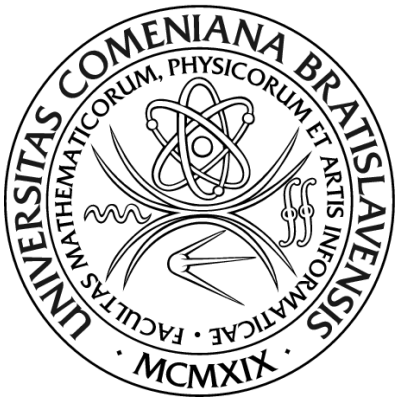
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

MODELOVANIE POČTU UDALOSTÍ

2011 **Peter Kertys**

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



MODELOVANIE POČTU UDALOSTÍ

Bakalárska práca

Študijný program:	Ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor:	9.1.9. aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Školiteľ:	RNDr. Beáta Stehlíková, PhD.
Evidenčné číslo:	38932630-da9a-4aef-bc9d-d4bf216bc203

Bratislava, 2011 **Peter Kertys**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Peter Kertys
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov : Modelovanie počtu udalostí

Cieľ : Naštudovať modely, ktoré sa dajú použiť na modelovanie počtu udalostí. Zamerať sa pritom na Poissonov regresný model, vysvetliť formuláciu modelu, interpretáciu parametrov, spôsob ich odhadovania, testovanie hypotéz. Ukázať odhadovanie a testovania modelu vo zvolenom softvéri a vysvetliť výstup, ktorý softvér dáva. Tieto poznatky aplikovať na samostatnú analýzu zvoleného modelu.

Vedúci : RNDr. Mgr. Beáta Stehlíková, PhD.

Dátum zadania: 29.10.2009

Dátum schválenia: 31.05.2011

.....
doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Dátum potvrdenia finálnej verzie práce, súhlas s jej odovzdaním (vrátane spôsobu sprístupnenia)

2.6.2011 .....
vedúci práce

ČESTNÉ PREHLÁSENIE

Prehlasujem, že predloženú bakalársku prácu som vypracoval samostatne s použitím uvedenej literatúry.

V Bratislave, 3.6.2011 .....

podpis autora práce

POĎAKOVANIE

Chcem sa úprimne poďakovať školiteľke mojej bakalárskej práce, RNDr. Beáte Stehlíkovej, PhD., za jej dôsledné vedenie, konštruktívne pripomienky, ale najmä za jej prívetivý prístup, flexibilitu a ochotu.

Ďakujem aj mojej rodine a priateľom za vytrvalé povzbudzovanie, trpezlivosť a psychickú podporu.

V neposlednom rade ďakujem aj pánu Bohu. Ďakujem taktiež aj Tomášovi Vroňskému a Jane Mózešovej za ich neoceniteľnú štylistickú pomoc počas celého písania práce.

Abstrakt

KERTYS, Peter – Modelovanie počtu udalostí.

[Bakalárska práca]

Kertys Peter – Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky. Školiteľ: RNDr. Beáta Stehlíková, PhD.

Bratislava FMFI UK, 2011, počet strán = 48

Predmetom predkladanej bakalárskej práce je aplikácia matematických poznatkov o Poissonovom rozdelení na vytvorenie Poissonovho regresného modelu, ktorý predpovedá budúci počet udalostí. V práci sa najskôr predvedú rôzne modely pre modelovanie počtu udalostí. Následne sa predstaví Poissonovo a negatívne binomické rozdelenie. V tejto časti sú taktiež opísané ich základné charakteristiky a odvodené stredné hodnoty a disperzie. Ďalej sa práca venuje podrobnému popisu Poissonovho regresného modelu, odhadovaniu parametrov, ich vlastnostiam a testovaniu hypotéz o modely. Dôležitú časť tvorí aj vytvorenie a vysvetlenie modelu v štatistickom softvéri R. Podstatou práce je nakoniec samotná praktická aplikácia poznatkov o Poissonovej regresii na modelovanie počtu úmrtí následkom dopravných nehôd vo Švédsku v rokoch 1986-2004. Vychádzame z analýzy a modelov štatistického inštitútu dopravy vo Švédsku, ktorú podrobne overíme a podrobíme dôkladnému preskúmaniu a testovaniu. Taktiež sa venujeme porovnaniu predpovedaných výsledkov s realitou.

Kľúčové slová: počet udalostí, Poissonovo rozdelenie, negatívne binomické rozdelenie, Poissonova regresia, Poissonov model v R, úmrtia následkom nehôd, smrteľné dopravné nehody, Švédska doprava 1986-2004

Abstract

KERTYS, Peter – Models for count data.

[Bachelor´s work]

Kertys Peter – Comenius University Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics. Thesis supervisor: RNDr. Beáta Stehlíková, PhD.

Bratislava FMFI UK, 2011, 48 pgs.

The theme of proposed work is the application of mathematical knowledge about Poisson distribution in order to create Poisson regression model, which should predict further count data. At the beginning we offer description of different models for count data. Furthermore the Poisson and negative binomial distribution are introduced. Main characteristics, derived mean and variance of these distributions are described too in this section. As follows we are aimed on precise description of Poisson regression model, estimating parameters, their attributes and testing model hypothesis. Creating and implementation of the Poisson model into statistic software R is a significant part as well. The main part is a practical application of information known about Poisson regression in spite of modeling traffic fatalities in Sweden during period 1986-2004. The keystone of this part is study on traffic fatalities made by Swedish National Road and Transport Research Institute. A proper analysis and revise of this study is then followed by model testing. Finally we would compare predicted results with the reality.

Keywords: count data, Poisson distribution, negative binomial distribution, Poisson regression, Poisson model in R, death caused by traffic accident, traffic fatalities, Sweden traffic in 1986-2004

Obsah

Obsah viii
Úvod 1
1 Modely s diskrétnou závislou premennou 2
2 Modelovanie počtu udalostí 4
2.1 Poissonovo rozdelenie 4
2.2 Negatívne binomické rozdelenie 7
2.3 Rozdiely medzi Poissonovým a negatívne binomickým rozdelením 11
2.4 Modelovanie väčšieho počtu núl 12
2.5 Ďalšie modely 13
3 Poissonov regresný model 14
3.1 Formulácia a špecifikácia 14
3.2 Odhadovanie parametrov 16
3.3 Testovanie hypotéz 19
4 Poissonov regresný model v softvéri R 23
4.1 Príprava dát 23
4.2 Modelovanie 24
4.3 Negatívne binomický model 25
5 Model počtu úmrtí následkom dopravných nehôd vo Švédsku 27
5.1 Švédska štúdia 27
5.2 Modelovanie počtu úmrtí následkom nehody podľa štúdie 28
5.3 Alternatívne modely 34
5.4 Testovanie modelov 35
5.5 Negatívne binomický model 36
Záver 38
Literatúra 39

Úvod

Túto tému som si zvolil pre jej spojitosť s blízkou budúcnosťou. Oddávna ľudia premýšľali, ako predpovedať nastávajúce udalosti a s rozvojom matematiky sa tieto sny stávajú skutočnosťou. Dnes sa už bežne modelujú ekonomické a sociologické javy pomocou rôznych modelov, z ktorých hlavná časť má svoj pôvod v dobre popísanej minulosti.

Modelovanie počtu udalostí sa od ostatných modelov líši hlavne svojou diskrétnosťou, t. j. celočíselnosťou. Vďaka tejto vlastnosti možno model počtu udalostí použiť v mnohých oblastiach reálneho života.

V tejto práci sa budeme venovať hlavne Poissonovmu regresnému modelu, ktorý je zároveň najviac používaným modelom pre počet udalostí. Nemáme za cieľ predložiť množstvo informácii o modeli, ale chceme ho predstaviť v jednoduchej, prístupnej forme a predkladáme aj niektoré základné dôkazy. Zhrnieme základné poznatky s prepojením na praktické použitie. Porovnanie predpovedí z minulosti o udalostiach, ktoré už medzitým nastali, nám následne ukáže vhodnosť takéhoto modelovania a celkový prínos modelu.

Dalo by sa povedať, že čitateľ by mohol použiť túto prácu ako návod pre vytvorenie svojho vlastného modelu.

1 Modely s diskrétnou závislou premennou

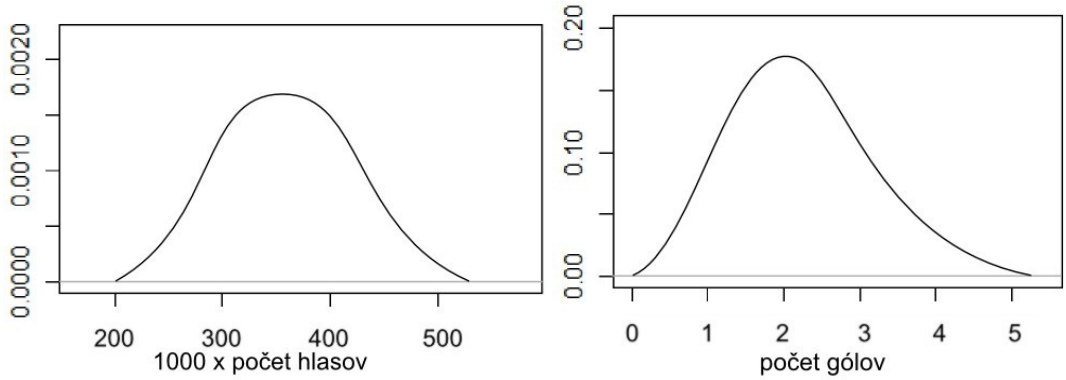
Pre kvalitné popísanie bežných javov a ich vývinu v budúcnosti sa často vytvárajú matematické modely. Jedným z najpoužívanejších štatistických a ekonometrických modelov je *lineárna regresia*. Závislú premennú *Y* modelujeme pomocou jednej, alebo viacerých vysvetľujúcich premenných $x_1,\ldots,x_k$. Predpokladáme, že závislosť má tvar

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon,$$

pre nejaké koeficienty $\beta_0,\ldots,\beta_k$. Náhodná premenná ε vyjadruje odchýlku skutočnej hodnoty *Y* od jej strednej hodnoty. Klasickým predpokladom je normálne rozdelenie ε. Akokoľvek aj pri oslabení tohto predpokladu, čo je bežné, ostáva *Y* pri daných hodnotách $x_1,\ldots,x_n$ spojitou náhodnou premennou. O regresnom modeli sa dá viac dočítať v knihách od Greenea [1], Montgomeryho [2], alebo Fishera [3].

Reálny svet je však v mnohých prípadoch diskrétny. Ako príklad nám môžu slúžiť výsledky volieb, keď kandidát dostane nakoniec presný počet hlasov, alebo počet ochorení, keď presne 21 ľudí ochorie, prípadne, keď sa za rok udeje presne 4200 dopravných nehôd.

V niektorých prípadoch môžeme pre počet udalostí použiť aproximáciu spojitým rozdelením. Na obrázku 1 vľavo je ukážka hustoty, ktorou by sme modelovali počet hlasov, ktorá nejaká strana dostane vo voľbách. Je to spojité rozdelenie, pravdepodobnosť každej konkrétnej hodnoty je nula, ale keďže nemá praktický význam rozlišovať napríklad medzi 420 000 a 420 200 hlasmi, nemusí to byť problém. Prakticky nás zaujímajú pravdepodobnosti určitých intervalov. (V skutočnosti by v tomto prípade mohlo byť lepšie modelovať percento počtu hlasov, ktoré strana dostala). Na obrázku 1 vpravo je spojité rozdelenie navrhnuté pre počet gólov vo futbalovom zápase. Na rozdiel od predchádzajúceho prípadu, tento model použiteľný nie je, lebo pri týchto malých počtoch spojitý model nie je dobrou aproximáciou.



obrázok 1: Príklady rozdelení

Modely pre diskrétne závislé premenné môžeme rozdeliť do niekoľkých skupín podobne ako Green [1]. Ich použitie ilustrujeme na príkladoch.

1. Modelovanie počtu udalostí

Vo všeobecnosti je to premenná, ktorej vývin nadobúda nezáporné hodnoty 0, 1, 2, ... Môžeme napríklad modelovať počet patentov za rok, počet predaných áut v krajine za mesiac, alebo počet vymeškaných dní žiakmi.

Tieto modely majú široké využitie vo viacerých odboroch a disciplínach ako zdravotníctvo, školstvo, poisťovníctvo, ale aj ekonómia.

2. Kvalitatívne modely

Sem patria modely, kde závislá premenná je kódovaním pre nejaký kvalitatívny výstup. Modely môžu byť rôzneho charakteru:

- Binárne modely – napríklad účasť poslancov v parlamente: „nie“ bude kódované ako 0 a „áno“ ako 1
- Modely poradia – napríklad názory v ankete: 0 pre „silne proti“, 1 „proti“, 2 „neutrálny“, 3 „za“, 4 „silne za“. Tieto čísla sú stupňovaním a hodnota je skôr poradím ako kvalitatívnym ukazovateľom.
- Kategorické modely – napríklad typ zamestnania: nech 0 je nezamestnaný, 1 študent, 2 inžinier, 3 právnik, atď. Tieto dáta sú zväčša kategórie, neudávajúce počet ani poradie.

Každý z týchto typov modelov si vyžaduje iný prístup k modelovaniu. My sa v tejto práci budeme zaoberať prvou triedou modelov – modelovaním počtu udalostí.

2 Modelovanie počtu udalostí

V nasledujúcej časti predstavím základné rozdelenia používané pri modelovaní počtu udalostí. Tieto rozdelenia sú spracované predovšetkým podľa [4], iné zdroje sú uvedené na príslušnom mieste. Niektoré základné tvrdenia uvedené v [4] bez dôkazu sú doplnené o dôkazy. Podrobnejšie popíšem hlavne dve najviac používané rozdelenia, Poissonovo a negatívne binomické.

2.1 Poissonovo rozdelenie

Je rozdelením, ktoré sa pri modelovaní počtu udalostí používa najviac. Jeho výhodou je relatívna jednoduchosť a pritom efektívnosť.

2.1.1 Definícia

Nech Y je náhodná premenná s diskrétnym rozdelením, ktoré je definované na množine celých nezáporných čísel $\{0,1,2,\ldots\}$. Y má Poissonovo rozdelenie s parametrom $\lambda > 0$ (označujeme ako $Y \sim Poisson(\lambda)$), ak jeho rozdelenie pravdepodobnosti je

$$P(Y = y)\big|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0,1,2,\ldots \ .$$

2.1.2 Stredná hodnota a disperzia

Strednú hodnotu náhodnej premennej $Y \sim Poisson(\lambda)$ je rovná parametru λ:

$$E(Y) = \sum_{y=0}^{\infty} yP(Y = y) = \sum_{y=0}^{\infty} y\frac{\lambda^{y}}{y!}e^{-\lambda} = e^{-\lambda}\sum_{y=1}^{\infty} y\frac{\lambda^{y}}{y!} = e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^{y}}{(y-1)!} =$$

$$= \lambda e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda \ .$$

Podobne vypočítame $E(Y^2)$:

$$E(Y^2)=\sum_{y=0}^{\infty} y^2 P(Y=y)=\sum_{y=0}^{\infty} y^2 \frac{\lambda^y}{y!}e^{-\lambda}=\lambda e^{-\lambda}\sum_{y=1}^{\infty} y\frac{\lambda^{y-1}}{(y-1)!}$$

$$=\lambda e^{-\lambda}\left(\sum_{y=1}^{\infty}(y-1)\frac{\lambda^{y-1}}{(y-1)!}+\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!}\right)$$

$$=\lambda e^{-\lambda}\left(\lambda\sum_{y=2}^{\infty}\frac{\lambda^{y-2}}{(y-2)!}+\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!}\right)$$

$$=\lambda e^{-\lambda}\left(\lambda\sum_{i\geq 0}^{\infty}\frac{\lambda^{i}}{i!}+\sum_{j=0}^{\infty}\frac{\lambda^{j}}{j!}\right)=\lambda e^{-\lambda}\left(\lambda e^{\lambda}+e^{\lambda}\right)=\lambda(\lambda+1)=\lambda^2+\lambda$$

Disperzia je potom

$$D(Y)=E(Y^2)-\left[E(Y)\right]^2=\lambda^2+\lambda-\lambda^2=\lambda\,.$$

Pre modelovanie pomocou tohto rozdelenia je veľmi dôležitý vzťah, že stredná hodnota je rovná disperzii.

2.1.3 Poissonovo rozdelenie a modelovanie počtu udalostí

Uvažujme udalosti, ktoré náhodne nastávajú v čase a definujme stochastický proces $\{X(t), t\in T\}$ ako počet udalostí, ktoré sa uskutočnia pred časom *t*. Uvedomme si, že potom pre $s<t$ udáva $X(t)-X(s)$ udáva počet udalostí v intervale (s,t).

Proces počtu udalostí sa nazýva stacionárnym ak rozdelenie počtu udalostí v akomkoľvek časovom intervale záleží iba od dĺžky intervalu, t. j. pre každé časy $0\leq t_1\leq t_2$ a pre každé $s>0$ (označuje dĺžku intervalu) platí

$$X(t_2+s)-X(t_1+s)\approx X(t_2)-X(t_1)\,.$$

Hovoríme, že proces počtu udalostí má nezávislé prírastky, ak počet udalostí v disjunktných intervaloch je nezávislý (prírastky procesu *X* vyjadrujú počet udalostí, ktoré nastali).

Pomocou týchto vlastností teraz definujeme Poissonov proces, pričom definujeme rozdelenie jeho prírastkov (t. j. popíšeme, ako nastávajú udalosti, ktorých výskyt tento proces počíta). Poissonov proces je spojitý proces počtu udalostí v čase so stacionárnymi a nezávislými prírastkami. Môžeme to teda formálne zhrnúť ako

1. Pravdepodobnosť, že sa objaví udalosť počas intervalu $(t,t+\Delta)$ je stochasticky nezávislá na počte udalostí pred časom *t*.

2. Pravdepodobnosti počtu udalostí počas intervalu $(t,t+\Delta)$ závisia iba od dĺžky tohto intervalu. Označme $Y(t,t+\Delta)$ počet udalostí, ktoré nastanú počas intervalu $(t,t+\Delta)$. Poissonov proces je charakterizovaný tým, že

$$P\{Y(t,t+\Delta)=1\}=\lambda\Delta+o(\Delta) \text{ a } P\{Y(t,t+\Delta)=0\}=1-\lambda\Delta+o(\Delta).$$

Pripomeňme si, že $o(\Delta)$ je funkcia, pre ktorú platí $[o(\Delta)/\Delta]\to 0$ ak $\Delta\to 0$.

Pravdepodobnosť nastania udalosti počas určitého časového intervalu je proporčná k dĺžke intervalu a proporčný faktor je konštanta nezávislá od *t*. Všimnime si, že potom

$$P\{Y(t,t+\Delta)>1\}=1-P\{Y(t,t+\Delta)=0\}-P\{Y(t,t+\Delta)=1\}=o(\Delta).$$

Označme ďalej $p_k(t+\Delta)=P\{Y(0,t+\Delta)=k\}$ ako pravdepodobnosť, že sa udeje *k* udalostí na intervale dĺžky $(t+\Delta)$ (zo stacionarity vyplýva, že táto pravdepodobnosť závisí iba od dĺžky intervalu). Hodnota pre $\{Y(0,t+\Delta)=k\}$ môže byť vyjadrená ako kombinácia $(k+1)$ možností:

$$\{Y(0,t)=k\}\wedge\{Y(t,t+\Delta)=0\},$$
$$\{Y(0,t)=k-1\}\wedge\{Y(t,t+\Delta)=1\},\ldots,\{Y(0,t)=0\}\wedge\{Y(t,t+\Delta)=k\}$$

Z nezávislosti vyplýva, že pravdepodobnosť každého vyššie uvedeného výstupu sa rovná súčinu jednotlivých pravdepodobností dvoch častí. Napríklad

$$P_1[\{Y(0,t)=k\}\cap\{Y(t,t+\Delta)=0\}]=p_k(t)(1-\lambda\Delta+o(\Delta))=p_k(t)(1-\lambda\Delta)+o(\Delta),$$

a podobne

$$P_2[\{Y(0,t)=k-1\}\cap\{Y(t,t+\Delta)=1\}]=p_{k-1}(t)(\lambda\Delta+o(\Delta))=p_{k-1}(t)\lambda\Delta+o(\Delta),$$
$$P[\{Y(0,t)=k-j\}\cap\{Y(t,t+\Delta)=j\}]=o(\Delta),$$

pre $j\geq 2$. Nakoniec, keďže tieto uvažované možnosti sú disjunktné, pravdepodobnosť ich zjednotenia je daná súčtom ich pravdepodobností, a teda dostaneme

$$p_k(t+\Delta)=p_k(t)(1-\lambda\Delta)+p_{k-1}(t)\lambda\Delta+o(\Delta), \text{ teda}$$
$$\frac{p_k(t+\Delta)-p_k(t)}{\Delta}=-\lambda(p_k(t)-p_{k-1}(t))+\frac{o(\Delta)}{\Delta}.$$

Ak spravíme limitu $\Delta\to 0$, dostaneme

$$\frac{dp_k(t)}{dt}=-\lambda(p_k(t)-p_{k-1}(t)).$$

Podobne

$$\frac{dp_0(t)}{dt}=-\lambda(p_0(t)-p_{-1}(t))=-\lambda(p_0(t)-0)=-\lambda p_0(t).$$

Touto diferenciálnou rovnicou s počiatočnou podmienkou $p_0(0)=1$ (keďže na intervale dĺžky nula nenastane žiadna udalosť s pravdepodobnosťou 1) získame pre $\lambda > 0$

$$p_0(t) = \exp(-\lambda t).$$

Matematickou indukciou teraz dokážeme, že pre každé *n* platí $p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$.

Pre výpočet $p_n(t)$ použijeme indukčný predpoklad, že toto tvdrdenie platí aj pre $k-1$ a dostaneme:

$$p_k(t) = \exp(-\lambda t) p_n(0) + \int_0^t \exp(-\lambda(t-s))\lambda \frac{(\lambda s)^{n-1}}{(n-1)!} \exp(-\lambda s) ds =$$

$$= \exp(-\lambda t) \int_0^t \frac{(\lambda s)^{n-1}}{(n-1)!} ds = \frac{(\lambda t)^n}{n!} e^{-\lambda t},$$

kde sme využili, že $p_n(0)=1$ pre $k \geq 1$.Potom pre každé *t*, tvorí $p_n(t)$ Poissonovo rozdelenie.

2.2 Negatívne binomické rozdelenie

Je hlavnou alternatívou voči Poissonovmu rozdeleniu. Výhodou tohto rozdelenia je pridaný ďalší parameter umožňujúci väčšiu flexibilitu v modelovaní disperzie. Ako sme videl, v prípade Poissonovho rozdelenia, stredná hodnota sa zhoduje s disperziou. Pri analýze reálnych dát sa však často stretávame s tým, že disperzia je väčšia (napríklad rozptyl volebných výsledkov býva takmer vždy väčší ako ich stredná hodnota, preto boli indické voľby v roku 2004 modelované cez toto rozdelenie – viď [5]). Negatívne binomické rozdelenie nám umožňuje túto charakteristiku zachytiť pomocou pridaného parametra θ.

2.2.1 Definícia

Náhodná premenná *Y* má negatívne binomické rozdelenie s parametrami $\alpha \geq 0$ a $\theta \geq 0$, (označujme ako $Y \sim Negbin(\alpha, \theta)$), ak rozdelenie pravdepodobnosti je

$$P(Y=y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{y}, \quad y = 0, 1, 2, \ldots,$$

kde $\Gamma(\bullet)$ je Gama funkcia definovaná vzťahom $\Gamma(s) = \int_0^\infty z^{s-1} e^{-z} dz$ pre $s > 0$.

Klasický binomický koeficient $\binom{n}{k}$ sa pomocou gama funkcie zovšeobecní aj pre n mimo množiny prirodzených čísel, pozri [6]. Potom pravdepodobnosť udalosti $Y = y$ môžeme zapísať ako $P(Y=y)=\binom{\alpha+y-1}{y}p^{\alpha}(1-p)^{y}$, kde sme označili $p=\left(\frac{1}{1+\theta}\right)$.

2.2.2 Stredná hodnota a disperzia

Najprv dokážeme, že platí rovnosť

$$\sum_{k=1}^{\infty}\binom{\alpha+k}{k}p^{\alpha+1}(1-p)^{k}=1\ .$$

Pre zovšeobecnené binomické koeficienty platí $\binom{n}{k}=(-1)^{k}\binom{k-n-1}{n}$, pozri [6].

Z toho dostaneme $\binom{-1-a}{k}=(-1)^{k}\binom{k+a}{n}$, resp. $\binom{k+a}{n}=(-1)^{k}\binom{-1-a}{k}$.

Využitím zovšeobecnenej binomickej vety (viď [6]) dostaneme

$$\sum_{n=0}^{\infty}\binom{k+a}{n}(1-p)^{k}=\sum_{n=0}^{\infty}\binom{-1-a}{k}(-1)^{k}(1-p)^{k}1^{(-1-a)-k}=p^{-1-a}\ ,$$

čo po dosadení potvrdzuje predpokladanú rovnosť.

Strednú hodnotu vyrátame ako:

$$\begin{aligned}
E(Y)&=\sum_{y=0}^{\infty}y\binom{\alpha+y-1}{y}p^{\alpha}(1-p)^{y}=\sum_{y=1}^{\infty}y\frac{(\alpha+y-1)!}{y\cdot(y-1)!(\alpha-1)!}\cdot\frac{\alpha}{\alpha}p^{\alpha}(1-p)^{y}=\\
&=\sum_{y=1}^{\infty}\alpha\binom{\alpha+y-1}{y-1}p^{\alpha}(1-p)^{y+1-1}=\\
&=\sum_{y=1}^{\infty}\alpha(1-p)\binom{\alpha+y-1}{y-1}p^{\alpha}(1-p)^{y-1}=\\
&=\alpha(1-p)\sum_{y=1}^{\infty}\binom{\alpha+y-1}{y-1}p^{\alpha}(1-p)^{y-1}=\\
&=\frac{\alpha(1-p)}{p}\sum_{y=1}^{\infty}\binom{\alpha+y-1}{y-1}p^{\alpha+1}(1-p)^{y-1}=\\
&=\frac{\alpha(1-p)}{p}=\\
&=\alpha\theta
\end{aligned}$$

.

Podobne $E(Y^2)$ bude:

$$E(Y^2)=\sum_{y=0}^{\infty} y^2\binom{\alpha+y-1}{y}p^{\alpha}(1-p)^{y}=\sum_{y=1}^{\infty} yy\frac{(\alpha+y-1)!}{y\cdot(y-1)!(\alpha-1)!}\cdot\frac{\alpha}{\alpha}p^{\alpha}(1-p)^{y}=$$

$$=\sum_{y=1}^{\infty}\alpha y\binom{\alpha+y-1}{y-1}p^{\alpha}(1-p)^{y+1-1}=\sum_{y=1}^{\infty}\alpha(1-p)y\binom{\alpha+y-1}{y-1}p^{\alpha}(1-p)^{y-1}=/(j=y-1)$$

$$=\frac{\alpha(1-p)}{p}\sum_{j=0}^{\infty}(j+1)\binom{\alpha+j}{j}p^{\alpha+1}(1-p)^{j}=\alpha\theta\left[\sum_{j=0}^{\infty} j\binom{\alpha+j}{j}p^{\alpha+1}(1-p)^{j}+1\right]=$$

$$=\alpha\theta\left[\sum_{j=1}^{\infty} j\frac{(\alpha+j)!}{j\cdot(j-1)!(\alpha)!}\frac{(\alpha+1)}{(\alpha+1)}p^{\alpha+1}(1-p)^{j}+1\right]=$$

$$=\alpha\theta\left[\sum_{j=1}^{\infty}\binom{\alpha+j}{j-1}(\alpha+1)p^{\alpha+1}(1-p)^{j}+1\right]=$$

$$=\alpha\theta\left[\sum_{j=1}^{\infty}\alpha\binom{\alpha+j}{j-1}p^{\alpha+1}(1-p)^{j}+\sum_{j=1}^{\infty}\binom{\alpha+j}{j-1}p^{\alpha+1}(1-p)^{j}+1\right]=/(k=j-1)$$

$$=\alpha\theta\left[\alpha\sum_{k=0}^{\infty}\binom{\alpha+k+1}{k}p^{\alpha+1}(1-p)^{k+1}+\sum_{k=0}^{\infty}\binom{\alpha+k+1}{k}p^{\alpha+1}(1-p)^{k+1}+1\right]=$$

$$=\alpha\theta\left[\alpha\sum_{k=0}^{\infty}\frac{(1-p)}{p}\binom{\alpha+k+1}{k}p^{\alpha+2}(1-p)^{k}+\sum_{k=0}^{\infty}\frac{(1-p)}{p}\binom{\alpha+k+1}{k}p^{\alpha+2}(1-p)^{k}+1\right]=$$

$$=\alpha\theta\left[\alpha\frac{(1-p)}{p}\sum_{k=0}^{\infty}\binom{\alpha+k+1}{k}p^{\alpha+2}(1-p)^{k}+\frac{(1-p)}{p}\sum_{k=0}^{\infty}\binom{\alpha+k+1}{k}p^{\alpha+2}(1-p)^{k}+1\right]=$$

$$=\alpha\theta\left[\alpha\frac{(1-p)}{p}+\frac{(1-p)}{p}+1\right]=\alpha\theta[\alpha\theta+\theta+1]$$

Disperzia je potom:

$$D(Y)=E(Y^2)-[E(Y)]^2=\alpha\theta(\alpha\theta+\theta+1)-\alpha^2\theta^2=\alpha\theta(\theta+1)=E(Y)(\theta+1).$$

Keďže $\theta\geq 0$, tak disperzia negatívne binomického rozdelenia presahuje svoju strednú hodnotu.

2.2.3 Parametrizácia

Toto rozdelenie sa používa v rôznych parametrizáciách. Označme

$$\lambda=\alpha\theta,$$

kde λ je stredná hodnota. Rozdelenie sa zväčša používa v 2 nasledovných formách, ktoré popísali už Cameron a Trivedi v roku 1986 (viď [7]).

- Negbin I: V tomto prípade sa uvažuje transformácia $\alpha=\lambda/\theta$. Potom disperzia bude

$$D(Y) = \lambda(1+\theta),$$

a pravdepodobnosť vyjadrená pomocou parametrov λ, θ má tvar

$$P(Y = y) = \frac{\Gamma(\lambda/\theta + y)}{\Gamma(\lambda/\theta)\Gamma(y+1)}\left(\frac{1}{1+\theta}\right)^{\lambda/\theta}\left(\frac{\theta}{1+\theta}\right)^{y}.$$

- Negbin II: Tu uvažujme transformácia $\theta = \lambda/\alpha$. Potom

$$D(Y) = \lambda + \alpha^{-1}\lambda^2.$$

Pravdepodobnosť vyjadrená pomocou parametrov λ, θ má tvar

$$P(Y = y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(\frac{\lambda}{\alpha+\lambda}\right)^{y}.$$

V praxi sa používajú aj mnohé iné parametrizácie spomínané v knihách venujúcich sa diskrétnym rozdeleniam (porovnaj [8]).

2.2.4 Vznik negatívne binomického rozdelenia

Budeme postupovať podľa [9]. Nech náhodná premenná *X* má Poissonovo rozdelenie s parametrom λ. Na veľkosť (hodnotu) parametra λ vplýva veľa faktorov. Preto ho môžeme považovať za náhodnú premennú. Ak predpokladáme, že má gama rozdelenie s parametrami $(k; q/p)$, t. j. jeho hustota je

$$f(\lambda) = \left(\frac{p}{q}\right)^{k}\frac{1}{\Gamma(k)}\lambda^{k-1}e^{\frac{-\lambda p}{q}}, \quad \lambda > 0,\ 0 < p < 1,\ q = 1-p,$$

tak náhodná premenná *X* má negatívne binomické rozdelenie s parametrami $(k; p)$. Pravdepodobnosť toho, že $X = x$ potom bude

$$P(X = x) = \int_0^{\infty}\frac{e^{-\lambda}\lambda^{x}}{x!}\left(\frac{p}{q}\right)^{k}\frac{1}{\Gamma(k)}\lambda^{k-1}e^{\frac{-\lambda p}{q}}d\lambda = \left(\frac{p}{q}\right)^{k}\frac{1}{\Gamma(k)}\frac{1}{x!}\int_0^{\infty}e^{-\lambda\left(1+\frac{p}{q}\right)}\lambda^{k+x-1}d\lambda =$$

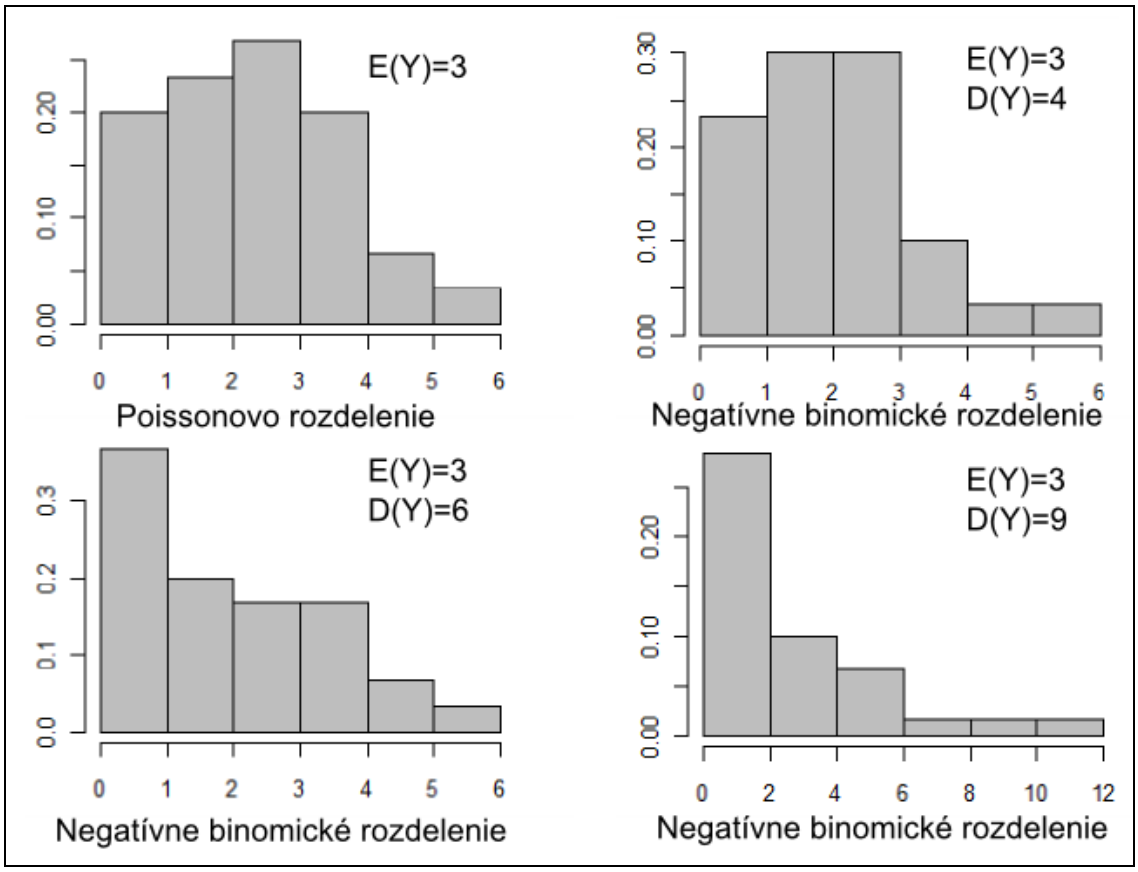
$$= \frac{q^k p^k}{x!\Gamma(k)}\Gamma(k+x) = \binom{k+x-1}{x}p^k(1-p)^x, \qquad x = 0,1,2,\ldots,$$

čo je pravdepodobnostné rozdelenie negatívne binomického rozdelenia.

2.3 Rozdiely medzi Poissonovým a negatívne binomickým rozdelením

Už sme spomenuli, že pri danej strednej hodnote je disperzia Poissonovho rozdelenia jednoznačne určená (a rovná strednej hodnote), kým v prípade negatívne binomického rozdelenia môže nadobúdať rôzne hodnoty (väčšie ako stredná hodnota).

Na obrázku 2 je Poissonovo rozdelenie so strednou hodnotou 3 a niekoľko negatívne binomických rozdelení pre porovnanie s rovnakou strednou hodnotou, ktoré sa líšia disperziou – v jednotlivých prípadoch sa rovná 4, 6, 9. Porovnanie Poissonovho rozdelenia so strednou hodnotou $E(Y) = 3$ a negatívne binomického je podľa obrázka 2 nasledovné:



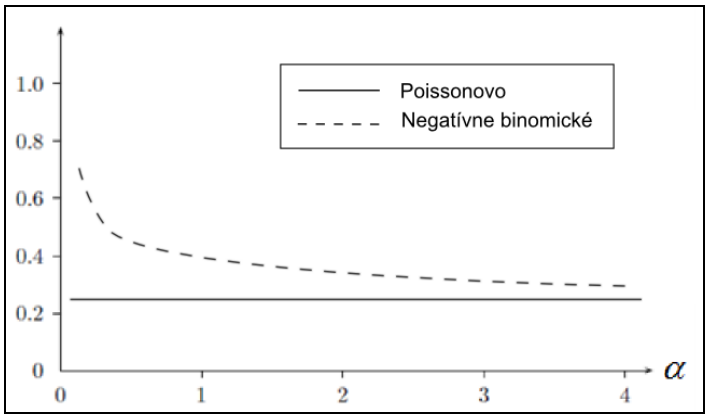
obrázok 2: Porovnanie Poissonovho a negatívne binomického rozdelenia

Ďalším rozdielom medzi Poissonovým a negatívne binomickým rozdelením je pravdepodobnosť toho, že modelovaná premenná nadobudne nulovú hodnotu. Z pravdepodobnostných rozdelení Poissonovho a negatívne binomického rozdelenia vyplýva pre nulový počet udalostí, že

$$P_{poiss}(Y=0)=e^{-\lambda} \qquad \text{a} \qquad P_{negbin}(Y=0)=\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}=\frac{1}{\left(1+\frac{\lambda}{\alpha}\right)^{\alpha}}.$$

Ak teda máme danú strednú hodnotu, v prípade Poissonovho rozdelenia je pravdepodobnosť, že $Y=0$ už určená. V prípade negatívne binomického rozdelenia závisí od druhého parametra. Zoberme teda strednú hodnotu $\lambda=1$, na obrázku 3 vidíme závislosť pravdepodobnosti nulovej udalosti v závislosti od parametra negatívne binomického rozdelenia. Táto pravdepodobnosť je porovnaná s pravdepodobnosťou tejto udalosti pre Poissonovo rozdelenie s rovnakou strednou hodnotou.

Platí (pozri [10]) $\lim_{\alpha\to\infty}\left(\frac{\alpha+\lambda}{\alpha}\right)^{-\alpha}=e^{-\lambda}$, a teda $\lim_{\alpha\to\infty} f_{nb}(0)=\lim f_{psn}(0)$. Pre konečné α však platí (pozri tiež [11]) $\left(1+\frac{\lambda}{\alpha}\right)^{\alpha}<e^{\lambda}$, čo znamená $P_{negbin}(0)>P_{poiss}(0)$.



obrázok 3:Podmienené priblíženie negatívne binomického rozdelenia k Poissonovmu

2.4 Modelovanie väčšieho počtu núl

Tieto modely slúžia na popísanie dát, ktoré obsahujú veľa nulových údajov. Dobré porovnanie týchto modelov je v [11]. Vo všeobecnosti môže nula v dátach znamenať buď skutočnú 0-ovú nameranú hodnotu, alebo nenameranú hodnotu.

- Ako príklad nám môže slúžiť počet detí ženy. Buď mala 0 detí, lebo bola neplodná, alebo mala 0 detí, lebo sa tak rozhodla.
- Iný príklad je modelovanie počtu chytených rýb deťmi, ktoré sa na výlete snažili chytať ryby na rôznych miestach v rôznych podmienkach. Nie všetky deti sa však do chytania rýb zapojili.

Tieto modely sa potom odhadujú pomocou dvoch rovníc, jednou pre počet udalostí vrátane nuly a druhou pre neuskutočniteľné udalosti - nadbytočné nuly.

V tomto modeli sa dopĺňa ešte jedna binárna premenná *c*, pomáhajúca vysvetliť 0-ové počty udalostí:

$$y = \begin{cases} 0 & ak\ c = 1 \\ y^* & ak\ c = 0 \end{cases}$$

Ak pravdepodobnosť, že $c = 1$ označíme ako ω, tak rozdelenie pravdepodobnosti Y môže byť napísané ako

$$P(Y = y) = \omega d + (1 - \omega) g(y), \quad y = 0, 1, 2, \ldots$$

kde $d = \begin{cases} 1\ ak\ y = 0 \\ 0\ ak\ y = 1,2,\ldots \end{cases}$ a $g(y)$ je pravdepodobnosť bežného Poissonovho, alebo negatívne binomického rozdelenia.

Na vyššie spomínanom príklade s rybami by sme v prvom kroku modelovali pravdepodobnosť, že dieťa nechytí rybu. Pre ostatné deti potom použijeme Poissonovo rozdelenie (v ktorom tiež môžu vzniknúť nejaké nuly, ak dieťa síce chytať bude, ale nepodarí sa mu chytiť žiadnu).

2.5 Ďalšie modely

Stručne spomenieme niektoré ďalšie typy modelov:

- Model, kde sa kombinujú dva modely a to počet nulových udalostí a kladný počet udalostí s Poissonovým, geometrickým, alebo negatívne binomickým rozdelením.
- Poissonov, alebo negatívne binomický model, kde sa z dát vyberú nulové hodnoty.
- Krížený model, ktorý vznikajú zložitejším spájaním a parametrizáciou modelu pre väčší počet núl.

Podrobnosti sa dajú nájsť v [4].

3 Poissonov regresný model

Poissonov regresný model je základný model pre modelovanie počtu udalostí, rovnako ako všeobecný lineárny regresný model pre modelovanie reálnych spojitých dát. Vlastnosti modelu, problémy a niektoré z príkladov vychádzajú opäť od Wienkelmana [5]. Prvýkrát sa model spomína v roku 1982 Gilbertom [12], neskôr ho podrobnejšie rozvinuli Cameron a Trivedi v roku 1986 [8]. Model dostal meno podľa základného predpokladu a to, že skúmaná veličina má Poissonovo rozdelenie. Pre svoju relatívnu jednoduchosť, pri ktorej si zachováva robustnosť, je zároveň aj najpoužívanejším modelom pri modelovaní počtu udalostí.

Poissonova regresia vychádza z toho, že závislá celočíselná a kladná premenná Y má Poissonovo rozdelenie. Má jeden parameter λ, ktorý je strednou hodnotou a taktiež disperziou rozdelenia skúmanej premennej. Logaritmus očakávanej hodnoty je modelovaný ako lineárna kombinácia neznámych parametrov. Tento model je tiež niekedy známy pod menom log-lineárny model [13].
Uveďme niekoľko príkladov konkrétneho použitia Poissonovej regresie

- McCullagh a Nelder [14] pozorovali nehody určitých lodí spôsobené vlnami na základe dát poisťovacej spoločnosti. Modelovali počet nehôd odhliadnuc od stupňa a rozsahu poškodenia.
- Švédska štúdia v roku 2005 [15], ktorej sa podrobnejšie venujeme ďalej modelovala počet úmrtí následkom dopravných nehôd za rok na základe dát o obyvateľoch a cestnej premávky z rokov 1986-2004 a predpovedali modelom počet úmrtí v roku 2007.
- Modelovanie počtu narodených detí na Fidži závisiaci od dĺžky trvania manželstva, miestu bydliska a vzdelania matky (viď [16]).

3.1 Formulácia a špecifikácia

Model Poissonovej regresie sa vzťahuje pravdepodobnostné rozdelenie závislej premennej Y v závislosti od regresorov X. Nech k je počet regresorov (zvyčajne zahŕňajúc konštantu), X je potom vektor dimenzie $k \times 1$. Nakoniec n je počet pozorovaní v súbore.
Štandardný Poissonov regresný model používa 3 nasledujúce predpoklady:

- Predpoklad 1 (Poissonovo rozdelenie): Predpokladáme, že skúmaná premenná *Y* má Poissonovo rozdelenie. Pravdepodobnosť, že nastane *y* udalostí za jednotku času, ak parameter rozdelenia je $\lambda > 0$, je potom

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0,1,2,\ldots.$$

- Predpoklad 2 (logaritmická forma): Parameter λ sa dá vyjadriť v tvare

$$\lambda = \exp(x'\beta),$$

kde β je $k \times 1$ vektor parametrov a x je $k \times 1$ vektor regresorov obsahujúci konštantu. To znamená, že logaritmus parametra λ je lineárnou kombináciou regresorov.

- Predpoklad 3 (nezávislosť): Pozorovania $(y_i, x_i), \quad i = 1,\ldots,n$ sú nezávislé a identicky rozdelené.

Hlavné výhody tohto tvaru strednej hodnoty sú (podľa [5]):

- Automaticky zachováva nezáporný rozsah závislej premennej.
- Poskytuje jednoduchú interpretáciu koeficientov v zmysle semi-elasticity – aký percentuálny vplyv má zmena jednej nezávislej premenne, pri zafixovaní ostatných premenných, na odhad $E(Y)$, t. j. λ.
- Ako uvidíme neskôr, vedie k jednoduchým výpočtom funkcie vierohodnosti a odhadovaniu parametrov.

Existujú však aj prípady, kedy takýto model nie je vhodný. Opäť podľa [5] uvedieme niekoľko takýchto situácii:

- **Nepozorovaná heterogenita dát.** Hodnoty regresorov $x_1,\ldots,x_k$ prostredníctvom funkcie $\lambda = \exp(x'\beta)$ určujú intenzitu, s akou sa udalostí objavia. Ak by v modeli bola ešte nejaká ďalšia premenná ovplyvňujúca intenzitu, ktorú by sme však nevedeli pozorovať, čelili by sme problému s nepozorovanou heterogenitou dát.
- **Chyba merania.** Najprirodzenejším pohľadom a modelovaním chyby merania je vychádzať z inej premennej odvodenej od *x* a to napríklad $z = x + \varepsilon$, kde sa o náhodnej premennej ε predpokladá, že je nezávislá od *x*, má strednú hodnotu 0 a kovariančnou maticou Ω. Guo a Li [17] dobre popísali dôsledky takéhoto

nastavenia. Zistili, že takto upravená premenná z nie je veľmi vhodná pre Poissonovu regresiu a narúša konzistentnosť modelu.

- **Závislosť.** Pôvodný model predpokladá o jednotlivých pozorovaniach, že sú nezávislé. V realite sa ale často stáva, že tento predpoklad je narušený, napríklad v prípade časových radov, keď hodnoty premenných v čase nie sú nezávislé. Dá sa to riešiť pomocou modelovania samotného procesu závislosti.
- **Model je zostavený iba z časti skutočných dát.** Stáva sa to napríklad vtedy, ak y^* je celkový počet udalostí (ktorý nepoznáme) a y nahlásený počet udalostí. Toto nenahlásenie môže vznikať z rôznych dôvodov (náhodné nenahlásenie, logistické nenahlásenie a podmienené nenahlásenie).
- Podobne ako v kapitole o pravdepodobnostných rozdeleniach môže byť problémom väčšia disperzia (pripomeňme, že v Poissonovom modeli je rovná strednej hodnote), väčší počet núl, atď.

Nebudeme sa venovať všetkým týmto problémom a potrebným modifikáciám Poissonovho modelu. V našej práci sa budeme venovať otázke disperzie.

3.2 Odhadovanie parametrov

3.2.1 Výpočet odhadov parametrov

Najskôr vysvetlíme princíp metódy maximálnej vierohodnosti (pozri napr. [18]).

Majme n pozorovaní, ktoré chceme modelom vysvetliť. Pre údaje typu (y_i, x_i), kde y je vysvetľovaná premenná, x je vektor regresorov a β je vektor parametrov, združená funkcia hustoty bude súčin jednotlivých podmienených hustôt (v spojitom prípade), resp. pravdepodobností (v diskrétnom prípade) – tu označujeme obe symbolom $f(y)$

$$f(y_1,\ldots,y_n|x_1,\ldots x_n;\beta)=\prod_{i=1}^{n} f(y_i|x_i;\beta).$$

Táto funkcia sa označuje ako funkcia vierohodnosti (likelihood function) a píše sa

$$L = L(\beta; y_1,\ldots,y_n,x_1,\ldots x_n),$$

teda chápeme ju ako funkciu parametrov β.

Potom odhad metódou maximálnej vierohodnosti bude argument maxima tejto funkcie, t. j.

$$\hat{\beta} = \arg\max L(\beta; y_1, \ldots, y_n, x_1, \ldots x_n).$$

Pre zjednodušenie výpočtu sa tento súčin zjednodušuje logaritmickou transformáciou (zachováva argument maxima) na logaritmickú funkciu vierohodnosti (log-likelihood function) $\ell = \log L$. Táto transformácia zároveň umožňuje využitie centrálnej limitnej vety pri skúmaní vlastností získaného odhadu. V prípade Poissonovho regresného modelu bude (pozri [5])

$$\ell(\beta; Y, X) = \log \prod_{i=1}^{n} P(Y = y|x, \beta) = \sum_{i=1}^{n} \log P(Y = y|x, \beta) =$$

$$= \sum_{i=1}^{n} -\exp(x_i'\beta) + y_i x_y'\beta - \log y_i! \quad .$$

Hodnota β, pre ktorú sa nadobúda maximum (označíme ju $\hat{\beta}$) sa vyráta klasickou prvou deriváciou funkcie ℓ, položenou do rovnosti s nulou. V Poissonovom regresnom modeli, tak získame *k* derivácii prislúchajúce k jednotlivým $\beta_0, \beta_1, \ldots, \beta_k$. Tieto hodnoty β_i budeme mať v jednom vektore označovanom ako *gradient*, alebo *skóre*. Tento vektor derivácii vyrátame ako

$$s_n(\beta; y, x) = \frac{\partial \ell(\beta; y, x)}{\partial \beta} = \sum_{i=1}^{n} [y_i - \exp(x_i'\beta)] \times x_i,$$

kde n je počet údajov. Riešime teda rovnicu

$$s_n(\hat{\beta}; y, x) = 0.$$

Na tomto mieste si ešte označme rezíduá $\hat{u}_i$ ako rozdiel predpovedaných hodnôt od nameraných a to $\hat{u}_i = y_i - \hat{E}(y_i|x_i) = y_i - \exp(x_i'\hat{\beta})$.

V získanom systéme rovníc pre vektor s_n ďalej podľa podmienok prvého rádu pre nájdenie maxima pre konštantné x_i musí platiť $\sum_{i=1}^{n} \hat{u}_i = 0$ a pre nekonštantné regresory podmienky ortogonality $\sum_{i=1}^{n} \hat{u}_i x_j = 0, \quad j = 2, \ldots, k$.

Pre overenie, či sme naozaj získali maximum potrebujeme zrátať aj druhú deriváciu funkcie vierohodnosti ℓ. Táto druhá derivácia nám dá Hessián

$$H_n(\beta; y, x) = \frac{\partial^2 \ell(\beta; y, x)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i',$$

ktorý je záporne definitný pre všetky hodnoty β. Funkcia ℓ je teda konkávna a tak platí, že získaný odhad $\hat{\beta}$ je skutočne maximom vierohodnostnej funkcie.

3.2.2 Vlastnosti odhadov

Vlastnosti odhadu $\hat{\beta}$ sú sumarizované v nasledujúcom konvergenčnom vzťahu, ktorý formulovali a dokázali pomocou centrálnej limitnej vety Amemiya [19] a Cramer [20]

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,I(\beta)^{-1;}\right),$$

kde $\xrightarrow{d}$ je konvergencia podľa distribúcie a Fisherova informačná matica $I(\beta)$ je rovná zápornej očakávanej strednej hodnote Hessiánu založenom na pozorovaniach s použitím skutočného parametru β

$$I(\beta)=-E\left[\frac{\partial^2\ell(\beta;y,x)}{\partial\beta\,\partial\beta'}\right]_{\beta}.$$

Odhad metódou maximálnej vierohodnosti je asymptoticky nevychýlený, lebo rozdelenie, ku ktorej konverguje je centralizované na skutočnú hodnotu parametru β. Odhad je ďalej asymptoticky efektívny, keďže jeho disperzia je rovná inverznej Fisherovej informácii, teda Cramér-Raovej dolnej hranici pre nevychýlený odhad.

Keďže tieto asymptotické vlastnosti v striktnom zmysle platia len pre nekonečné vzorky dát, v praxi sa často predpokladá, že platia len približne. Približný odhad distribúcie $\hat{\beta}$ je daný vzťahom

$$\hat{\beta}\sim N\left(\beta,[nI(\beta)]^{-1}\right).$$

Medzi ďalšie dôležité vlastnosti odhadu patrí aj robustnosť. Platí, že odhad pre β je konzistentný pokiaľ stredná hodnota λ je dobre vysvetlená, aj keď predpoklad Poissonovho rozdelenia nie je splnený [4].

3.2.3 Odhad kovariančnej matice

Pre vytvorenie kovariančnej matice potrebujeme najprv zrátať Fisherovu informačnú maticu $I(\beta)$, ktorá závisí na neznámom β. Jeho hodnota je v princípe neznáma, preto môžeme spraviť iba odhad variančnej matice.

Ako sme uviedli, $I(\beta)$ sa rovná zápornej očakávanej strednej hodnote Hessiánu založenom na pozorovaniach a β. Už sme ukázali ako vyrátať Hessián. Môžeme ho

teda použiť pre výpočet $\hat{I}(\beta)$, t. j. aproximácie Fisherovej informačnej matice. Všeobecne použiteľný odhad variančnej matice použitím vzťahu $\hat{I}(\beta) = -nH_n(\hat{\beta})$ je

$$\hat{V}ar(\hat{\beta}) = -\left[H_n(\hat{\beta})\right]^{-1}.$$

3.3 Testovanie hypotéz

Keďže odhadujeme Poissonov regresný model metódou maximálnej vierohodnosti, na testovanie hypotéz môžeme použiť niektorý z troch štandardných testov – test pomerom vierohodnosti, Waldov test a test Lagrangeových multiplikátorov (LM test). Príklady hypotéz, ktoré môže byť zaujímavé testovať sú rôzne lineárne a nelineárne hypotézy o parametroch, alebo Poissonovo verzus nejaké iné všeobecnejšie rozdelenie, prípadne negatívne binomické rozdelenie. Niektoré testy popíšeme ďalej. Príklady mnohých iných testov sú zhrnuté v [21].

Uvažujme hypotézu, ktorá predstavuje platnosť reštringovaného modelu. Nech $\hat{\ell}_r$ je hodnota funkcie vierohodnosti vyrátaná pre reštringovaný model a $\hat{\ell}_u$ je hodnota pôvodnej funkcie vierohodnosti a nech k je počet reštrikcií. **Test pomerom vierohodnosti** je založený na porovnaní týchto dvoch hodnôt. Ak platí H_0, teda reštrikcia je správna, je štatistika

$$LR = -2\left(\hat{\ell}_r - \hat{\ell}_u\right) \sim \chi^2_{(k)},$$

kde $\chi^2_{(k)}$ je chí-kvadrát rozdelenie s k stupňami voľnosti. Ak hodnota LR prekročí stanovenú kritickú hodnotu, zamietame nulovú hypotézu (podľa [1]). Nevýhodou tohto testu je, že vyžaduje pripraviť zvlášť dva odhady a modely, ktoré sa potom porovnávajú.

LM test (test Lagrangeovými multiplikátormi) na rozdiel od predchádzajúceho a aj Waldovho testu nevyžaduje rátať alternatívny model. Je známy aj pod menom skóre test, lebo využíva vektor skóre, t. j. vektor prvých derivácii logaritmickej funkcie vierohodnosti. Majme $\log(L_u)$ ako logaritmickú funkciu vierohodnosti nereštringovaného modelu. Potom $\hat{\theta}_u$ je riešením podmienok prvého rádu

$$\left.\frac{\partial \log(L_u)}{\partial \theta}\right|_{\theta_u} = \left.\frac{\partial \ell_u}{\partial \theta}\right|_{\theta_u} = 0 .$$

Ak do tohto vektora skóre dosadíme odhad reštringovaného modelu θ_r, tak nedostaneme presne nulový vektor. Ak ale hypotéza o reštringovanom modeli platí, nemal by byť od nulového vektora „príliš vzdialený“. Hypotézu H_0 zamietame ak je hodnota skóre „ďaleko“ od nuly. Formálne (viď [1]) pri *k* reštrikciách

$$LM = \left(\frac{\partial \ell(\hat{\theta}_r)}{\partial \hat{\theta}_r}\right)' \left[I(\hat{\theta}_r)\right]^{-1} \left(\frac{\partial \ell(\hat{\theta}_r)}{\partial \hat{\theta}_r}\right) \sim \chi^2_k .$$

V súvislosti s Poissonovou regresiou uveďme zaujímavý príklad LM testu. Je ním jednoduchý test, ktorým vieme po odhadnutí Poissonovej regresie testovať, či by nebol vhodnejší nejaký všeobecnejší model na modelovanie väčšej disperzie (podľa [1] a [8]). Pre dobré porovnanie Poissonovho verzus negatívne binomického modelu slúži LM štatistika

$$LM = \left[\frac{\sum_{i=1}^{n} \hat{w}_i \left[(y_i - \hat{\lambda}_i)^2 - y_i\right]}{\sqrt{2\sum_{i=1}^{n} \hat{w}_i \hat{\lambda}_i^2}}\right]^2 .$$

Váhy $\hat{w}_i$ závisia na predpokladanom alternatívnom rozdelení. Pre negatívne binomicky model budú váhy rovné 1. Nasledovne potom štatistika pre tento model bude jednoducho v tvare

$$LM = \frac{(e'e - n\bar{y})^2}{2\hat{\lambda}'\hat{\lambda}} .$$

Hlavná výhoda tohto testu je, že stačí odhadnúť iba Poissonov model. Ak platí hypotéza, že dáta majú Poissonovo rozdelenie, LM štatistika má chí-kvadrát rozdelenie s 1 stupňom voľnosti.

Test dobrej zhody vo všeobecnosti porovnáva Poissonov model so skutočnými hodnotami (nakoľko dobre ich model popisuje), respektíve s jednoduchým modelom - strednou hodnotou (či je prínos modelu v porovnaní s takýmto jednoduchým modelom dostatočne veľký).

Jeden možný prístup vychádza z Pearsonovej štatistiky

$$P = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}.$$

Motiváciou zostavenia takejto štatistiky je rovnosť (pozri [1])

$$E\left[\sum_{i=1}^{n}(y_i - \lambda_i)^2 / \lambda_i\right] = \sum_{i=1}^{n} E\left[(y_i - \lambda_i)^2 / \lambda_i\right] = \sum_{i=1}^{n} \frac{1}{\lambda_i} E\left[(y_i - \lambda_i)^2\right] =$$
$$= \frac{1}{\lambda_i}\sum_{i=1}^{n}\lambda_i = \frac{n\lambda_i}{\lambda_i} = n \quad ,$$

kde sme využili, že $E\left[(y_i - \lambda_i)^2\right] = D[y_i] = \lambda_i$. V praxi väčšinou presnú hodnotu λ_i nepoznáme, máme k dispozícii len odhad. Preto je P porovnávané s $n-k$, kde k je počet odhadovaných parametrov (pozri [1]). Ak P prekročí kritickú hodnotu χ^2_{n-k} rozdelenia, hypotézu o vhodnosti Poissonovho modelu zamietame.

Cameron a Windmeijer (viď [22]) rozoberali použitie pseudo R^2 na určenie dobrej zhody v rámci tried modelov počtu udalostí. V prípade Poissonovej regresie nie je taký priamočiary predpis R^2 ako pre lineárny model. V knihe [1] sa uvádza niekoľko možností, nie všetky však majú vlastnosti, ktoré by sme od takejto miery kvality modelu očakávali. Príkladom môže byť

$$R_p^2 = \frac{\sum_{i=1}^{n}\left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}\right]^2}{\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]^2}.$$

Tento odhad má dobrú vlastnosť, že porovnáva odhad modelu s modelom založenom len na konštantnom člene. Nevýhodou je jeho možnosť zápornej hodnoty a klesanie ak vynecháme v modeli nejakú premennú (porovnaj [1]). Lepší prístup (podľa [24]) je

$$R_d^2 = \frac{\sum_{i=1}^{n}\left[y_i \log\left(\hat{\lambda}_i / \bar{y}\right) - \left(y_i - \hat{\lambda}_i\right)\right]}{\sum_{i=1}^{n} y_i \log\left(y_i / \bar{y}\right)}.$$

Vieme však, že ak Poissonov model obsahuje konštantný člen, súčet rezíduí je nulový (viď kapitola 3.2.1). Teda sčítanec $\sum_{i=1}^{n}\left(y_i - \hat{\lambda}_i\right)$ je nulový a zostane

$$R^2 = \frac{\sum_{i=1}^{n} y_i \log\left(\hat{\lambda}_i / \bar{y}\right)}{\sum_{i=1}^{n} y_i \log\left(y_i / \bar{y}\right)}.$$

Dá sa dokázať (pozri [1]), že odhad je ohraničený nulou a jednotkou a rastie s počtom pridaných regresorov do modelu.

Pre ďalšie testovanie modelov existuje ešte široká paleta ďalších nástrojov. Napríklad aj simulácie Monte-Carlo, Vuongov, či Hausmanov test. Podrobnejšie sú tieto aj predtým uvedené testy rozobraté takmer vo všetkých knihách venujúcich sa ekonometrickým odhadom a modelom, napríklad [23],[24] a [25].

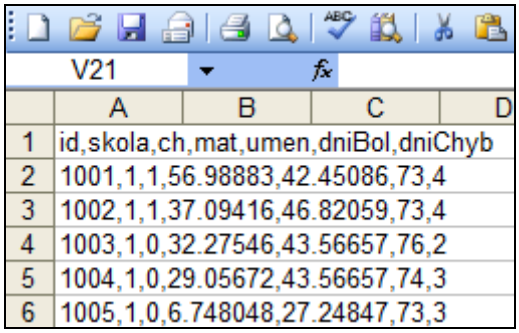
4 Poissonov regresný model v softvéri R

Pre spomínané modely existuje široká paleta nástrojov a programov, v ktorých sa relatívne jednoducho dajú zrátať. Podstatná časť práce potom záleží na pozorovateľovi, ktorý musí vedieť správne získané výsledky vysvetliť a interpretovať. Najrozšírenejším a zároveň najdostupnejším programom pre tieto modely je softvér R, v ktorom sme aj my naše modely.

Keďže tento program je voľne dostupný a na jeho vývoji sa neustále pracuje, ponúka širokú paletu prídavných balíkov a funkcií, ktoré vedia efektívne zrátať všetky vyššie spomínané modely a testy. V nasledujúcej časti uvedieme len tie jednoduchšie príklady modelov a simulácii spomenuté v publikácii o používaní programu R na modelovanie poštu udalostí Achimom Zeileisom (viď [12]).

4.1 Príprava dát

Použitie softvéru R vysvetlíme na konkrétnom príklade. Budeme pracovať s dátami z [24]. Samotné dáta sú v riadku oddelené čiarkami (pozor na desatinné čiarky, respektíve nastavenie rozhrania v R) v súbore s koncovkou „.csv“. V prvom riadku sú názvy premenných, tiež oddelené čiarkami. Viď obrázok 4.


V21

	A	B	C	D
1	id,skola,ch,mat,umen,dniBol,dniChyb			
2	1001,1,1,56.98883,42.45086,73,4			
3	1002,1,1,37.09416,46.82059,73,4			
4	1003,1,0,32.27546,43.56657,76,2			
5	1004,1,0,29.05672,43.56657,74,3			
6	1005,1,0,6.748048,27.24847,73,3			

obrázok 4: Ukážka pripravených dát

Pre načítanie dát použijeme tieto príkazy:

```
p<-read.table("K:/BAKALARKA/chybanie/chybanie.csv", sep=",", header = TRUE)
attach(p)
names(p)
```

Príkaz *attach()* vytvorí vektory veličín podľa názvov v prvom riadku. Pomocou príkazu *names()* si zobrazíme aké názvy premenných máme v dátach. Použitím knižnice **fields** (inicializujeme ju príkazom *library(fields)*) a príkazom *stats()* môžeme získať základný prehľad o modelovaných dátach. Výstup je zobrazený na obrázku 5.

```
> stats(p)
                     id      skola          ch        mat       umen    dniBol    dniChyb
N             316.0000 316.000000 316.0000000 316.000000 316.000000 316.00000 316.000000
mean         1576.3386   1.496835   0.4873418  48.751011  50.063794  74.65823   5.810127
Std.Dev.      502.3638   0.500783   0.5006325  17.880756  17.939211  11.46743   7.449003
min          1001.0000   1.000000   0.0000000   1.007114   1.007114  13.00000   0.000000
Q1           1079.7500   1.000000   0.0000000  37.725360  40.150260  70.00000   1.000000
median       1158.5000   1.000000   0.0000000  48.943760  50.000000  76.00000   3.000000
Q3           2078.2500   2.000000   1.0000000  61.043880  61.043880  84.00000   8.000000
max          2157.0000   2.000000   1.0000000  98.992890  98.992890  86.00000  45.000000
missing values    0.0000   0.000000   0.0000000   0.000000   0.000000   0.00000   0.000000
```

obrázok 5: Výstup načítaných dát a základné charakteristiky dát

Pred vytváraním modelu je potrebné vysvetliť premenné, ktoré chceme použiť v modeli. Z dát, ktoré máme k dispozícii, nás zaujímajú (pre modelovanie *dniChyb)*:

dniChyb ~ počet dní za rok, kedy žiak chýbal,

ch ~ „dummy“ premenná určujúca pohlavie žiaka (0 značí dievča, 1 chlapca),

mat ~ udáva počet bodov vo vstupnom teste z matematiky

umen ~ udáva počet bodov vo vstupnom teste z umenia

4.2 Modelovanie

Na začiatku si musíme uvedomiť akú funkcionálnu formu má model mať.

Poissonov regresný model má exponenciálny tvar $\lambda = \exp(x'\beta) = e^{(x'\beta)}$, teda $\ln(\lambda) = x'\beta$.

Aby sme mohli tento model v R odhadnúť, potrebujem vymenovať premenné, ktoré tvoria vektor *x*.

Preto napríklad model v tvare $dniChyb = a \times b^{mat} \times c^{umen} \times d^{ch}$ zapíšeme ako

$$\ln(dniChyb) \sim \ln a + mat \times \ln b + umen \times \ln a + ch \times \ln d\ .$$

Ak označíme

$$\ln(dniChyb) \sim \beta_0 + \beta_1 \times mat + \beta_2 \times umen + \beta_3 \times ch\,,$$

koeficienty modelu teda budú

$$a = e^{\beta_0}\,,\quad b = e^{\beta_1}\,,\quad c = e^{\beta_2}\,,\quad d = e^{\beta_3}\,.$$

Model odhadnem pomocou funkcie *glm().* Vstupný príkaz pre vytvorenie Poissonovho modelu „m1“ bude

m1<–glm(dniChyb~1+mat+lnumen+ch, family=poisson),

kde „*dniChyb*“ je názov závislej premennej a „*mat*“, „*umen*“ a „*ch*“ sú názvy nezávislých premenných, teda vektor regresorov x vo všeobecnej formulácii modelu.

Pre zobrazenie výsledkov modelu sa použije príkaz *summary()* poskytujúci základné charakteristiky modelu – odhady koeficientov, ich signifikantnosť a ďalšie štatistiky.

```
> summary(m1)

Call:
glm(formula = dniChyb ~ 1 + mat + lnumen + ch, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9579  -2.5939  -1.1073   0.8308  11.6659

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.804379   0.123401  22.726  < 2e-16 ***
mat         -0.009393   0.001599  -5.873 4.28e-09 ***
lnumen      -0.118063   0.039104  -3.019  0.00253 **
ch          -0.360073   0.047977  -7.505 6.13e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2409.8  on 315  degrees of freedom
Residual deviance: 2269.5  on 312  degrees of freedom
AIC: 3138.9

Number of Fisher Scoring iterations: 6
```

obrázok 6: Výstup z R pre príklad Poissonovho modelu

Na obrázku 6 vidieť hodnoty vyrátaných koeficientov v stĺpci „Estimate“ ako hodnoty už spomínaných koeficientov $\beta_0, \beta_1, \beta_2, \beta_3$. Podľa kódov signifikancie vidíme, že všetky koeficienty sú v tomto príklade významné, menej hviezdičiek značí 0,1%, 1% a 5% hladinu významnosti. Pre samostatné zobrazenie koeficientov slúži príkaz *print()*. Predpovedané hodnoty môžeme získať pomocou príkazu *predict()*:

fitted<-predict(m1, type="response").

Ak chceme s týmito hodnotami ďalej pracovať v inom softvéri, dáta sa dajú exportovať pomocou príkazu *write()*:

write(round(fitted), file="K:/BAKALARKA/data.csv", ncolumns=1).

4.3 Negatívne binomický model

Tento model uvádzame ako hlavnú alternatívu k Poissonovmu modelu. Použije sa ak sme vyvrátili hypotézu o rovnosti disperzie so strednou hodnotou. Pri zadávaní modelu použijeme príkaz z knižnice **MASS** a to *glm.nb()*. Tento model nám zráta koeficienty taktiež v logaritmickej forme a navyše pridá hodnotu ďalšieho parametra θ pre lepší odhad disperzie. Vzťah medzi disperziou a odhadnutým parametrom θ, podľa tohto modelu v programe R, je $D(Y) = E(Y) + \frac{[E(Y)]^2}{\theta}$, viď kapitola 2.2.3.

Model, bude vyzerať rovnako ako pri Poissonovom modeli, vyrobíme ho príkazom

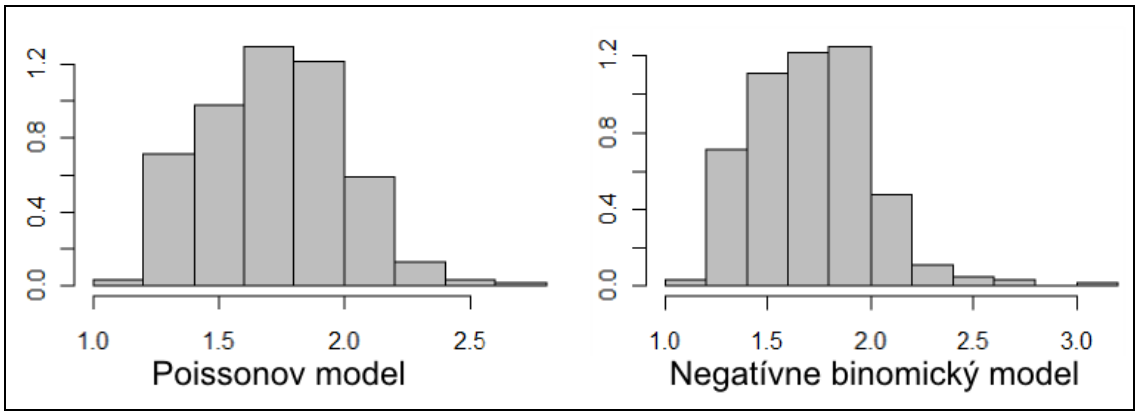
m2<-glm.nb(dniChyb~1+mat+lnumen+ch, link="log").

Výstup tohto konkrétneho modelu je potom na obrázku 8. Odhad pre parameter θ vyšiel 0,7627, čo dáva, pri $E(dniChyb) = 5{,}810127$, $D(dniChyb) = 50{,}07075$. Zároveň nám vyšiel koeficient pri premennej „*mat*“ ako nesignifikantný.

```
> summary(m2)

Call:
glm.nb(formula = dniChyb ~ 1 + mat + lnumen + ch, link = "log",
    init.theta = 0.7626734251)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0347  -1.0607  -0.4424   0.3273   3.0404

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.129105   0.444286   7.043 1.88e-12 ***
mat         -0.006471   0.004710  -1.374  0.16947
lnumen      -0.235799   0.135135  -1.745  0.08100 .
ch          -0.405053   0.139323  -2.907  0.00365 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7627) family taken to be 1)

    Null deviance: 373.71  on 315  degrees of freedom
Residual deviance: 356.90  on 312  degrees of freedom
AIC: 1776.1

Number of Fisher Scoring iterations: 1


              Theta:  0.7627
          Std. Err.:  0.0725

 2 x log-likelihood:  -1766.1420
```

obrázok 7: Výstup z R pre príklad negatívne binomického modelu

Odhady modelovanej premennej môžeme vyvolať tým istým príkazom *predict()*. Hodnoty pravdepodobností si môžeme zobraziť príkazom *hist(predict(m2), col="grey", freq=FALSE))*. Atribút „*freq=FALSE*“ nám zobrazí pravdepodobností, inak by zobrazilo početnosť. Na obrázku 8 je porovnanie pravdepodobností dvoch vyššie popísaných modelov.



obrázok 8: Porovnanie pravdepodobností Poissonovho a negatívne binomického modelu

5 Model počtu úmrtí následkom dopravných nehôd vo Švédsku

V nasledujúcej časti sa budeme venovať konkrétnej analýze vybraných častí zo „Štatistiky nehôd a dopravy a iných súvisiacich vplyvov vo Švédsku," ktorá bola publikovaná v roku 2005 (viď [16]). Štúdia sa venuje kompletnej analýze úmrtí následkom dopravných nehôd po rok 2004.

5.1 Švédska štúdia

Publikácia najskôr predkladá štatistické dáta z obdobia rokov 1950 až 2004. V tejto časti sa venuje počtom obyvateľov, úmrtí, narodení, dopravných nehôd, úmrtí následkom dopravných nehôd, vplyvu práce na nehody a popisu dopravných nariadení v tom období. Štúdia tieto údaje porovnáva na medzinárodnej, ako aj škandinávskej úrovni.

Ďalej sa štúdia venuje štatistickým trendom v rokoch 1996 až 2004. Táto časť má hlavne vysvetľujúci charakter, ktorý dáva do súvisu spomínané okolnosti. Pre toto obdobie ponúka údaje o úmrtiach, rôznych typoch zranení, ich následkov a vplyvov vzhľadom na zabezpečenie políciou, firmami a zdravotníctvom. Možno tu taktiež nájsť údaje o počte vozidiel, vodičov, ich zručností, používania bezpečnostných opatrení, vplyve alkoholu, rýchlostných limitov a bezpečnostných opatrení. Smrteľné nehody potom podľa policajných údajov porovnáva k miere premávky, počtu áut, ľudí a zranení, tiež sa venuje porovnaniu s inými krajiny a mierou nehodovosti na 100 000 obyvateľov.

V tretej časti sa štúdia venuje detailnému popisu vyššie predložených štatistík. Vysvetľuje vplyvy polície, zákonov a trendov dopravy s konkrétnym prepojením na spomínané čísla. Sú tu taktiež spomenuté základné štatistické ukazovatele ako priemerná nehodovosť za mesiac, alebo percentuálne vplyvy počasia, prípadne vyťaženosť ciest, vek vozidiel. K záveru je spomenutá predpoveď vývinu populácie. Kľúčovým prvkom tejto časti je „Predpoveď úmrtí následkom dopravných nehôd pre rok 2007," ktorú budeme ďalej analyzovať.

V poslednej kapitole štúdie sú spomenuté všetky ďalšie dopravné štatistiky a dodatky, ktoré boli použité pri analýze dát, ako i všetky podkladové dáta.

5.2 Modelovanie počtu úmrtí následkom nehody podľa štúdie

V prvých dvoch častiach tejto kapitoly popíšeme modely štúdie [16] a navrhnuté predikcie pomocou nich získané. Štúdia za zamerala na predikciu pre rok 2007, ktorý bol vrcholným obdobím strednodobého plánu znížiť počet úmrtí následkom dopravných nehôd pod hranicu 360. Na nasledujúcich riadkoch bude citovaná predpoveď zo štúdie. Uvedené grafy sme taktiež získali vlastnými výpočtami, ktoré následne vysvetlíme.

5.2.1 Použité premenné

Pre modelovanie počtu úmrtí následkom dopravných nehôd boli použité tieto premenné a údaje:

Mrtvy ~ počet úmrtí spôsobených dopravnými nehodami za rok,

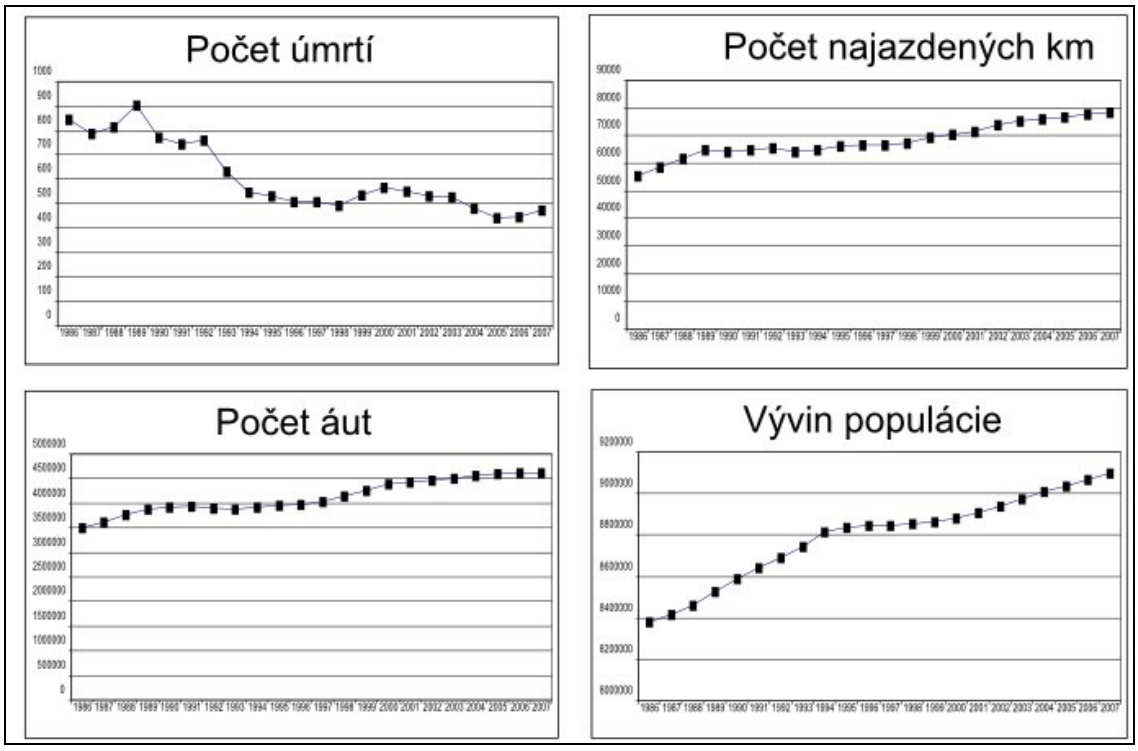
Rok ~ premenná pre potreby modelu, očíslovanie poradia rokov začínajúc 1-kou pre rok 1986 a končiac 22 pre rok 2007,

Prem ~ počet miliónov najazdených km vozidlami vo Švédsku za rok,

Ludi ~ počet obyvateľov Švédska na konci roka,

Aut ~ počet áut vo Švédsku na konci roka.

Vývin týchto premenných je sumárne zhrnutý v obrázku 9:



obrázok 9: Vývin premenných zo štúdie

5.2.2 Odhad modelov a výpočet predikcií pre rok 2007

Všetky modely majú tvar:

$$Očakávaný\ počet\ mŕtvych\ = a \times b^{Rok} \times Prem^{\ c},$$

kde Rok je rovný 1,2,3,... a Prem je miera premávky – počet 1 000 000 najazdených kilometrov. Modely budeme odhadovať pomocou Poissonovej regresie.

Model I

Model I bol vytvorený na základe dát z obdobia rokov 1986 až 2004, a pre predpoveď po rok 2007 bola predpokladaná zvýšená miera premávky o 1% ročne.

Výstup odhadovaného modelu je na obrázku 10:

```
> summary(m1)

Call:
glm(formula = Mrtvy ~ 1 + Rok + lnPrem, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.37506  -1.15446   0.08351   1.37046   3.87050

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -16.87192    3.49782  -4.824 1.41e-06 ***
Rok          -0.06336    0.00454 -13.955  < 2e-16 ***
lnPrem        2.15535    0.31870   6.763 1.35e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 546.248  on 18  degrees of freedom
Residual deviance:  73.515  on 16  degrees of freedom
AIC: 236.58

Number of Fisher Scoring iterations: 4
```

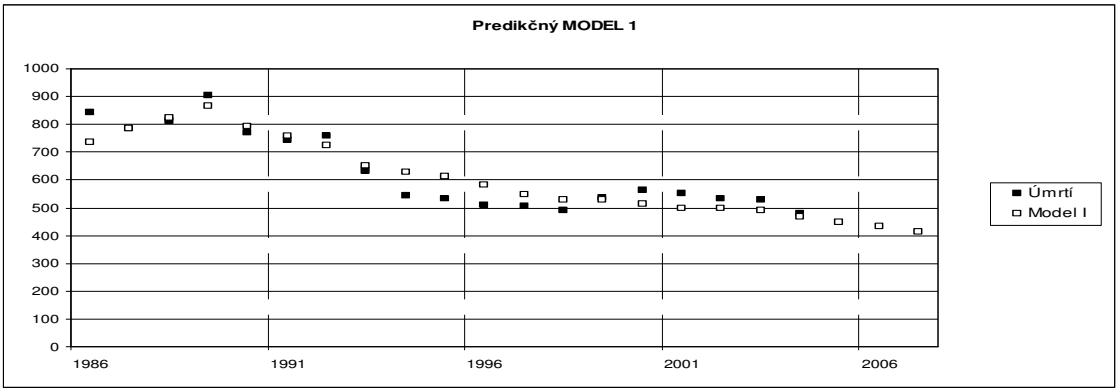
obrázok 10: Výstup z R pre Model I

Všetky koeficienty vyšli signifikantné. Očakávaný počet mŕtvych je teda:

$$Očakávaný\ počet\ mŕtvych\ = e^{-16,87192} \times e^{-0,06336\ \times Rok} \times Prem^{\ 2,15535},$$

kde už je rovnica v tvare podľa kapitoly 4.2.

Pre výpočet predpovede po rok 2007 bola predpokladaná zvýšená miera premávky o 1% ročne. Grafické porovnanie predpovedí s reálnymi počtami, ako aj predpovede pre niekoľko nasledujúcich rokov sú na obrázku 11.



obrázok 11: Porovnanie reality a Modelu I s predpoveďou po rok 2007

Ako vidno na obrázku 11, skutočné počty boli v rokoch 1994–1998 nižšie ako očakávané (predpovedané). V roku 1999 sa trend zmenil a potom boli počty vyššie ako očakávané. Po dosadení predpokladanej miery premávky (rastúcej každoročne o 1%) v roku 2007 na úrovni 78418 dostaneme rovnicu

$$Očakávaný\ počet\ mŕtvych = e^{-16,87192} \times e^{-0,06336 \times 22} \times 78418^{2,15535}$$

Výsledkom je 413 predpokladaných úmrtí následkom dopravných nehôd.

Model II

Model II bol vytvorený len na základe trendu počas rokov 1986 až 1998.

Výstup odhadovaného modelu je na obrázku 12:

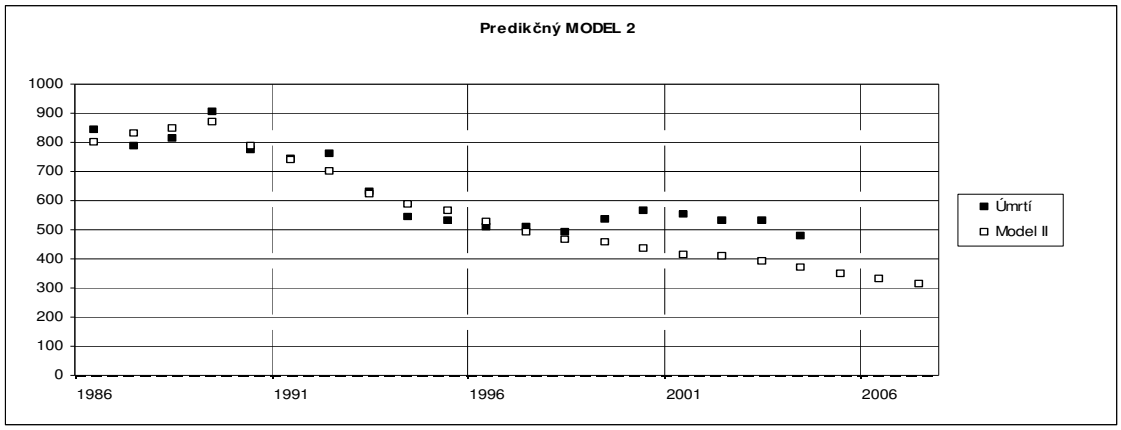
```
> summary(m2)

Call:
glm(formula = Mrtvy ~ 1 + Rok + lnPrem, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8205  -1.2742   0.1285   1.1326   2.1921

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.535416   3.547481  -4.097 4.18e-05 ***
Rok          -0.077348   0.005034 -15.366  < 2e-16 ***
lnPrem        1.950180   0.323141   6.035 1.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 391.958  on 12  degrees of freedom
Residual deviance:  21.134  on 10  degrees of freedom
AIC: 135.52

Number of Fisher Scoring iterations: 3
```

obrázok 12: Výstup z R pre Model II

Aj pre model II, ktorý vychádzal iba z údajov z rokov 1986 až 1998 vyšli všetky koeficienty opäť signifikantné a preto má Model II tvar

$$Očakávaný\ počet\ mŕtvych = e^{-14,535416} \times e^{-0,07734 \times Rok} \times Prem^{1,950180}.$$

Pre výpočet predpovede po rok 2007 bola predpokladaná zvýšená miera premávky o 1% ročne. Grafické porovnanie predpovedí s reálnymi počtami, ako aj predpovede pre niekoľko nasledujúcich rokov sú na obrázku 13:



obrázok 13: Porovnanie reality a Modelu II s predpoveďou po rok 2007

Po dosadení predpokladanej miery premávky (rastúcej každoročne o 1%) v roku 2007 na úrovni 78418 dostaneme rovnicu

$$Očakávaný\ počet\ mŕtvych = e^{-14,535416} \times e^{-0,07734 \times 22} \times 78418^{1,950180}.$$

Výsledkom predpovede je 413 úmrtí. Skutočné a predpovedané hodnoty sedia veľmi dobre pre celé obdobie rokov 1986 až 1998, ale pre roky 1999 až 2004 sú už skutočné hodnoty značne vyššie ako predpovedané.

Model III

Model III (rovnako ako Model I) bol tiež robený z dát pre roky 1986-2004 a doplnený násobiacim vylepšujúcim faktorom („dummy“ premenná pre roky 1999-2007). „Dummy“ premenná mala pre roky 1986-1998 hodnotu $d = 0$ a pre roky 1999-2007 hodnotu $d = 1$. Dôvodom je lepšie vysvetľovanie počtu mŕtvych.

Výstup pre Model III z R je na obrázku 14:

```
Call:
glm(formula = Mrtvy ~ 1 + Rok + lnPrem + d, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1122  -1.0948   0.1873   0.8738   2.1093

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.539675   3.470259  -4.478 7.54e-06 ***
Rok          -0.077212   0.005016 -15.392  < 2e-16 ***
lnPrem        2.040925   0.316187   6.455 1.08e-10 ***
d             0.238500   0.035276   6.761 1.37e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 546.248  on 18  degrees of freedom
Residual deviance:  27.492  on 15  degrees of freedom
AIC: 192.56

Number of Fisher Scoring iterations: 3
```

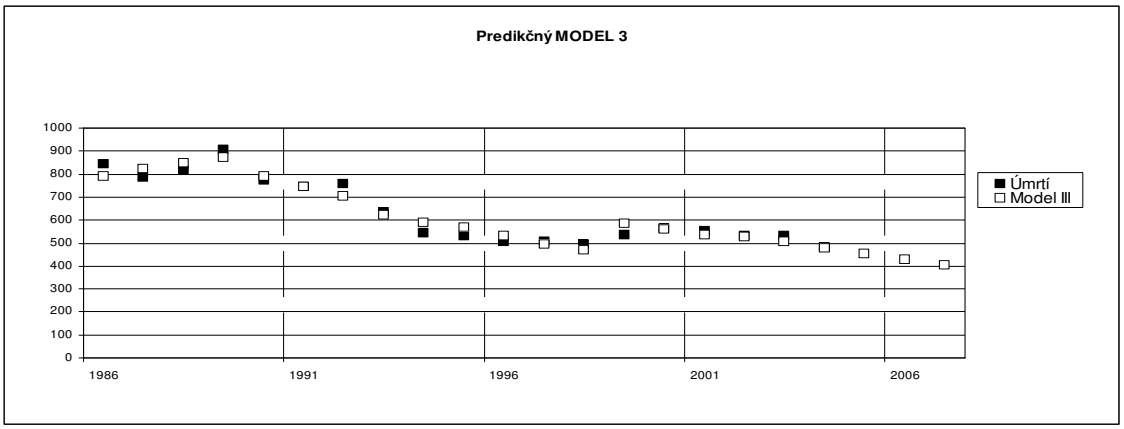
obrázok 14: Výstup z R pre Model III

Aj v tomto modeli vyšli všetky koeficienty signifikantné a Model III má tvar

$$Očakávaný\ počet\ mŕtvych = e^{-15,539675} \times e^{-0,077212 \times Rok} \times Prem^{2,040822} \times e^{0,238506 \times d}.$$

Doplňujúci faktor v tomto prípade násobí model pre roky 1999-2007 približne číslom 1,27.

Grafické porovnanie predpovedí s reálnymi počtami, ako aj predpovede pre niekoľko nasledujúcich rokov sú na obrázku 15:



obrázok 15: Porovnanie reality a Modelu III s predpoveďou po rok 2007

Očakávaná hodnota v roku 2007 (za predpokladu rastu premávky 1%, na úrovni 78418 v roku 2007) nám vyšla 404.

V tabuľke 1 sú zhrnuté výsledky modelov a taktiež aj použité premenné.

Rok	Mrtvy	Model I	Δ I	Model II	Δ II	Model III	Δ III
1986	844	736	108	799	45	789	55
1987	787	785	2	830	-43	823	-36
1988	813	824	-11	850	-37	847	-34
1989	904	865	39	870	34	872	32
1990	772	792	-20	788	-16	789	-17
1991	745	757	-12	741	4	743	2
1992	759	726	33	700	59	702	57
1993	632	651	-19	621	11	622	10
1994	545	627	-82	589	-44	590	-45
1995	531	613	-82	565	-34	568	-37
1996	508	581	-73	528	-20	531	-23
1997	507	549	-42	492	15	494	13
1998	492	528	-36	465	27	468	24
1999	536	530	6	458	78	586	-50
2000	564	514	50	436	128	560	4
2001	551	497	54	415	136	533	18
2002	532	500	32	409	123	527	5
2003	529	489	40	393	136	507	22
2004	480	469	11	371	109	479	1
2005		450		350		453	
2006		431		330		427	
2007		413		311		404	

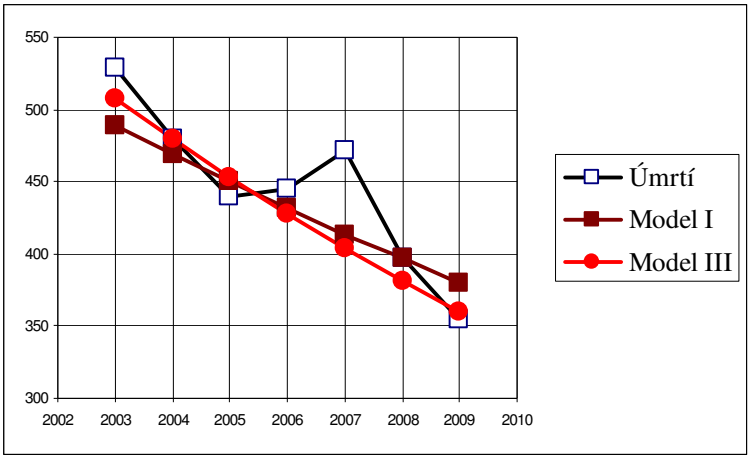
tabuľka 1: Číselné porovnanie predpovedí modelov I, II a III s realitou

5.2.3 Porovnanie predikcií s realitou

Podľa štatistík [26] meraných Európskou úniou bol vo Švédsku vývin počtu úmrtí následkom nehôd relatívne podobný, ako predpovedala štúdia.

Rok	Úmrtí	Model I	Model III
2003	529	489	507
2004	480	469	479
2005	440	450	453
2006	445	431	427
2007	471	413	404
2008	397	396	381
2009	355	380	360

tabuľka 2: Dáta modelov I a III



obrázok 16: Grafické porovnanie vývinu úrmtí

Ako vidieť na obrázku 16 hodnota z roku 2004 skúmaná štúdiou v roku 2004, keď robili odhad na rok 2007 sa z nevysvetlených príčin vymyká predchádzajúcemu a nastávajúcemu trendu. Pre výpočet Modelu I a III pre roky 2004 až 2009 sme predpokladali rast dopravy o 1% ročne.

5.3 Alternatívne modely

Model IV

Pokúsme sa nájsť iný model, vyskúšame použiť aj ďalšie relevantné premenné, ktoré máme k dispozícii. Model bude mať tvar *Očakávaný počet mŕtvych* $= a \times b^{Rok} \times Prem^{\ c} \times Ludi^{d} \times Aut^{e}$. V tomto modeli vyšli signifikantné iba koeficienty pri Roku a počte Ľudí. Ak by sme robili model v tvare *Očakávaný počet mŕtvych* $= a \times b^{Rok} \times Ludi^{c} \times Aut^{d}$, vyjdú nám všetky koeficienty už signifikantné a model má tvar:

$$Očakávaný\ počet\ mŕtvych\ = e^{46{,}343237} \times e^{-0{,}053485\ \times \mathrm{Rok}} \times Ludi^{\ -5{,}546274} \times Aut^{3{,}238902}.$$

Tento model však nie je dobre použiteľný, lebo len ťažko odôvodniť negatívny vplyv rastúceho počtu ľudí (pri zafixovaní ostatných parametrov) na pokles počtu úmrtí následkom nehôd. Pre rovnaký dôvod nie je použiteľný ani model v tvare *Očakávaný počet mŕtvych* $= a \times b^{Rok} \times Ludi^{c} \times Prem^{d}$. Testovali sme aj ostatné kombinácie, avšak tie sa nedali použiť z rovnakých, už spomínaných, dôvodov. Ako príklad majme model:

$$Očakávaný\ počet\ mŕtvych\ = a \times b^{Rok} \times Prem^{c} \times Ludi^{d} \times D^{e},$$

kde D je dummy premenná s hodnotou 0 pre roky 1986-1998 a 1 pre 1999-2004. Výstup tohto modelu je na obrázku 17:

```
Call:
glm(formula = Mrtvy ~ 1 + Rok + lnPrem + lnLudi + d, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7115  -1.0001   0.1450   0.7236   2.4386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 57.02669   38.55609   1.479  0.13912
Rok         -0.05496    0.01277  -4.304 1.68e-05 ***
lnPrem       2.15393    0.32120   6.706 2.00e-11 ***
lnLudi      -4.63062    2.45094  -1.889  0.05885 .
d            0.15342    0.05705   2.689  0.00716 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 546.248  on 18  degrees of freedom
Residual deviance:  23.922  on 14  degrees of freedom
AIC: 190.99

Number of Fisher Scoring iterations: 3
```

obrázok 17: Zlý príklad modelu

Tvar modelu s koeficientmi je:

$$Očakávaný\ počet\ úmrtí = e^{57,02669} \times e^{-0,05496\ \times \mathrm{Rok}} \times Prem^{\ 2,15393} \times Ludi^{-4,63062} \times e^{0,15342\ \times D}$$

Podľa obrázka 17 vidieť, že časť premenných vyšla nesignifikantná a signifikantný koeficient pri premennej „*Ludi*“ vyšiel záporný. Aj tento model je teda nevhodný. Dospeli sme takto k takému istému záveru ako švédska štúdia s Modelom III, na základe ktorého predpovedali vývin po rok 2007. Model III je preto najvhodnejší.

Model V

Pre tento model použijeme novšie dáta, bude vychádzať z údajov z rokov 1989-2009 a budeme sa ním snažiť predpovedať počet úmrtí v roku 2010. Použijeme už osvedčený tvar $Mrtvy = a \times b^{Rok} \times Prem^{\ c} \times d^{D}$, kde D je „dummy“ premenná $D = 0$ pre roky 1989-1998 a $D = 1$ pre roky 1999-2009. Po výpočte bude mať model tvar

$$Očakávaný\ počet\ mŕtvych = e^{-29,487049} \times e^{-0,085777\ \times \mathrm{Rok}} \times Prem^{\ 3,281921} \times e^{0,174919\ \times D}.$$

Pre rok 2010 tak dostaneme hodnotu

$$e^{-29,487049} \times e^{-0,085777\ \times 22} \times 79994^{3,281921} \times e^{0,174919\ \times 1} \doteq 348.$$

Keď budú skutočné dáta k dispozícii, je možné túto predpoveď overiť.

5.4 Testovanie modelov

Existuje mnoho literatúry, ktorá sa podrobne venuje testovaniu regresných modelov (vrátane Poissonovho modelu) v softvéri R (napr. [27], [29],). My sa tu budeme venovať hlavne testovaniu hypotéz spomenutých v teoretickej časti v kapitole 3.3. Budeme testovať Modeli I a III.

- **Pseudo R^2.** Táto hodnota určuje celkovú kvalitu modelu. Čím je táto hodnota bližšie k 1, tým lepšie, lebo to dokazuje, že modelovaná veličina sa málo líši od skutočnosti. Ak pridávame do modelu premennú, musí byť nárast štatistiky dostatočne vysoký. Hodnotu zrátame ako

$$R^2 = \frac{\sum_{i=1}^{n} y_i \log\left(\hat{\lambda}_i / \bar{y}\right)}{\sum_{i=1}^{n} y_i \log\left(y_i / \bar{y}\right)}.$$

  Pre Model I vyšla hodnota 0,8654185, čo je už relatívne blízko k jednotke. Pre Model III vyšla hodnota 0,9496716, čo je jemný nárast.

- **Test dobrej zhody.** V tomto teste porovnávame Pearsonovu štatistiku Modelu I s 5% kritickou hodnotou χ^2_{16} (26,296). Štatistiku vyrátame ako

$$P = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}.$$

Pre Model I dostaneme vysokú hodnotu 73,18174, pre Model III máme hodnotu 27,42053, ktorú porovnávame s 5% kritickou hodnotou χ^2_{15} (24,996), lebo tam máme 4 parametre. Táto štatistika taktiež vyvracia vhodnosť Poissonovho modelu pre tieto dáta.

- **LM test Poissonovho rozdelenia.** Hlavným cieľom je zistiť, či je Poissonov model vhodný. Ak nie, treba použiť alternatívu – negatívne binomický model. Uvažujme hypotézu, že Poissonovo rozdelenie je správne. Potom spomínaná LM štatistika:

$$LM = \frac{\left(e'e - n\bar{y}\right)^2}{2\hat{\lambda}'\hat{\lambda}}$$

nám dá, za predpokladu platnosti nulovej hypotézy, hodnotu s rozdelením chí-kvadrát s $n-k$ stupňami voľnosti. Pre Model I vyšla hodnota 180,4964. Táto hodnota je ďaleko aj za 5% kritickou hodnotou χ^2_{16} (26,296). Pre Model III vyšla 180,4987. Túto hodnotu sme však porovnávali s 5% kritickou hodnotou χ^2_{15} (24,996).). Náš predpoklad, že modelovaný počet úmrtí má Poissonovo rozdelenie je teda podľa týchto testov v oboch modeloch nesprávny.

Výsledky týchto testov sú zhrnuté v tabuľke 3 a 4.

	pseudo R^2	n-k	$5\%\ \chi^2_{n-k}$	LM test	verdikt	Zhoda	verdikt
Model I	0,865	16	26,296	180,496	zamietame	73,182	zamietame

tabuľka 3: Výsledky testov pre Model I

	pseudo R^2	n-k	$5\%\ \chi^2_{n-k}$	LM test	verdikt	Zhoda	verdikt
Model I	0,950	15	24,996	180,499	zamietame	27,421	zamietame

tabuľka 4: Výsledky testov pre Model III

Ako vidieť v oboch prípadoch zamietame Poissonovo rozdelenie, t .j. ani jeden model nie je vhodný pre Poissonovu regresiu.

5.5 Negatívne binomický model

Keďže LM test aj test dobrej zhody signalizovali nevhodnosť predpokladu o Poissonovom rozdelení, spravíme ešte **Model IIIB**, ktorý predpokladá negatívne

binomické rozdelenie. Použijeme dáta z rovnakého obdobia ako v prípade Modelu III (t. j. z rokov 1986 až 2004) aby sme mohli porovnať výsledky týchto modelov.

Model IIIB má taký istý tvar ako Model III.

Výstup z R so všetkými potrebnými údajmi pre Model IIIB je na obrázku 18:

```
Call:
glm.nb(formula = Mrtvy ~ 1 + Rok + lnPrem + d, init.theta = 1357.429066,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7768  -0.8885   0.1512   0.7303   1.7062

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.58906    4.34090  -3.591 0.000329 ***
Rok          -0.07716    0.00616 -12.526  < 2e-16 ***
lnPrem        2.04534    0.39551   5.171 2.32e-07 ***
d             0.23814    0.04205   5.663 1.49e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1357.429) family taken to be 1)

    Null deviance: 369.246  on 18  degrees of freedom
Residual deviance:  18.558  on 15  degrees of freedom
AIC: 192.86

Number of Fisher Scoring iterations: 1


              Theta:  1357
          Std. Err.:  1357

 2 x log-likelihood:  -182.856
```

obrázok 18: Výstup z R pre Model IIIB

Všetky koeficienty vyšli aj v tomto modeli signifikantné a model bude mať teda tvar

$$Očakávaný\ počet\ mŕtvych = e^{-15,58906} \times e^{-0,07716 \times \mathrm{Rok}} \times Prem^{2,04534} \times e^{0,23814 \times \mathrm{d}}$$

Disperzie modelu pri odhadnutom parametri $\theta = 1357$ je podľa vzťahu uvedeného v kapitole 2.2 $D(Mrtvy) = 633,2105 + \frac{[633,210]^2}{1357} = 928,6826$.

V tabuľke 5 je porovnanie výsledkov predpovedaných úmrtí týchto modelov.

Rok	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*
Model III	789	823	847	872	789	743	702	622	590	568	531	494
Model IIIB	788	823	847	872	788	743	702	622	590	568	531	495
Rozdiel	1	0	0	0	1	0	0	0	0	0	0	-1
Rok	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*
Model III	468	586	560	533	527	507	479	453	427	404	381	360
Model IIIB	468	587	560	533	527	507	479	453	428	404	382	361
Rozdiel	0	-1	0	0	0	0	0	0	-1	0	-1	-1

tabuľka 5: Číselné porovnanie Modelov III a IIIB

Oba modely teda takmer identicky predpovedajú vývin počtu úmrtí následkom dopravných nehôd v období rokov 1986 až 2009.

Záver

Ako sme videli, predložené modely počtu udalostí, resp. Poissonov regresný model má skutočne adekvátne a relatívne pravdivé uplatnenie v praxi. Konkrétny príklad švédskej štúdie je taktiež výsledok ich dôsledného sledovania dopravnej situácie a celkového prehľadu o problematike. Takéto schopné inštitúcie, ktoré verejne informujú o svojej dôkladnej práci nám môžu byť vzorom.

Vďaka pokročilej technike a softvérom dnes už nie je problém vskutku okamžite zrátať aj veľký model s mnoho premennými. Na nás však potom ostáva hlavné bremeno modelovania, a to dobre vysvetliť, čo znamenajú premenné v modeli a aký majú vplyv na predpokladanú realitu. Keďže sme používali logaritmickú formu modelovania, pôjde o nejaký percentuálny vplyv nezávislých premenných na modelovaný odhad. Tieto informácie môžu prakticky využiť vlády a inštitúcie pri určovaní a dosahovaní cieľov na zníženie počtu ochorení, úmrtí, alebo prípadne havárií o 5, 10, či 15%.

Napokon je dobré si uvedomiť a mať na pamäti, že akékoľvek modelovanie založené na minulosti nebude nikdy úplne presne predpovedať budúcnosť, lebo žijeme v rýchlom dynamicky sa meniacom svete, do ktorého vstupuje mnoho viac, alebo menej náhodných faktorov. Dôsledky takýchto udalostí možno badať každý deň a bolo by naivné spoliehať sa iba na modely. Ako hovorí vývojár [28] o predpovedaní budúcnosti pomocou zložitých modelov: „ľudia budú musieť vždy urobiť dôležité kroky, ktoré počítač nezvládne.“

Otázka blízkej budúcnosti nadobúda vďaka modelom konkrétnejšiu podobu, ktoré však nenahrádzajú aktuálne plné žitie prítomnosti.

Literatúra

[1] GREENE William, Econometric Analysis, Prentice Hall 5th edition, 2002, ISBN-10: 0130661899

[2] MONTGOMERY Douglas, Introduction to Linear Regression Analysis, Wiley-Interscience, 4th edition, 2006, ISBN 10: 0471754951

[3] FISHER Ronald, Statistical Methods For Research Workers, 1925, ISBN: 0050021702

[4] WINKELMAN Rainer, Econometric Analysis of Count Data, Springer-Verlag 5th edition, 2008, ISBN: 3540776486

[5] BHATTACHARYA Kaushik, Emergence of Independent Candidates:A Negative Binomial Regression Model of an Indian Parliamentary Election, Südasien Institut (SAI), 2010, ISBN: 1617-5069

[6] Wolfram Mathorld, http://mathworld.wolfram.com/BinomialCoefficient.html, 31.5.2011, 16:16

[7] CAMERON Colin A. – TRIVEDI Pravin K., Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests, Journal of Applied Econometrics, strany 29-53, John Wiley & Sons, 1986

[8] JOHNSON Norman L. – KEMP Adrienne W. – KOTZ Samuel, Univariate Discrete Distributions, John Wiley and Sons, 3rd edition, 2005, ISBN: 0471272469

[9] WIMMER Gejza, Diskrétne jednorozmerné rozdelenia pravdepodobnosti, MATFYZPRESS, 2000, ISBN: 80-85863-60-X

[10] NEUBRUNN Tibor – VENCKO Jozef, Matematická analýza I, 1992, ISBN 80 – 223 – 0050 – 0

[11] ZEILEIS Achim – KLEIBER Christian – JACKMAN Simon, Regression models for count data in R, 2008, http://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf 30.5.2011, 12:21

[12] GILBERT Christopher L., Econometric models for discrete (integer valued) economic processes, in: E.G. Charatsis (ed.) Selected Papers on Contemporary Econometric Problems, The Athens School of Economics and Business science, 1982

[13] RODRÍGUEZ Germán, Generalized Linear Models, 2010 http://data.princeton.edu/wws509/notes/c4.pdf, 29.5.2011,13:46

[14] MCCULLAGH P. – NELDER J.A., Generalized linear models, Chapman & Hall/CRC, 2nd edition, 1989, ISBN: 0412317605

[15] VTI, Basic statistics for accidents and traffic and other background variables in Sweden, VTI - N27A-2005, Version 2005-06-30, 2005, http://www.vti.se/EPiBrowser/Publikationer%20-%20English/N27A-2005.pdf, 20.3.2011, 14:52

[16] LITTLE Roderick J.A., Generalized Linear Models for Cross-Classified Data from the WFS, International Statistical Institute in The Hague, 1978

[17] GUO Jie – LI Tong, Poisson regression models with errors-in-variables: implication and treatment, Department of Economics, Indiana University, 2002

[18] KUBÁČEK L. – PÁZMAN A., Štatistické metódy v meraní, Veda, 1979

[19] AMEMIYA T., Advanced Econometrics, Harvard University Press, 1985, ISBN: 0674005600

[20] CRAMER J.S., Econometric Applications of Maximum Likelihood Methods, Cambridge Univerity Press, 1986, ISBN: 0521253179

[21] ROTHERMEL G. – HARROLD M.J., Analyzing regression test selection techniques, Journal IEEE Transactions on Software Engineering, IEEE Press Piscataway, str. 529 - 551 1996, ISSN: 0098-5589

[22] CAMERON Adrian – WINDMEIJER Frank, R-squared Measures for Count Data Regresssion Models with Applications to Health-care Utilization, Working Paper No. 93–24, University of California, 1993

[23] DURBIN J., KOOPMAN S.J., Monte Carlo Maximum Likelihood Estimation for non-Gaussian State Space Models, Biometrika, 1996

[24] BRUIN J., Introduction to SAS, UCLA: Academid Technology Services, Statistical Consulting Group, 2006, http://www.ats.ucla.edu/stat/stata/ado/analysis/, 29.2011, 20:21

[25] YAFFE Robert, A primer for Panel Data Analysis, 2005

[26] EUROSTAT, http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=road_ac_death&lang=en, 20.5.2011, 10:21

[27] BRUIN J., Introduction to SAS, UCLA: Academid Technology Services, Statistical Consulting Group, 2006, http://www.ats.ucla.edu/stat/R/dae/, 30.5.2011, 23:00

[28] GREEN Travis, webstránka Technet CZ, rozhovor zo dňa 27.5.2011, http://technet.idnes.cz/google-kdyz-mate-hodne-dat-muzete-predpovidat-budoucnost-i-chut-piva-1i8-/sw_internet.aspx?c=A110523_100056_sw_internet_pka, 2.6.2011, 16:52