FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Filip Hauptfleisch

Mixed Poisson models for claim counts

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis:	RNDr. Michal Pešta, Ph.D.
Study programme:	Mathematics
Study branch:	General Mathematics

Prague 2017

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague, May 18, 2017 signature of the author

Title: Mixed Poisson models for claim counts

Author: Filip Hauptfleisch

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The thesis summarizes the theory of mixed Poisson models. Poisson distribution is one of the popular distributions in modelling count data, but its use is limited because it requires equidispersion. Because of this we introduce both continuous and finite mixtures. From continuous mixtures the main representative is the negative binomial model, which arises as Poisson Gamma mixture, while from discrete models we deal mainly with zero-inflated models and hurdle models. For these models we use the maximum likelihood estimates of their parameters. In the end we apply these models to fit automobile insurance data from Australia, where we use MLE to fit Poisson regression, negative binomial regression and Poisson hurdle regression.

Keywords: Poisson mixed model, negative binomial model, Poisson hurdle model, claim counts, Poisson regression

I would like to thank my supervisor RNDr. Michal Pešta, Ph.D. for his advices and for his patience with my questions.

Contents

Introduction

This bachelor's thesis is dealing with modelling claim counts, which is important in the insurance industry for classifying risk of policyholders and for setting paid premium correctly. The basic model for count data is the Poisson distribution, but the practical use of Poisson distribution is limited by its restriction on the mean variance equality and by the fact that there are often too many zeroes in real-life data. To overcome this, we introduce mixed Poisson models which can be used in a broader spectrum of applications. The broader use is compensated by more parameters to estimate.

The estimation of parameters will be carried out by the method of maximum likelihood, which achieves good asymptotic properties under regularity conditions.

To deal with the problem of overdispersion we present the concept of continuous mixtures and most notably the negative binomial, which can be derived as Poisson gamma mixture. Then we introduce regression based on this model.

In order to analyse data with the excess of zeroes, finite mixtures are discussed with the main focus on zero inflated Poisson distribution and Poisson hurdle distribution. Both these models are two component mixtures, where in zero inflated model we assume two processes generating zero counts and in hurdle models we assume that the process generating zeroes is different from that generating higher counts.

Then we deal with the problem of choosing the correct model for our data as well as choosing right regressors to be included in the model.

In the last chapter we use the introduced methods on real-life data from automobile insurance dataset from Australia and model claim numbers of policyholders.

1. Poisson distribution

1.1 Definition

Poisson distribution is a discrete distribution, which is widely used in practical applications. X is Poisson distributed ($X \sim Po(\lambda)$) if

$$\mathsf{P}\left[X = k\right] = e^{\lambda} \cdot \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}, \lambda \in \mathbb{R}_+$$

and it has the following properties:

$$\mathsf{E}\left[X\right] = \lambda, \qquad\qquad \operatorname{var}\left[X\right] = \lambda.$$

Thus Poisson distribution is equidispersed, i.e. mean and variance are the same. This property can be limiting in practical use, since real-life data often violate this assumption .

1.2 Origin

There are two main ways from which Poisson distribution can be derived and which have an intuitive idea behind them.

1.2.1 Limit of series of Bernoulli trials

First of the ways is getting Poisson distribution as a limit of series of Bernoulli trials, which is the same as a limit of binomial distribution.

Bernoulli trial Y is a trial which has two possible outcomes, one with probability p, second with $1 - p$. The outcomes are usually noted 1 and 0. So formally

$$\begin{aligned} \mathsf{P}\left[Y = 0\right] &= 1 - p, \\ \mathsf{P}\left[Y = 1\right] &= p, \\ \mathsf{P}\left[Y = k\right] &= 0, \quad k > 1. \end{aligned}$$

If we define $X = \sum_{i=1}^{n} Y_i$, where Y_i are Bernoulli trials with the same p, then X has Binomial distribution with coefficients n and p, so

$$P\left[X = k\right] = \binom{n}{n-k} \cdot p^k \cdot \left(1 - p\right)^{n-k}.$$

When we take limiting form for $n \to \infty$, but with the probability p_n as function of n getting smaller, so that $np_n = \lambda$ is fixed, we get

$$\begin{aligned} \lim_{n \to \infty} P\left[X = k\right] &= \lim_{n \to \infty} \binom{n}{n-k} p_n^k \left(1 - p_n\right)^{n-k} = \\ &= \lim_{n \to \infty} \binom{n}{n-k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \\ &= \lim_{n \to \infty} \frac{n\left(n-1\right)\left(n-2\right)\cdots\left(n-k+1\right)}{n^k} \lambda^k \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} = \\ &= e^{-\lambda} \frac{\lambda^k}{k!}, \end{aligned}$$

which is the probability of Poisson distribution with parameter λ.

Since binomial distribution is being interpreted as the number of successes in n trials with a probability of success in one trial of p, the limiting form can have an interpretation that there is a lot of trials ($n \to \infty$), but with very small probability p (so that $p_n n = \lambda$).

This intuitive interpretation can be used to approximately fit our case of claim counts. We can take the number of policyholders as n where each of them has probability of an accident and thus filling a claim of p. It is intuitively seen that p tends to be rather small (just a small percentage of people have an accident) while the total number of policyholders tends to be large.

1.2.2 Poisson process

Second way of getting Poisson distribution is from the Poisson process.

Let $\{X_t, t \in T\}$ be a stochastic count process. It is called Poisson process, if all of the following requirements are fulfilled.

- For $k \in \mathbb{Z}$, $h > 0$ and $s \geq t$: $\mathsf{P}\left[X_{t+h} - X_t = k\right] = \mathsf{P}\left[X_{s+h} - X_s = k\right]$, i.e. the process has stationary increments.
- For any integer k and any numbers $0 \leq t_0 < t_1 < t_2 < ... < t_k$ random variables $X_{t_1} - X_{t_2}, X_{t_2} - X_{t_3}, ..., X_{t_k-1} - X_{t_k}$ are independent, i.e. the process has independent increments.
- $\mathsf{P}\left[X_h = k\right] = \begin{cases} 1 - \lambda h + o\left(h\right), & k = 0 \\ \lambda h + o\left(h\right), & k = 1 \\ o\left(h\right), & k \in \mathbb{N} \setminus \{1\} \end{cases}$

  for $h \to 0$.

Here we will consider only real Poisson process, that is $T = \mathbb{R}$.

As proven in Denuit et al. [2007], when $\{X_t, t \in \mathbb{R}\}$ is Poisson process with parameter λ, then the number of events during time period s is Poisson distributed with mean $s\lambda$, formally

$$\left(X_{k+s} - X_k\right) \sim Po\left(s\lambda\right).$$

Because of this, the Poisson distribution can be interpreted as number of occurrences of certain event during a period of time. That is once more useful when thinking about claim counts, since we are interested in modelling claim counts during different periods of time.

1.3 Maximum likelihood estimate for Poisson distribution

In this thesis, we will be estimating only parameters of discrete distributions, so the discussion here will be limited to these distributions. One of the most widely used techniques for parameter estimation is the maximum likelihood estimate (MLE), which can be used if the data come from a distribution with probability mass function (PMF) $p\left(x, \boldsymbol{\theta}\right)$ where p is a known function and $\boldsymbol{\theta}$ is the parameter to estimate.

Let $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^\top$ be a vector of independent random variables from distributions with probability mass functions $p_i(x, \boldsymbol{\theta})$ where the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m)^\top$ is a vector of parameters to estimate. Then we define the maximum likelihood function as the joint probability mass function of $\boldsymbol{X}$ as follows:

$$L(\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p_i(x_i, \boldsymbol{\theta}).$$

So $L\left(\boldsymbol{X}, \boldsymbol{\theta}^{(\mathbf{0})}\right)$ is the probability of $\boldsymbol{x}$ occurring if X_i was from distribution with probability mass function $f\left(x, \boldsymbol{\theta}^{(\mathbf{0})}\right)$. Now it makes sense to maximize the probability and thus the maximum likelihood function. Since the product can be hard to deal with, we usually work with logarithm of the maximum likelihood function, which we denote $l(\boldsymbol{X}, \boldsymbol{\theta})$, i.e.

$$l(\boldsymbol{X}, \boldsymbol{\theta}) = \ln(L(\boldsymbol{X}, \boldsymbol{\theta})) = \sum_{i=1}^{n} \ln(p_i(x_i, \boldsymbol{\theta})).$$

Since logarithm is a strictly increasing function, $l(\boldsymbol{X}, \boldsymbol{\theta})$ is maximized at the same point $\hat{\boldsymbol{\theta}}$ as $L(\boldsymbol{X}, \boldsymbol{\theta})$. This point is called the maximum likelihood estimate and we will denote it by hat above the parameter name. To find the maximum, we usually solve the following system of maximum likelihood equations:

$$\frac{\partial l(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\boldsymbol{X}, \hat{\boldsymbol{\theta}}\right) = 0. \tag{1.1}$$

Estimate for Poisson distribution

Now we will look at Poisson distribution. But as discussed before, when dealing with claim counts, it is typical to have the observations for different periods of time. Let $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ be a vector of random observations with $X_i \sim Po(\omega_i \lambda)$, where ω_i is the exposure of the observed X_i. Then, we have the log-likelihood function

$$l(\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(e^{-\omega_i \lambda} \frac{(\omega_i \lambda)^{x_i}}{x_i!}\right) = \sum_{i=1}^{n} \left(-\omega_i \lambda + x_i \ln(\omega_i) + x_i \ln(\lambda) - \ln(x_i!)\right).$$

And after differentiating with regard to λ we get

$$\frac{\partial l(\boldsymbol{X}, \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \left(-\omega_i + \frac{x_i}{\lambda}\right).$$

Putting this equation equal to 0 gives

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \omega_i}$$

as the maximum likelihood estimate of λ. Here we can note that if $\omega_i = 1$, for $i = 1, \dots, n$, we get $\hat{\lambda} = \bar{\boldsymbol{X}}$, the sample mean.

Asymptotic properties

Under regularity conditions, some asymptotic properties can be shown to hold. The following regularity conditions were stated in Casella and Berger [2002] [p. 516]:

1. $X_1, X_2, \ldots, X_n$ are iid.

2. The parameter θ is identifiable, that is, if $\theta \neq \theta'$ then $f(x \mid \theta) \neq f(x \mid \theta')$.

3. The densities $f(x \mid \theta)$ have common support and $f(x \mid \theta)$ is differentiable in θ.

4. The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

5. For every $x \in \mathbb{X}$, the density $f(x \mid \theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ and $\int f(x \mid \theta)\,\mathrm{d}x$ can be differentiated three times under the integral sign.

6. For any $\theta_0 \in \Omega$ there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left|\frac{\partial^3}{\partial\theta^3} \ln g(x \mid \theta)\right| \leq M(x) \text{ for all } x \in \mathbb{X}, \theta_0 - c < \theta < \theta_0 + c$$

   with $E_{\theta_0}[M(\boldsymbol{X})] < \infty$.

And the following theorem from Casella and Berger [2002][Theorem 10.1.6] gives us asymptotic properties of MLEs.

Theorem 1. *Let X_i be iid with density $f(x \mid \theta)$. Let $\hat{\theta}$ denote the MLE of θ. Let $g(\theta)$ be a continuous function, then, under the regularity conditions (1)-(6), $\hat{\theta}$ is a consistent estimate of θ and*

$$\sqrt{n}\left[g(\hat{\theta}) - g(\theta)\right] \rightarrow n[0, v(\theta)],$$

where $v(\theta)$ is the Cramér-Rao Lower bound, so that $\hat{\theta}$ is fully efficient.

1.4 Poisson Regression

Now let us assume we have n observations $(X_i, \boldsymbol{Y_i})$ which are iid. The variable X_i is our response variable and $\boldsymbol{Y_i} = (Y_{i,1}, \ldots Y_{i,m})^\top$ is a vector of regressors. And let us assume that

$$\mathsf{P}[X_i = x \mid \mathbf{Y_i} = \mathbf{y_i}] = e^{-\lambda(\mathbf{y_i}, \boldsymbol{\beta})} \frac{\lambda(\mathbf{y_i}, \boldsymbol{\beta})^{x_i}}{x_i!},$$

where $\lambda(\mathbf{y_i}, \boldsymbol{\beta})$ is a non-negative function of $\mathbf{y_i}$ and a vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ of parameters to estimate.

Most common choice of $\lambda\left(\mathbf{y_i}, \boldsymbol{\beta}\right)$ is the log-linear model: $\lambda\left(\mathbf{y_i}, \boldsymbol{\beta}\right) = e^{\mathbf{y_i}^\top \boldsymbol{\beta}}$. To estimate $\boldsymbol{\beta}$, we can use the same approach as before and use MLE. In this case, when we have observations $\mathbf{X}$, the maximum likelihood function is

$$L\left(\mathbf{X}, \boldsymbol{\beta}\right) = \prod_{i=1}^{n} e^{e^{\mathbf{y_i}^\top \boldsymbol{\beta}}} \frac{\left(e^{\mathbf{y_i}^\top \boldsymbol{\beta}}\right)^{x_i}}{x_i!}$$

and

$$l\left(\mathbf{X}, \boldsymbol{\beta}\right) = -\sum_{i=1}^{n} e^{\mathbf{y_i}^\top \boldsymbol{\beta}} + \sum_{i=1}^{n} x_i \mathbf{y_i}^\top \boldsymbol{\beta} + \sum_{i=1}^{n} \ln\left(x_i!\right).$$

After differentiating with respect to β_k we get

$$\frac{\partial l\left(\mathbf{X}, \boldsymbol{\beta}\right)}{\partial \beta_k} = \sum_{i=1}^{n} x_i y_{i,k} - e^{\boldsymbol{y_i}^\top \boldsymbol{\beta}} y_{i,k},$$

which can be written as vector in the following manner:

$$\frac{\partial l\left(\mathbf{X}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} x_i \boldsymbol{y_i} - e^{\boldsymbol{y_i}^\top \boldsymbol{\beta}} \boldsymbol{y_i}.$$

From this we get the MLE estimate $\hat{\boldsymbol{\beta}}$ as the solution of the following equation:

$$\sum_{i=1}^{n} \boldsymbol{y_i}\left(x_i - e^{\hat{\boldsymbol{\beta}} \boldsymbol{y_i}}\right) = 0.$$

But the solution has to be found numerically.

Regularity of Poisson regression

For Poisson regression we can introduce regularity conditions which are easier to check. Here we will present the conditions on stochastic regressors as stated in Fahrmeir and Kaufmann [1986].

1. The pairs $\left(X_i, \boldsymbol{Y_i}\right)$ are iid as pairs of random variables.
2. The matrix of second moments of $\boldsymbol{Y_i}$ exists and is positive definite:
$$0 < \mathsf{E}\left[\boldsymbol{Y_i}^\top \boldsymbol{Y_i}\right] < \infty.$$
3. $\mathsf{E}\left[e^{\boldsymbol{Y_i}^\top \boldsymbol{\beta}}\right] < \infty$ for all $\boldsymbol{\beta}$, i.e. the moment generating function exists.

Under these conditions the asymptotic results hold as was shown in Fahrmeir and Kaufmann [1986]. The normality can be written as

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N\left[\mathbf{0}, \boldsymbol{J}\left(\boldsymbol{\beta}_0\right)^{-1}\right],$$

where $\boldsymbol{\beta}_0$ is the true value of parameter $\boldsymbol{\beta}$ and $\boldsymbol{J}\left(\boldsymbol{\beta}_0\right)$ is the Fisher information matrix. In order to use this finding, we need consistent estimate of $\boldsymbol{J}\left(\boldsymbol{\beta}_0\right)$, which can be achieved by taking $\boldsymbol{J}(\hat{\boldsymbol{\beta}})$.

If the model is misspecified, but $\mathsf{E}\left[X\right] = e^{\boldsymbol{y}^\top \boldsymbol{\beta}}$ holds and the variance is finite, then the MLE estimate is still consistent, but the asymptotic normality does not work with the variance matrix stated above. To have asymptotic normality preserved even when the model is misspecified, we have to use sandwich form of variance matrix in the form of $A^{-1}BA^{-1}$, where, in the case of Poisson regression,

$$A = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} e^{\boldsymbol{y}_i^\top \boldsymbol{\beta}_0} \boldsymbol{y}_i^\top \boldsymbol{y}_i$$

and

$$B = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} v_i \boldsymbol{y}_i^\top \boldsymbol{y}_i,$$

where v_i is the real variance of X_i. When the model specification is correct, $A = B$ and so $A^{-1}BA^{-1} = A^{-1}$. In order to use this result, we have to get consistent estimate of the variance matrix by taking consistent estimates of $\boldsymbol{\beta}_0$ and v_i. The process of getting these results is described in Cameron and Trivedi [2013][pp. 31–34].

2. Mixed Poisson models

There are several reasons why we need better models than Poisson distribution. The first source of this need may come from overdispersion, the property of variance exceeding the mean in our dataset. Since Poisson distribution is equidispersed, it cannot sufficiently cover overdispersed data. In this case Mixed Poisson models can explain unobserved heterogenity. That is, in our data there may be some groups, which have different parameters, but we cannot observe to which group which observation belongs. In case of regression, it means that there are still some unobserved parameters left that influence the output. In our case of insurance this would mean that not all of the policyholders have the same probability of causing an accident and that they come from different groups with different means, where this variability cannot be explained by regressors. We will be dealing with this problem mainly in Section 2.1.

Another common problem with Poisson distribution is that in real datasets there is often excess of zeroes compared to expected number of zeroes from Poisson distribution with the same mean. In non-life insurance, this can be seen as unwillingness to report claims with low damage, as not to increase the insurance rate. This problem will be dealt with in Subsections 2.2.1 and 2.2.2.

2.1 Continuous mixtures

We say that X follows mixed Poisson distribution with mixing distribution Λ, where Λ has values in $\Theta \in \mathbb{R}_+$ and has a density g, if X has probability mass function of the following form:

$$p(x) = \int_\Theta e^{-\theta} \frac{\theta^x}{x!} g(\theta) \mathrm{d}\theta. \tag{2.1}$$

We note this mixed Poisson distribution as MP(g).

It follows that the conditional probability

$$p(x \mid \theta = \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

is Poisson distributed.

2.1.1 Properties

Continuous mixtures have the following properties:

$$\begin{aligned}
&\mathsf{E}\left[X \mid \lambda\right] = \lambda, \\
&\mathsf{E}\left[X\right] = \mathsf{E}\left[\mathsf{E}\left[X \mid \lambda\right]\right] = \mathsf{E}\left[\lambda\right], \\
&\mathrm{var}[X] = \mathrm{var}[\mathsf{E}\left[X \mid \lambda\right]] + \mathsf{E}\left[\mathrm{var}[X \mid \lambda]\right] = \mathrm{var}[\lambda] + E[\lambda].
\end{aligned}$$

From the last expression it is clear that MP(g) has higher variance than mean, so it is overdispersed.

Another interesting property can be derived from Shaked two crossing theorem as written in Denuit et al. [2007][p. 26] and proven in Shaked [1980].

Theorem 2. *Let $X \sim \mathrm{MP(g)}$ with $\mathsf{E}[X] = \lambda$, then there exist two integers $0 \leq k_0 < k_1$, such that*

$$\begin{aligned}
\mathsf{P}[X = x] &\geq e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots, k_0, \\
\mathsf{P}[X = x] &\leq e^{-\lambda}\frac{\lambda^x}{x!}, \quad k = k_0 + 1, \ldots, k_1, \\
\mathsf{P}[X = x] &\geq e^{-\lambda}\frac{\lambda^x}{x!}, \quad x > k_1.
\end{aligned}$$

This theorem allows us to compare mixed Poisson distributions with standard Poisson distributions with the same mean. It tells us that MP(g) has higher probability of zero and also thicker tail than Poisson distribution.

Another theorem, which gives us useful property under some conditions is stated in Holgate [1970]:

Theorem 3. *Let $f(\lambda)$ be the probability density function of a positive, unimodal absolutely continuous random variable. Then the nonnegative integer-valued random variable with probability function*

$$p_n = (n!)^{-1}\int_0^\infty e^{-\lambda}\lambda^n f(\lambda) d\lambda,$$

where $n \geq 0$, is a unimodal lattice variable.

This theorem helps when estimating using MLE, since it tells us that under the assumptions of the theorem the likelihood function of the model has just one maximum and so that the maximum is global. So in the case of using numerical methods to maximize the likelihood we can be sure we are approaching the global maximum.

2.1.2 Negative binomial

We say that random variable X follows negative binomial distribution with parameters p and r, if the PMF is

$$\mathsf{P}[X = k] = \binom{x + r - 1}{k} p^k (1 - p)^r.$$

Negative binomial as Poisson mixture

Negative binomial distribution can be obtained as Poisson mixture distribution, if we take Gamma distribution as the mixing distribution. Random variable Y follows Gamma distribution with parameters α and β ($Y \sim Gam(\alpha, \beta)$) if it has the following density:

$$f_g(y) = \frac{y^{\alpha-1}\beta^\alpha e^{-\beta y}}{\Gamma(\alpha)},$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x$, and if $\alpha \in \mathbb{N}$ then $\Gamma(\alpha) = (\alpha - 1)!$.

When we use $Gam(\alpha,\beta)$ for the mixture distribution we get from 2.1 for $X \sim MP(f_g)$

$$\begin{aligned}
\mathsf{P}[X=x] &= \int_0^\infty e^{-\lambda}\frac{\lambda^x}{x!}\frac{\lambda^{\alpha-1}\beta^\alpha e^{-\beta\lambda}}{\Gamma(\alpha)}d\lambda = \\
&= \frac{\beta^\alpha}{x!\Gamma(\alpha)}\int_0^\infty e^{-\lambda(1+\beta)}\lambda^{x+\alpha-1}d\lambda = \\
&= \frac{\beta^\alpha}{x!(\alpha-1)!}(x+\alpha-1)!\frac{1}{(1+\beta)^{x+\alpha}} = \\
&= \binom{x+\alpha-1}{x}\left(\frac{1}{1+\beta}\right)^x\left(\frac{\beta}{1+\beta}\right)^\alpha,
\end{aligned}$$

if α is an integer. The result is a Negative binomial distribution with parameters α and $\frac{1}{1+\beta}$.

Since negative binomial distribution has two parameters, it gives us more control over the shape of the distribution than Poisson distribution, while having closed form. It has

$$\mathsf{E}[X] = \frac{\alpha}{\beta} \qquad \mathrm{var}[X] = \frac{(1+\beta)\alpha}{\beta^2}.$$

But this parametrization of negative binomial distribution does not allow us to easily get back to Poisson distribution. For we can use better parametrization, commonly referred to as NB2. In Cameron and Trivedi [2013][pp. 117–119] this parametrization is derived from Poisson when defining $\mathsf{E}[X] = \lambda\omega$, where ω is gamma distributed with $\mathsf{E}[\omega] = 1$. The resulting probability mass function is as follows:

$$\mathsf{P}[Y=y] = \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}}\left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^y,$$

This parametrization is also more general than the standard parametrization since it allows α to be real. This parametrization has the mean and variance

$$\mathsf{E}[Y] = \lambda, \qquad \mathrm{var}[Y] = \lambda(1+\alpha\lambda).$$

As mentioned before, we have the property that Poisson distribution is the limiting case, meaning

$$\lim_{\alpha\to 0_+}\mathsf{P}[Y=y] = \mathsf{P}[X=y],$$

where X is Poisson distributed with mean λ.

Negative binomial regression

From now on we will be working with the latter parametrization NB2. Now we will introduce regressors. We will assume the same mean function as in Poisson regression:

$$\lambda(\boldsymbol{y_i}) = e^{\boldsymbol{y_i}^\top\boldsymbol{\beta}},$$

where $\boldsymbol{y_i}$ is the vector of regressors and the parameter α not depending on the regressors. From theorem 3 we have that when we maximize the likelihood, we get the global maximum. As proven in Cameron and Trivedi [2013][p. 81] the log-likelihood function is

$$l(\alpha,\lambda)=\sum_{i=1}^{n}\left[\left(\sum_{j=0}^{x_i-1}\ln(j+\alpha^{-1})\right)-\ln(x_i!)-(x_i\alpha^{-1})\ln(1+\alpha e^{\boldsymbol{y_i}^\top\boldsymbol{\beta}})+\right.$$
$$\left.+\,x_i\ln(\alpha)+x_i\boldsymbol{y_i}^\top\boldsymbol{\beta}\right].$$

From which the following maximum likelihood equations can be derived

$$\sum_{i=1}^{n}\frac{x_i-\lambda(\boldsymbol{y_i})}{1+\alpha\lambda(\boldsymbol{y_i})}\boldsymbol{y_i}=\mathbf{0},$$

$$\sum_{i=1}^{n}\left[\frac{1}{\alpha^2}\left(\ln(1+\alpha\lambda(\boldsymbol{y_i}))-\sum_{j=0}^{x_i-1}\frac{1}{j+\alpha^{-1}}\right)+\frac{x_i-\lambda(\boldsymbol{y_i})}{\alpha(1+\alpha\lambda(\boldsymbol{y_i}))}\right]=0.$$

Again, under mild conditions on the vector of regressors we have consistency and asymptotic normality.

2.1.3 Other continuous mixtures

Among other Poisson continuous mixtures are the Poisson-Inverse Gaussian and the Poisson-Lognormal, which do not have closed forms and will not be discussed in this thesis. Further information about these models can be found in Denuit et al. [2007].

2.2 Discrete mixtures

A special case of mixture distributions are the finite mixtures. Random variable X has a discrete (finite) mixture distribution if its probability mass function can be written as

$$\mathsf{P}[X=x]=p(x\mid\boldsymbol{q},\boldsymbol{\theta})=\sum_{i=1}^{n}q_ip_i(x\mid\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ is a parameter of PMF p_i and $\boldsymbol{q}$ is a vector of mixing weights as defined in Denuit et al. [2007][pg. 23]. This allows us to have different p_i come from different distributions.

2.2.1 Zero-inflated models

First of these discrete models we will discuss are the zero-inflated distributions. If we encounter excess of zeroes in our data, this mixed distribution is one of the ways to deal with it. Random variable X has a zero-inflated distribution if

$$\mathsf{P}[X=k]=\begin{cases}p+(1-p)\,\mathsf{P}[Y=0], & k=0\\ p\,\mathsf{P}[Y=k], & k\in\mathbb{N}\end{cases}$$

where Y is a count random variable and $p\in(0,1)$. That is, X is zero with probability p and it is distributed according to the distribution of Y with probability $1-p$.

Zero-inflated Poisson distribution (ZIP)

One of the possibilities for Y is $Y \sim Po(\lambda)$. That is

$$\mathsf{P}[X = k] = \begin{cases} p + (1-p)e^{\lambda}, & k = 0 \\ pe^{-\lambda}\frac{\lambda^k}{k!}, & k \in \mathbb{N}. \end{cases}$$

Then X has the following properties:

$$\begin{aligned} &\mathsf{E}\left[X\right] = 0 + \sum_{i=1}^{\infty}(1-p)i\,\mathsf{P}[Y = i] = (1-p)\,\mathsf{E}\left[Y\right] = (1-p)\lambda, \\ &\mathsf{E}\left[X^2\right] = (1-p)E[Y^2] = (1-p)(\lambda + \lambda^2), \\ &\mathrm{var}[X] = \mathsf{E}\left[X^2\right] - (\mathsf{E}\left[X\right])^2 = (1-p)\lambda + (1-p)\lambda^2. \end{aligned}$$

We used the property that expected value does not rely on $\mathsf{P}[X = 0]$. From this we can see that ZIP is overdispersed.

MLE for ZIP

Now to estimate the parameters (p, λ), we will be using the maximum likelihood estimate. We will note $Z = \sum_{i=1}^{n} I\{X_i = 0\}$ the number of zeroes among our observations, where I is the indicator function, that is $I\{x_i = 0\} = 1$, if $x_i = 0$ and $I\{x_i = 0\} = 0$ otherwise. The maximum likelihood function is

$$L(p, \lambda, \boldsymbol{x}) = (p + (1-p)\,\mathsf{P}[Y = 0])^Z \prod_{j=1}^{n}((1-p)\,\mathsf{P}[Y = x_i])^{1-(I\{x_i=0\})}.$$

And as shown in Johnson et al. [2005][p. 353], the resulting maximum likelihood equations are

$$\hat{p} + (1-\hat{p})e^{\hat{\lambda}} = \frac{Z}{n} \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\lambda}(1-\hat{p}).$$

Other zero-inflated models

It is possible to take any discrete distribution and make a zero inflated version. Another popular and widely used model is a Negative binomial zero-inflated model.

2.2.2 Hurdle models

Hurdle models are another type of two component finite mixture. The idea behind them is that there is a hurdle which has to be overcome. This seems plausible in our task of modelling number of claim counts, since a policyholder may not want to report a claim when the damage is low, but after the first claim this behaviour may change. In the simple case of zero-truncated distributions, it means that if the hurdle is not passed, we will observe zero, if it is passed, we observe a variable distributed according to zero-truncated distribution. That is, if Y has PMF $\mathsf{P}[Y = k]$, then zero-truncated distribution will look like

$$\mathsf{P}[Y = k \mid X > 0] = \frac{\mathsf{P}[Y = k \wedge Y > 0]}{\mathsf{P}[Y > 0]} = \frac{\mathsf{P}[Y = k]}{1 - \mathsf{P}[Y = 0]} \qquad \text{for } k > 0.$$

So hurdle model has distribution

$$\mathsf{P}[X=k]\begin{cases} p_0, & k=0 \\ \frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{P}[Y=k], & k\in\mathbb{N}. \end{cases}$$

With the expected value and variance

$$\begin{aligned} \mathsf{E}\,[X] &= \sum_{k=0}^{\infty}\mathsf{P}[X=k]k = \sum_{k=1}^{\infty}\frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{P}[Y=k]k = \frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{E}\,[Y], \\ \mathsf{E}\,[X^2] &= \sum_{k=0}^{\infty}k^2\,\mathsf{P}[X=k] = \sum_{k=1}^{\infty}\frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{P}[Y=k]k^2 = \frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{E}\,[Y^2], \\ \operatorname{var}[X] &= \mathsf{E}\,[X^2]-(\mathsf{E}\,[X])^2 = \frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{E}\,[Y^2]-\left(\frac{1-p_0}{1-\mathsf{P}[Y=0]}\right)^2(\mathsf{E}\,[Y])^2. \end{aligned}$$

Then with the use of indicator function $I_{\{x_i=0\}}$ we can express the probability mass function of iid vector with hurdle distribution as

$$\mathsf{P}[X=k]=\prod_{i=1}^{n}(p_0)^{I_{\{x_i=0\}}}\left(\frac{1-p_0}{1-\mathsf{P}[Y=0]}\,\mathsf{P}[Y=k]\right)^{1-I_{\{x_i=0\}}},$$

which is also the likelihood function. For Y we can use the Poisson, Negative binomial or any other count distribution.

Poisson hurdle distribution

Now we will assume $Y\sim Po(\lambda)$.

The log-likelihood is

$$\begin{aligned} l(p_0,\lambda,\boldsymbol{x}) &= \sum_{i=1}^{n}\Bigg[I_{\{x_i=0\}}\ln(p_0)+ \\ &\quad +(1-I_{\{x_i=0\}})\left((\ln(1-p_0)-\ln(1-e^{-\lambda})+\ln\left(e^{-\lambda}\frac{\lambda^k}{k!}\right)\right)\Bigg]= \\ &= \sum_{i=1}^{n}I_{x_i=0}\ln(p_0)+(1-I_{x_i=0})(\ln(1-p_0))+ \\ &\quad +\sum_{i=1}^{n}(1-I_{X_i=0})(-\ln(1-e^{-\lambda})-\lambda+k\ln(\lambda)-\ln(k!), \end{aligned} \tag{2.2}$$

where the first part is a function of p_0 and does not depend on λ and the second part is a function of λ and does not depend on p_0. In this case, we can maximize those two parts independently.

In the simple case without taking regressors into account, the hurdle model as specified here and the zero inflated models are just reparametrizations of the same distribution. The zero-inflated model is a hurdle model with $p_0=p+(1-p)e-\lambda$. But when dealing with regression, the models differ.

Hurdle Poisson regression

When dealing with hurdle model and regressors we have to specify two link functions, since we are predicting two models. One of them is the model controlling the phase of passing the hurdle, thus controlling the number of zeroes, and the second one is the Poisson truncated distribution. For the Poisson part of the model we will use the log-link as before in Section 1.4. For p_0 it is common to use the logit link function, i.e.

$$\lambda(\boldsymbol{y_i}, \boldsymbol{\beta}) = e^{\boldsymbol{y_i}^\top \boldsymbol{\beta}},$$

$$p_0(\boldsymbol{y_i}, \boldsymbol{\gamma}) = \frac{e^{\boldsymbol{y_i}^\top \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{y_i}^\top \boldsymbol{\gamma}}}.$$

After putting this into log-likelihood 2.2, we can numerically maximize the log-likelihood and find the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$.

2.2.3 Other discrete mixtures

In other discrete mixtures we usually reflect models of the following form:

$$\mathsf{P}[X = x] = \sum_{i=1}^{k} q_k Y_i,$$

where $Y_i \sim Po(\lambda_i)$. In this case, when k is known, we can use the expectation-maximization (EM) algorithm to estimate the coefficients as in Wang et al. [1996].

2.3 Choosing a model and tests

The next difficult task is choosing the right model for our data. One thing is that the model should have an explanation for the data we are using, and the second thing is that the model should fit the data well.

2.3.1 Choosing a model

Tests

For nested models (where one model is a special case of another) we can use likelihood ratio test. Let us assume we have two models $f(\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $f(\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is fixed. If f follows the regularity conditions, then under the null hypothesis

$$\text{LR} = 2(L(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - L(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_0)) \sim^{as} \chi^2(k - l),$$

where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are maximum likelihood estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. This can be used for testing the models against each other.

In our case, we would like to test if the data come from negative binomial distribution against the alternative that it comes from Poisson distribution. In this case we encounter a problem - for the NB to become Po, the dispersion parameter α has to be zero and thus it has to be on the boundary of the parameter space. This means that it no longer complies with the regularity conditions. In this case the test statistics is asymptotically 1/2 at zero and 1/2 at $\chi^2(k - l)$, so to test at level η we will use $\chi^2_{1-2\eta}(k-l)$. This approach was used in Cameron and Trivedi [2013][p. 90] and proven in Chernoff [1954].

Information criteria

A different way of choosing a model is using information criteria. Since with every introduced regressor the log-likelihood function is expected to be somewhat higher, we cannot rely merely on log-likelihood when comparing models. For this we can use Akaike information criterion (AIC) which penalizes the number of parameters in the model. If we are estimating parameter $\boldsymbol{\theta}$ we note $k = \dim(\boldsymbol{\theta})$. Then the AIC has the form

$$\text{AIC} = -2\left(l(\hat{\boldsymbol{\theta}}_n) - k\right),$$

where lower values are preferred. Another option is to use the Bayesian information criteria (BIC), which penalizes dimension of estimated parameter even more. It is given by

$$\text{BIC} = -2\left(l(\hat{\boldsymbol{\theta}}_n) - \ln(n)k\right)$$

and again, lower values are preferred. These information criteria are defined for example in Denuit et al. [2007][p. 43].

2.3.2 Choosing regressors

Here we will be mainly interested in a test concerning null hypothesis $H_0 : \beta_k = 0$ against $H_1 : \beta_k \neq 0$. The test statistics we are going to present first is the so-called t-statistics $T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$, where $\hat{\sigma}_{\hat{\beta}_j}$ is the standard deviation of $\hat{\beta}_j$. This test statistics is under the regularity conditions and null hypothesis asymptotically $N(0,1)$ distributed. This can be used to form confidence interval and to carry out the test. This result follows from theorem 1.

Another way is using the likelihood ratio test, as formulated in Subsection 2.3.1. Then with the same hypothesis as above, we take the test statistics

$$\text{LR} = 2(L(\boldsymbol{x}, \hat{\boldsymbol{\beta}}) - L(\boldsymbol{x}, \hat{\beta}_0, \ldots, \hat{\beta}_{k-1}, 0, \hat{\beta}_{k+1}, \ldots, \hat{\beta}_n)),$$

which, under the regularity conditions and null hypothesis, is asymptotically χ^2_1 distributed.

Another way is to use the information criteria defined in Subsection 2.3.1 and compare the values of AIC and BIC of the model with the selected regressor and without it.

3. Use on real-life data

In this chapter we will show the use of the methods introduced in the preceding chapters. For all our computations we will be using **?** and the main parts of code used are in Appendix A. The tables in this section were made with the help of package *stargazer*.

3.1 Australia car insurance

First, we will look at data from Australian car insurance from the years 2004 and 2005. The dataset *ausprivauto0405* containing $67,856$ observations is from the *CASdatasets* R package and it was first used in de Jong and Heller [2008]. The data structure is shown in the table 3.1. Most of the data columns are in the form of factors. A factor variable is such that there is a fixed number k of categories (levels) of the variable and each observation belongs to one of those categories. In order to use these factors in our models, we choose one level as base level and for the remaining levels we will introduce new variables, where each of them takes value 0 or 1, 1 if the observation belongs to the corresponding level, 0 otherwise. When all of the corresponding variables are 0, the observation belongs to the base category. As the base level we will choose the most numerous level.

We will be modelling distribution of the number of claims (which we denote Nb), which is in our data in the column ClaimNb. The mean of claim number in our dataset is 0.073, the standard deviation 0.278, so sample mean is 0.077. It may seem that the data is not too overdispersed. However, we need to consider what are we trying to estimate. Since the number of claims is for different time periods for different policyholders, we need to take this into account. Thus, if we denote exposure as ω we will be dealing with models with mean $\mu = \omega\lambda$. So to explore more our data, we will be looking at rates $r = \frac{Nb}{\omega}$, instead of claim numbers.

mean of r: 0.21   sd: 2.88   sample variance: 8.30

Here we can see that the rate is highly overdispersed.

Column name	**Column data**
Exposure	Exposure of the policyholder in years
VehValue	Value of vehicle in thousands of AUD
VehAge	A factor of 4 levels indicating age of vehicle
VehBody	A factor of 13 levels indicating the car type
DrivAge	A factor of 6 levels indicating the age of the driver
ClaimNb	Number of claims - our response variable
ClaimAmount	Sum of claim amounts

Table 3.1: Data columns in Australian insurance dataset

3.1.1 Models without regressors

First we will fit the Poisson model to our data with the function *glm()*, even though we can see the overdispersion is likely to be a problem. As MLE of λ we get

$$\hat{\lambda}_n = \frac{\sum_{i=1}^n Nb_i}{\sum_{i=1}^n \omega_i} \doteq 0.16.$$

Fitted frequencies are in the table 3.2.

Now we will fit the data negative binomial distribution with no regressors using the package *MASS* and function *glm.nb()*. As MLE we get

$$\hat{\alpha}_n \doteq 0.44, \qquad \hat{\lambda}_n \doteq 0.64.$$

As was shown, the negative binomial distribution is equal to Poisson distribution when $\alpha = 0$. Thus, we can perform likelihood ratio test (2.3.1)

$$H_0 : \alpha = 0 \quad \text{against} \quad H_1 : \alpha \neq 0$$

with the test statistics

$$\mathrm{LR} = 2(l(\hat{\alpha}, \hat{\lambda}) - l(0, \hat{\lambda})) \doteq 103.62.$$

Here we encounter the problem mentioned in 2.3.1, so we will take the value of χ^2_{p-q} in $1-\alpha/2$ instead of $1-\alpha$. The critical value of χ^2_{p-q} in $1-\alpha/2$, for $\alpha = 5\%$ is 3.841, so we can strongly reject the hypothesis that the data come from Poisson distribution.

Now we have to think about the meaning of the model - does it make sense to use the negative binomial in this case? Since we are modelling claim numbers of policyholders, it is obvious that the driving skills among people differ and therefore introducing the random term in the Poisson model is logical.

The negative binomial was one way of improving the data. Another one could be the hurdle model. As discussed in Chapter 2, we could expect that some policyholders may not fill a claim when the value is not high enough, thus the process generating zeroes could be different from the process behind claim numbers. We used the function *hurdle()* from the package *pscl*.

We will not consider zero-inflated models here, since, as already mentioned in 2.2.2, the zero-inflated model is the same as hurdle model in the case of absence of regressors. Another reason is that in our data the excess of zeroes is not particularly high and there does not seem to be an intuitive explanation for zero inflated model.

Number of accidents					
	1	**2**	**3**	**4**	**5**
Sample frequency	63232	4333	271	18	2
Poisson	63094	4591	167	4	0
Negative binomial	63235	4325	278	17	1
Poisson hurdle	63232	4313	294	15	1

Table 3.2: Frequencies of fitted models

Now in the table 3.2 there are the expected frequencies of our models as well as the actual sample frequency. It seems that both the negative binomial and Poisson hurdle models fit our data well. To choose between them we can look at the AIC or BIC as suggested in Subsection 2.3.2.

$$\text{AIC(Po)} \doteq 34943.67 \qquad \text{AIC(NB)} \doteq 34899.59 \qquad \text{AIC(HPo)} \doteq 34891.31$$

$$\text{BIC(Po)} \doteq 34952.80 \qquad \text{BIC(NB)} \doteq 34917.84 \qquad \text{BIC(HPo)} \doteq 34909.56$$

The lowest value for both the AIC and BIC is for the hurdle Poisson model so it seems as the most reasonable choice.

3.1.2 Models with regressors

So far, we have been looking at the response variable and we have chosen the Poisson hurdle model, but we have to consider the purpose of our model. Simple model with no regressors allow us just to predict the numbers of claim frequencies, but it does not allow us to differentiate between the policyholders according to their risk of claims, which is the main reason of modelling in insurance, as to set the premium correctly and fairly for everyone. For this, we will introduce regressors to our models.

Poisson regression

First we will consider Poisson regression, since the regressors can explain some of the overdispersion, which we observed in the last section. For the regression we will use the robust standard error as in Section 1.4 by using the package *sandwich*. The estimated coefficients are in table B.1. From this summary, it seems that gender may not be significant with p-value of 0.52. We will perform the likelihood ratio test for this parameter as introduced in Subsection 2.3.2. We denote β_k the parameter corresponding to gender.

$$H_0 : \beta_k = 0, \text{ against } H_1 : \beta_k \neq 0$$

$$\text{LR}_{Gen} = 2(L(\hat{\boldsymbol{\beta}}) - L(\hat{\beta}_0, \ldots, \hat{\beta}_{k-1}, 0, \hat{\beta}_{k+1}, \ldots, \hat{\beta}_n))$$

$$\text{LR} \doteq 0.58$$

The critical value at the level of 5% is $c \doteq 3.84$, and since $\text{LR} < c$ we cannot decline the hypothesis that gender is not significant. Together with higher AIC in model without gender, we will remove this parameter from our model. For standard likelihood ratio test we are using function *lrtest()* from package *lmtest*.

Next parameter of interest will be the vehicle amount, since it has a high p-value as well. We will perform the likelihood test again, with the resulting value of test statistics $\text{LR}_{VA} \doteq 1.74$ so once more, we cannot decline the hypothesis that vehicle amount is not significant. The AIC of model without vehicle amount is also lower, so we will use the model without it.

Now, all other variables are at least partly significant, but since they are factors, not all levels are necessarily also significant. We will continue as before with the only difference that we will not test models with and without some parameter, but with two levels of a factor merged. As before, first we look at the

p-value of the first test given in Subsection 2.3.1 and afterwards we will try the LR test. The highest p-value of 0.79 has the level minibus of factor vehicle body. If we make model with the levels "Sedan" and "Minibus" merged and run the likelihood test, we find that we cannot reject the hypothesis that the level is insignificant. Also when comparing AIC we find that the new model with less levels has lower AIC and this leads us to drop this level out of our model. We continue by performing the same actions for the levels "Truck", "Convertible", "Hardtop", "Panel van", "Roadster" and "Station wagon".After that we also merge the levels "working people" with "older work. people" from the variable "drivAge", because of insignificance of "working people". After this procedure all the remaining coefficients are significant and we reduced the number of parameters by 8, while the log-likelihood was lowered just by 2.

The coefficients of the resulting model are in the table 3.3. We will look at the interpretation of the coefficients together with the negative binomial model, since the mean function remains the same and the coefficients are similar and thus the effects are similar.

The resulting fitted frequencies from Poisson regression are in the table 3.4. From this table, it does not seem like the regressors managed to resolve the problems of Poisson equidispersion condition - we still see too few zeroes, too many ones, and too few higher values. This could indicate the presence of mixed distribution, since mixed Poisson distributions have the property from Shaked's two crossing theorem 2.

		Poisson		Negative binomial	
		coef	Robust SE	coef	SE
	(Intercept)	-1.847	0.032	-1.845	0.032
VehAge	Oldest cars	-0.077	0.040	-0.075	0.039
	Young cars	0.128	0.039	0.128	0.039
	Youngest cars	0.086	0.039	0.084	0.044
VehBody	Bus	0.912	0.304	0.906	0.335
	Coupe	0.411	0.121	0.413	0.121
	Hatchback	-0.079	0.033	-0.077	0.034
	Motorized caravan	0.560	0.256	0.561	0.268
	Utility	-0.208	0.064	-0.209	0.064
DrivAge	Old people	-0.236	0.045	-0.236	0.045
	Oldest people	-0.226	0.056	-0.228	0.056
	Young people	0.077	0.039	0.076	0.039
	Youngest people	0.244	0.049	0.247	0.050
	θ	—	—	2.252	0.414
Observations		67, 856		67, 856	
Log Likelihood		$-17,392.43$		$-17,372.72$	
Akaike Inf. Crit.		$34,810.85$		$34,773.44$	

Table 3.3: Resulting coefficients of Poisson and negative binomial models

Negative binomial regression

Next we will fit the negative binomial regression to our data. The coefficients of the model with all regressors are in Table B.2. To choose the parameters to include in our model we proceed in the same way as in the case of Poisson distribution and we get the same results, so we end up with the same coefficients. The coefficients of the resulting model are in Table 3.3. Now we will run a test of Poisson regression model against negative binomial one. When we denote the dispersion parameter of NB α, we formulate the test as follows:

$$H_0 : \alpha = 0 \text{ against } H_1 : \alpha \neq 0$$

$$\text{LR} = 2(L(\hat{\alpha}, \hat{\boldsymbol{\beta}}) - L(0, \hat{\boldsymbol{\beta}}))$$

$$\text{LR} \doteq 39.41$$

which is once more asymptotically a distribution with one half in zero and one half chi-squared with one degree of freedom. Thus for testing at level η we will take $\chi^2_{1-2\eta}(1)$ instead of $\chi^2_{1-\eta}$. For $\eta = 5\%$ it means the critical value is $\chi^2_{0.9} \doteq 2.71$, which is highly exceeded by the value of our test statistics and so we can reject the hypothesis that our data is distributed according to Poisson regression model.

Next we will look at the coefficients and their meaning. All the coefficients in the negative binomial model are quite similar to those of Poisson model and the link function for mean is the same so we will analyse them at the same time. The first thing to notice is that all our remaining regressors are factors. This means, that if we have factor F with k levels and corresponding parameters $\beta_p, \beta_{p+1}, \dots \beta_{p+k-2}$ (there is one less parameter than there are levels for the base level, see 3.1), the factor having a value of i means that the predicted mean will be e^{β_i} times higher or lower than that of the base level with other regressors remaining the same. When the parameter is smaller than zero, the policyholder with corresponding factor level has smaller predicted mean than that with base level and when it is higher than zero they have higher predicted mean.

From the coefficients we can see that the expected mean of accidents of oldest cars is 0.92 times that of old cars, while young cars have the expected mean 1.14 times that of old cars, so young cars are expected to have more claims than the old ones, which may be surprising. Another trend can be seen in the age of policyholder, where it seems that younger people (where youngest people have 1.3 times the mean of the base level) have much higher mean of claim numbers than older people (old people have 0.8 times the mean of base level). With the vehicle types, not many levels are left as significant and even with those one has to be careful when making inference. For example in our dataset buses are

Table 3.4: Fitted frequencies

	0	1	2	3	4
Sample frequency	63,232	4,333	271	18	2
Poisson regression	63,164	4,456	226	9	0
NB regression	63,253	4,284	297	21	2
Hurdle Poisson regression	63,232	4,316	292	16	1

expected to have 2.5 times higher mean than base level, but before making any broader inference we have to note that we have only 40 observations of buses in our dataset.

Poisson hurdle regression

As in the models with no regressors another way of improving the Poisson regression is by considering Poisson hurdle regression. In the case of regression, we will use the log link for the Poisson part of the distribution and logit link for the binomial part, which handles the zero counts as stated in 2.2.2. The coefficients of the model with all regressors are shown in the tables B.3 and B.4. We perform the same process of regressor selection as with Poisson distribution and negative binomial, but in this case we have two parts of our model and we select the regressors for each model separately. We present the resulting coefficients in the table 3.5. It is interesting to note that in the Poisson truncated count part there are just two regressors remaining, both of which are levels of vehicle body factor. This is not very surprising, since when we look at our dataset sample frequencies (3.4), we notice that the number of policyholders with two or more accidents is very low compared to those with zero or one accident. This suggests that when we may not have enough data to estimate the count probability according to regressors.

		coef	SE	P(>|z|)
Poisson truncated count model				
(Intercept)		−1.465	0.063	0
VehBody	Hatchback	−0.297	0.137	0.030
	Roadster	1.568	0.945	0.097
Zero logit model				
	(Intercept)	−1.859	0.035	0
VehAge	Proboldest cars	−0.095	0.042	0.024
	Probyoung cars	0.123	0.041	0.003
	Probyoungest cars	0.082	0.046	0.078
	ProbBus	1.083	0.380	0.004
	ProbCoupe	0.466	0.131	0.0004
	ProbHardtop	0.177	0.095	0.063
	ProbMotorized caravan	0.679	0.291	0.020
	ProbStation wagon	0.080	0.037	0.029
	ProbUtility	−0.170	0.068	0.012
DrivAge	old people	−0.235	0.048	0.000
	oldest people	−0.234	0.059	0.0001
	young people	0.076	0.041	0.064
	youngest people	0.279	0.053	0.000

Table 3.5: Final Poisson hurdle model

Also in the case of Poisson hurdle model the interpretation of the coefficients is difficult, since the model consists of two parts. Because of this one regressor has two roles - one in predicting the zero count and another one in predicting the expected mean. Also because the mean has the form

$$\mathsf{E}\left[X_i\right] = \frac{1 - \frac{e^{\boldsymbol{y}_i^\top \gamma}}{1+e^{\boldsymbol{y}_i^\top \gamma}}}{1 - e^{-e^{\boldsymbol{y}_i \boldsymbol{\beta}}}} e^{\boldsymbol{y}_i \boldsymbol{\beta}}$$

there is no simple way of telling what do the regressors do.

Finally we will compare our regression models according to log-likelihood, AIC and BIC.

$$\begin{array}{ccc} l(Po) \doteq -17392.43 & l(NB) \doteq -17372.72 & l(PoH) \doteq -17362.69 \\ \mathrm{AIC}(Po) \doteq 34810.85 & \mathrm{AIC}(NB) \doteq 34773.44 & \mathrm{AIC}(PoH) \doteq 34759.38 \\ \mathrm{BIC}(Po) \doteq 34929.48 & \mathrm{BIC}(NB) \doteq 34901.2 & \mathrm{BIC}(PoH) \doteq 34914.51 \end{array}$$

Now we have to choose the right model for our data. In this case it is not clear which model is the best - according to AIC we should choose Poisson hurdle model and according to BIC the negative binomial model. Another question is the purpose of the model. If interpretability of effects of individual regressors is important to us, negative binomial may be a better choice.

Conclusion

Mixed Poisson distributions are widely used in many studies because of their numerous useful features and intuitive interpretability. Because of its wide spread the theory behind mixed Poisson models is well developed. We summarized the theory of the simpler models in the first two chapters, but since the topic is quite broad, we omitted many mixed Poisson models, which all have their use in many situations, like truncated models and more general finite mixtures, as well as continuous mixtures with other mixing distribution than gamma distribution.

In the third chapter we used the introduced theory on real-life data from Australian motor insurance. We compared the performance of the mixed Poisson models and tried to get the best fit possible. During this process we were dealing with choosing the right regressors and omitting the non-significant ones by using testing techniques introduced before as well as comparing the final Poisson regression model with the negative binomial regression model and the Poisson hurdle model. We did not get to one best model, since the negative binomial one and the Poisson hurdle were quite close in the concerns of fitting the data with the hurdle model slightly better, but with the negative binomial the simpler model with easier interpretation.

A. R code

Here we will present the important parts of code used to generate the models. If we used the same code structure many times we have included just first one or two steps. All the code is written in **?**.

The packages used are

```
library(sandwich)
library(MASS)
library(pscl)
library(CASdatasets)
library(stargazer)
```

A.1 Models without regressors

```
#Poisson model with no regressors
AusPoNoReg <- glm(ClaimNb˜1+ offset(log(Exposure)),
        family="poisson",data=aus)
AusPoNoRegPred <- predict(AusPoNoReg,type = "response")
AusPoNoRegFreq <- 1:6
for(number in AusPoNoRegFreq){
        AusPoNoRegFreq[number]=
                sum(dpois(x=number-1,AusPoNoRegPred))
}

#Negative binomial model with no regressors
AusNBNoReg<-glm.nb(ClaimNb˜1+ offset(log(Exposure)),data=aus)
AusNBNoRegPred <- predict(AusNBNoReg,type="response")
AusNBNoRegFreq<-1:6
for(number in AusNBNoRegFreq){
        AusNBNoRegFreq[number]=sum(dnbinom(x=number-1,
                size=AusNBNoReg$theta,mu= AusNBNoRegPred))
}
AusNBNoRegFreq

#Poisson hurdle model with no regressors
AusHPNoReg <- hurdle(ClaimNb˜1 +
        offset(log(Exposure)),data=aus)
AusHPNoRegProb<-predict(AusHPNoReg, type="prob", at=0:5)
AusHPNoRegFreq<-colSums(AusHPNoRegProb)


#Likelihood ratio test of NB against Poisson on level 5%
testStat1=2*(logLik(AusNBNoReg)-logLik(AusNoReg))
critValue = qchisq(0.9,df=1)
```

A.2 Models with regressors

A.2.1 Poisson regression model

```
#Model with all the regressors
summary(PoMod <- glm(ClaimNb~VehAge+VehBody+VehValue+DrivAge + Gender
         +offset(log(Exposure)),family="poisson",data=aus))
#Removing some regressors
summary(PoMod1 <- glm(ClaimNb~VehAge+VehBody + VehValue+ DrivAge
        + offset(log(Exposure)),family="poisson",data=aus))
lrtest(PoMod,PoMod1)

summary(PoMod2 <- glm(ClaimNb~VehAge+VehBody + DrivAge
        + offset(log(Exposure)),family="poisson",data=aus))
lrtest(PoMod1,PoMod2)



aus1<- aus

#Now follow two steps of removing levels from a factor
levels(aus1$VehBody)<-c("Sedan","Bus","Convertible","Coupe",
        "Hardtop","Hatchback","Sedan","Motorized caravan","Panel van",
        "Roadster","Station wagon","Truck","Utility")
summary(PoMod2.2 <-glm(ClaimNb~VehAge+VehBody + DrivAge+
        offset(log(Exposure)),family="poisson",data=aus1))
lrtest(PoMod2,PoMod2.2)
aus1<- aus

levels(aus1$VehBody)<-c("Sedan","Bus","Convertible","Coupe",
        "Hardtop","Hatchback","Sedan","Motorized caravan","Panel van",
        "Roadster","Station wagon","Sedan","Utility")
summary(PoMod2.3 <-glm(ClaimNb~VehAge+VehBody + DrivAge
        + offset(log(Exposure)),family="poisson",data=aus1))
lrtest(PoMod2.3,PoMod2.2)

#Getting the robust SE for poisson model PoFinal
robustVarianceMatrix = vcovHC(PoFinal,type="HC0")
robustStdErr = sqrt(diag(robustVarianceMatrix))
summaryPoFinal = cbind(Estimate = coef(PoFinal),
        ‘Robust SE‘ = robustStdErr, ‘Pr(>|z|)‘ = 2 *
pnorm(abs(coef(PoFinal)/robustStdErr), lower.tail = FALSE),
        LL = coef(PoFinal) - 1.96 *
robustStdErr, UL = coef(PoFinal) + 1.96 * robustStdErr)
```

A.2.2 Negative binomial regression model

```
#Choosing regressors to include
summary(NBMod <- glm.nb(ClaimNb~VehAge+VehValue+VehBody+Gender+DrivAge
        + offset(log(Exposure)),data=aus))
summary(NBMod1 <- glm.nb(ClaimNb~VehAge+VehValue+VehBody+DrivAge
        + offset(log(Exposure)),data=aus))
summary(NBMod2 <- glm.nb(ClaimNb~VehAge+VehBody+DrivAge
        + offset(log(Exposure)),data=aus))
lrtest(NBMod,NBMod1)
lrtest(NBMod1,NBMod2)

#Two steps of choosing level to include
aus1<- aus
levels(aus1$VehBody)<-c("Sedan","Bus","Convertible","Coupe",
        "Hardtop","Hatchback","Sedan","Motorized caravan","Panel van",
        "Roadster","Station wagon","Truck","Utility")
summary(NBMod2.1 <- glm.nb(ClaimNb~VehAge+VehBody+DrivAge+
        offset(log(Exposure)),data=aus1))
lrtest(NBMod2.1,NBMod2)

aus1<- aus
levels(aus1$VehBody)<-c("Sedan","Bus","Convertible","Coupe",
        "Hardtop","Hatchback","Sedan","Motorized caravan","Panel van",
        "Roadster","Station wagon","Sedan","Utility")
summary(NBMod2.2 <- glm.nb(ClaimNb~VehAge+VehBody+DrivAge+
        offset(log(Exposure)),data=aus1))
lrtest(NBMod2.2,NBMod2.1)
```

```
# Test for Poisson against Negative binomial final models
LRFinal = 2 * (logLik(NBFinal) - logLik(PoFinal))
critValue = qchisq(0.9,df=1)
```

A.2.3 Hurdle Poisson regression model

```
# First removing regressors unsignificant for both parts of our model
hurdleModelPo <- hurdle(ClaimNb~VehAge+VehValue+VehBody
        +Gender+DrivAge + offset(log(Exposure)) ,data=aus,link =
        "logit",dist = "poisson")
hurdleModelPo1 <- hurdle(ClaimNb~VehAge+VehValue+VehBody
        +DrivAge + offset(log(Exposure)) ,data=aus,link =
        "logit",dist="poisson")
hurdleModelPo2 <- hurdle(ClaimNb~VehAge+VehBody+DrivAge
        + offset(log(Exposure)) ,data=aus,link =
        "logit",dist="poisson")
lrtest(hurdleModelPo,hurdleModelPo1)
lrtest(hurdleModelPo1,hurdleModelPo2)


# And first step of removing levels from factors
aus2<-aus2BK
levels(aus2$VehBody)<-c("Sedan","Bus","Sedan","Coupe",
        "Hardtop","Hatchback","Minibus","Motorized caravan",
        "Panel van","Roadster","Station wagon","Truck","Utility")
levels(aus2$VehBodyProb)<-c("Sedan","Bus","Convertible",
        "Coupe","Hardtop","Hatchback","Minibus","Motorized caravan",
        "Panel van","Sedan","Station wagon","Truck","Utility")
summary(hurdleModelPo2.1 <- hurdle(ClaimNb~VehAge+VehBody+DrivAge
        + offset(log(Exposure))|VehAgeProb+VehBodyProb+DrivAgeProb +
        offset(log(Exposure)) ,data=aus2,link = "logit",dist="poisson"))
lrtest(hurdleModelPo2.1,hurdleModelPo2)
```

B. Tables

		Coefficient	Robust SE	P(>|z|)
(Intercept)		−1.903	0.048	0
VehAge	oldest cars	−0.059	0.042	0.161
	young cars	0.111	0.040	0.006
	youngest cars	0.056	0.049	0.254
VehBody	Bus	0.924	0.305	0.002
	Convertible	−0.746	0.590	0.206
	Coupe	0.413	0.124	0.001
	Hardtop	0.090	0.090	0.316
	Hatchback	−0.051	0.039	0.187
	Minibus	−0.060	0.152	0.692
	Motorized caravan	0.536	0.260	0.040
	Panel van	0.071	0.130	0.584
	Roadster	0.359	0.701	0.609
	Station wagon	0.012	0.044	0.782
	Truck	−0.039	0.097	0.685
	Utility	−0.196	0.069	0.004
VehValue		0.024	0.016	0.134
DrivAge	old people	−0.219	0.050	0.000
	oldest people	−0.203	0.061	0.001
	working people	0.028	0.042	0.500
	young people	0.088	0.044	0.044
	youngest people	0.259	0.053	0.000
Gender	Male	−0.023	0.031	0.457
Observations			67, 856	
Log Likelihood			−17, 388	
Akaike Inf. Crit.			34, 824	

Table B.1: Poisson regression with all regressors

		Coefficient	SE	P(>|z|)
(Intercept)		−1.903	0.049	0
VehAge	oldest cars	−0.057	0.042	0.174
	young cars	0.111	0.041	0.006
	youngest cars	0.052	0.049	0.288
VehValue		0.025	0.018	0.154
VehBody	Bus	0.918	0.336	0.006
	Convertible	−0.750	0.596	0.209
	Coupe	0.414	0.123	0.001
	Hardtop	0.088	0.093	0.343
	Hatchback	−0.050	0.039	0.201
	Minibus	−0.065	0.156	0.678
	Motorized caravan	0.536	0.271	0.048
	Panel van	0.067	0.128	0.602
	Roadster	0.344	0.603	0.568
	Station wagon	0.010	0.045	0.814
	Truck	−0.043	0.095	0.651
	Utility	−0.198	0.068	0.004
Gender	Male	−0.023	0.031	0.460
DrivAge	old people	−0.220	0.050	0.000
	oldest people	−0.205	0.060	0.001
	working people	0.029	0.042	0.495
	young people	0.088	0.044	0.047
	youngest people	0.262	0.054	0.000
	θ	2.252	0.414	
Observations			67,856	
Log Likelihood			−17,369	
Akaike Inf. Crit.			34,786	

Table B.2: Negative binomial with all regressors

		Coefficient	SE	P(>|z|)
(Intercept)		−1.351	0.190	0
VehAge	oldest cars	0.188	0.164	0.253
	young cars	0.205	0.156	0.188
	youngest cars	0.031	0.197	0.874
VehValue		0.009	0.074	0.905
VehBody	Bus	0.159	0.988	0.872
	Convertible	−14.700	3,228.095	0.996
	Coupe	0.286	0.386	0.459
	Hardtop	−0.645	0.419	0.124
	Hatchback	−0.449	0.156	0.004
	Minibus	−0.682	0.714	0.339
	Motorized caravan	−0.060	1.001	0.952
	Panel van	0.020	0.415	0.962
	Roadster	1.509	0.970	0.120
	Station wagon	−0.279	0.172	0.105
	Truck	−0.058	0.333	0.861
	Utility	−0.289	0.270	0.285
Gender	Male	−0.090	0.118	0.448
DrivAge	old people	−0.312	0.203	0.124
	oldest people	−0.127	0.227	0.576
	working people	−0.047	0.158	0.766
	young people	0.010	0.164	0.950
	youngest people	−0.098	0.216	0.650

Table B.3: Poisson hurdle model with all regressors - count part

		Coefficient	SE	P(>|z|)
(Intercept)		−1.893	0.052	0
VehAge	oldest cars	−0.074	0.044	0.098
	young cars	0.107	0.043	0.013
	youngest cars	0.052	0.052	0.316
VehValue		0.027	0.019	0.151
VehBody	Bus	1.060	0.381	0.005
	Convertible	−0.733	0.610	0.230
	Coupe	0.430	0.134	0.001
	Hardtop	0.143	0.099	0.147
	Hatchback	−0.023	0.042	0.576
	Minibus	−0.030	0.163	0.852
	Motorized caravan	0.618	0.294	0.036
	Panel van	0.075	0.138	0.589
	Roadster	0.030	0.749	0.968
	Station wagon	0.032	0.047	0.498
	Truck	−0.044	0.102	0.667
	Utility	−0.196	0.072	0.007
Gender	Male	−0.018	0.033	0.581
DrivAge	old people	−0.218	0.053	0.000
	oldest people	−0.214	0.063	0.001
	working people	0.036	0.045	0.422
	young people	0.096	0.047	0.042
	youngest people	0.300	0.058	0.000
Observations			67, 856	
Log Likelihood			−17, 353	
Akaike Inf. Crit.			34, 797	

Table B.4: Poisson hurdle model with all regressors - zero part

Bibliography

A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data.* Cambridge University Press, 2013.

George Casella and Roger L. Berger. *Statistical Inference.* Wadsworth Group, 2002.

Herman Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, pages 573–578, 1954.

Piet de Jong and Gillian Z. Heller. *Generalized Linear Models for Insurance Data.* Cambridge University Press, 2008.

Michel Denuit, Xavier Maréchal, Sandra Pitrebois, and Jean-François Walhin. *Actuarial Modelling of Claim Counts.* John Wiley & Sons Ltd, 2007.

Ludwig Fahrmeir and Heinz Kaufmann. Asymptotic inference in discrete response models. *Statistische Hefte*, 27:179–205, 1986.

P. Holgate. The modality of some compound poisson distributions. *Biometrika*, 57(3):666–667, 1970.

Norman L. Johnson, Adrienne W. Kemp, and Samuel Kotz. *Univariate discrete distributions.* John Wiley & Sons, Inc., 2005.

Moshe Shaked. On mixtures from exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):192–198, 1980.

Peiming Wang, Martin L. Puterman, Iain Cockburn, and Nhu Le. Mixed poisson regression models with covariate dependent rates. *Biometrics*, 52(2):381–400, 1996.

List of Tables