$

Druhá rovnice říká, že ve směru vzniku pojistných událostí máme trend ι a také, že vývojový faktor mezi obdobím $j-1$ a j je $\gamma_j + \iota$. Je to jen matematické potvrzení toho, že inflace se odráží v obou dalších směrech.
Hlavní rozdíl mezi modelem CL a následujícím CCI modelem je, že CL nemodeluje inflaci.

2.2.4 Separační model (SM)

Tyto modely umožňují odhadnout inflaci ve škodních nárocích přímo z pozorovaných dat. Nekumulativní data závisí na roce vzniku i, na zpoždění j a také na parametru vyjadřující inflaci v období $i+j$, tj. lze je modelovat pomocí 3 různých parametrů

$$x(i,j) = e_i d_j \lambda_{i+j}, \tag{2.9}$$

kde tedy e_i vyjadřuje závislost na období vzniku škody, d_j na zpoždění a λ_{i+j} na kalendářním období. Také to znamená, že v rámci logaritmů lze rovnici vyjádřit jako

$$y(i,j) = \alpha + \sum_{s=1}^{j} \gamma_s + \sum_{t=1}^{i+j} \iota_t \tag{2.10}$$

kde opět γ_s odpovídají vývojovým faktorům a ι_t inflaci mezi obdobím $t-1$ a t.
SM modeluje změny v inflaci mezi každými dvěmi následujícími kalendářními obdobími a také významné změny na základě vývojového trendu mezi dvěmi vývojovými období.

2.2.5 Deterministic Development Factor Family (DDF) - Třída modelů založená na deterministických vývojových faktorech

Každý model ze souboru DDF může být vysvětlen jako verze modelu CC, která obsahuje trendy ve vývojových a kalendářních období i období vzniku škody

$$y(i,j) = \alpha_i + \sum_{s=1}^{j} \gamma_s + \sum_{t=1}^{i+j} \iota_t. \tag{2.11}$$

Model má parametr α_i jako základní hladinu pro období vzniku i, mezi každými dvěmi vývojovými obdobími je trend γ_s a mezi kalendářními ι_t.

Všechny modely popsané výše tedy náleží do skupiny modelů DDF.
Všechny obsahují

i) trendy ve směru vývojových období - γ_j (vývojové faktory)

ii) počáteční hodnoty pro každé období vzniku - α_i

iii) trendy během kalendářních období - ι_{i+j} (inflace)

Kapitola 3

Extended Link Ratio Family (ELRF)

3.1 Úvod

Tímto názvem bude v diplomové práci označena skupina modelů, které jsou založeny na vývojových faktorech definovaných v kapitole 1, což odpovídá proložení dat regresní přímkou (definice 3) bez absolutní složky. Za předpokladu normality dat (data s normálním rozdělením $N(\mu, \sigma^2)$) se odvodí výsledky pro obecnější soubor modelů - ELRF, které obsahují trendy v letech vzniku během každého vývojového roku (tj.trendy ve sloupcích vývojového trojúhelníku). Obecně se dá říci, že modely založené na vývojových faktorech mají několik nedostatků a často ani optimální model z ELRF není vhodný pro reálná data. Navíc modely založené na logaritmech nekumulativních dat mají větší předpovídací schopnost (tj.doplnění trojúhelníku na čtverec) než optimální ELRF model.

3.2 Výpočet vývojových faktorů pomocí regrese

Pro tuto kapitolu označme kumulativní data ve vývojovém roce $j-1$ jako $x(i)$ pro období vzniku $i = 1, 2, \ldots, n$ a $y(i)$, příslušná kumulativní data ve vývojovém roce j. Geometricky lze vývojový faktor $y(i)/x(i)$ interpretovat jako směrnici přímky procházející počátkem a bodem $(x(i), y(i))$. Tzn., že každý vývojový faktor je trend. Vzhledem k tomuto faktu je metoda vývojových faktorů založena na regresi

$$y(i) = bx(i) + \varepsilon(i), \qquad Var[\varepsilon(i)] = \sigma^2 x(i)^\delta \quad \text{pro} \quad i = 1, 2, \ldots, n. \tag{3.1}$$

Parametr b odpovídá směrnici „nejlepší“ přímky procházející počátkem a bodem $(x(i), y(i)), i = 1, 2, \ldots, n$. Odchylka $y(i)$ od přímky závisí na $x(i)$

pomocí funkce $x(i)^\delta$, kde δ označuje váhu parametru. Nejčastější případy diskutované ve článku Barnett, Zehnwirth [2]:

i) $\delta = 0$
Tento případ odpovídá regresní přímce pomocí metody nejmenších čtverců, která prochází počátkem (neboli váženému průměru, kde váhy jsou druhé mocniny objemu škod).

ii) $\delta = 1$
Odhad β parametru b metodou nejmenších čtverců dostaneme jako

$$\beta = \frac{\sum x(i) \cdot y(i)/x(i)}{\sum x(i)}.$$

Jedná se o průměrný vážený vývojový faktor (např. u Chain-Ladder).

iii) $\delta = 2$
Odkud parametr b pomocí nejmenších čtverců má tvar

$$\beta = \frac{1}{n} \sum \frac{y(i)}{x(i)}.$$

Je vidět, že jde o jednoduchý aritmetický průměr vývojových faktorů.

Vidíme tedy, že změnou parametru δ dostáváme různé „faktorové“ metody. Jednou z výhod užívání regrese pro odhad vývojových faktorů je možnost získat směrodatnou odchylku parametrů i směrodatnou odchylku předpovědí do budoucna.
Důležitým předpokladem pro nás budou standardizovaná rezidua

$$\varepsilon(i)/(\sigma x(i)^{\frac{\delta}{2}}),$$

kde $i = 1, 2, \ldots, n$ s normálním rozdělením se střední hodnotou 0 a rozptylem 1. Normální rozdělení se ověří zkoumáním několika grafů, např. Normal Probability Plot, Box Plot nebo histogramu vážených standardizovaných reziduí. Další předpoklad, který by měl být testován:

$$E[y(i) \mid x(i)] = bx(i). \tag{3.2}$$

Tento předpoklad se poměrně rychle otestuje z grafu závislosti $y(i)$ na $x(i)$. Velmi často bývá zjištěno, že přímka neprochází počátkem. Rovnice může být napsána také ve tvaru:

$$E[y(i) - x(i) \mid x(i)] = (b - 1)x(i). \tag{3.3}$$

Jak je řečeno ve článku Barnett, Zehnwirth [2] vysoké hodnoty jsou přeceněny

a malé podhodnoceny, což značí, že odhad $E[y \mid x] = bx$ je vychýlený.
Vzhledem k ekvivalenci rovnic (3.2) a (3.3), jsou rezidua kumulativních dat také rezidua dat nekumulativních. Kumulativní hodnota ve vývojovém roce $j-1$ může jen zřídka předpovědět budoucí přírůstek.
Přejděme teď na obecnější regresní model, který neprochází počátkem

$$y(i) = a + bx(i) + \varepsilon(i), \qquad Var[\varepsilon(i)] = \sigma^2 x(i)^\delta. \tag{3.4}$$

Jestliže průsečík a by byl nenulový a nezahrnuly bychom ho do modelu, potom odhad parametru b by nebyl nestranný. Poslední rovnici modelu můžeme přepsat jako

$$y(i) - x(i) = a + (b-1)x(i) + \varepsilon(i), \tag{3.5}$$

kde levá strana rovnice odpovídá přírůstku ve vývojovém roce j. Předpokládejme následující situace:

i) $b > 1$, $a = 0$
Abychom předpověděli průměrný přírůstek ve vývojovém roce j, bereme kumulativní hodnotu x v roce $j-1$ a násobíme ji $(b-1)$

ii) $b = 1$, $a \neq 0$
Tato kombinace značí, že $x(i)$ nemá žádnou váhu v předpokladu budoucího přírůstku. Odhad parametru a je vážený průměr přírůstků z vývojového období j.

Jestliže $b = 1$, potom graf přírůstků $y(i) - x(i)$ v závislosti na $x(i)$ by měl zobrazovat náhodné fluktuace kolem hodnoty a. Korelace je v tomto případě nulová.
Jestliže nekumulativní data v sobě obsahují trend v období vzniku škod, odhad parametru b nebude nulový, tzn., že společně s parametrem a budou mít předpovídací schopnost. Za těchto okolností můžeme do rovnice pro přírůstky přidat trend závisející na roce vzniku škody:

$$y(i) - x(i) = a_0 + a_1 i + (b-1)x(i) + \varepsilon(i), \qquad Var[\varepsilon(i)] = \sigma^2 x(i)^\delta \tag{3.6}$$

Budeme užívat následující označení pro tyto tři parametry:
a_0 - průsečík
a_1 - trend
b - vývojový faktor (směrnice).
Pro reálná data škod, která obsahují konstantní trend během vývojových

let, bude parametr a_1 významnější než b. Často také bývá parametr b nevýznamný. To znamená, že trend bude mít větší předpovídací schopnost než vývojový faktor a zbývající předpovídající schopnost faktoru po odečtení trendu bude nevýznamná. Příklady modelu ze souboru ELRF odpovídající rovnici (3.6):

- Chain-Ladder vývojové faktory
  $a_0 = a_1 = 0, \delta = 1$

  $$y(i) - x(i) = (b-1)x(i) + \varepsilon(i), \qquad Var[\varepsilon(i)] = \sigma^2 x(i)^\delta$$

- Cape Cod - s nenulovým průsečíkem
  Předpokládá se, že vývojový faktor je roven jedné a trend je nulový. Metoda Cape Cod odhaduje vážený průměr přírůstků (váhy závisející na δ) v každém vývojovém roce. Předpovědi jsou také založeny na váženém průměru přes období vzniku škod v každém vývojovém roce. Model může být zapsán jako:

  $$y(i) - x(i) = a_0 + \varepsilon(i), \qquad Var[\varepsilon(i)] = \sigma^2 x(i)^\delta.$$

- Trend s vývojovým faktorem rovným 1
  Model odhaduje vážený trend (závisející na δ) s parametry a_0 a a_1 během období vzniku škod pro každý vývojový rok

  $$y(i) - x(i) = a_0 + a_1 + \varepsilon(i), \qquad Var[\varepsilon(i)] = \sigma^2 x(i)^\delta.$$

  Předpovědi jsou také založeny na váženém trendu během období vzniku škod pro každý vývojový rok.

3.3 Shrnutí

Zabývali jsme se dvěma případy, které se mohou vyskytnout v reálných datech: přírůstky pro konkrétní vývojový rok mají nulový trend nebo případ, kdy přírůstky mají konstantní trend. V obou případech vývojové faktory jsou často nevýznamné, tím pádem chybí předpovídací schopnost. V praxi se ale nejčastěji vyskytují data, ve kterých se trend mění v čase (v trojúhelníku je to trend po diagonálách - inflace). To znamená, že když se podíváme postupně do každého vývojového roku, trend se mění s různými roky vzniku škod.
Modely popsané pomocí rovnice (3.6) se mohou využít k identifikaci změny inflace, ale nemohou tuto změnu konkrétně určit a ani s ní předpovídat do budoucna. ELRF modely tvoří přechod k modelům, které také obsahují právě

parametr inflace - Probabilistic Trend Family (PTF).
Je důležité připomenout, že ELRF modely uvažují normální rozdělení standardizovaných reziduí. Jestli je tento předpoklad pravdivý, pak odhady regresních parametrů jsou optimální. Na druhou stranu jestliže předpoklad splněn není, potom odhady mohou být velice slabé. Tento předpoklad normality je jen zřídka pravdivý pro data škodních rezerv. Ve skutečnosti standardizovaná rezidua dosahují větších výchylek na kladné poloose než-li na záporné. Jestliže by byl předpoklad normality správný, potom graf by měl být zhruba symetrický okolo nulové přímky.
Užíváním regrese v ELRF můžeme říci, že pro každý typ reálných škodních trojúhelníků jsou metody vývojových faktorů nedostačující. Analýza přírůstků v logaritmickém měřítku zároveň s použitím inflace má větší předpovídací schopnost, jak je popsáno v následujících kapitolách.

Kapitola 4

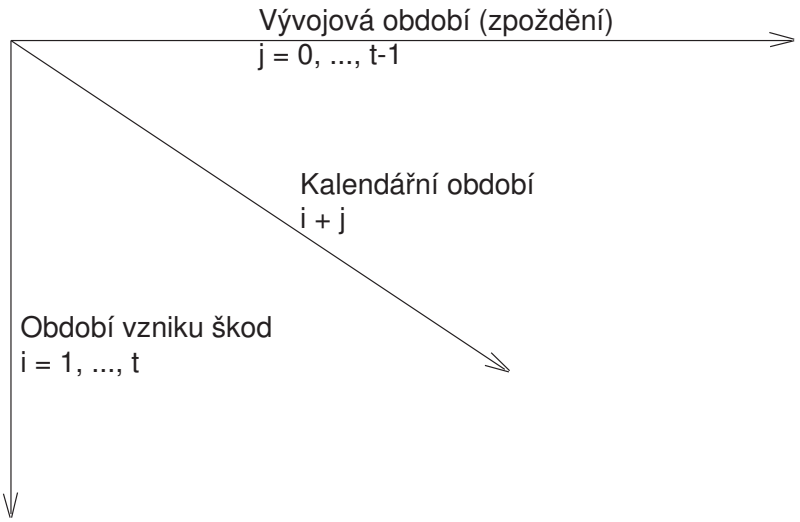
Regresní metody užívané pro výpočet rezerv na pojistná plnění

4.1 Úvod

V této sekci se budeme zabývat modelem, který popisuje měnící se trendy; takové trendy se na originálních datech nedají dobře analyzovat, protože obecně nejsou lineární (vezmeme-li inkrementální data, ty se mohou například měnit v rámci procent - nejdříve v prvních dvou letech 3% nárůst hodnot a v dalších letech 6% pokles). Z obecné zkušenosti a článků Barnett, Zehnwirth [2] a Zehnwirth [6] to jsou právě logaritmy nekumulativních dat, které dávají lineární trend. A na těchto logaritmech budeme také zkoumat změny v trendech. Modely tohoto typu vyžadují hlubší znalost vývoje škodních trojúhelníků. Výhodou pak je, že umožňují snadnější přístup k informacím, které lze ze škodního trojúhelníku získat.

4.2 Vlastnosti trendu škodního trojúhelníku

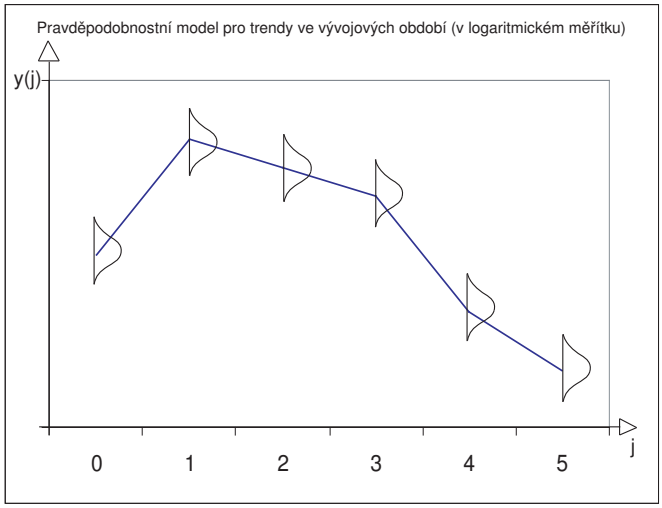
Trendy ve vývojovém trojúhelníku se vyskytují ve 3 geometrických směrech, tak jak to ukazuje následující obrázek.



Vývojová období jsou označena písmenem $j, j = 0, 1, \ldots, s-1$; období vzniku písmenem $i, i = 1, 2, \ldots, s$ a kalendářní období písmenem $t, t = 1, 2, \ldots, s$, kde $t = i + j$. Ze vztahu mezi těmito směry vidíme, že pouze dva z nich jsou nezávislé. Vývojová období a období vzniku jsou dva směry, které jsou na sebe ortogonální. To znamená, že trend v jednom z těchto směrů není přenesen do druhého a naopak. Trend ve směru kalendářních období (tj. po diagonálách) není ortogonální ani k jednomu z těchto směrů a znamená to, že se odráží jak v období vzniku tak ve vývojových období. A také naopak, trendy z období vzniku a vývojových období se odrážejí v trendech kalendářních období. Hlavní myšlenkou je možnost modelování trendů ve všech z těchto tří směrů.
V této části nechť hodnota $y(i, j)$ je logaritmus nekumulativní hodnoty $Y_{i,j}$, tedy $y(i, j) = ln(Y_{i,j})$, z roku vzniku i ve vývojovém roce j.
Nejprve předpokládejme jediný rok vzniku, tj. jedinou hodnotu parametru i. Parametr α je předpokládaná hodnota dat v prvním vývojovém období tzn. s nulovým zpožděním. Trendy ve vývojovém období se modelují přidáváním parametrů, které reprezentují předpokládanou změnu trendu mezi po sobě následujícími roky. Následně se modeluje rozptyl dat (odchylka od trendu) pomocí náhodné chyby s nulovou střední hodnotou a normálním rozdělením. Vyjádřeno rovnicí:

$$y(j) = \alpha + \sum_{k=1}^{j} \gamma_k + \varepsilon_j, \tag{4.1}$$

a znázorněno následujícím obrázkem



Pro tento pravděpodobnostní model α není přesná hodnota y pozorovaná s nulovým zpožděním, ale je to střední hodnota $y(0)$. Tudíž $y(0)$ má normální rozdělení se střední hodnotou α a rozptylem σ^2. Stejným způsobem γ_j není pozorovaný trend mezi vývojovými roky $j-1$ a j, ale je to střední hodnota trendu mezi těmito roky $E[y(j) - y(j-1)] = \gamma_j$. Parametry stochastického modelu odpovídají středním hodnotám náhodných veličin. Ovšem model (pomocí logaritmů) obsahuje normální rozdělení pro každý vývojový rok, kde střední hodnoty normálních rozdělení jsou dány pomocí parametru α a parametry trendu $\gamma_1, \gamma_2, \ldots$. Založeno na modelu daném rovnicí (4.1), náhodná veličina Y_j má logaritmicko-normální rozdělení s následujícími vlastnostmi:

$$\text{Medián} = exp\left[\alpha + \sum_{j=1}^{d} \gamma_j\right] \tag{4.2}$$

$$\text{Střední hodnota} = \text{medián} \cdot exp\left[\frac{1}{2}\sigma^2\right] \tag{4.3}$$

$$\text{Směrodatná odchylka} = \text{střední hodnota} \cdot \sqrt{(exp[\sigma^2] - 1)}. \tag{4.4}$$

Stochastický model pro Y_j zahrnuje logaritmicko-normální rozdělení pro každý vývojový rok, ve kterém jsou mediány dány rovnicí (4.2) a střední hodnoty rovnicí (4.3). Trend, který graficky obsahuje úsečky je pouze jednou částí modelu. Hlavní částí modelu je rozložení kolem takových trendů.
Rovnice (4.2) poznamenává, že exponenciela ze střední hodnoty logaritmů nám dává medián v normálním měřítku. K předpovědím se obvykle více užívá střední hodnota než medián. Náhodná složka měřená pomocí směrodatné odchylky u logaritmicko-normálního rozdělení je důležitá složka předpovědi. Jestliže tímto způsobem vypočítáme předpokládané hodnoty logaritmů vý-

vojových faktorů na nekumulativních datech, dostaneme

$$E\left[ln\frac{Y_j}{Y_{j-1}}\right] = E[\gamma_j + \varepsilon_j - \varepsilon_{j-1}] = \gamma_j. \tag{4.5}$$

To znamená, že parametry trendu podporují tento nový model.
Popsaný model zahrnuje pouze jeden rok vzniku škod. Je tedy nutné ještě zavést parametry pro trendy v kalendářních období a období vzniku. Nechť ι_t je střední hodnota inflace mezi roky t a $t+1$. Takže soubor modelů může být napsán pomocí rovnice:

$$y(i,j) = \alpha_i + \sum_{k=1}^{j} \gamma_k + \sum_{t=1}^{i+j} \iota_t + \varepsilon_{i,j}. \tag{4.6}$$

Tento soubor modelů nazýváme Probabilistic Trend Family (PTF). Střední hodnota trendu mezi $y(i,j-1)$ a $y(i,j)$ je $\gamma_j + \iota_{i+j}$ a střední hodnota trendu mezi $y(i-1,j)$ a $y(i,j)$ je $\alpha_{i+1} - \alpha_i + \iota_{i+j}$.
Jestliže hodnoty reziduí $\varepsilon_{i,j}$ mají normální rozdělení se střední hodnotou nula a nemají konstantní rozptyl, potom se takové změny rozptylu musí také modelovat.
Obvykle hodnota parametrů v období vzniku nevykazuje s postupem času velké změny. Na druhou stranu parametry ve směru vývojových let odrážejí trend v po sobě následujících období, který je obvykle v logaritmech lineární. Často se také data pozdějších vývojových období vyjadřují pouze jedním parametrem, protože se v nich nevyskytují výraznější změny. Většinou stačí jen několik parametrů nutných pro vyjádření trendu v datech. Z toho vyplývá, že optimální model pro konkrétní vývojový trojúhelník by měl být co nejjednodušší. Takový model by měl dát jasnější představu o tom, jak se přírůstky s postupem času vyvíjí. V reálných datech se může vyskytovat nemálo změn a je zřetelné, že, i přes různá rozšíření regresních modelů ELRF daných rovnicí (3.6), tyto modely nedávají hodnoty odpovídající těm skutečným. Statistiké modelování se tedy zabývá trendy a jejich variabilitou ve všech třech směrech.

4.3 Proložení dat regresní přímkou

Regrese se užívá nejen k odhadu trendů ve výplatě škod a odhadu rozdělení plateb $y(i,j)$ ve škodním trojúhelníku (kde právě odhadnuté trendy odpovídají střední hodnotě těchto rozdělení v logaritmickém měřítku), ale také k proložení dat ve vývojovém trojúhelníku regresní přímkou. Vzhledem k tomu, že se trend v kalendářních období přenáší do obou dalších směrů - období vzniku škody a vývojového období, zobrazení dat jednoho směru nevypovídá o trendech v daných datech. Definujme reziduum jako rozdíl mezi

pozorovanou hodnotou a odhadnutou hodnotou pomocí modelu:
$reziduum = y - y'$. Na rezidua může být také nahlíženo jako na odchylky od regresní přímky. Pro ilustraci bude uvedeno několik příkladů.
Předpokládejme trojúhelník založený na CC modelu definovaném v kapitole 2.2. Data v takovém modelu označena výrazem *CC data* jsou vyjádřena následovně

$$\text{CC data} = \text{CC trendy} + \text{náhodná chyba.} \tag{4.7}$$

Jestliže je CC model aplikován na taková data, potom rezidua budou mít následující tvar:

$$\begin{aligned} \text{Reziduum} &= \text{CC data - odhadnutá data pomocí CC trendů} \\ &= \text{odhad chyby} \end{aligned}$$

Reziduum je tedy zbytek po odečtení odhadnuté hodnoty pomocí daného modelu od hodnoty pozorované.
Nyní předpokládejme data taková, že:

$$\text{Data} = \text{CC data} + 10\% \text{ trend v kalendářním období..}$$

Jestliže na tato data aplikuji CC model, potom tvar reziduí vypadá:

$$\begin{aligned} \text{Reziduum} &= \text{Data - odhadnutá data pomocí CC trendů} \\ &= \text{odhad chyby} + 10\% \text{ trend v kalendářním období.} \end{aligned}$$

Grafické zobrazení závislosti reziduí na kalendářním období by odpovídalo rostoucímu trendu (+ náhodná složka). Po odstranění CC trendů z dat zůstává 10% trend v kalendářním období plus náhodné fluktuace. Jestliže se odhaduje průměrný trend v reziduích, je dosaženo odhadu asi 10%, což odpovídá trendu dodanému do dat. Jestliže je na data aplikován CCI model definovaný opět v části 2.2, odhaduje se trend reziduí v kalendářních období ze zmíněného CC modelu. Rezidua v závislosti na kalendářním období budou náhodná po odečtení všech odhadnutých trendů.
Předpokládejme data vytvořená dle následující formule

$$\begin{aligned} \text{Data} = \text{CC data} &+ 10\% \text{ trend mezi 1 až 8 kalendářním období} \\ &+ 20\% \text{ trend mezi 8 až 14 kalendářním období.} \end{aligned}$$

Když na taková data aplikujeme CC model, potom rezidua vypadají

$$\begin{aligned} \text{Reziduum} &= \text{Data - odhad pomocí CC modelu} \\ &= \text{odhad chyby} + 10\% \text{ (1 až 8)} + 20\% \text{ (8 až 14).} \end{aligned}$$

Rezidua vzhledem ke kaledářním období následují dva trendy mezi 1. a 8. rokem a strmější mezi 8. a 14. rokem. Jestliže teď na data aplikujeme CCI

model, odhadujeme trend reziduí v kalendářních období. Průměrný trend je mezi 10% a 20%. Regresní přímka proložená rezidui má zlom v období 8. To vede k odhadování dvou trendů kalendářních období. Projekce inflace do trendů v obou dalších směrech (vývojového období a období vzniku) nám říká, že inflaci můžeme získat až po odstranění trendů z vývojových období. Platí to i naopak, po odstranění inflace lze získat trend období vývojových. Regrese je tedy velmi silný nástroj na separování trendů ve všech třech směrech od náhodných fluktuací.

4.4 Změny parametrů

Kdybychom vzali vztah trendů ve vývojových letech, letech vzniku a kalendářních letech, modely s více parametry v jednom směru mohou být postiženy problémem multikolinearity.

4.4.1 Multikolinearita

Pro přímku jednoduché lineární regrese $y_i = \alpha + \beta x_i + \varepsilon$ pro všechna i se pomocí metody nejmenších čtverců odhaduje absolutní složka α a směrnice β. Odpovídá to minimalizaci přes α a β součtu

$$S^2 = min \sum^{N} (y_i - \alpha - \beta x_i)^2.$$

Rovnice se řeší tak, že parciální derivace dle α a β se položí rovné nule:

$$\text{dle } \alpha \quad -2 \sum^{N} (y_i - \alpha - \beta x_i) \;=\; 0 \tag{4.8}$$

$$\text{dle } \beta \quad -2 \sum^{N} x_i (y_i - \alpha - \beta x_i) \;=\; 0 \tag{4.9}$$

Ekvivalentně lze rovnici (4.8) rozepsat jako

$$\sum^{N} y_i - N\alpha - \beta \sum^{N} x_i \;=\; 0.$$

Označíme-li $\bar{y} = \frac{1}{N} \sum^{N} y_i$ a podobně i pro $\bar{x} = \frac{1}{N} \sum^{N} x_i$, dostaneme

$$\bar{y} - \alpha - \beta \bar{x} \;=\; 0. \tag{4.10}$$

Z rovnice (4.9) roznásobením dostaneme

$$\sum^{N} x_i y_i - N\alpha\bar{x} - \beta \sum^{N} x_i^2 \;=\; 0. \tag{4.11}$$

Rovnice (4.10) a (4.11) se nazývají *normální rovnice* a jejich řešením jsou odhady parametrů α a β pomocí metody nejmenších čtverců:

$$\alpha = \bar{y} - \beta\bar{x}$$

$$\beta = \frac{\sum^{N} x_i y_i - N\bar{x}\bar{y}}{\sum^{N} x_i^2 - N\bar{x}^2}.$$

Pro model ze souboru DDF definovaný v části 2.2.5 ve formě:
$y(i,j) = \alpha_i + \sum_{s=1}^{j} \gamma_s + \sum_{t=2}^{i+j} \iota_t + \varepsilon$ mající P parametrů dostaneme P lineárních homogenních rovnic pro P neznámých. Avšak jako výsledek neortogonality kalendářních let na zbylé dva směry jsou některé normální rovnice nadbytečné a tedy nemají jednoznačné řešení. Jestliže máme téměř stejné rovnice, pak odhady budou mít velká rezidua a model by byl nestabilní. Stabilita modelu je popsána v následujících kapitolách.

4.4.2 Řešení problému multikolinearity

Multikolinearita vztažená na modely s pevným počtem parametrů může sloužit jako forma informací. Může indikovat, že ve vývojovém trojúhelníku není dostatečné množství informací pro odhad více parametrů v kalendářním období a období vzniku (zejména pro pozdější období vzniku). Další interpretací je, že regresní proměnné nejsou nutně nezávislé, což by opět zkreslilo výsledky.
Jestliže se do modelu přidá pro poslední nastalou škodu další parametr α, tento parametr odpovídá jedinému období vzniku škody. Znamená to, že danému období vzniku škody se přisuzuje plná kredibilita a ostatním období vzniku kredibilita nulová na základě odhadu parametru α. Lepší způsob spočívá v tom, že se nenulová kredibilita přiřadí dřívějším období a následně poslednímu období vzniku nepřipadne plná kredibilita.
Důležitá je změna parametru α (na rozdíl od přidávání dalších parametrů), aby se právě multikolinearitě předešlo. Z článku Zehnwirth [6] je to důvod pro užití exponenciálního vyrovnávání ve směru let vzniku (vyrovnávání pomocí exponenciálního vážení do minulosti). Jednoduché exponenciální vyrovnávání se používá pro lokálně konstantní trend, což odpovídá právě změně parametru α. Tento přístup je nezbytný, velmi hodnotný a zvyšuje stabilitu modelu. Zejména jestliže během posledních let vzniku škod jsou nějaké výkyvy. To znamená, že multikolinearita může vést k regresním modelům

s neměnnými parametry, které 1) jsou nestabilní a 2) mají velké chyby předpovědi.
Exponenciální vyrovnávání je velmi často užíváno ve smyslu předpovídání časových řad. Na následující argumenty a příklady se lze dívat dvěma způsoby. Daty zde mohou být například

i) data o prodeji během časového úseku, nebo

ii) přírůstky zaplacených škod s nulovým zpožděním pro všechna období vzniku škod.

Konstantní hladina průměru

Předpokládejme časovou posloupnost pozorování $y_1, y_2, \ldots, y_N$ takovou, že

$$y_t = \alpha + \varepsilon_t \qquad t = 1, 2, \ldots, N, \tag{4.12}$$

kde α značí konstantní průměr a ε_t je posloupnost nekorelovaných chyb s konstantním rozptylem.
Model (4.12) vyjadřující data je nejjednodušší regresní model, který má pouze jeden parametr. To znamená, že roky $t, t = 1, 2, \ldots, N$ jsou homogenní a tedy stabilní. Označme symbolem $\hat{y}_{(N)+1}$ předpověď hodnoty y_{N+1} založenou na základě N známých pozorování. Je-li α známé, nejlepší předpovědí do budoucna bude

$$\hat{y}_{(N)+1} = \alpha.$$

Je-li parametr α neznámý, použijí se pro odhad $\hat{\alpha}$ minulá data $(y_1, \ldots, y_N)$. Pomocí metody nejmenších čtverců, tj. minimalizací přes α součtu $\sum_t^N (y_t - \alpha - \varepsilon_t)^2$ dostaneme

$$\hat{\alpha} = \sum_t \frac{y_t}{N},$$

tudíž předpověď o jedno období dopředu na základě N známých pozorování bude mít tvar

$$\hat{y}_{(N)+1} = \bar{y}.$$

Odhad můžeme také odvodit pomocí rekurentního vzorce

$$\begin{aligned} \hat{y}_{(N+1)+1} &= \frac{\sum_t^N y_t + y_{N+1}}{N+1} \\ &= \frac{(N+1)\bar{y} - \bar{y} + y_{N+1}}{N+1} \\ &= \hat{y}_{(N)+1} + \frac{(y_{N+1} - \hat{y}_{(N)+1})}{N+1} \end{aligned} \tag{4.13}$$

která říká, že předpovědi od času $n+1$ mohou být vyjádřeny jako lineární kombinace předpovědí do času n a nejpozdějšího pozorování.
Při předpokladu $\hat{\alpha} = \bar{y}$ mají všechna pozorování stejné váhy. Z pohledu škodních rezerv se tedy předpokládá homogenita let vzniku.

Nestabilní hladina průměru - každý rok má svůj vlastní parametr

Mějme model

$$y_t = \alpha_t + \varepsilon_t,$$

kde hladina průměru α_t se rychle mění v po sobě následujících časových úsecích. Každý rok t má svůj vlastní parametr α_t. Zde je nejlepší předpovídat pomocí poslední známé hodnoty

$$\hat{y}_{(N)+1} = y_N.$$

Dává se plná váha posledním pozorováním a nulová váha všem minulým hodnotám. Z pohledu škodních rezerv jsou období vzniku heterogenní. Každý jednotlivý parametr pro období vzniku je odhadován každou jednotlivou hodnotou toho daného období.

Lokálně konstantní hladina průměru

Často se vyskytují situace, kde daný průměr je lokálně konstantní. Předpoklad o tom, že hodnoty minulých období by měly mít stejnou váhu jako hodnota posledního období, je pro nás omezující. Kdyby se předpovědi odhadovaly pomocí poslední hodnoty (minulým obdobím by se přiřadily nulové váhy), ztratila by se informace o minulosti úplně. Je tedy lepší hodnotu vah geometricky snižovat se stářím daného pozorování - exponenciální vyrovnávání. Může se tedy dostat:

$$\hat{y}_{(N)+1} = Ky_N + K(1-K)y_{N-1} + K(1-K)^2 y_{N-2} + \ldots, \quad \text{kde } K \in \langle 0, 1\rangle$$

Pro dostatečně velká N by se předcházející rovnice mohla zapsat jako

$$\begin{aligned} \hat{y}_{(N)+1} &= \hat{y}_{(N-1)+1} + K(y_N - \hat{y}_{(N-1)+1}) \\ &= (1-K)\hat{y}_{(N-1)+1} + Ky_N. \end{aligned}$$

To také značí kredibilitní formuli.

4.5 Odhad parametrů a předpověď rozdělení

K určení regresního modelu je potřeba vytvořit vektor hodnot obsahující nezávislé proměnné odpovídající každému pozorování y.

Nechť $y(i,j) = ln[Y_{i,j}]$ je logaritmus inkrementální hodnoty a nechť β' je řádkový vektor obsahující parametry regresního modelu

$$\beta' = (\alpha_1, \alpha_2, \ldots, \alpha_k, \gamma_1, \ldots, \gamma_l, \iota_1, \ldots, \iota_m).$$

Model má

i) k různých parametrů α, kde α_1 odpovídá období vzniku $1, 2, \ldots, i_1$; α_2 období vzniku $i_1 + 1, \ldots, i_2$ etc.

ii) l různých parametrů γ, kde γ_1 je trend během vývojových období $0, 1, \ldots, j_1$; γ_2 je trend během vývojových období $j_1 + 1, \ldots, j_2$ etc.

iii) m různých parametrů ι, kde ι_1 je inflací během kalendářních období $0, 1, \ldots, t_1$; ι_2 je inflací během kalendářních období $t_1 + 1, \ldots, t_2$ etc.

Argumenty k, l, m mohou nabývat také hodnoty nula.
Vezmeme pro vysvětlení nejjednodušší případ - trojúhelník se stejným počtem řádků i sloupců, kde k odpovídá počtu období vzniku (v našem případě to bude $1, \ldots, s$), l počtu vývojových období (také rovno s) a m počet tomu odpovídajících kalendářních období $t = i + j$, kde $i \in \langle 1, s \rangle, j \in \langle 0, s - 1 \rangle$. Obecně vektor vztahující se k nezávislým proměnným má tvar:

$$x'(i,j) = (\delta_{\alpha 1}, \delta_{\alpha 2}, \ldots, \delta_{\alpha k}, \delta_{\gamma 1}, \ldots, \delta_{\gamma l}, \delta_{\iota 1}, \ldots, \delta_{\iota m});$$

a pro náš případ je to nula-jedničkový vektor:

$$x'(i,j) = (\delta_{\alpha 1}, \delta_{\alpha 2}, \ldots, \delta_{\alpha s}, \delta_{\gamma 0}, \ldots, \delta_{\gamma(s-1)}, \delta_{\iota 1}, \ldots, \delta_{\iota t}),$$

kde každý parametr δ je definován následovně

$$\begin{aligned}
\delta_{\alpha z} &= 1, \quad \text{když } z = i \\
&= 0 \quad \text{pro } z \neq i; \\
\delta_{\gamma z} &= 1 \quad \text{když } z \leq j \\
&= 0 \quad \text{pro } z > j \\
\text{a } \delta_{\iota z} &= 0, \quad \text{když } z < i \\
&= 1, \quad \text{když } i \leq z \leq i + j \\
&= 0, \quad \text{když } z > i + j.
\end{aligned}$$

Pozorování y se nyní srovná do sloupcového vektoru

$$Y = (y(1,0), \ldots, y(1,s-1), y(2,1), \ldots, y(2,s-2), \ldots, \ldots, y(s,0))$$

a odpovídající nula-jedničkové vektory do regresní matice

$$\mathbb{X} = (x'(1,0), \ldots, x'(s,0)).$$

Rovnice pro pozorovaná data nyní vypadá

$$Y = \mathbb{X}\beta + \varepsilon,$$

kde β je vektor parametrů a vektor ε obsahuje nezávislé náhodné chyby s normálním rozdělením se střední hodnotou 0 a rozptylem σ^2.

Tvrzení 1. *$\hat{\beta}$ odhad vektoru β pomocí metody nejmenších čtverců dostaneme*

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y.$$

Důkaz: Minimalizací $min(Y - \mathbb{X}\beta)^2$ dostaneme

$$\begin{aligned} 2(Y - \mathbb{X}\hat{\beta})\mathbb{X}' &= 0 \\ \mathbb{X}'Y - \mathbb{X}'\mathbb{X}\hat{\beta} &= 0 \\ \hat{\beta} &= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y. \end{aligned}$$

Tvrzení 2. *Pro odhad $\hat{\beta}$ platí:*

$$\begin{aligned} E\hat{\beta} &= \beta \\ Var\hat{\beta} = V(\hat{\beta}) &= \hat{\sigma}^2(\mathbb{X}'\mathbb{X})^{-1}. \end{aligned}$$

Důkaz:

$$E\hat{\beta} = E(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'E\mathbb{Y} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}\beta = \beta$$

$$\begin{aligned} V(\hat{\beta}) &= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'(Var\mathbb{Y})\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} \\ &= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'(\sigma^2\mathbb{Y})\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}. \end{aligned}$$

$V(\hat{\beta})$ je kovarianční matice odhadu $\hat{\beta}$.

4.5.1 Předpověď rozdělení

V této části přejdeme k dolnímu trojúhelníku, jak je znázorněno obrázkem v kapitole 1. Je třeba poukázat znovu na to, že regresní model je model založený na pravděpodobnosti (stochastický) a že modely ze souboru DDF definované v části 2.2.5 předpokládají normální rozdělení logaritmů hodnot ve škodním trojúhelníku. Pro budoucí kalendářní období se předpokládá, že průměrná inflace následující po období s včetně, bude ι_s. Což je také odhad trendu mezi obdobím $s-1$ a s. Jestliže ι_s není součástí modelu, pak $\iota_s = 0$. Předpokládejme také, že směrodatná odchylka trendu je i směrodatná odchylka odhadu - označme $se(\iota_s)$. Čím větší je směrodatná odchylka $se(\iota_s)$ při konstantní trendu ι_s, tím větší budou (střední) hodnoty zaplacených škod.

Vektor odhadů parametrů pro dolní trojúhelník nyní obsahuje stejný počet parametrů α a γ, ale pouze jeden parametr ι.
Hodnota nezávislé proměnné ve vektoru $x'(i,j)$ odpovídající parametru ι_s je tedy $i+j-s =$ počet kalendářních období od období s do období $i+j$. Toho lze také dosáhnout sečtením nula-jedničkových vektorů odpovídajících příslušným parametrům ι_s. Ostatní proměnné budou vypadat stejně jako při odhadování modelu daného rovnicí (4.6). Odhad $\hat{y}(i,j)$ hodnoty $y(i,j)$ pro $i+j \geq s$ je dán rovnicí

$$\hat{y}(i,j) = x'(i,j)\hat{\beta}.$$

Nechť $\hat{Y}$ je vektor předpovědí $\hat{y}$, $\mathbb{X}$ matice složená z vektorů x' a ε vektor náhodných chyb s rozdělením $N(0,\sigma^2)$:

$$\hat{Y} = \hat{\beta}\mathbb{X} + \varepsilon.$$

Odhad kovarianční matice $\hat{Y}$ lze odvodit jako:

$$\begin{aligned} Var(\hat{Y}) &= \mathbb{X}'Var(\hat{\beta})\mathbb{X} + Var(\varepsilon) \\ V(\hat{Y}) &= \mathbb{X}'V(\hat{\beta})\mathbb{X} + \hat{\sigma}^2 I, \end{aligned}$$

kde I značí jednotkovou matici. Veličina $\hat{\sigma}^2$ je odhadem rozptylu v modelu, kde $\mathbb{X}'V(\hat{\beta})\mathbb{X}$ je funkcí rozptylu $\hat{\beta}$ reprezentující náhodnost parametru.
$V(\hat{\beta})$ je funkcí $\hat{\sigma}^2$ a to znamená, že náhodnost parametrů a náhodost v celém modelu jsou vzájemně propojeny. Obvykle čím menší je σ^2, tím menší bude rozptyl parametrů.
Vektor Y bude mít vícerozměrné normální rozdělení se střední hodnotou $\hat{Y}$ a kovarianční maticí $V(\hat{Y})$. Jestliže se aplikuje standardní regrese, mohou být odhadnuty hodnoty y z vícerozměrného normálního rozdělení v dolním trojúhelníku (pro kalendářní období větší než s). Nejlepší odhad $\hat{y}$ veličiny y minimalizuje střední kvadratickou chybu $E[(y-f(y))^2]$ přes všechny statistiky $f(.)$, kde $f(.)$ je funkcí vektoru dat Y.
V rámci získání rozdělení inkrementálních a kumulativních dat, je třeba vědět vztah mezi vícerozměrným logaritmicko-normálním a normálním rozdělením a také statistické teorie o rozptylu součtu náhodných veličin. Střední hodnoty logaritmicko-normálního rozdělení jsou nejlepším odhadem příslušných přírůstků.

4.6 Postup modelování

Data škodních trojúhelníků se mohou rozložit na složku trendovou a složku náhodných fluktuací kolem takových trendů.

$$\text{Logaritmy plateb} = \text{Trend} + \text{Náhodné fluktuace}.$$

Jiný pohled na statistický model je považovat *trendy* za matematické vyjádření vývoje dat a *náhodné fluktuace* za charakteristiky nevysvětlené právě trendovou složkou.
Kritéria dobrého modelu s vysokou předpovídací schopností:

- Jednoduchost
- Vhodnost
- Stabilita (v následující kapitole).

4.6.1 Jednoduchost modelu

Princip jednoduchosti spočívá ve výběru ze dvou modelů se stejnou vypovídací schopností; ten jednodušší ze dvou modelů je vybrán. Také jednodušší model, který lépe vysvětluje hlavní charakteristiky dat, je většinou preferován před složitějším modelem s menším rozptylem reziduí (to znamená větším koeficientem determinace R^2, který je definován v dodatku). Je nutné připomenout, že statistika R^2 neměří předpovídací schopnost modelu.

Předpokládejme dva modely, které generují data:

1) $y_t = \mu + \varepsilon_t$,
kde ε_t má normální rozdělení $N(0, \sigma^2)$ a předpokládá se, že poměr μ/σ^2 je velký. Zde statistika $R^2 = 0$ a čím bude σ^2 menší, tím budou předpovědi založené na tomto modelu správnější.

2) $y_t = \alpha + \beta t + \varepsilon_t$,
kde $\varepsilon_t \sim N(0, \sigma^2)$. Předpokládá se relativně velké σ^2 a hodnota R^2 rovna 85%. Předpovědi založené na tomto modelu budou mít větší odchylku než předpovědi z modelu 1. Chyby předpovědí tedy nejsou funkcí R^2.

Při určování modelu mohou nastat případy:
Podparametrizování (model obsahuje méně parametrů, než by měl obsahovat) - obsahuje odlišné parametry narozdíl od optimálního modelu.
Přeparametrizování (model obsahuje více parametrů, než je nutné) - model je detailnější, než by bylo třeba.
Jestliže model obsahuje více parametrů, pak dopad této skutečnosti bude mít jiný vliv, než kdyby model obsahoval parametrů méně. Přeparametrizování vede k velkým chybám v předpovědi. Předpovídané hodnoty jsou extrémně senzitivní vůči náhodné složce v pozorování. Náhodná složka u přeparametrizování se může chovat jako část trendu (systematicky a nenáhodně). Na druhou stranu nedostatek parametrů vede k určitému vychýlení více než k nestabilitě.

Na jednoduchém příkladu uveďme problém přeparametrizování.
Mějme nějaké roční hodnoty, například objemy prodeje, generované pomocí rovnice

$$Y_t = 1 + 2t + 3t^2 + \varepsilon_t,$$

kde ε_t je náhodná složka s rozdělením $N(0, \sigma^2)$ a Y_t vyjadřuje počet prodaných pojistek v čase t.
Je třeba určit prodej pro rok 2007. Mohl by se použít lineární model:

$$\hat{Y}_t = \beta_0 + \beta_1 t + \varepsilon_t. \tag{4.14}$$

Takový model ale neobsahuje náhodná rezidua, takže je odmítnut.
Na druhou stranu kvadratický model

$$\hat{Y}_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t \tag{4.15}$$

nám dává rezidua jevící se jako náhodná. Navíc statistika R^2 má větší hodnotu a parametry jsou významné.
Jestliže zkusíme model pátého stupně:

$$\hat{Y}_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \beta_5 t^5 + \varepsilon_t, \tag{4.16}$$

takový model nebude mít žádná rezidua a $R^2 = 100\%$. Ale je třeba si dávat pozor, protože takový model nemá význam ve smyslu předpovídání. Jestliže se změní jediná hodnota, pak předpovědi budou úplně jiné. Navíc jestliže se použijí k předpovědi pro rok 2007 data z roku 2005 a potom pro upřesnění data z roku 2006, dochází se k rozdílným výsledkům. Data v tomto případě nejsou nestabilní, nestabilní je model. Model je příliš senzitivní na náhodnou složku a měl by být ovlivňován pouze složkou systematickou - trendy.

4.6.2 Akaikeho informační kritérium

Při rozhodování o vhodnosti modelů by mělo být speciálně rozhodováno o jednoduchosti daného modelu. Akaikeho informační kritérium (AIC) je jak funkcí S^2 (statistika definovaná v dodatku) tak funkcí počtu parametrů v modelu. AIC je užíváno při rozhodování mezi dvěma modely, i když nejsou do sebe vnořené (nemají stejné parametry). Obecně se AIC udává ve tvaru

$$AIC = -2I(\Theta) + 2P, \tag{4.17}$$

kde $I(\Theta)$ je logaritmicko-věrohodnostní funkce (viz.Dodatek) v bodě daném odhadnutými parametry a P je počet parametrů.
Pro modely DDF z části 2.2.5 se obecná rovnice (4.17) zjednodušuje na

$$AIC = N\left[ln\left(2\pi\frac{S^2}{N}\right)+1\right]+2P, \qquad \text{kde} \tag{4.18}$$

N = počet pozorování
S^2 je reziduální součet čtverců
a P označuje počet parametrů
Cílem je získat model, který má relativně minimální kritérium AIC. Musíme mít ale na paměti, že AIC může diferencovat mezi jakýmikoliv dvěma modely nezávisle na tom, že budou mít některé parametry stejné.

4.6.3 SSPE

Mějme časovou řadu $z_1, z_2, \ldots, z_n$, kde $\hat{z}_{(t)+1}$ dává předpověď hodnoty z_{t+1} založenou na hodnotách $z_1, z_2, \ldots, z_t$. To znamená předpověď založenou na informaci až do času t. Chyba předpovědi o jedno období je dána

$$\hat{\varepsilon}_t(1) = Z_{t+1} - \hat{Z}_{(t)+1}.$$

Zápis $\hat{\varepsilon}_t(1)$ vyjadřuje chybu předpovědi o jedno období vypočítanou z minulých dat až do času t. Odhady parametrů modelu jsou založeny na datech $Z_1, Z_2, \ldots, Z_t$. K výpočtu chyb $\hat{\varepsilon}_t(1)$ by měl být model odhadnut několikrát. Součet čtverců chyb předpovědí o jedno období, označován jako statistika SSPE (squared one-step-ahead prediction errors), je dán formulí

$$SSPE = \sum_{t=t_0}^{N} \varepsilon_t^2(1). \tag{4.19}$$

Čas t_0 je vybrán tak, že přesahuje maximální počet parametrů v modelu alespoň o jeden. Výpočet statistiky SSPE může zabrat hodně času, i když je k dispozici dobrý software, protože je nutné odhadovat podmnožiny z dat $Z_1, Z_2, \ldots, Z_t$, kde $t = t_0, t_0 + 1, \ldots, N - 1$. V praktické části diplomové práce budu tedy využívat Akaikeho informační kritérium.
Porovnání hodnot statistik R^2, AIC a $SSPE$, jestliže daty prokládáme polynom stupně 1 až 6:

- R^2 se zvětšuje s dodáním více parametrů
- AIC se zmenšuje, když se přechází od polynomu stupně 1 k polynomu stupně 2 a následně hodnota AIC narůstá (pro většinu případů)
- $SSPE$ se chová ve většině případů stejně jako AIC

Polynom stupně většího než 2 bude předpovídat „hůře“ než polynom stupně 2. Hodnota statistiky SSPE by měla být spíše menší než větší. Ale existují i další aspekty testování modelu, jako jsou významnost parametrů, předpoklady rozdělení, rezidua a počet parametrů. Na tyto testy je nutno pohlížet jako na doplňující ne jako na směrodatné.

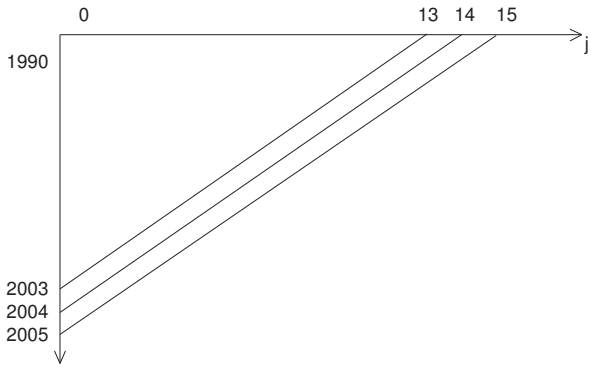
Kapitola 5

Ověření správnosti modelu a stabilita

Při sestavování modelu je potřeba si ohlídat, zda odhadnutý model lze použít i pro data mimo konkrétní vzorek, na kterém se právě tento model odhadoval. Odstraní-li se část, například poslední kalendářní období, je provedeno testování. Takovému testování předpovědí se říká ověření správnosti modelu. Ověřování správnosti se vztahuje také na stabilitu modelu. Škodní data předpovídaná s i bez pomoci posledního kalendářního období by se neměla lišit. Cílem je tedy určit model, který je stabilní s postupem času, za předpokladu stabilních trendů během časového období.

5.1 Ověřování správnosti

Mějme škodní trojúhelník:



Byl určen model pomocí dat získaných do roku 2005. Předpovídal by ten samý model určený v roce 2002 přírůstky let 2003, 2004 a 2005?
V roce 2002 by byly parametry určeny z menšího vzorku dat a udělala by se předpověď každého rozdělení logaritmů dat let 2003, 2004 a 2005. Důležité je

ověření, zda se na pozorované hodnoty let 2003, 2004 a 2005 může nahlížet jako na vzorek z předpovídaného rozdělení. Speciálně, nechť $\hat{y}$ je hodota předpovědi y pro roky 2003, 2004 nebo 2005, potom

$$\hat{\varepsilon} = y - \hat{y} \tag{5.1}$$

se bude nazývat reziduum nebo chyba předpovědi.
Chyby předpovědi se testují na:

i) náhodnost ve všech třech směrech - zpoždění v placení, období vzniku škody a kalendářním období

ii) náhodnost vzhledem k předpovězeným hodnotám $\hat{y}$

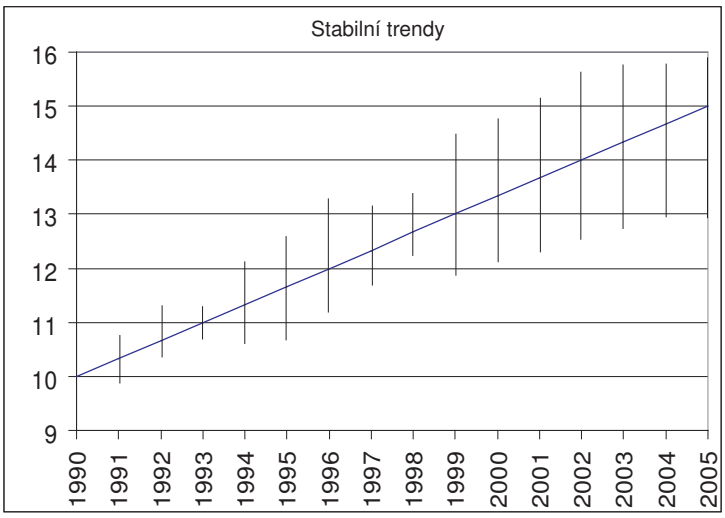
iii) normalitu (mají-li normální rozdělení $N(0, \sigma^2)$

5.2 Stabilita modelu

Jestliže se vrátíme k předcházejícímu příkladu, je na místě otázka, zda doplnění trojúhelníku na čtverec v roce 2002 by se lišilo od doplnění v roce 2005. Odpověď bude negativní, jestliže trendy v datech jsou stabilní (zejména v kalendářních letech).
Uveďme čtyři příklady (ne)stability trendů, jak jsou popsány v článku [2]. Pro ilustraci je u každého případu obrázek. Ve skutečnosti je ovšem možností mnohem více.

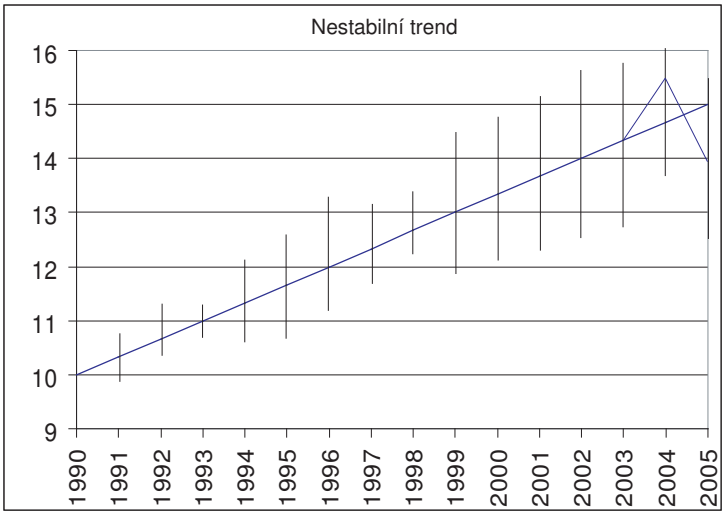
Příklad 1

Předpokládejme stabilní trend a jeho odhad roven $30\,\% \pm 2\,\%$ (obrázek). Stabilita trendu se zjistí tak, že po odstranění dat posledních období se odhady trendů významně neliší od těch, které byly dělány se všemi daty. Například po odstranění posledních dvou let by odhad trendu byl $29\,\% \pm 2,2\,\%$. Jako druhá možnost je odhadnout trend mezi roky 2003 a 2005 a dále šetřit, jestli se daný trend změnil.



Příklad 2

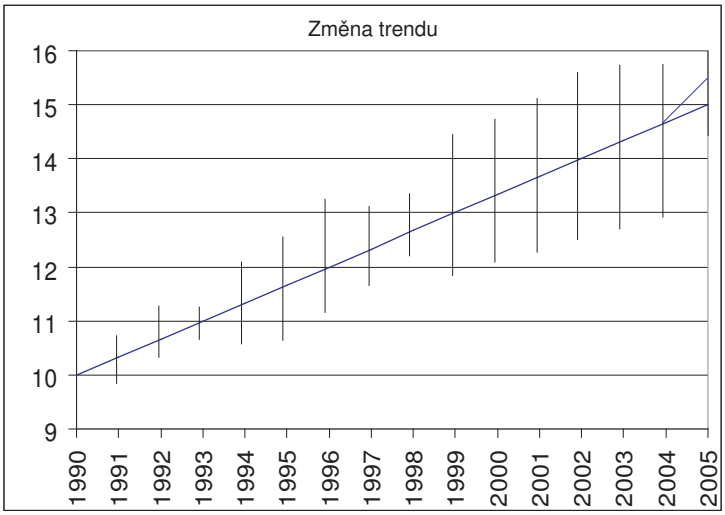
Předpokládejme relativně stabilní trend do roku 2003. Opět odhad může být $30\,\% \pm 2\,\%$. Trend mezi roky 2003 a 2004 by byl větší $40\,\%(\pm 1\,\%)$ a mezi roky 2004 a 2005 bude $-10\,\%(\pm 1,1\,\%)$ (obrázek). Otázkou tedy bude, jaký trend bude předpovězen bez jakékoliv další informace. Nebylo by nejspíš správné předpovídat poslední známý trend (čili $-10\,\%(\pm 1,1\,\%)$). Nejméně špatné řešení by bylo odhadnout trend pomocí střední hodnoty $30\,\%$ se směrodatnou odchylkou $2\,\%$, což znamená trendem známým do roku 2003. Rozhodnutí by však mělo být na aktuárovi, jeho úsudku a zkušenosti.



Příklad 3

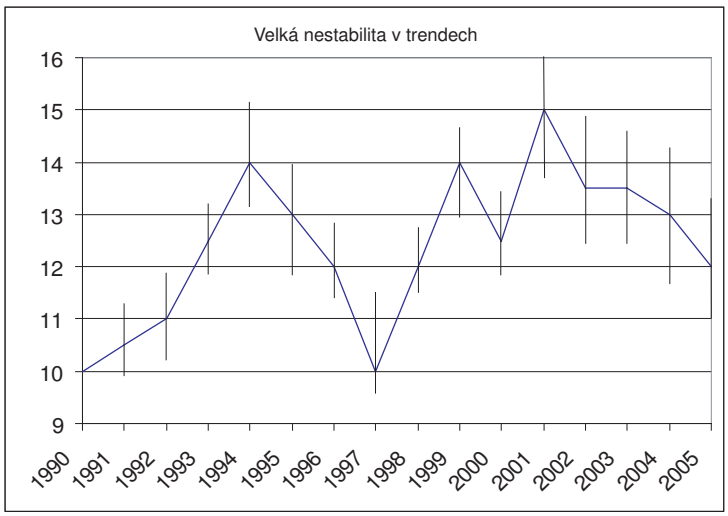
Mějme opět data se stabilním trendem do roku 2004. Nechť se tento trend pohybuje opět kolem $30\,\% \pm 2\,\%$. Mezi roky 2004 a 2005 se trend výrazně

zvýší na $40\,\% \pm 3\,\%$. Předpovědi do budoucna budou záviset na vysvětlení takového výkyvu dat. Jestliže by tento jev byl pouze dočasný, lze do budoucna předpokládat trend stejný jako do roku 2004 $30\,\% \pm 2\,\%$. Na druhou stranu jestliže trend bude vysvětlen stálou změnou, potom trend nastolený v posledním období bude pravděpodobně pokračovat. Rozhodnutí o vývoji do budoucna je v takovém případě těžší než v případě nestabilního trendu.



Příklad 4

Předpokládejme velkou nestabilitu trendu - mezi roky se neustále měnící. V tomto případě je nejlepší vypočítat vážený průměr trendů stanovených pro minulá období $\hat{\imath}$ a k nim vážený rozptyl $\hat{\sigma}^2$ a pro předpověď použít trend se střední hodnotou $\hat{\imath}$ a směrodatnou odchylkou $\hat{\sigma}$. Bude-li $\hat{\sigma}$ velké, předpověď střední hodnoty bude větší než medián předpovědi a směrodatná odchylka rozdělení bude relativně velká.



Kapitola 6

Identifikace modelu a předpoklady do budoucna

Mějme výsledný odhadnutý model je jednoduchý model ze skupiny PTF, který odděluje trendy od náhodných fluktuací a navíc určuje, zda trendy v kalendářních letech jsou stabilní.
Příklad: model CCI dle části 2.2.3 s konstantním vývojem v závěru vývojových období, který udává stabilní trend v kalendářních období. Takový model by byl vyhovující a dával stabilní předpovědi do budoucna v případě, že by se odhad budoucích škod dělal i bez dat z posledních let.

6.1 Identifikace modelu

Určení optimálního statistického modelu probíhá následujícím způsobem:

Krok 1: Prvotní analýza spočívá v určení heterogenity dat. (Následně mohou být určeny typy heterogenity.)

Krok 2: Specifikování modelu na základě *Kroku* 1.

a) Buď se začíná s modelem, který má pouze jeden parametr v každém směru a postupně se přidávají parametry trendů ve vývojových letech a následně parametry inflace na základě analýzy reziduí u modelu s jedním parametrem v každém směru. Záleží na tom, který z trendů bude mít větší vliv na změny.

b) Nebo se začne s plným modelem a vybírá se postupně nejvýznamnější parametr - forward kroková regrese.

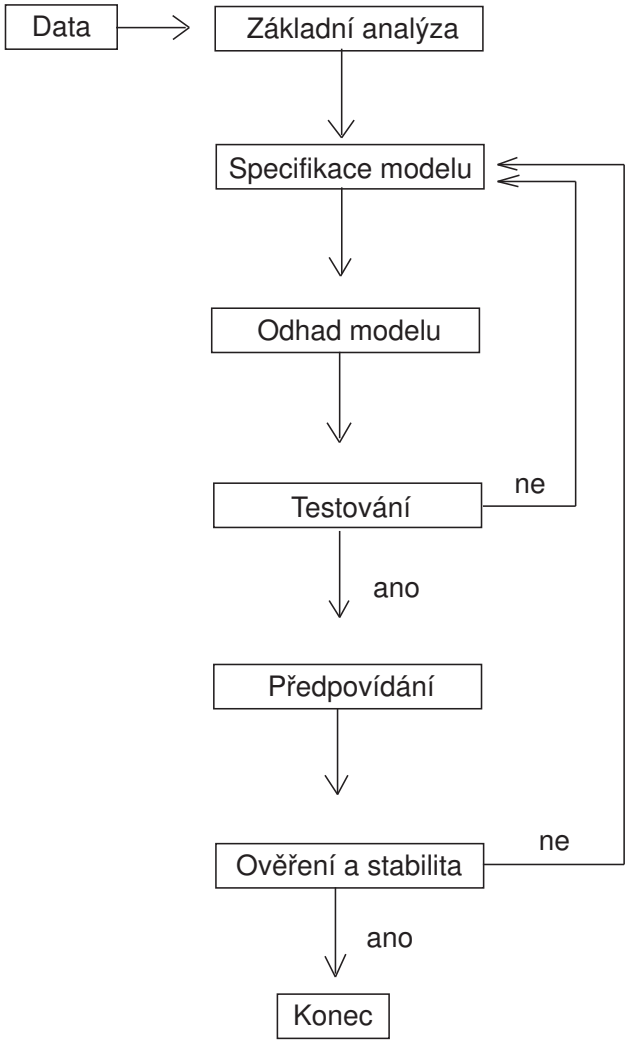
Krok 3: Odhadnutí parametrů daného modelu.

Krok 4: Kontrola, že data splňují předpoklady dané modelem. Při nesplnění předpokladů je nutné se vrátit ke *Kroku* 2 a cyklus se opakuje.

Krok 5: Nejlépe vyhovující model je testován na stabilitu trendů a je ověřena jeho správnost. Jestliže některé kritérium nevyhovuje, je opět nutné vrátit se ke *Kroku* 2 a zopakovat celý cyklus.

Krok 6: Jsou učiněny předpoklady do budoucna založené na *Kroku* 5 a tím pádem stanoveny předpovědi a rezidua.

Krok 7: Konec a uložení modelu.



6.2 Předpoklady do budoucna

Jestliže inflace je stabilní v pozdějších datech, potom předpoklad do budoucna je relativně jednoznačný. Jestliže odhad ι pro posledních sedm let je $\hat{\iota} \pm se(\hat{\iota})$, potom pro budoucnost předpokládáme střední hodnotu trendu $\hat{\iota}$ se směrodatnou odchylkou $se(\hat{\iota})$. Pozor, nepředpokládá se konstantní trend. Model obsahuje variabilitu (nestálost) trendu v budoucnu. Jestliže je na druhou stranu trend nestabilní, předpoklady do budoucna závisí na vysvětlení této nestability. Je nemožné vyčíslit všechny možné případy. Čím více má aktuár zkušeností s touto statistickou metodou, tím lépe je schopen zformulovat předpoklady pro budoucnost při nestabilních trendech.

Kapitola 7

Zpracování dat

Teorii PTF modelů jsem aplikovala na data získané z pojišťovny. Na následujícím listu je trojúhelník zaplacených škod a trojúhelník stanovených rezerv na pojistná plnění v automobilovém pojištění. Určení parametrů modelu pro inkrementální trojúhelník - regresní část - bylo provedeno na logaritmech dat pomocí programu STATISTICA. Následující předpovědi do budoucna - doplnění přírůstkového trojúhelníku na čtverec - byly provedeny v programu Microsoft Excel, s pomocí programu STATA8 a StatTransfer7. Výpočet a zdrojový kód je obsažen na přiloženém CD.

7.1 Možnost I. - Plný model

Modelování dat vycházející z plného modelu

$$y(i,j) = \alpha_i + \sum_{k=1}^{j} \gamma_k + \sum_{t=1}^{i+j} \iota_t + \varepsilon_{i,j}$$

forward krokovou regresí nejprve eliminuje počet parametrů a posléze se vyberou parametry odpovídající zvolené hladině p.

7.1.1 Regresní část - modelování pomocí trendů

Pro logaritmy inkremetálních dat trojúhelníku zaplacených škod je nutno určit matici vysvětlovacích proměnných $\mathbb{X}$. Tyto vysvětlující proměnné jsou „indikátory“ času. Sloupce tedy určují parametry trendu nejdříve ve směru vývojových období označené písmenem $\gamma_j, j = 1, 2, \ldots, s-1$, kde s je pro získaná data rovno 24, a následují parametry inflace $\iota_t, t = i + j$, kde $i = 0, 1, \ldots, s-1$. Parametr α_i je nahrazen užitím regresního modelu neprocházejícího počátkem. Pro názornost uvedu první a poslední řádky a sloupce matice $\mathbb{X}$. V prvním sloupci jsou logaritmy zaplacených škod (hodnoty trojúhelníku logaritmů zaplacených škod jsou po sloupcích dány pod sebe).

Y	*Intercept*	γ_1	γ_2	γ_3	γ_4	...	γ_{23}	ι_1	ι_2	ι_3	ι_4	...	ι_{23}
10,380199	1	0	0	0	0	...	0	0	0	0	0	...	0
10,59942686	1	0	0	0	0	...	0	1	0	0	0	...	0
10,62356617	1	0	0	0	0	...	0	1	1	0	0	...	0
10,8215251	1	0	0	0	0	...	0	1	1	1	0	...	0
10,92714557	1	0	0	0	0	...	0	1	1	1	1	...	0
10,94949777	1	0	0	0	0	...	0	1	1	1	1	...	0
10,89073083	1	0	0	0	0	...	0	1	1	1	1	...	0
11,00031809	1	0	0	0	0	...	0	1	1	1	1	...	0
10,95482611	1	0	0	0	0	...	0	1	1	1	1	...	0
⋮	⋮					⋮						⋮	
4,538735949	1	1	1	1	1	...	0	1	1	1	1	...	1
5,555314059	1	1	1	1	1	...	0	1	1	1	1	...	0
5,68711294	1	1	1	1	1	...	0	1	1	1	1	...	0
5,283674227	1	1	1	1	1	...	0	1	1	1	1	...	1
5,561845454	1	1	1	1	1	...	0	1	1	1	1	...	0
5,54476994	1	1	1	1	1	...	0	1	1	1	1	...	1
6,802397217	1	1	1	1	1	...	1	1	1	1	1	...	1

Pro výpočet regresních parametrů v programu STATISTICA bude matice $\mathbb{X}$ mít tedy rozměry (300×46) - 300 hodnot v trojúhelníku, 23 parametrů γ a 23 parametrů ι.

Pomocí vícenásobné krokové regrese dostaneme pouze významné parametry modelu na hladině p, kterou si předem zvolíme. Pro $p = 5\%$ budou významné pouze parametry s hladinou p menší než 0,05.

Parametr	Odhad	hladina p
Intercept	10,54720	0,000000
γ_6	-0,33051	0,014675
γ_{13}	-0,30629	0,073818
γ_3	-0,65785	0,000000
γ_9	-0,35362	0,000807
γ_2	-1,09304	0,000000
γ_{20}	-0,83019	0,000000
γ_{14}	-0,43990	0,014418
γ_4	-0,36834	0,004264
γ_1	0,48581	0,000064
γ_{11}	-0,30010	0,028949
γ_{23}	1,33343	0,002176
γ_7	-0,34392	0,004520
γ_5	-0,29607	0,024670
γ_{15}	-0,27538	0,062463
γ_{12}	-0,25979	0,114186
ι_2	0,43345	0,095900
ι_6	-0,37284	0,042887
ι_{12}	0,28095	0,026874
ι_{18}	-0,20263	0,051472
ι_{19}	0,14234	0,165487
ι_{11}	-0,20918	0,125220
ι_7	0,21887	0,196533

Výčet významných parametrů je tedy:

$$Intercept, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_9, \gamma_{11}, \gamma_{14}, \gamma_{20}, \gamma_{23}, \iota_6, \iota_{12}.$$

Model bude založen na výše uvedených významných parametrech. Po provedení klasické regrese s maticí $\mathbb{X}$ obsahující právě jen dané parametry se odhady samotných pozorování dostanou pomocí rovnice:

$$\hat{y} = \mathbb{X}b,$$

kde b je vektor odhadů parametrů.

7.1.2 Správnost modelu

Adekvátnost modelu jsem testovala na datech redukovaných o jeden poslední známý rok, čili o 4 období. Pro odhad jsem použila tedy pouze data za roky 2000 až 2004. Výsledky analýzy nezamítají navržený model.

7.1.3 Simulační část

Odhadnuté hodnoty v logaritmickém měřítku jsou tvořeny součtem příslušných odhadů parametrů. Avšak střední hodnoty odhadovaných dat v původním netransformovaném měřítku se nezískají pomocí exponenciely

$$\hat{Y}_{i,j} = exp\,(\hat{y}(i,j)),$$

která dává pouze odhad mediánu. Odhadovaná data se získají pomocí rovnice

$$\hat{Y}_{i,j} = exp\,(\hat{y}(i,j) + \frac{1}{2}\hat{\sigma}^2_{i,j}),$$

kde

$$\hat{\sigma}^2_{i,j} = \hat{\sigma}^2\,(\mathbb{I} + \mathbb{X}^\star(\mathbb{X}'\mathbb{X})^{-1})\mathbb{X}^{\star'})_{i,j},$$

kde matice $\mathbb{X}$ odpovídá horní části trojúhelníku a matice $\mathbb{X}^\star$ dolní části trojúhelníku.

Pomocí odhadů parametrů se doplní trojúhelník logaritmů dat na čtverec. Pro parametry ι_t se nepředpokládá náhlá změna. Pro $t > 24$, což odpovídá 4.čtvrtletí roku 2005, bude ι_t rovna nule.

7.2 Možnost II. - Základní model

Modelování dat vycházející ze základního modelu

$$y(i,j) = \alpha + \beta i + \gamma j + \iota t + \varepsilon_{i,j},$$

kde se předpokládá, že platí $y(i,j+1)$–$y(i,j) \sim \gamma$, pro každé j a $y(i+1,j)$–$y(i,j) \sim \beta$.

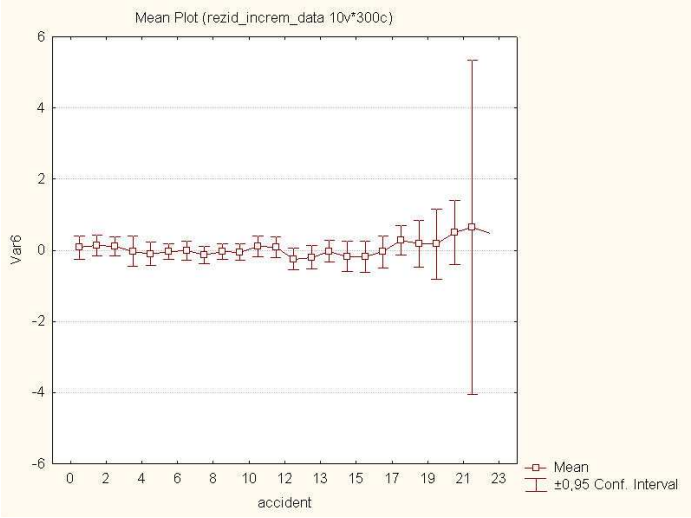
7.2.1 Regresní část

Analýza vychází z hledání zlomů v grafu reziduí základního modelu, pro který v tomto případě matice $\mathbb{X}$ vypadá takto (sloupce γ, β, ι):

Y	α	γ	ι	β
10,380199	1	0	0	0
10,59942686	1	0	1	1
10,62356617	1	0	2	2
10,8215251	1	0	3	3
10,92714557	1	0	4	4
10,94949777	1	0	5	5
10,89073083	1	0	6	6
11,00031809	1	0	7	7
10,95482611	1	0	8	8
⋮	⋮	⋮	⋮	⋮
4,538735949	1	20	23	3
5,555314059	1	21	21	0
5,68711294	1	21	22	1
5,283674227	1	21	23	2
5,561845454	1	22	22	0
5,54476994	1	22	23	1
6,802397217	1	23	23	0

Následující grafy (7.1), (7.2), (7.3) zobrazují rezidua v závislosti na každém směru - vývojovém období neboli zpoždění, kalendářním období a v neposlední řadě na období vzniku. Je vidět, že ve směrech období vzniku (graf (7.1)) a kalendářních období (graf (7.2)) nenastává výraznější trend.

Naopak ve směru vývojových období (graf (7.3)) se nevyskytuje pouze jeden „zlom“. Proto je na místě zabývat se především parametry ve směru vývojových období. První změna trendu nastala mezi prvním a třetím obdobím, následná patrná tendence je mezi třetím a osmým obdobím a osmé až 24 vývojové období bude reprezentovat také jeden parametr. Z grafu (7.4) se dá vyčíst ještě konstantní trend mezi osmým a třináctým obdobím - připojí se navíc jeden parametr (graf (7.5)). Výkyv dat z dřívějších období vzniku v posledním kalendářním roce pravděpodobně nebude pokračovat a v grafu se již neobjevují trendy. Výsledný identifikovaný model bude mít nulové trendy ve směru období vzniku a kalendářních období a 5 trendů ve vývojových období.

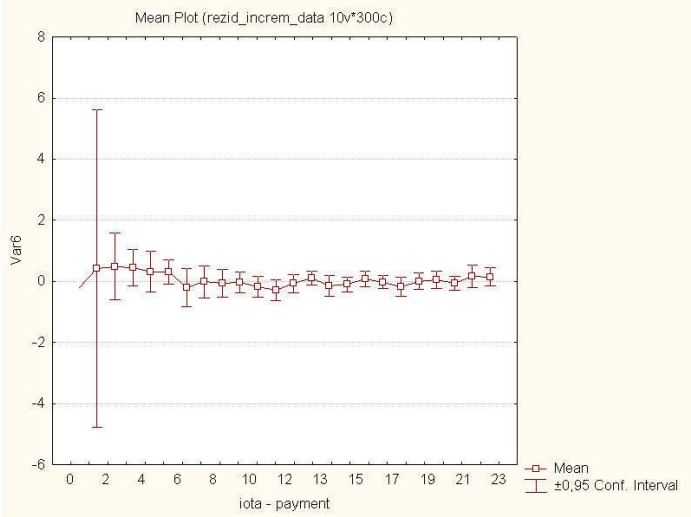


Obrázek 7.1: Graf reziduí vůči období vzniku

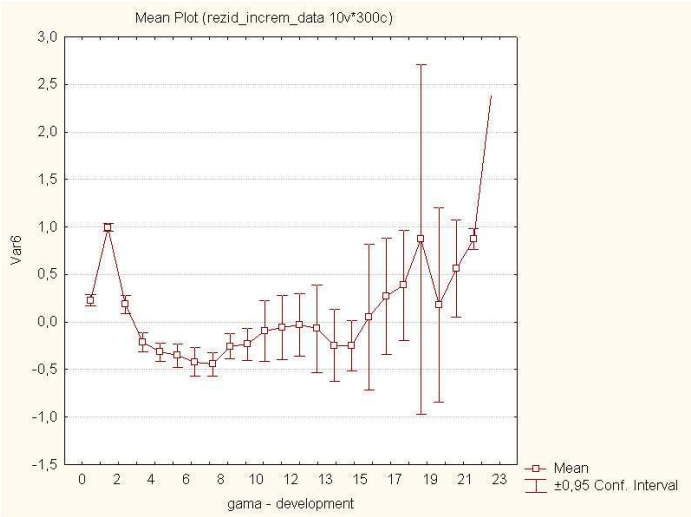
Parametr	Odhad	hladina p
Intercept	10,82609	0,000000
$\gamma(0-1)$	0,44250	0,000302
$\gamma(1-3)$	-0,90847	0,000000
$\gamma(3-8)$	-0,25950	0,000000
$\gamma(8-13)$	-0,24270	0,000000
$\gamma(13-24)$	-0,15928	0,000000

7.2.2 Simulační část

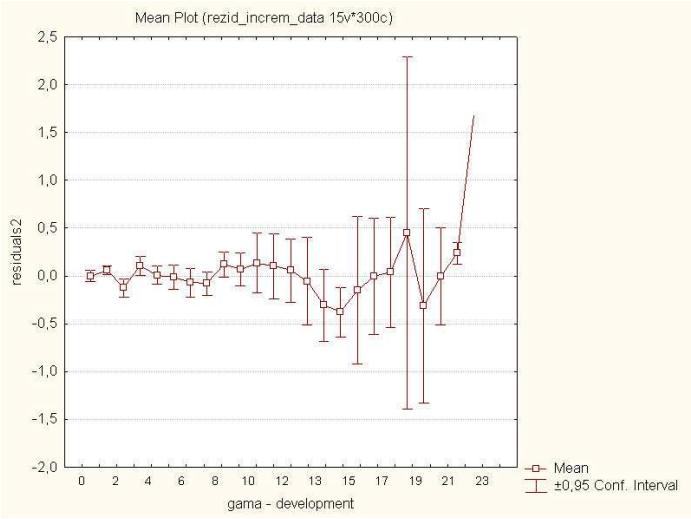
Odhad budoucích hodnot probíhá na stejné bázi jako u plného modelu.



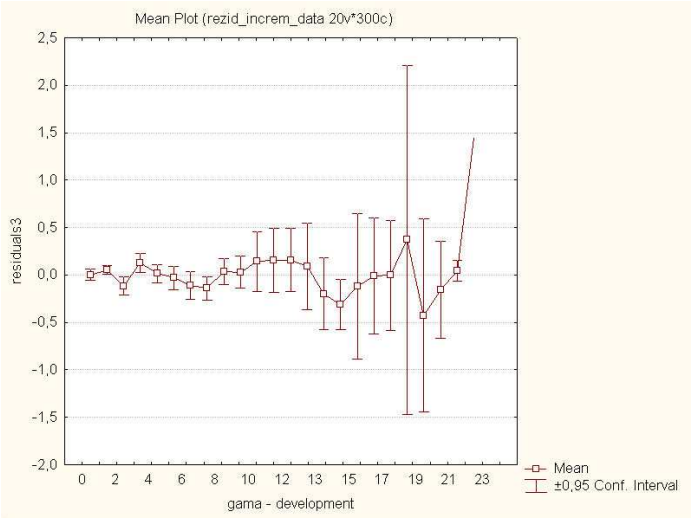
Obrázek 7.2: Graf reziduí vůči kalendářnímu období



Obrázek 7.3: Graf reziduí vůči zpoždění



Obrázek 7.4: Graf reziduí vůči zpoždění - 4 parametry ve směru vývojových období



Obrázek 7.5: Graf reziduí vůči zpoždění - 5 parametrů ve směru vývojových období

Závěr

V této diplomové práci jsme se zabývali různými technikami výpočtu rezerv na pojistná plnění pomocí vývojových trojúhelníků. Začali jsme u jednoduchých deterministických modelů užívajících vývojové faktory. Tyto modely ovšem nemodelují možné budoucí výkyvy v datech. Je tedy na místě přejít ke složitějším metodám užívajících pravděpodobnosti, které dávají přesnější odhady rezerv na budoucí plnění. Tomu odpovídá náhodná složka (chyba), která je v každém takovém modelu obsažena.
Hlavním tématem tedy byla skupina modelů Probabilistic Trend Family využívající regresi, u které byl popsán odhad parametrů pomocí regresního modelu a postup aplikace modelu na zadaná data vývojového trojúhelníku.
Výsledky a výstupy z programů pro dva možné postupy modelování budoucích rezerv jsou obsaženy na přiloženém CD.
Bylo by zajímavé srovnání modelů PTF s jinými metodami výpočtu rezerv na pojistná plnění, ale to je již nad rámec této diplomové práce.

Dodatek A

Statistické termíny

Nechť $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ je vektor náhodných veličin, pro něž platí $\mathbf{Y} = \mathbb{X}\beta + \mathbf{e}$, kde $\mathbb{X}$ je matice typu $\mathsf{n} \times \mathsf{k}\, k < n$, $\beta = (\beta_1, \ldots, \beta_k)'$ je vektor neznámých parametrů a $\mathbf{e} = (e_1, \ldots, e_n)'$ náhodný vektor splňující $E\,\mathbf{e} = 0, var\,\mathbf{e} = \sigma^2 I$ (I je jednotková matice). Parametry $\beta_1, \ldots, \beta_k$ se odhadují metodou nejmenších čtverců, tj.

$$min[(\mathbf{Y} - \mathbb{X}\beta)'(\mathbf{Y} - \mathbb{X}\beta)].$$

Tyto odhady se označí $\mathbf{b} = (b_1, \ldots, b_k)$. Vektor $\hat{\mathbf{Y}} = \mathbb{X}\mathbf{b}$ může být považován za nejlepší aproximaci vektoru $\mathbf{Y}$.

Statistika S^2 nebo SSE

Veličině S^2 se říká reziduální součet čtverců a je definována jako

$$S^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}).$$

Statistika R^2

Koeficient determinace R^2 je dán vzorcem

$$R^2 = 1 - \frac{S^2}{S_T},$$

kde

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Logaritmicko-věrohodnostní funkce

Nechť θ je jednorozměrný parametr, nechť $L(x, \theta) = ln\, p(x, \theta)$ je funkce proměnné θ při pevném x, kde $p(x, \theta), \theta \in \Omega$ je sdružená hustota náhodného vektoru $\mathbb{X} = (X_1, \ldots, X_n)'$ při pevném x jakožto funkce θ. $X_1, \ldots, X_n$ je náhodný výběr z rovnoměrného rozdělení na intervalu $[0, \theta]$ pro $\theta > 0$.

Literatura

[1] Anděl, J.: *Základy matematické statistiky*, Preprint, Praha 2002

[2] Barnett, G., Zehnwirth, B.: *Best Estimates for Reserves* Proceedings of the CAS, Volume LXXXVII, str. 245-303 2000, www.casact.org

[3] Hoedemakers, T.: *Loss Reserving* Amsterdam 2006

[4] Jedlička, P., Kočvara, J., Strnad, J.: *Techniky výpočtu IBNR a jejich aplikace v pojištění odpovědnosti z provozu vozidel*, Seminář z aktuárských věd MFF UK, 2004, www.actuaria.cz

[5] Mandl, P., Mazurová, L.: *Matematické základy neživotního pojištění* MATFYZPRESS, Praha 1999

[6] Zehnwirth, B.: *Probabilistic Development Factor Models with Applications to Loss Reserve Variability, Prediction Intervals, and Risk Based Capital* Variability in Reserves Prize Program Papers (Volume 1) 1994, www.casact.org

Zaplacené škody

	1/2000	2/2000	3/2000	4/2000	1/2001	2/2001	3/2001	4/2001	1/2002	2/2002	3/2002	4/2002	1/2003	2/2003	3/2003	4/2003	1/2004	2/2004	3/2004	4/2004	1/2005	2/2005	3/2005	4/2005
1/2000	32 215	69 409	39 489	21 721	10 066	8 854	2 973	2 946	2 072	2 140	3 319	1 098	2 112	1 708	901	353	981	488	436	128	376	259	260	900
2/2000	40 112	74 786	33 054	16 487	14 822	4 806	5 348	3 826	2 646	1 868	1 589	2 032	2 107	5 155	628	653	522	319	680	1 124	343	295	256	
3/2000	41 092	83 483	40 242	22 279	8 018	7 902	6 700	6 267	2 768	2 112	2 043	1 815	774	1 297	1 244	601	2 539	1 596	455	198	197	197		
4/2000	50 087	88 739	32 441	12 248	8 175	7 045	6 773	3 671	2 858	1 618	5 633	906	763	555	1 131	581	1 020	245	182	5 173	94			
1/2001	55 667	81 436	20 863	15 368	8 477	6 309	3 885	2 330	4 265	2 818	1 469	2 637	3 811	347	202	788	854	389	936	432				
2/2001	56 925	83 382	31 759	15 906	9 458	5 958	4 992	3 342	4 301	3 191	3 060	2 012	2 112	1 152	654	404	208	562	437					
3/2001	53 677	100 820	29 789	18 364	10 368	6 787	6 762	4 381	4 022	2 725	4 583	3 181	847	1 054	598	399	212	1 029						
4/2001	59 893	103 568	35 378	14 337	8 559	7 670	6 978	4 111	4 321	4 526	1 767	1 166	1 353	898	614	466	223							
1/2002	57 230	81 148	24 508	11 118	9 659	8 506	4 723	3 684	3 853	4 038	2 035	3 106	2 549	1 119	976	986								
2/2002	50 672	84 023	24 947	13 791	11 161	5 905	5 627	4 364	3 841	4 413	1 395	1 719	1 688	1 586	1 043									
3/2002	52 653	86 458	32 151	18 139	10 688	10 813	7 467	4 504	3 527	3 054	7 258	4 352	921	891										
4/2002	53 074	91 691	34 088	13 991	12 415	10 139	8 306	3 290	2 727	2 648	3 004	4 514	1 842											
1/2003	46 144	77 629	19 156	11 633	7 274	8 758	5 077	4 723	3 982	3 188	1 294	829												
2/2003	47 707	74 671	27 034	13 642	10 278	5 186	4 007	3 222	5 130	2 247	1 793													
3/2003	47 019	89 285	24 433	15 788	9 639	7 826	4 961	4 619	4 564	3 476														
4/2003	50 764	79 046	24 871	11 377	11 229	6 669	2 787	3 552	4 369															
1/2004	52 814	74 164	18 083	10 830	6 729	5 681	5 326	5 441																
2/2004	51 187	72 553	26 935	11 795	11 320	6 444	5 310																	
3/2004	48 434	87 182	29 881	16 999	14 247	12 784																		
4/2004	53 005	86 535	25 869	14 315	10 302																			
1/2005	51 494	82 740	22 889	10 894																				
2/2005	49 302	75 497	34 929																					
3/2005	53 074	84 875																						
4/2005	64 286																							

Rezervy

	1/2000	2/2000	3/2000	4/2000	1/2001	2/2001	3/2001	4/2001	1/2002	2/2002	3/2002	4/2002	1/2003	2/2003	3/2003	4/2003	1/2004	2/2004	3/2004	4/2004	1/2005	2/2005	3/2005	4/2005
1/2000	73 033	71 734	60 384	69 973	57 726	55 507	45 420	45 585	45 833	36 902	41 297	41 220	44 281	38 582	36 280	34 538	34 460	40 990	46 595	47 234	45 691	52 636	53 226	47 863
2/2000	95 813	67 800	73 756	51 977	43 181	38 488	35 635	33 106	28 313	28 386	34 231	37 833	28 132	18 450	17 264	15 161	18 592	20 395	19 137	18 289	18 291	19 749	18 223	
3/2000	114 778	113 636	74 530	61 986	56 776	63 892	58 255	36 507	37 103	44 540	31 630	31 360	29 511	21 522	15 433	21 934	21 644	25 567	25 087	24 490	26 211	25 659		
4/2000	115 542	88 125	64 709	55 749	48 761	47 555	37 234	38 080	50 873	48 805	40 490	39 458	30 441	23 609	26 069	29 783	35 269	35 172	38 486	39 231	35 639			
1/2001	120 044	83 836	63 401	48 075	40 364	29 996	26 906	31 857	27 922	29 955	29 498	24 252	15 881	20 899	22 535	21 610	25 179	19 793	18 465	20 257				
2/2001	115 097	82 877	60 065	46 568	34 036	31 394	37 320	34 460	33 333	33 000	28 546	25 067	26 003	33 942	39 088	38 983	39 032	38 325	32 095					
3/2001	130 232	98 749	64 168	44 045	36 081	45 021	34 170	38 248	35 674	32 650	31 079	35 826	46 937	45 872	49 695	51 962	60 813	64 583						
4/2001	115 912	86 011	51 884	41 443	41 959	35 330	37 185	38 141	33 686	33 709	35 405	38 984	32 970	32 164	29 845	36 195	29 544							
1/2002	99 664	51 388	37 022	45 153	39 220	37 704	34 069	39 458	42 775	40 694	47 917	52 172	53 087	52 546	53 388	46 715								
2/2002	81 652	54 152	51 591	39 852	32 696	28 008	27 824	25 628	25 412	29 109	44 835	42 346	43 542	42 515	45 100									
3/2002	100 741	86 740	58 277	45 705	40 007	40 815	54 077	66 691	59 866	63 387	54 328	49 149	71 946	78 456										
4/2002	102 364	78 966	55 649	48 369	44 065	39 793	38 896	45 929	40 457	71 585	72 197	76 339	99 877											
1/2003	89 975	55 767	41 720	74 145	72 236	69 629	92 695	93 222	89 018	87 094	98 839	86 497												
2/2003	80 802	56 030	68 254	51 475	50 144	56 544	46 319	54 951	48 609	48 656	50 634													
3/2003	105 023	84 290	78 898	70 047	74 194	91 265	90 764	89 581	87 923	83 734														
4/2003	97 384	70 675	52 025	48 189	50 696	60 686	74 753	64 717	94 263															
1/2004	93 909	60 113	46 875	48 062	46 541	53 096	71 949	73 695																
2/2004	107 106	94 766	80 694	79 592	85 762	70 822	67 769																	
3/2004	136 497	120 824	104 732	130 919	142 596	159 133																		
4/2004	147 473	119 530	102 435	69 341	60 096																			
1/2005	107 114	117 158	99 105	141 538																				
2/2005	119 615	112 177	98 225																					
3/2005	123 763	96 318																						
4/2005	117 812																							

Odhadnuté hodnoty

		1/2000	2/2000	3/2000	4/2000	1/2001	2/2001	3/2001	4/2001	1/2002	2/2002	3/2002	4/2002	1/2003	2/2003	3/2003	4/2003	1/2004	2/2004	3/2004	4/2004	1/2005	2/2005	3/2005	4/2005
Poz	1	32 215	69 409	39 489	21 721	10 066	8 854	2 973	2 946	2 072	2 140	3 319	1 098	2 112	1 708	901	353	981	488	436	128	376	259	260	900
Předp		54 758	89 506	30 509	15 549	11 162	8 013	4 887	3 508	3 508	2 442	2 442	1 393	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	2	40 112	74 786	33 054	16 487	14 822	4 806	5 348	3 826	2 646	1 868	1 589	2 032	2 107	5 155	628	653	522	319	680	1 124	343	295	256	
Předp		54 224	90 398	30 521	15 549	11 162	6 808	4 887	3 508	3 508	2 442	2 442	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	3	41 092	83 483	40 242	22 279	8 018	7 902	6 700	6 267	2 768	2 112	2 043	1 815	774	1 297	1 244	601	2 539	1 596	455	198	197	197		
Předp		54 224	89 539	30 813	15 549	9 483	6 808	4 887	3 508	3 508	2 442	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	4	50 087	88 739	32 441	12 248	8 175	7 045	6 773	3 671	2 858	1 618	5 633	906	763	555	1 131	581	1 020	245	182	5 173	94			
Předp		54 224	89 501	30 508	13 343	9 487	6 811	4 887	3 508	3 508	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	5	55 667	81 436	20 863	15 368	8 477	6 309	3 885	2 330	4 265	2 818	1 469	2 637	3 811	347	202	788	854	389	936	432				
Předp		54 224	89 501	25 920	13 216	9 578	6 811	4 887	3 508	3 858	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	6	56 925	83 382	31 759	15 906	9 458	5 958	4 992	3 342	4 301	3 191	3 060	2 012	2 112	1 152	654	404	208	562	437					
Předp		54 224	76 041	25 920	13 216	9 487	6 876	4 887	3 858	3 858	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	7	53 677	100 820	29 789	18 364	10 368	6 787	6 762	4 381	4 022	2 725	4 583	3 181	847	1 054	598	399	212	1 029						
Předp		46 068	76 041	25 920	13 210	9 483	6 808	5 428	3 859	3 859	2 687	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	8	59 893	103 568	35 378	14 337	8 559	7 670	6 978	4 111	4 321	4 526	1 767	1 166	1 353	898	614	466	223							
Předp		46 068	76 041	25 920	13 210	9 483	7 487	5 376	3 896	3 859	2 687	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	9	57 230	81 148	24 508	11 118	9 659	8 506	4 723	3 684	3 853	4 038	2 035	3 106	2 549	1 119	976	986								
Předp		46 068	76 041	25 920	13 210	10 429	7 487	5 376	3 859	3 896	2 687	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	10	50 672	84 023	24 947	13 791	11 161	5 905	5 627	4 364	3 841	4 413	1 395	1 719	1 688	1 586	1 043									
Předp		46 068	76 041	25 920	14 528	10 429	7 487	5 376	3 859	3 859	2 712	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	11	52 653	86 458	32 151	18 139	10 688	10 813	7 467	4 504	3 527	3 054	7 258	4 352	921	891										
Předp		46 068	76 041	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 712	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	12	53 074	91 691	34 088	13 991	12 415	10 139	8 306	3 290	2 727	2 648	3 004	4 514	1 842											
Předp		46 068	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 547	1 532	1 532	577	577	577	577	577	577	237	237	237	900
Poz	13	46 144	77 629	19 156	11 633	7 274	8 758	5 077	4 723	3 982	3 188	1 294	829												
Předp		50 663	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 547	1 532	577	577	577	577	577	577	237	237	237	900
Poz	14	47 707	74 671	27 034	13 642	10 278	5 186	4 007	3 222	5 130	2 247	1 793													
Předp		50 663	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 547	577	577	577	577	577	577	237	237	237	900
Poz	15	47 019	89 285	24 433	15 788	9 639	7 826	4 961	4 619	4 564	3 476														
Předp		50 663	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	583	577	577	577	577	577	237	237	237	900
Poz	16	50 764	79 046	24 871	11 377	11 229	6 669	2 787	3 552	4 369															
Předp		50 662	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	577	583	577	577	577	577	237	237	237	900
Poz	17	52 814	74 164	18 083	10 830	6 729	5 681	5 326	5 441																
Předp		50 662	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	577	577	583	577	577	577	237	237	237	900
Poz	18	51 187	72 553	26 935	11 795	11 320	6 444	5 310																	
Předp		50 662	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	577	577	577	583	577	577	237	237	237	900
Poz	19	48 434	87 182	29 881	16 999	14 247	12 784																		
Předp		50 662	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	577	577	577	577	583	577	237	237	237	900
Poz	20	53 005	86 535	25 869	14 315	10 302																			
Předp		50 662	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	583	237	237	237	900
Poz	21	51 494	82 740	22 889	10 894																				
Předp		50 662	83 625	28 505	14 528	10 429	7 487	5 375	3 859	3 859	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	240	237	237	900
Poz	22	49 302	75 497	34 929																					
Předp		50 661	83 625	28 505	14 528	10 429	7 487	5 375	3 858	3 858	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	240	237	900
Poz	23	53 074	84 875																						
Předp		50 661	83 625	28 505	14 528	10 429	7 487	5 375	3 858	3 858	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	240	900
Poz	24	64 286																							
Předp		50 661	83 625	28 505	14 528	10 429	7 487	5 375	3 858	3 858	2 686	2 686	1 532	1 532	1 532	577	577	577	577	577	577	237	237	237	909