# Pricing of Non-Life Insurance Products

*Author:*

Joeri Ronald Frank Deckers

*Thesis Committee:*

Dr. Pasquale Cirillo
drs. Eric Brandenburg AAG
drs. Alexander Sels AAG
Prof. dr. Cornelis W. Oosterlee

A thesis presented for the degree
Master of Science
in Applied Mathematics and Financial Engineering

Applied Mathematics
Faculty EEMCS
Technical University of Delft
The Netherlands
11-08-2017

# Abstract

A medium size Dutch insurance company with third-party car insurance products initiated questions on whether the premium can be based on a statistical analysis where the expected future liabilities are taken into account. These questions are as follows:

- Which statistical models can be used to base the premiums on expected future liabilities?
- Are there enough data available to predict future liabilities accurately enough?
- How can the 'best' model be chosen?
- How can the models be implemented?
- What are the results when using these models for the third-party car products?

After a practical introduction about insurances in society, the thesis starts with theory that can be used to answer the research questions. This analysis showed that generalized linear models are very useful models for the pricing of non-life insurance products. However, there are some disadvantages to these models which could be avoided by other models, such as hierarchical generalized linear models.

We will explore several methods to determine if enough data is available to obtain credible enough estimates. One of these methods can be applied before implementing a generalized linear model.

Choosing the 'best' model is a non-trivial subject. Several statistical tests to choose which risk factors should be included in the model and how they should be included are discussed. These include tests for adding risk factors as random or fixed effects, but also which definition of an risk factor should best be used. This includes whether they should be added as a variate, as a factor or added dynamically. In addition, several statistical methods to choose the distribution that has the 'best' fit for the observations for both the number of claims and the losses are discussed. These include graphical comparison methods, but also hypothesis testing.

To answer the question how the models can be implemented, we will use the statistical programming language `R`. Algorithms that are used by some packages to calculate the estimates of the models are discussed, as well as several features of these algorithms. Codes are provided in the supplementary section of the thesis.

Next, a statistical analysis is performed for the third-party car products of the insurance company. The performed theoretical analysis is applied in prac-

tice on the available data, and unknowns were calculated. Then an analysis is performed to determine which distribution 'best' fits the number of claims and which distribution 'best' fits the losses. The data was subdivided into several risk factors, such as age and region, and was analyzed again. A generalized linear model with a Poisson and log-link assumption was implemented for the number of claims, and a generalized linear model with a Gamma and log-link assumption was implemented for the losses. How and if the risk factors should be added was evaluated using a bottom-up approach. Initially, the models were applied without allowing interaction between risk factors, and subsequently the models were applied again, this time allowing interaction between risk factors to determine if this improved the models.

Other models that may lead to a better fit for the data are also implemented. These include generalized linear mixed models, which do not assume that the observations are independent and assume a Normal distribution for the risk factors that are added as random effects. Also, a pure premium model in which the Tweedie family is used was applied.

The study showed that the preferred models to calculate the pure premium are a generalized linear model with Negative Binomial and log-link assumption for the number of claims and a generalized linear model with Gamma and log-link assumption for the losses. Due to overdispersion of the observations for the number of claims, the Negative Binomial proved to be a better choice of distribution leading to a better fit of the model for the number of claims. The Normal and Pareto distribution were too symmetric and too right-skewed for the observations, respectively. The pure premium model showed a worse fit, when compared to the model for the number of claims. Furthermore, the effect of the risk factors on the risk profile of a risk group were very clear when a two-stage regression approach was used. The hierarchical models were not better models, because the estimates were less accurate.

The results for the different models were then compared with the currently used pricing system of the company and the expected outcomes of the data analysis. This leads to recommendations for the insurance company, including recommendations for pricing in general but also specific recommendations for the pricing system of the third-party insurance product.

The full thesis contains confidential information, therefore, a public version was provided in which the insurance company is anonymous. The full thesis was made available to the thesis committee.

# Contents

# A Practical Introduction

This introduction starts with a general introduction about insurances in society. Then, the importance of a statistical analysis in pricing non-life insurance products is described.

In our daily lives all kinds of accidents can happen and mostly they do not happen on purpose. Sometimes the accidents lead to very large losses. This could lead to undesirable debts for the responsible party. The party that has to reimburse the damage can be compensated by the insurer, within the boundaries of the contract. These losses are compensated by the premiums earned by the insurer, which have to be paid by the policyholder regardless of whether there is a claim or not. This means that the policyholders that pay more premium than they obtain from the particular contract pay indirectly for the large losses of others, *dispersion of risk*. With the insurance contract, the risk is transferred from the policyholder to the insurer. When the insurer has a large number of similar policies, his portfolio becomes more predictable and behaves like the expected value of the portfolio. This last point only holds true if not a substantial number of policyholders ends or enters a contract. This is a result of the law of large numbers, which will be discussed in section 12.9.

Before going into some features of a good premium, two concepts will be discussed, moral hazard and adverse selection.
*Moral hazard* refers to changes in the behaviour if, for example, a person is not directly at risk for their actions. In the special case of insurance, moral hazard can occur in two ways. Firstly, accidents may occur more often. The reason for this is, for example, that a person is more likely to be less careful with a car if this person knows that he or she is insured against damages of the car. Secondly, a person is more likely to claim a reimbursement when a damage occurs to try to obtain some money from the insurance company even when this might be unreasonable.
*Adverse selection* refers to a group B that mainly wants to do business with company A, but which leads to bad risks for company A. For example, suppose that for a particular car insurance people that are younger than 30 generate more and larger losses on average than people that are older than 30. Furthermore, let the premium policy of the insurance company for the particular product be such that for the group of people that are younger than 30 and the group of people that are older than 30 the same average premium is charged. It is then more likely that drivers younger than 30 want to buy insurances at the insurance company since the price will on average be lower than what they claim and get from the insurance company. (Here, we are not considering the costs of the insurance company. Minimizing the costs of the insurance company such that the premiums can be lower is another topic which is not in the scope of this thesis.) Hence, the company attracts bad risks for this product. Furthermore, the drivers that are older than 30 are likely to terminate their insurance

contract at this particular insurance company since on average they pay more than they claim. Therefore, the company also pushes off good risks for the product. Note that good or bad risks not only depend on the total losses, but also the premium asked. When the premium asked of drivers younger than 30 is more than the average loss for this group, than it is not bad to attract young drivers to your company for the particular product. However, it is still bad to push off the group of drivers that is older than 30 because of too large a premium.

When are premiums good and when are they bad? Some features of a good premium are mentioned below. Note that defining the reinsurance policy, the policy to minimize the costs of the company, the pricing policy and the policy conditions by the company are all important in deciding what a good premium is for a particular product. However, the features below are features that a good premium should have in general.

1. The premium is competitive with premiums for similar products used by companies with similar features, such as the reinsurance policy, the number of policyholders, costs and pricing policy.
2. The premium protects from moral hazard. Although the policy conditions can also protect from moral hazard, for example by not paying the claim when it is proved that there was reckless behaviour involved, the influence of the used pricing system should not be neglected. For example, to have discounts on the premium according to the claim behaviour. That is, to have larger discounts when there are no claims for a longer period of time.
3. The premium protects from adverse selection. The premiums asked of a person should, in a perfect world, meet the exact losses that the person will cause. However, there will always be problems with *asymmetric information*, that is, a person will probably have more and possibly important information about his or her situation which is not available for the insurance company. Furthermore, estimating the exact loss that a person will generate is a non-trivial subject, which is a main topic of this thesis.
4. The premium leads to a healthy loss/profit ratio on the balance sheet of the insurance company.

Now some arguments will be posed for the use of a research method based on statistical analysis, which was used in this thesis, when a product is priced, as opposed to (solely) using a research method based on price comparison with other companies.

1. A statistical study gives a better insight in the risk profile of the policyholders of a particular insurance company. The research per risk group is done much more in depth in a statistical study than in a general comparison with other companies and only calculating unknowns such as the loss ratios and claim frequencies. For example, in a statistical study, also the loss distribution is evaluated which gives a better insight in the risk profile. In addition, the statistical significance of the risk groups on the

loss distribution and claim frequencies can be investigated, which leads to more and better conclusions of the risk profile of the risk groups. Also the correlation between risk groups can be more precisely investigated, which can lead to better pricing conclusions. One can imagine that discounting two risk groups that are very closely correlated can lead to discounting more or less the same group twice, which might lead to discounting a large group of people too much.

2. The pure premium is calculated per risk level, which should be asked to cover the losses of the own policyholders. There can always be risk groups that are not taken into account by an insurance company and other companies while they are of significant impact. This means that using the pure premium is a safer approach, because without explicitly taking it into account as a variable in the model, it is taken into account by the model anyway since the claim and loss behaviour is captured in the data. The premiums used by the other companies can be misleading for the own risks.

3. Comparisons with other companies lead to the question to what extent the companies really are comparable.

   (a) What are the choices of the other companies regarding the premium? Do they want a low premium to obtain a large number of policyholders or do they want a higher profit per policy?

   (b) What are the policy conditions of the product, what exactly is insured and under which conditions will the company pay the claim?

   (c) What is the reinsurance policy of the other companies? Is the reinsurance such that the other company can more easily ask low premiums?

   (d) What is the risk profile of the policyholders of the other company?

   When relying on the market comparison, answering these questions is important. However, answering these questions can be challenging, maybe even impossible. A study based on a statistical analysis does not require answers to these questions since the insurance company can make decision based on the risk profiles of the policyholders and the risk appetite of the insurance company. In this way, an insurance company can be more autonomous.

Monitoring the market is very important nonetheless. The premium in comparison to other insurance companies will have a large impact on the number of policyholders for the product, more specifically for different risk groups. When all other companies are much more expensive, it is wise to follow their lead. On the other hand, depending on the risk appetite, reinsurance policy and product policy, being relatively expensive might not be preferable. Having more policyholders means being able to better monitor the product, being able to cover large losses better, and being able to predict the total losses better. Furthermore, the cost per policyholder is lower. When the estimates are not credible

enough, for example, when the number of claims is not sufficient, the pricing can definitely not solely depend on the pure premium calculation given by the models that are described. In that case, these premiums are then more a guideline. On the other hand, when the estimates are credible enough and the pure premiums are evaluated and they are realistic per group, the pure premiums can be leading in determining the premiums. In this case the market research can be done for adjustments on the premiums determined by the pure premiums and not the other way around.

# Part I

# (Hierarchical) Generalized Linear (Mixed) Models and Model Selection

# Chapter 1

# Generalized Linear Models

## 1.1 Overview of Chapter 1

In chapter 1 generalized linear models are described. These models are widely used in the pricing of non-life insurance products. In general they are regression models and pricing is often based on these models as a two-stage regression model. The claim numbers and losses can be modelled separately, that is, with two different generalized linear models.
Furthermore, some features of generalized linear models will be discussed. Afterwards, some advantages and disadvantages of generalized linear models in actuarial practice will be discussed.

## 1.2 Characteristics

A generalized linear model is a model of the form

$$g(\mu_i) = \sum_j x_{ij}\beta_j. \tag{1.1}$$

The following three characteristics hold for (non-hierarchical) generalized linear models:

1. The stochastic component of the model states that the observations are *independent* random variables $Y_i$, $i \in 1, ..., n$, with a density in the exponential dispersion family.
2. The systematic component of the model attributes to every observation a *linear predictor* $\eta_i = \sum_j x_{ij}\beta_j$, linear in the parameters $\beta_1, ..., \beta_p$. The $x_{ij}$ are called covariates.

3. The link function links the expected value $\mu_i$ of $Y_i$ to the linear predictor by $\eta_i = g(\mu_i)$.

## 1.3 Stochastic Component

A density in the *exponential dispersion family* has the following form

$$f(y'_{ij}; \theta_{ij}; \phi) = e^{[y_{ij}\theta_{ij} - b(\theta_{ij})]/a(\phi) + c(y_{ij}, \phi)}. \tag{1.2}$$

For different functions of $a(.)$, $b(.)$ and $c(.)$ there are several known densities. The function $a(\phi)$ can be of the form $a(\phi) = \phi$. The *dispersion parameter* is denoted by $\phi$. Suppose that $\phi$ is known then the density function becomes

$$f(y_{ij}; \theta_{ij}) = a^*(\theta_{ij}) b^*(y_{ij}) e^{y_{ij} Q(\theta_{ij})}, \tag{1.3}$$

where $Q(\theta) = \frac{\theta}{a(\phi)}$, $a^*(\theta) = e^{-b(\theta)/a(\phi)}$ and $b^*(y) = e^{c(y,\theta)}$.
Formula 1.2 can be applied to describe two parameter families such as the Normal and Gamma distribution. Formula 1.3 can be applied to describe one parameter distributions, such as the Poisson distribution.
Furthermore, a Poisson distribution where $\phi$ is not equal to one is called a *quasi-Poisson* distribution.
Distributions that belong to the exponential dispersion family are allowed, such as the Normal, Poisson, Binomial and Gamma distributions.
Note that in practice, weights have to be taken into account since some cells will typically contain more policies than others. In this case, deciding not taking weights into account, disregards the fact that observations in cells with many policies have been measured with much more precision than the ones in practically empty cells. Weights are often the natural or *exposure weights*. For example, let the claim frequency be the quantity of interest. Suppose that the observations are all claims over one particular year. Then the exposure is the number of policies that are in force throughout the year. Note that in some cases policies are in force during a part of the year, in these cases the exposure is the sum of the number of policies that are in force the whole year and the fractions of the year for the policies that are only in force a part of the year.

## 1.4 Likelihood

The log likelihood function is given by $l(\phi; \theta; y) = log f_Y(y; \theta; \phi)$. Also, given the exponential dispersion family, the two relations $E(\frac{\partial l}{\partial \theta}) = 0$ and $E(\frac{\partial^2 l}{\partial \theta^2}) + E(\frac{\partial l}{\partial \theta})^2 = 0$ can be derived. This leads to the following relation for the expected value $\mu_{ij}$ of $Y_{ij}$ and the variance of $Y_{ij}$

$$E(Y_{ij}) = \mu_{ij} = b'(\theta_{ij}) \tag{1.4}$$

$$Var(Y_{ij}) = \mathbb{V}(Y_{ij}) = b''(\theta_{ij})a(\phi). \tag{1.5}$$

However, the variance can also be described as a function of the average, that is, $Var(Y_{ij}) = a(\phi)V(\mu_{ij})$. The function $V(.)$ is called the variance function. Note that including weights leads to $Var(Y_{ij}) = a(\phi)/w_{ij}V(\mu_{ij})$.

Suppose that the distribution is not known, but that the first two moments are given, then the so-called *quasi-likelihood* function can be used instead of the likelihood function.
Let $Y_1, Y_2, ..., Y_n$ be independent random variables with expected value $E(Y) = \mu$ and variance $Var(Y) = V(\mu)$ for a certain known function $V(.)$. The so-called quasi-likelihood function is defined as (when the dispersion parameter is equal to one and weights are not included)

$$Q(\mu) = \sum_{i=1}^{n} \int_{t=y_i}^{\mu_i} \frac{y_i - \mu}{V(\mu)} d\mu. \tag{1.6}$$

In many ways this function behaves the same as the likelihood function. If $\mu$ is estimated in a similar way, by maximizing the quasi-log-likelihood, then, typically, results will be found with similar optimality conditions.

For the maximum likelihood estimates for the parameters in the generalized linear models the following property holds [17]. Note that we use an iterative algorithm, such as the iteratively reweighted least squares method which is explained in section 14.1.1.

**Property 1.4.1.** *For the Maximum Likelihood Estimates, $\vec{\beta^*}$,*

1. *$\vec{\beta^*}$ is an asymptotically unbiased and consistent estimator of $\vec{\beta}$.*
2. *$\mathbb{V}(\vec{\beta^*}) \rightarrow (X^T W X)^{-1}\phi$ consistently, as the iteratively estimated $\vec{\beta^*}$ converges to the true $\vec{\beta}$, where $W = diag(w_i, ..., w_n)$ with weights $w_1 = [\phi g'(\mu_i)V(\mu_i)]^{-1}$, and matrix $X$ with the covariates vectors.*
3. *$\vec{\beta^*} \rightarrow^d N(\vec{\beta}, (XWX)^{-1}\phi)$, i.e. it converges in distribution with the iterative algorithm.*

## 1.5 Link Function

The *link function* $g(.)$ links the stochastic and systematic component, where $g(.)$ is monotone and differentiable. Often used link functions are the logarithmic or identity function. Hence, the models can be additive but also multiplicative.
The link function influences the estimates made. Adding also to the previous section, the link function has an influence on the bias of the maximum likelihood estimates [21]. The canonical link function has some nice properties [2], however one may judge that under another link function, the estimates are better.

### 1.5.1 Canonical Link Function

Each of the distributions of the exponential family have their own natural link function which is called the *canonical link function*. Some examples are the identity function for the Normal distribution, the logarithmic function for the Poisson distribution and the reciprocal function for the Gamma distribution.

**Definition 1.5.1.** The canonical link function is the link function that has the property that the natural (also called canonical) parameter $\theta$ coincides with the linear predictor $\eta$.

This holds if the link function is the inverse function of $\mu(\theta) = b'(\theta)$.

**Property 1.5.1.** *If in any generalized linear model with covariates $x_{ij}$ and canonical link $g(.)$, the fitted value for observations $i = 1, ..., n$ under a maximum likelihood estimation is $\mu_i^* = g^{-1}(\eta_i^*) = g^{-1}(\sum_{j=1}^{p} x_{ij}\beta_j^*)$, the following equations hold:*

$$\sum_i w_i y_i x_{ij} = \sum_i w_i \mu_i^* x_{ij}, j = 1, ..., p \tag{1.7}$$

*Proof.* An extremum of the total log-likelihood based on the entire set of observations $Y_1 = y_1, ..., Y_n = y_n$ satisfies the conditions

$$\sum_i \frac{\partial}{\partial \beta_j} l(\beta_1, ..., \beta_p; y_i) = 0, j = 1, ..., p \tag{1.8}$$

Applying the chain rule and using the $\theta \equiv \eta$ property of the canonical link leads to:

$$\frac{\partial}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{\partial \eta}{\partial \beta_j} \tag{1.9}$$

Using the dispersion parameter $\phi$, the density of exponential dispersion family and the equation $\mu(\theta) = b'(\theta)$, it can be shown that the relation for the observations, with $i = 1, ..., n$:

$$\frac{\partial l}{\partial \beta_j} = \frac{w_i(y_i - \mu_i)x_{ij}}{\phi}, j = 1, ..., p \tag{1.10}$$

holds. Now summing over all $i = 1, ..., n$, the observations, the log-likelihood of the whole sample $y_1, ..., y_n$ is obtained. Now setting the normal equations equal to zero, this directly leads to maximum likelihood equations of the form to be proved. □

Note that if the $x_{ij}$ are dummies with a characterized membership of a certain group like a row or column of a table. Furthermore, the $y_i$ are averages of $w_i$ independent identically distributed observations, on the left hand side there is the observed total, and on the right the fitted total.

**Property 1.5.2.** *In a generalized linear model, if the canonical link* $\theta_i \equiv \eta_i = \sum_j x_{ij}\beta_j$ *is used, the quantities* $S_j = \sum_i w_i Y_i x_{ij}$, $j = 1, ..., p$, *are a set of sufficient statistics for* $\beta_1, ..., \beta_p$.

*Proof.* The joint density of $Y_1, ..., Y_n$ can be factorized as

$$f_{Y_1,...,Y_n}(y_1, ..., y_n; \beta_1, ..., \beta_p) = g(s_1, ..., s_p; \beta_1, ..., \beta_p)h(y_1, ..., y_n), \tag{1.11}$$

where $s_j = \sum_i w_i y_i x_{ij}$, $j = 1, ..., p$ and suitable functions $g(.)$ and $h(.)$. Indeed,

$$\begin{aligned} & f_{Y_1,...,Y_n}(y_1, ..., y_n; \beta_1, ..., \beta_p) \\ & = \Pi_{i=1}^n exp(\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i; \phi/w_i) \\ & = exp(\sum_i \frac{y_i \sum_j x_{ij}\beta_j - b(\sum_j x_{ij}\beta_j)}{\phi/w_i})exp(\sum_i c(y_i; \frac{\phi}{w_i})) \\ & = exp(\frac{1}{\phi}[\sum_j \beta_j \sum_i w_i y_i x_{ij} - \sum_i w_i b(\sum_j x_{ij}\beta_j)]exp(\sum_i c(y_i; \frac{\phi}{w_i})). \end{aligned} \tag{1.12}$$

Now, the functions $g(.)$ and $h(.)$ can be derived immediately. □

Background information about statistics can be found in section 12.3.
The two properties stated above can be very convenient in actuarial practice. The observations may have been aggregated into a table of which only the marginals (row and column sums) are available. Also, the policies could have been grouped into cells, to save time and space. Moreover, if the canonical link function is used, the marginal totals, as well as the cell totals, apparently are sufficient statistics. Hence, knowing only their outcomes, the maximum likelihood estimates can still be determined.

## 1.6 Parameters as Variates and Factors

The parameters to be estimated in the generalized linear model of your choice, can be estimated as a *variate* or a *factor*. In the case of a factor the risk factor levels $i = 1, ..., I$ are used as labels. Which means that every $\alpha_i$ is an arbitrary number. This implies that every $\alpha_i$ needs to be estimated. On the other hand, if variates are used, then for the level $i$ there is $i * \alpha$, which implies that only $\alpha$ needs to be estimated. In practice, the question whether such a relation is indeed more or less valid, arises immediately and has to be discussed. On the other hand, using the risk characteristic, when divided into more than two levels, has consequences for the complexity of the model.

## 1.7 Full and Null Models

The generalized linear model that only uses a constant for the systematic component is called the *null model*. This model implies that every observation has the same distribution and that the weighted average is the best estimator for every $\mu_i$. On the other extreme, there is the *full model*. This model implies that every unit of observation $i$ has its own parameter. Intuitively one may immediately think that the null model is not accurate enough to be of any use. On the other hand, the full model may be too complicated and has too many parameters that have to be estimated to be of any practical use because of computational issues. And indeed, often the optimal model is somewhere in between. This trade-off has to be made.

## 1.8 Offset

The *offset* is defined as an additional model variable with coefficient 1. Often the offset is included when dealing with Poisson counts, where the canonical log-link function is used. Then the fit looks like

$$log(\eta_i) = \sum_j x_{ij}\beta_j + log(\epsilon_i) \tag{1.13}$$

where $\epsilon$ is the offset. Note that when using weights, claim frequencies can be obtained. In this case these will still have a Poisson distribution since the claim frequency modelling, with weighted exposure, is equivalent to claim count modelling with log-linked exposure.

## 1.9 Generalized Linear Models in Actuarial Practice

After describing the generalized linear model, some advantages and disadvantages for the use in actuarial practice are discussed. First of all, we state to what extent the generalized linear model is a generalization of the linear model. The generalization is in two directions [2]. The random variables involved are assumed to be Normal in a linear model with a variance independent of the mean. In the generalized linear model however, the observed values, need to have a density in the exponential dispersion family. In actuarial practice this is a very convenient generalization, since the number of claims and losses are often assumed to be for example Poisson and Gamma respectively. On the other hand, in the linear model, the response is assumed to be linear in the covariates on the identity scale. For the generalized linear model the scale can be different from the identity scale. Often we see a scale which involves logarithm in actuarial practice, since this leads to having a multiplicative model rather than an

additive one.
Although the generalized linear models have some advantages that can be very convenient in actuarial practice in comparison to the linear models, they still have disadvantages. Two disadvantages will be stated which may not be valid for other statistical models that are discussed in this thesis. First note that the independent assumption of the observations given their explanatory variables still holds. This assumption need not be valid and can therefore be a disadvantage. Also, either zero or full credibility is given to the data and blending is not possible. There are four other disadvantages stated by the GIRO APT working party from the Institute and faculty of Actuaries in the UK [7]. These disadvantages are as follows. Firstly, model predictions depend on the mixture of rating factors in the data. Secondly, maximum likelihood estimate of prediction is lower than mean of prediction distribution. Thirdly, the link function could bias the model prediction and significantly change the lower and upper bound of prediction. Finally, model diagnostics is only relevant in the segments where the model is used.

# Chapter 2

# Hierarchical Generalized Linear Models

## 2.1 Overview of Chapter 2

In chapter 2 the hierarchical generalized linear model is discussed. First the reason why the hierarchical generalized linear models can be useful in actuarial practice is discussed. Then, the difference between the hierarchical generalized linear model and the generalized linear model is explained, that is, to what extent the hierarchical generalized linear model is a generalization of the generalized linear model. At the end some features about likelihood estimates in the case of hierarchical generalized linear models are explained.

## 2.2 Introduction

In the case of hierarchical generalized linear models the assumption of the observations being independent from each other given the predictor values does no longer have to be valid. The situation of the observations being dependent on each other given the predictor values, also in actuarial practice, can certainly occur. Hierarchical models are often made with the use of Bayesian inference. In this case there is a distribution chosen a priori of the parameter(s) that has to be estimated in the distribution of the particular random variable of interest. And afterwards, the distribution is updated by computation using the data to get the posterior distribution. Using the conjugate prior family, where the posterior and prior distribution are of the same parametric form is desirable due to some properties that hold in that case. Also a likelihood approach can be used. The results of the likelihood-based analysis can be used, for instance to choose

starting values for the chains and to check the reasonableness of the results. In an actuarial context, an important advantage of the Bayesian approach is that it yields the posterior predictive distribution of quantities of interest.
However, although there are advantages of the hierarchical generalized linear model in actuarial practice, there is an increase of the complexity of both the model and the calculations. The discussion on the trade-off between complexity, credibility of the estimates and performance of the models arises immediately.

## 2.3 Characteristics

The characteristics of the hierarchical generalized linear models are different in some ways. Firstly, note that the stochastic component is different in the sense that the observations do not have to be independent. Secondly, note that the systematic component is different in the sense that a stochastic component is added to the linear predictor. This results in $\eta_i = \sum_j (x_{ij}\beta_j + z_{ij}u_j)$ where there is a distribution parametrized by $U$ over the parameters of $\vec{u}$. The random effects do not only determine the correlation structure between observations on the same subject, they also take into account heterogeneity among subjects, due to unobserved characteristics. Furthermore, note that for the purpose of completeness, if the random effect $U$ follows a Normal distribution, then the hierarchical generalized linear model is also called a *Generalized Linear Mixed Model*.

## 2.4 Structure of Hierarchical Generalized Linear Models

In the previous section it was already mentioned that the hierarchical generalized linear model provides an extension for the generalized linear model by the addition of random factors. The structure is as follows

$$\begin{aligned} y_i \mid u &\sim f_{Y_i|u}(y_i \mid u), \\ f_{Y_i|u}(y_i \mid u) = exp(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} &- c(y_i;\phi)), \\ u &\sim f_U(u), \end{aligned} \tag{2.1}$$

where $f_U(u)$ denotes the density function for the random effect. When $f_U(u)$ is a Normal density function, then we have a generalized linear mixed model, but in general there can be a non-Normal distribution for the random effect.
Note that there can be dependence now, specified by the covariance. The following equalities hold for the mean, variance and covariance, that depend on

the link function.

$$\begin{aligned} E[y_i] &= E[E[y_i \mid u]] \\ &= E[\mu_i] \\ &= E[g^{-1}(\sum_j (x_{ij}\beta_j + z_{ij}u_j))] \end{aligned}$$

$$\begin{aligned} Var(y_i) &= Var(E[y_i \mid u]) + E[Var(y_i \mid u)] \\ &= Var(\mu_i) + E[\phi V(\mu_i)] \\ &= Var(g^{-1}[\sum_j (x_{ij}\beta_j + z_{ij}u_j)]) + E[\phi V(g^{-1}[\sum_j (x_{ij}\beta_j + z_{ij}u_j)]) \end{aligned}$$

$$\begin{aligned} Cov(y_i, y_j) &= Cov(E[y_i \mid u], E[y_i \mid u]) + E[Cov(y_i, y_j \mid u)] \\ &= Cov(\mu_i, \mu_j) \\ &= Cov(g^{-1}[\sum_j (x_{ij}\beta_j + z_{ij}u_j)], g^{-1}[\sum_j (x_{ij}\beta_j + z_{ij}u_j)]) \end{aligned} \tag{2.2}$$

To get more insight into particular cases which will be applied, the further derivation by implementing the link function will be useful.

## 2.5 Maximum Likelihood Estimation

For hierarchical generalized linear models there are some other features for the maximum likelihood estimates. In the next section the restricted maximum likelihood estimates are proposed. Calculating maximum likelihood estimates often involves an integral that does not have a closed-form solution. A closed-form solution can be found for conjugate distributional specifications. Also, calculating the integral often has to be done for every iteration. There are several approximation techniques that are discussed in chapter 14.

### 2.5.1 Restricted Maximum Likelihood Estimation

For hierarchical generalized linear models, the standard maximum likelihood estimation can give biased estimates for the variance of (random) components. This is due to the fact that in the estimation procedure (by taking the derivative of the log-likelihood function with respect to both parameters, which can be vectors, and set to zero), the existence of fixed effects is ignored, and the degrees of freedom used in deriving the estimators do not adjust to this. The restricted maximum likelihood estimation method can be a solution in the sense that it can give unbiased estimates for the random effects. The restricted maximum

likelihood estimation method works by first obtaining regression residuals for the observations modelled by the fixed effects portion of the model, ignoring at this point any random components. Then the question what is the statistical model for these residuals should be asked. There is no more fixed effect part, because all fixed effects are taken out when the residuals are taken. Aspects of the random effects part and error part remain. Taking residuals changes the covariance structure, but that is taken into account. This method works in a very convenient way for generalized linear mixed models. However for hierarchical generalized linear models, where the distribution of the random component belongs to the family of conjugate Bayes distributions for an exponential family, the Laplace approximations of the marginal likelihood function is proposed. Restricted maximum likelihood estimation makes use of a different likelihood function than simple likelihood (in particular, it does not even depend on the fixed effects coefficients), so its achieved likelihood is also different. Comparing nested models that only differ in the random terms can be done by using the restricted maximum likelihood or the ordinary likelihood. Comparing models that differ in fixed effects terms can be done by using ordinary likelihood.

# Chapter 3

# Calculating the Premium

## 3.1 Overview of Chapter 3

The (hierarchical) generalized linear (mixed) models have been described, however the calculation of the risk premium when using these models is not yet derived. In this chapter the calculation of the pure premium is discussed. However, when the data is not sufficient enough to implement reliable generalized linear models an alternative is presented. The alternative that is discussed, is calculating the credibility premium with the use of the Bühlmann-Straub model. Into the next sections, three different premiums are mentioned, namely the *relative premium*, the *pure premium* and the *quoted premium*. The relative premium is the premium for a given risk class relative to the intercept group. The intercept group is formed when all risk levels, for the risk factors chosen in the model, are at the standard level. The pure premium is directly given by the output of the models. And the quoted premium is the premium that the insurer brings to the market. The relative and pure premium can be calculated by combining the models for the claim severity and the claim frequency.

## 3.2 Calculating the Pure Premium

Suppose $\vec{\beta}$ is estimated such that it is the best fit for the equation

$$Y = X\vec{\beta} \tag{3.1}$$

in a fixed setting and that $\vec{\beta}$ and $\vec{u}$ have been estimated such that they give the best fit for the equation

$$Y = X\vec{\beta} + Z\vec{u} \tag{3.2}$$

in a model where random effects are also considered. The observation $y_i$ is the observation given the risk classes and where $\vec{\beta}$ and $\vec{u}$ are estimated such that they are the best fit for the equation

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} \tag{3.3}$$

in a purely fixed setting and

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} + u_1 + u_2 z_{i2} + u_3 z_{i3} + ... + u_p z_{ip} \tag{3.4}$$

in a model where random effects are also considered. The calculation of the pure premium, given the estimates $\vec{\beta^*}$ and $\vec{u^*}$, is then very straightforward. Suppose that $S$ is the severity and $N$ is the number of claims, where the link functions used are $h$ and $g$, respectively. Furthermore, suppose that the chosen variables for the severity model are denoted by $x^-$ and $z^-$ for the fixed and random factors respectively and $x$, $z$ for the count model. Then for the number of counts given the risk classes the following equation holds,

$$N_i = g^{-1}(\beta_1^* + \beta_2^* x_{i2} + \beta_3^* x_{i3} + ... + \beta_p^* x_{ip}) \tag{3.5}$$

and for the severity

$$S_i = g^{-1}(\beta_1^* + \beta_2^* x_{i2}^- + \beta_3^* x_{i3}^- + ... + \beta_p^* x_{ip}^-), \tag{3.6}$$

given that only fixed effects are considered. If random effects are included then the above equations are changed to

$$N_i = g^{-1}(\beta_1^* + \beta_2^* x_{i2} + \beta_3^* x_{i3} + ... + \beta_p^* x_{ip} + u_1^* + u_2^* z_{i2} + u_3^* z_{i3} + ... + u_p^* z_{ip}) \tag{3.7}$$

and

$$S_i = g^{-1}(\beta_1^* + \beta_2^* x_{i2}^- + \beta_3^* x_{i3}^- + ... + \beta_p^* x_{ip}^- + u_1^* + u_2^* z_{i2}^- + u_3^* z_{i3}^- + ... + u_p^* z_{ip}^-), \tag{3.8}$$

respectively. Now the pure premium, given the risk classes, is given by

$$PP_i = N_i * S_i. \tag{3.9}$$

## 3.3 Credibility Premium

In the case where a limited amount of data is available, *credibility theory* could be used to determine the *credibility premium*, which is given by

$$z_j \bar{X}_j + (1 - z_j)\bar{X} \tag{3.10}$$

where $\bar{X}_j$ is the average claim (severity) of group $j$ and $\bar{X}$ the overall mean of the data. The most difficult task is to determine $\vec{z}$, the *credibility factor* for each group $j = 1, ..., n$. The Bühlmann-Straub model is often used, but the classical approach can also be used.

### 3.3.1 Heterogeneous Portfolio

Consider the random variable $X_{jt}$, representing the claim figure of cell $j$ in year $t$, where $j \in \{1, 2, ..., J\}$ and $t \in \{1, 2, ..., T\}$. Suppose that all $X_{jt}$ are independent and $N(m_j, s^2)$ distributed, with possible unequal mean $m_j$ for each cell, but equal variance $s^2$>0. If the portfolio is homogeneous in the sense of the mean, in other words that it is reasonable to assume all $m_j$ are equal, then there is no reason not to ask the same premium for each contract. To test if the portfolio is heterogeneous or homogeneous, the so-called *variance ratio* or $F$ ratio can be used, which is defined as

$$F = \frac{MSB}{MSW} = \frac{\frac{1}{J-1}\sum_j T(\bar{X_j} - \bar{X})^2}{\frac{1}{T(J-1)}\sum_j \sum_t (X_{jt} - \bar{X_j})^2}. \tag{3.11}$$

It can be proved that the ratio $F$ has a $F(J-1, J(T-1))$ distribution. Furthermore, for this statistical technique of analysis of variance, a linear model explaining the responses $X$ from the group number $j$ as a factor can be defined [2].

### 3.3.2 Bühlmann-Straub Model

Suppose that $X_{jt}$ is the average claim (severity) for contract $j = 1, ..., J$ in year $t$. Assume that all $X_{jt}$ are independent and $N(m_j, s^2)$ distributed, with possible unequal mean $m_j$ for each cell, but equal $s^2 = a$. Note that if the portfolio is homogeneous, that is, the group means are all equal, then there is sufficient reason to ask the same premium for each contract (which could be the average loss for one year). However, when the portfolio is homogeneous, which means that the group means are not all equal, then there is reason to ask different premiums. Let the portfolio be heterogeneous. Assume that each $m_j$ is produced by 'white noise' similar to the one responsible for the deviations from the mean within each cell. Then the following model holds, which is a *variance components model*,

$$X_{jt} = m + \Xi_j + \Xi_{jt}, j = 1, ..., J, t = 1, ..., T, \tag{3.12}$$

with $\Xi_j$ and $\Xi_{jt}$ as independent random variables for which

$$E[\Xi_j] = E[\Xi_{jt}] = 0, Var[\Xi_j] = a, Var[\Xi_{jt}] = s^2. \tag{3.13}$$

In the Bühlmann-Straub model there are the following estimates of the (corresponding) structure parameters $a, s^2$ and $m$, which are unbiased [2]

$$\begin{aligned} m^* &= X_{ww}, \\ s^{2*} &= \frac{1}{J(T-1)} \sum_{j,t} w_{jt}(X_{jt} - X_{jw})^2, \\ a^* &= \frac{\sum_j w_{j\sum}(X_{jw} - X_{ww})^2 - (J-1)s^{2*}}{w_{\sum\sum} - \sum_j w_{j\sum}^2 / w_{\sum\sum}}, \end{aligned} \tag{3.14}$$

where

$$\begin{aligned} X_{jw} &= \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\sum}} X_{jt}, \\ X_{ww} &= \sum_{j=1}^{J} \frac{w_{j\sum}}{w_{\sum\sum}} X_{jw}, \\ w_{j\sum} &= \sum_{t=1}^{T} w_{jt}, \\ w_{\sum\sum} &= \sum_{j=1}^{J} w_{j\sum}. \end{aligned} \tag{3.15}$$

When these estimates have been calculated, one can determine the credibility premium by substituting the estimates in the following equation

$$z_j X_{jw} + (1 - z_j) X_{zw} \tag{3.16}$$

where

$$\begin{aligned} z_j &= \frac{a w_{j\sum}}{s^2 + a w_{j\sum}}, \\ z_\sum &= \sum_{j=1}^{J} z_j, \\ X_{zw} &= \sum_{j=1}^{J} \frac{z_j}{z_\sum X_{jw}}. \end{aligned} \tag{3.17}$$

Note that in section 15.3 there is a code included to calculate the credibility premiums for this model. Also, the credibility premiums with $a$, $m$ and $s^2$ known are compared with the credibility premiums with unbiased estimates $a^*$, $m^*$ and $s^{2*}$ as described above, for the same generated data.

# Chapter 4

# Model Choice

## 4.1 General Aspects of Model Choice

The question of which models should be chosen is important and non-trivial. To answer this question, one should answer several other questions. For example, whether if the sample size is sufficient to implement a model. Or which distribution fits the observations best. Also, which risk characteristics show the most significant changes for the response of interest. And, which levels should be chosen for the risk characteristics. There are many more questions that should be answered. In chapter 4, theory on how one could discriminate between models, to end with the 'best' model is presented.

Answering the question whether the amount of useful data is sufficient in order to be able to implement (hierarchical) generalized linear (mixed) models is very useful. Indeed, it is not useful to implement generalized linear model if the estimates are not credible. For evaluating whether the sample size is sufficient enough to implement generalized linear models, the method explained in section 4.5.2 may be very helpful.
Note that the distribution which fits the observations best has implications for the model that should be chosen. In section 4.2 some possible distributions and ways to discriminate between distributions are discussed.
Also, the choice of the link function will be discussed in section 4.3. Although the canonical link function is often chosen, other link functions can be preferred as well.
Furthermore, which risk characteristics should be added (if they should be added) as random or fixed variables in the model will be discussed. Note that if there are no random effects, then a generalized linear model is obtained, otherwise a hierarchical model is obtained. More on deciding if a risk characteristic should be added as a fixed or random variable can be found in section 4.4

Furthermore, some fitting criteria are given in section 4.6 to make a choice between which risk characteristics should be taken into account. The goodness of fit that a model provides is crucial in this case. To answer these questions, one may use *fitting criteria*, such as *residuals* and *log-likelihood ratio*. The purpose often is to reflect on the distance between fits and observations; large distances are intuitively not desirable. Furthermore, using for example deviance residuals, nested models can be compared. But the trade-off between complexity and the goodness of fit is also an interesting topic regarding the choice of the model. The use of *information criteria* is often recommended for this purpose.
This thesis will generally use a *bottom-up* approach, where the method starts with an empty/null model, or with a model which only includes a set of risk characteristics that should be included in any case. Then risk characteristics are added and it can be determined if they have a significant impact. The choice for this approach is made because of the choice of included factors is very clear. If a *top-down* approach is chosen, then the method starts with the full model and the risk characteristics are removed. A risk characteristic is in this case not included, or, to be more precise, removed, if the test shows little to no predictive power.
Apart from choosing the risk factors and which of the (hierarchical) (generalized) linear (mixed) models that should be chosen, the choice between a Bayesian approach, with for example (Markov Chain) Monte Carlo methods and a point-estimation approach, with for example a (restricted) maximum likelihood estimate should be made. The choice between a hierarchical or non-hierarchical model and the properties of the estimates are important. Fitting criteria can be also applied for this purpose. Next to these properties there is also the computational complexity that has to be considered.

## 4.2 Choice of Distribution

### Overview of Section 4.2

As mentioned before, the generalized linear model and the linear model, hierarchically modelled or not, differ in the assumption of the distribution of the observed data. Hence, the choice between a linear model or a generalized linear model simply comes down to deciding which distribution has the best fit for the observed data in both a statistical and explanatory sense. If this distribution is a Poisson or Gamma distribution, which often are distributions seen as the best fits for the claim frequency and severity respectively, then a generalized linear model should be chosen instead of a linear model. However, other distributions can be preferred.
Briefly some possibilities for the distributions of the severity model are discussed in section 4.2.1. Then, some possible comparison methods are discussed, such as the Zipf, mean-excess function, the discriminant moment-ratio plot, the Zenga and the $QQ$-plot. More detailed information can be found in [22].

In section 4.2.2, possibilities for modelling the claim frequency are discussed. The Poisson distribution in a generalized linear model is often chosen, where time independence is assumed. However, time dependence may occur. For example, in the winter more accidents could occur than in other periods of the year. Furthermore, when the variance divided by the mean shows a big difference from one, it might be wise to choose another model since the Poisson distribution assumes that the mean and variance are equal.

### 4.2.1 Loss Distribution

For the loss distribution, sometimes the Normal, Lognormal, Gamma or Pareto distribution are chosen. The Normal distribution is symmetric and often does not fit the right tail of the loss distribution. When it is chosen in the model, the problem may arise that large risks are not modelled correctly. The Lognormal, Gamma and, in a more extreme way, the Pareto distributions, generally show large outliers. The Gamma and Lognormal distributions are quite similar. However, the Gamma distribution tends to have a heavier left tail and a lighter right tail than the Lognormal distribution. An illustration can be found in figure 4.1.
Important to mention in the framework of generalized linear models, is that the Lognormal distribution does not belong to the exponential dispersion family. This means that transforming has to be done. A way to nonetheless fit a generalized linear model anyway with a Lognormal distribution for the observations is fitting a normal linear model with a log transformation of the responses. However, the result is not quite the same since $E(log(Y)) \neq log(E(Y))$ generally.

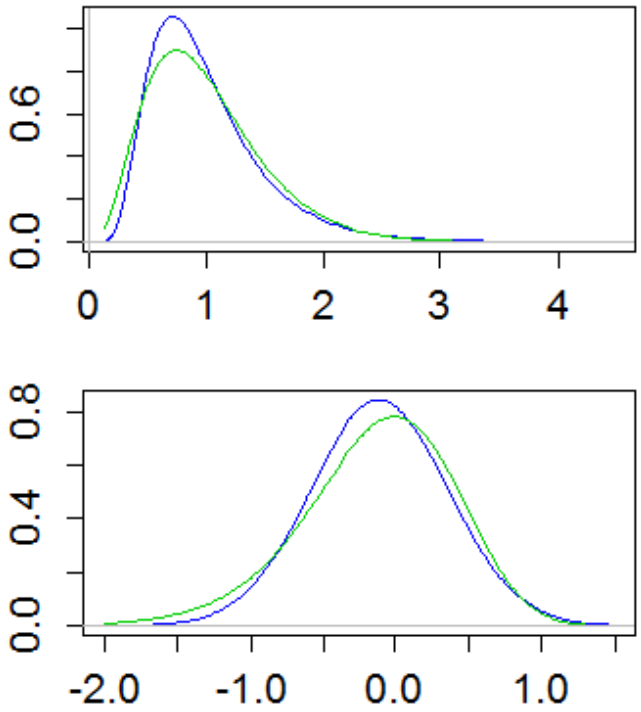


Figure 4.1: An example where both the Lognormal and Gamma distribution have a mean of 1 and a variance of 1/4. The top plot shows the density and the lower plot shows the densities of the logs, where the green colour indicates the Gamma and the blue colour the Lognormal distribution.

### Zipf Plot

Suppose that the cumulative distribution function is denoted as $F(x)$. The *survival function* is then defined as $\bar{F}(x) = 1 - F(x)$. In other words, the chance of obtaining larger values than $x$. The *Zipf* plot is the survival function on the $y$-axis and the losses on the $x$-axis, where both axis are on the log scale. If the data follow a power law (that is, a Pareto distribution), then one can expect to observe a negative linear trend in the Zipf plot. Figure 4.2 is made for the theoretical behaviour of the Zipf plot for some distributions.

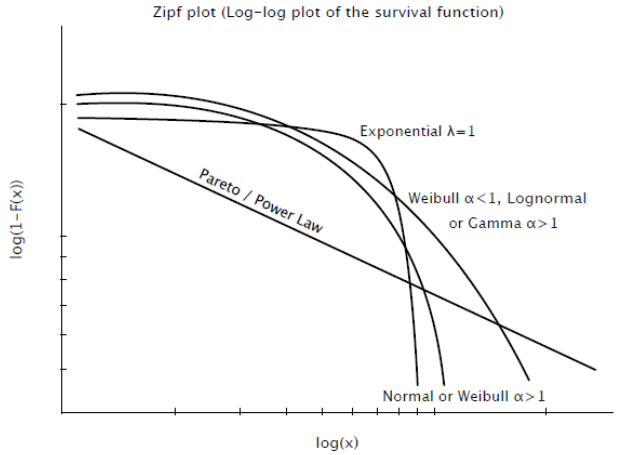


Figure 4.2: Zipf plot behaviour for some classical distributions [22].

Note that in practice, the observed data will generally not exactly follow one of the lines of figure 4.2. Therefore, discriminating between distributions might be a problem. Some other plots can be used to obtain more information on which theoretical distribution fits the data best.

### Mean Excess Function Plot

The *mean excess function* of $X$ (a random variable with distribution $F$) is defined as

$$e(u) = E[X - u \mid X>u] = \frac{\int_u^\infty (t-u)dF(t)}{\int_u^\infty dF(t)}, 0<u<x_F, \tag{4.1}$$

where $x_F = sup\{x \in \mathbb{R} : F(x)<1\}$. In practice, this comes down to

$$e_n(u) = \frac{\sum_{i=1}^n (X_i - u)}{\sum_{i=1}^n 1_{\{X_i>u\}}}, \tag{4.2}$$

for a sample of size $n$. In words, the sum of the exceedances over the threshold $u$ divided by the number of data points exceeding $u$. Also in this case, in figure 4.3 the theoretical behaviour of some distributions can be found for the mean excess function plot.

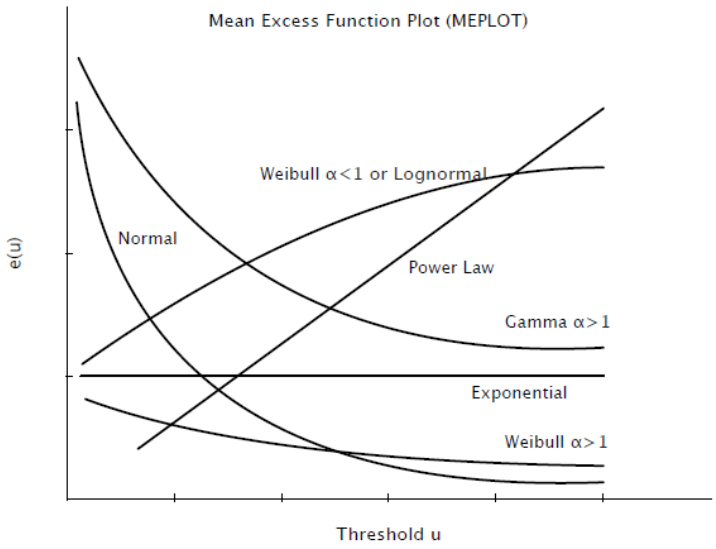


Figure 4.3: Shape of the mean excess function $e(u)$ for some classical distributions as a function of the threshold $u$ [22].

### The Discriminant Moment-ratio Plot

A moment-ratio plot is a graph in which a distribution is represented as a pair of standardized moments plotted on a single set of coordinated axes. In figure 4.4 some guidelines for the interpretation of these plots can be found.

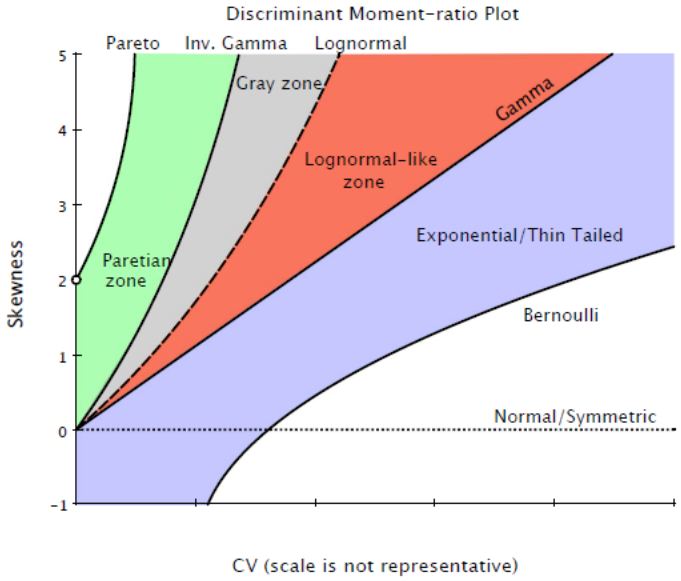


Figure 4.4: Guidelines for the interpretation of the Discriminant Moment-ratio Plot. Notice that the scale of the CV (coefficient of variation) axis is not representative, since it has been condensed [22].

For the Zipf and mean excess function plot, the corresponding figures can be made for the empirical data. Afterwards, the behaviour of the graphs can be studied and compared with figures 4.2 and 4.3. In this case it is suggested to use the point $(\gamma_2^*, \gamma_3^*)$ in the discriminant moment-ratio plot [22], where

$$y_2^* = \frac{\bar{X}}{\sigma_X^*} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \tag{4.3}$$

$$y_3^* = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_X^*}\right)^3. \tag{4.4}$$

The location of this point will indicate which underlying distribution is likely to be true. A description of what the location of $(\gamma_2^*, \gamma_3^*)$ indicates follows. If the point lies in the Lognormal-like zone then the underlying distribution is likely to be Lognormal or Gamma: the closer to the Lognormal line, the more likely it is Lognormal and vice versa for Gamma. Similarly, for the Paretian zone, the closer the point is to the Pareto line, the more likely the underlying distribution is Pareto I, and the more it moves away, the more likely that it is for example a Pareto II distribution. For the Exponential/thin tailed zone, the Weibull is a possible distribution, the Lognormal and the Pareto are very unlikely in this case. If the point falls in the grey zone then more analysis is

required (for example an analysis of the other plots). If the point is close to the Normal line, then a symmetric distribution, or a very thin tailed distribution, is likely to fit the observed data best. However, this is not likely for the loss distribution of a non-life insurance product.

### Maximum to Sum Plot

The maximum to sum plot indicates whether moment $p \in \mathbb{Z}_{>0}$ for the loss distribution is finite or not. This plot relies on the fact that, for a sequence $X_1, X_2, ..., X_n$ of nonnegative independent identically distributed random variables, if for $p = 1, 2, 3, ...$, $E[X^p] < \infty$, then $R_n = M_n^p / S_n^p \to 0$ as $n \to \infty$, where $S_n^p = \sum_{i=1}^n X_i^p$ and $M_n^p = max(X_1^p, X_n^p)$. This follows from the law of large numbers [24]. The maximum to sum plot is, in this case, given by plotting $R_n$ against $n$ and can indicate if Paretianity can be ruled out or not. Indeed, it can show the existence of the first four moments.

### Zenga Plot

The Zenga plot is based on the *Zenga curve* which can be expressed as

$$Z(u) = \frac{u - L(u)}{u[1 - L(u)]}, 0 < u < 1, \tag{4.5}$$

where $L(u) = \frac{1}{E[X]} \int_0^u F^{-1}(s)ds, u \in [0, 1]$, is the Lorenz curve. For the Zenga plot the behaviour of some distributions can be found in figure 4.5.

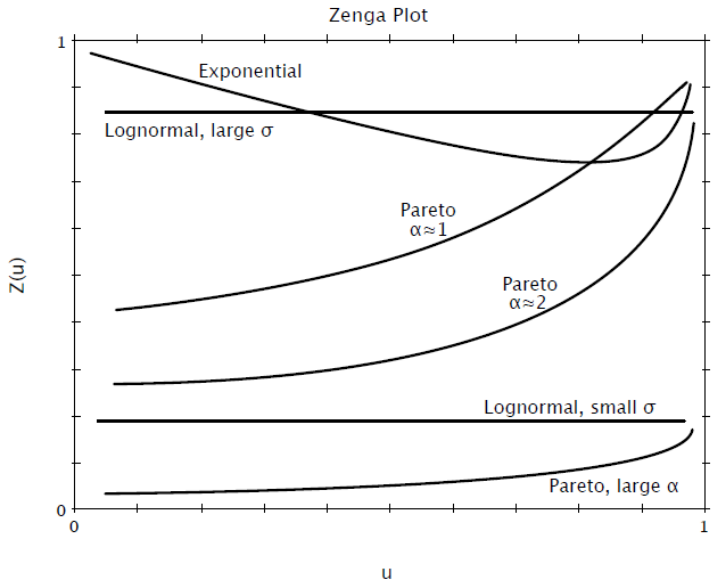


Figure 4.5: Zenga curve behaviour for some classical size distributions [22].

### QQ-Plot

A QQ-plot is a probability plot that compares two probability distributions by plotting their quantiles against each other. If the QQ-plot is similar, the point in the QQ-plot will approximately lie on the line $y = x$.

### Model Comparisons

Graphical methods have been discussed in this section to determine which distribution is most likely to fit the data. However, one can also observe the quality of the estimates of the model. The choice can be made to discriminate between models, where the distributions fitted on the data are both likely, but where the performance of the estimates are different, for example in the sense of standard error. If the accuracy of the estimates of the first model is much worse than for the second, then the second model can be the preferred model. Also other fitting criteria, which are discussed in section 4.6, can be used for this purpose. However, it is not preferable to choose the underlying distribution. Determining the underlying distribution by using the model comparison can lead to wrong decisions. This is due to computational issues for example. That is, the performance of a model is dependent on the algorithm used to calculate the values for the parameters.

### 4.2.2 Claim Frequency Modelling

The generalized linear model with a Poisson assumption is often chosen as the model for the number of claims for non-life insurance products [27]. Also, there are some elegant mathematical features for the Poisson distribution, such as the simplicity: there is only one parameter and the variance is equal to the mean. However, the Negative Binomial distribution can be preferred for the frequency model instead of the Poisson distribution when overdispersion is observed. Furthermore, the Poisson process can be *homogeneous* or *non-homogeneous*. The difference lies in the intensity. In the case of a homogeneous process the intensity is constant, and for the non-homogeneous case the intensity can be a function depending on, for example, time. This means that the non-homogeneous Poisson process, which has been used to model the number of medical malpractice claims in Italy [24], can better model phenomena like claim occurrences to be more likely to depend on the year. Furthermore, not only the Negative Binomial can be used, but also the *mixed* Poisson process when overdispersion is observed. In the mixed Poisson process the distribution is given by a mixture of Poisson processes, such that the process behaves like a homogeneous Poisson process [23].

## 4.3 Choice of Link Function

### 4.3.1 Limited Fluctuations Credibility

The canonical link-function has some elegant properties, but sometimes one may prefer a different link function [21]. In this section a method to compare models with different link functions is shown.
Let $\mu_i^*$ be the estimator of a parameter $m_i$. The *full credibility* is achieved when the distance between the estimator and the parameter is small enough with a large enough probability, that is

$$\mathbb{P}\{\mid \mu_i^* - m_i \mid \leq r m_i\} \geq \pi_i^{(g)}, \tag{4.6}$$

where $r \in (0, 1)$ is the estimation-error tolerance level and $\pi_i^{(g)}$ is the probability. Note that the tolerance level should be chosen.
The confidence coefficient $\pi_i^{(g)}$ indicates how credible the estimator is. That is, large values for the confidence coefficient indicate higher credibility of the estimates. As indicated, it depends on the link function $g$ and the covariates factor $\vec{x_i}$. For an arbitrary link function $g$, the confidence coefficient $\pi_i^{(g)}$ is given by

$$\pi_i^{(g)} = \Phi(\frac{Q_2}{s_i}) - \Phi(\frac{Q_1}{s_i}), \tag{4.7}$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution and $s_i$ is the standard deviation of the sum of the estimators. This is a function of the covariate vector(s) (there are multiple vectors in a mixed setting). Furthermore, $Q_1$ and $Q_2$ are given by

$$Q_1 = g[(1-r)m_i] - g(m_i) \tag{4.8}$$

and

$$Q_2 = g[(1+r)m_i] - g(m_i). \tag{4.9}$$

For the log-link the closed form with $Q_2 = ln(1+r)$ and $Q_1 = ln(1-r)$ is known [17]. Also it appears that the portfolio size influences the size of the confidence coefficient. Therefore, it is possible that the size of the portfolio, the number of observations, is too small to achieve a large enough confidence coefficient in a generalized linear (mixed) model setting.
A natural criterion for the choice of the link function is now

$$\pi_i^{*(g_1)} < \pi_i^{*(g_2)}, \tag{4.10}$$

where the estimator under the link function $g_1$ is less credible, and therefore that $g_2$ is a better choice in this sense. Note that $\pi_i^{*(g_1)}$ is notated, since, if the true parameter value $m_i$ is unknown, then $Q_1$ and $Q_2$ can be approximated by substituting the estimator $\mu_i^*$ for $m_i$.

## 4.4 Fixed or Random Factors

Choosing whether factors are added as random effects or as fixed effects is important in choosing whether a hierarchical generalized linear model or a generalized linear model should be chosen. Also, choosing which factors should be added as random effects or as fixed effects is important. This subject is under a lot of debate. Sometimes the reasoning not to use random effects is that if the covariates are correlated with the unit effects, there may be resulting bias in the parameter estimates. However, note that because there will in almost any case be some level of correlation between risk characteristics, there will be bias. Random effects can be a good fit for this correlation. Hence, as is often the case in statistical analysis, there should be a trade-off between how much bias is created against how much variance is introduced, by using fixed instead of random effects.
Often, the random effects will create more bias but less variance, which could lead to the estimate being closer to the real parameter value, depending also on how much data is provided. A little amount of data will generally lead to more variance for a model with fixed effects, especially if there are a lot of risk characteristics in the model, with a lot of factor levels. There should be an awareness of *sparse data*, which means that there is a greater number of parameters to be estimated than observations. However, apart from the problem with bias

created by the random effects, there is also a possible problem with the more complex structure and, with that, the more complex computational part. For this computational part, further explanation can be found in chapter 14.

### 4.4.1 Hausman test

The *Hausman test* can be used to decide whether a factor should be added as random or fixed. Although it is neither a necessary nor sufficient statistic for deciding between fixed and random effects [8], it is still a helpful tool in detecting whether estimates in the fixed effects are similar to estimates in the random effects. If this is indeed the case, then there is no to little correlation between the independent variable(s) and the unit effects. This test suggests a measure of the difference between two estimates:

$$H = (\mu^* - \beta^*)^T [Var(\beta^*) - Var(\mu^*)]^{-1} (\mu^* - \beta^*), \tag{4.11}$$

where $\beta^*$ is the estimate in the fixed effects model and $\mu^*$ is the estimate in the random effects model. After calculation of the measurement, the test uses the assumption of $H$ to be $\chi^2$ distributed, with the degrees of freedom equal to the number of regressors in the model. If $p<0.05$ then the random effects model in favour of the fixed effect model should be rejected since the two models are different enough. However, if $p>0.05$, then there is not such a clear interpretation. Hence, in that case there should be a further evaluation. In this thesis, a code is added for this test, which can be found in section 15.2.

## 4.5 Data Analysis

### 4.5.1 Variables

Before implementing models, it can be assessed which risk characteristics have a significant influence on, for example, the claim frequency and loss distribution. For example, one can make boxplots with on the horizontal x-axis the risk specifications and on the y-axis the logarithmic scale of the severities. Then one can see if the loss distributions for these risk specifications roughly follow the logarithmic distribution of the losses, when considering all observations.
It can happen that with a one-dimensional analysis one sees a significant impact of a risk factor on for example the claim frequency but that in combination with other risk characteristics, this impact vanishes. This can be explained by the correlation between risk characteristics. Therefore, the data analysis should not stop at a one-dimensional analysis. Tables can be made in which the claim frequency and the exposure is given for more risk characteristics and levels to see correlations.
Also, the *loss ratio* percentage can be determined, which is defined as 100 times

the total claim divided by the total premium. This can be done for every risk characteristic at all levels of the particular risk factor, but also for risk characteristics in combination with each other. When presented in a table, it can easily be judged at which risk level for the risk characteristic relatively less is paid (according to the corresponding claim total). If there is a percentage of more than 100, then there is a loss. In general, there is a loss when the combined ratio (which is the sum of the loss and cost ratio) is larger than 100 percent. This is undesirable. However, for other reasons, such as commercial reasons and keeping policyholders in the portfolio, it can be decided not to completely balance the loss ratio for each risk class at each level. Note that you actually overcharge another group in that case. If so, take into account that a 'wrong' rating system could lead to policyholders leaving, because they are overcharged while on the other hand it could attract bad risks. Tables can be made for the loss ratio percentage before fitting models and after, from which the premium can be determined from the fitted values. In this way, it can be determined whether the new premium is an improvement in the sense of balancing the losses.

### 4.5.2 In a Generalized Linear Model Tariff Setting

The amount of data typically gives an implication of how precise estimators are. Answering the questions on how much data is needed to get accurate enough estimates and on when estimates are accurate enough, are very important in practical situations. When the amount of data is truly too little, *credibility theory* can be used to calculate the premium instead of the (hierarchical) generalized linear models.

Assume that the estimates are calculated using a tariff calculation with a generalized linear model. A method to determine the number of claims that will be needed to get estimates that can be considered credible, depending on the number of risk characteristics and their levels, is as follows [16].
Let $c$ be the region around the true parameter value in which the estimate is allowed to lie and let $p$ be the probability that the estimate should be in the allowed region around the true parameter value. For example, suppose that the observations are assumed to be Poisson distributed and that the link function used is the canonical log-link. An estimate, which typically can be seen as the expected number of claims in a given time period for a particular risk profile, is given by $e^{\beta_1^*+\ldots+x_{ip}\beta_j^*}$. This estimate should not be lower than $(1-c)e^{\beta_1+\ldots+x_{ip}\beta_p}$ and not higher than $(1+c)e^{\beta_1+\ldots+x_{ip}\beta_j}$. This is equivalent to saying that the distribution of the exponent of the estimate, which follows approximately a Normal distribution, should not deviate from its expected value by more than $ln(1-c)$. The use of a $z$-score, $z_p$, seems now natural since it is defined by $\mathbb{P}\{\mid Z \mid \leq z_p\} = p$, where $Z$ follows the standard Normal distribution. Note that since the cumulative distribution function of a standard Normal distribution does not have an algebraic closed form, values are taken from well-known

tables. Now let $u$ be the reciprocal sum of the following elements. Let $a_i$ be the number of claims of risk level $i$ for risk characteristic $a \in \{1, ..., q\}$. Furthermore assume that the data are subdivided into $i \in \{1, ..., n\}$ risk levels for risk characteristic $a$. Let $j, k, r \in \{1, ..., n\}$ be such that $a_j \geq a_i$ and $a_k \leq a_r \leq a_i$ for all $i$. The elements that are needed for the calculation of $u$ consist of $a_k$, the element $\sum_{p \in R} a_p - a_q$, for all $q \in Q = \{1, ..., n\} \setminus \{j, k\}$, where $R = \{1, ..., n\}$, the previous elements should be determined for all risk characteristics $a \in \{1, ..., q\}$, and the element $\sum_{p \in R} a_p$. Note that $u$ is the upper bound of the variance of the estimate with the largest variance. Hence, if the inequality

$$u \leq \frac{ln(1-c)^2}{z_p^2} \tag{4.12}$$

is satisfied, the estimate is sufficiently precise. However, if this inequality does not hold, then $f$ is the factor by which each number of claims for a given risk profile should be multiplied to obtain sufficiently precise estimate and is given by

$$f = \frac{z_p^2 * u}{ln(1-c)^2}. \tag{4.13}$$

Often used constants for $p$, $c$ and $z_p$ are 0.95, 0.1 and 1.96, respectively.
Note that cells in which there is only a very small number of observed claims, have the biggest impact on how big $u$ is, indeed, these cells have a big impact on how precise the estimate is. However, in practice it might just be tolerated that some rather insignificant tariff segment is less accurately rated, provided that the accuracy of the important segments is sufficient. Testing if there is enough data to get accurate enough estimates without this particular cell can be done in this case. Also note that the advantage of this approach is that whether ones estimates are going to be precise enough can be evaluated without having to calculate estimates. Hence, this method can be applied before building the actual model.

## 4.6 Fitting Criteria

The following criteria can be taken into account after the implementation of the models, when the estimates have been determined. Note that this is also the case for the Hausman test.

### 4.6.1 Residuals

Suppose there are observation $y$ and fitted value $\mu$.

### Pearson Residuals

The *Pearson residual* is defined as

$$r^P = \frac{y - \mu}{\sigma}, \tag{4.14}$$

where $\sigma$ is the standard deviation, dependent on $\mu$ through $\sigma^2 = \phi V(\mu)$, with $V(.)$ the variance function; these residuals are simple, but often remarkably skewed.

### Deviance Residuals

This measure for the difference between vectors of fitted values and observations is widely used in the case of generalized linear models. It is defined as

$$r^D = sign(y - \mu)\sqrt{d}, \tag{4.15}$$

where $d$ is the contribution of the observation to the deviance (that is, the *likelihood ratio statistic*). Furthermore,

$$D = -2\phi log(\Lambda) \tag{4.16}$$

where $\Lambda$ is the likelihood ratio such that the maximized likelihood under the particular model is divided by the maximized likelihood of the *full* model. The *scaled deviance* is when both sides on the equations are divided by $\phi$.
In the case of *quasi-likelihood*, the *quasi-deviance* is given by

$$-2\phi log(\Lambda) = -2\phi q(\mu_1^*, ..., \mu_n^*), \tag{4.17}$$

where the resulting means are depending on the parameters of the systematic component.

### Analysis of Deviance

It is known that the scaled deviance is approximately $\chi^2$ distributed, with as degrees of freedom the number of observations minus the number of estimated parameters. Also, if one model is a sub model (*nested* model) of another model, it is known that the difference between the scaled deviances is a $\chi^2$-distribution. An example of a (sub) model A that is nested in model B is if the two models are the same but not in the sense that a factor of model B is replaced by a variate for model A. Recall that a variate was of form $i * \alpha$, where $i$ denotes the risk level and $\alpha$ is some real constant, hence when the value of $\alpha$ is determined, for every risk level the value is known. The choice of a factor results in the risk level $i$ being a label for the variable for that particular level. This means that for every risk level $i$ there can be arbitrary values, therefore, they can be the same as in the variate case, but it can also have other values. In this sense A is

nested in B. Another example could be, when for B some interactions between risk factors are allowed that are not allowed for model A. Model A is a *restricted* version of model B and B is the *relaxed* model, when A is nested in B.
The analysis used is that the nested/sub/restricted model will be chosen when the relaxed model is not a significantly better fit. The relaxed model is said not to be a significantly better fit if the gain in scaled deviance from choosing the restricted model exceeds the, say, 95 percent critical value of the $\chi^2(k)$ distribution (with the degrees of freedom equal to the extra parameters estimated that the relaxed model uses). The null-hypothesis that the extra parameters are actually equal to zero in the linear predictor is rejected or not rejected. The null-hypothesis is rejected if the gain in scaled deviance exceeds the critical value and it is not rejected if it does not. In other words, the relaxed model is chosen if the gain in scaled deviance exceeds the critical value and the restricted model is chosen if the gain in scaled deviance does not exceeds the critical value.
Note that this method cannot be used when models are not nested.

### 4.6.2 Information Criteria

To examine the balance of the complexity of the model together with the goodness of its fit, the information criteria can be used. The problem which often occurs is that a model has a very good fit but is not very useful because of its complexity, in this case choosing a less complex model with a less good fit can be a good idea, this trade-off has to be examined if the choice of the model is based on good reasoning. Note that in contrast with the analysis of deviance, the information criteria can be used to compare models that are not nested.

#### Akaike Information Criterion

The preferred model, which in the sense of the Akaike Information Criterion has the best balance, is the model with the lowest Akaike Information Criterion value. The Akaike Information Criterion is defined as

$$AIC = -2l + 2k, \tag{4.18}$$

where $k$ denotes the number of parameters and $l$ the maximized log-likelihood, both for the particular model.
The *relative likelihood* is strongly linked to the Akaike Information Criterion. Suppose that there are $R \in \mathbb{N}$ different candidate models and let the $j$th model (with $j \in \mathbb{N}_{\leq R}$) be the model that has the smallest value for the Akaike Information Criterion. The relative likelihood of model $i \in \mathbb{N}_{\leq R}$ is then given by

$$exp\{(AIC_{M_j} - AIC_{M_i})/2\}, \tag{4.19}$$

which can be interpreted as being proportional to the probability that the $i$th model minimizes the estimated information loss. In other words, suppose that

the value of the relative likelihood of model $i$ is given by $x \in \mathbb{R}$, then the $i$th model is $x$ times as probable as the $j$th model to minimize the expected information loss. If $x$ is very close to zero, say smaller than 0.05, then it is reasonable to say that model $i$ is omitted from further consideration. However, when $x$ is for example larger than 0.2 then it is not straightforward which model is in this sense better. Choosing between the $i$th and $j$th model has to be done by (also) using other fitting criteria in this case.
The sample size, proportional to the number of parameters used in the model, is important for the quality of using an Akaike Information Criterion test as described above. When the sample size, denoted by $n$, is not at least many times $k^2$, then the probability of the Akaike Information Criterion test leading to choosing a model that has too many parameters, in other words over fitting, can be significantly large. The $AICc$ has a greater penalty for a model having extra parameters, it is defined as

$$AICc = AIC + \frac{2(k+1)(k+2)}{n-k-2}. \tag{4.20}$$

The same test as for the Akaike Information Criterion can then be used.
Suppose that the full model is denoted by $F$ and a restricted model denoted by $R$. Then,

$$\Delta AIC = AIC_R - AIC_F \tag{4.21}$$

is based on the same statistical information as the $p$-value [20]. One can use this expression to compare or obtain decisions on which model to choose. A (certainly not unique and generally accepted) table of interpretations of $\Delta AIC$ is given in chapter 13.

### Bayesian Information Criterion

In the case of a Bayesian approach the Bayesian Information Criterion can be useful to give information about the balance between complexity and the fit of the model. Models with lower values of

$$BIC = -2l + log(n)k, \tag{4.22}$$

where $k$ denotes the number of parameters an $l$ the maximized log-likelihood of the particular model and $n$ the number of observations, in the sense of the Bayesian Information Criterion, have a better balance.
The *Bayes factor* is strongly linked with the Bayesian Information Criterion. Suppose that under model $M_1$, the prior density is given by $g_1(\theta)$ and the number of parameters is $p_1$, analogue for $g_2(\theta)$ and $p_2$ for model $M_2$, then the

following approximation can be made

$$B = \frac{\int_{\Theta_1} f(x;\theta)g_1(\theta)d\theta}{\int_{\Theta_2} f(x;\theta)g_2(\theta)d\theta}$$
$$\approx exp\{-\frac{1}{2}[-2log(\frac{sup_{\theta_1} f(x;\theta_1, M_1)}{sup_{\theta_2} f(x;\theta_2, M_2)}) - (p_2 - p_1)log(n)]\} = exp\{-\frac{1}{2}\Delta BIC\}. \tag{4.23}$$

This approximation is very useful because of the often complex form of the Bayes factor. The approximation is more accurate when the sample size is bigger.
A well-known table for the Bayes factor is given by Harold Jeffreys and can be found in chapter 13. Note that using the Bayes factor could lead to a different conclusion than when using the classical approach [1].

**Deviance Information Criterion**

The Deviance Information Criterion is a hierarchical modelling generalization of the Akaike Information Criterion and the Bayesian Information Criterion. It is especially useful in Bayesian model selection where the posterior distribution of the model is obtained by Markov Chain Monte Carlo simulation, which is explained in chapter 14. Models with lower values of

$$DIC = D(E[\theta]) + 2(E_\theta[D(\theta)] - D(E[\theta])) \tag{4.24}$$

are better in this sense.

### 4.6.3 Out of Sample Comparison of Models

To compare distinctively different models, tests for the predictive capability of a model when it is introduced to new data can be used. Two tests that are often used are the *Root Mean Squared Error*, which is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i^* - y_i)^2}{n}} \tag{4.25}$$

and therefore gives the mean squared error of prediction, and the Mean Absolute Deviance.
Some difficulties can occur with an out-of-sample comparison since data have to be generated or found. Deciding if the data are valid for the out-of-sample comparison can be difficult [18].

# Part II

# Practice

# Chapter 5

# Data Description

## 5.1 Overview of Chapter 5

In this chapter, the data structure of the third-party car insurance products will be investigated. Questions on what can be found in the data and which risk characteristics are analyzed are answered in this chapter. Before going into the next sections and chapters, general definitions are given. A *risk characteristic* is the general characteristic of a group of people that is investigated (for example, the age of the person). A *risk group* is the group of people that all belong to the same subdivision for this risk characteristic (for example, all people younger than 35). And a *risk level* is a subdivision of the risk characteristic (for example, the age of the person with an upper bound of 35).

## 5.2 Data Structure

There are two Excel-documents. The first document has insurance information, that is, the information about the product that is insured, the premium and the person that has the insurance. Let DOCV be this document. In the second document, one can find information about the claims, ranging from which product was the claim, who the owner of the product is, and the claim severity to the date of the claim.

### 5.2.1 DOCV

For the car insurance that is evaluated, the information that can be found in DOCV can be summarized as follows.

1. Information details of the person (or company) that is owner of the insurance product, such as relation number, gender, the Bonus-Malus class, date of birth and region where the person lives.
2. Information details of the insurance product for this particular policy code, such as premium, which product and when the product started and ended (if it ended).
3. Information details of the product that is insured, such as, the type of car, in which year the car was built and the weight of the car can be found here.

In table 5.1 there is an explanation of the risk characteristics that are investigated in the analysis, their abbreviations are also shown.

Table 5.1: The risk characteristics with the programming notations (which are also the abbreviations used) added. Explanations and more information can be found in chapter 7.

| Notation | Explanation |
|---|---|
| `BJ` | Age of the car. |
| `G` or `GEW` | Weight of the car. |
| `SVJ` or `SVJ.BM` | Bonus-Malus number of years. |
| `R` or REG | Region. |
| `GES` | Gender. |
| `VO` | Insured part. |
| `BRST` | Fuel used by the car. |
| `VERM` | Capacity of the car. |
| `JAREN.GN.SCHADE` | Years that there are no claims. |
| `TR.BM` | Bonus-Malus class. |
| `KM` | Mileage. |
| `ASS` | Assertivity of the car. |
| `LFD` | Age of the person. |

The covariate values, now in the form of dummy variables, are constructed as follows

$$x_{ip} = \begin{cases} 1, & \text{if the risk factor for } p \text{ is true} \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

For example, if the risk factor for $p$ is when the car is 'small', then the covariate just gives the value one in case the particular car falls in the defined category 'small'. This is done for all risk factors that are considered as possible interesting factors for the premium at all their levels. Then, the number of claims and insurances for each risk factor at each level is determined, not only to check if all contracts have been selected, but also to get an idea whether it is a good idea to include them.
Apart from making the covariate levels, the age is calculated, where the years

are integers and the months are divided by 12, which later on are combined to get a more accurate idea of the real age. This method is also used to calculate exposure factors for those policies that are only in force a part of the time.

### 5.2.2 DOCC

The claim data are in DOCC. Here the information about when the claim is made and how much the claim has cost the company can be found. Because DOCV has information about contracts of VO 48** (royal car insurances, see 13.5), the only claims that can be taken into account are claims on these contracts.
The data consider the policies that are not ended before 2016 and the claims can be found for these policies from 2007 until 2016. For a lot of claims in 2016 the losses are not reported. To avoid the consequences of recording delays, or even loss delays, the year 2016 is not taken into account. That is, also in the exposures.

## 5.3 Assumptions

Not only for the model used or the algorithm implemented assumptions are made, but also for the data. The data should be correct and complete. Assumptions are of the following kind.

1. The policy code for the claims and contracts are the same.
2. All claims for the given contracts are in the data for the given time period.
3. The given information, such as premium, date of birth, start-date and so on, are correct.

If these assumptions are not met, then the analysis can be in real danger.

# Chapter 6

# Distributions

## 6.1 Introduction

In this section the underlying distributions for the losses and number of claims are evaluated. Note that a Gamma or Lognormal distribution are often used for loss distributions of non-life insurance products. On the other hand, the Poisson distribution is often used for the frequency distribution. It is expected that also in this case these distributions fit the observed data well. However, just using these distributions, without investigating whether these distributions can model the observed losses and claim frequencies well, can lead to wrong indications of the risks. To investigate which theoretical distribution fits the losses and claim frequencies well, some comparison methods are used which are discussed in chapter 4.2. The distributions that seem to fit best will be chosen in making the models. The possibility that more than one distribution seems a good fit is certainly there. Although a distribution is chosen to build the models, there are still methods to evaluate if another distribution has a better fit, as explained in section 4.2.1. Please note that for the graphical comparisons the figures 4.2, 4.3, 4.4 and 4.5, for the Zipf, mean excess function, discriminant moment-ratio and Zenga plot, respectively, are used for the comparison with the theoretical distributions. The `R` codes that are used to make the plots can be found in section 15.5.

## 6.2 Loss Distributions

For the loss distribution, two distributions are evaluated. The distribution of the observed losses but also of the observed loss divided by the corresponding number of claims are investigated and compared. This choice is made since it is sometimes convenient, when the weight of the severity model, the number

of claims, is not statistical significant, not to take this variable into account. For example, when out-of-sample comparison with a generalized linear model assuming a distribution from the Tweedie family is done, it is convenient not to take the number of claims into account, since then a proxy needs to be developed [18]. Note that testing the statistical significance after the models have been built is still a wise decision.

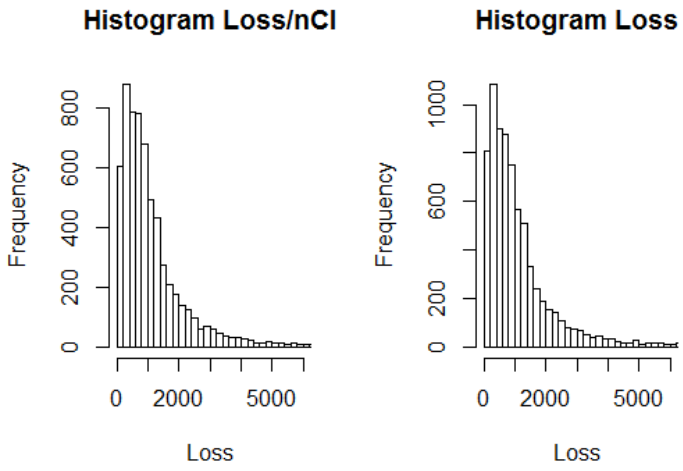


Figure 6.1: The histogram of the observed losses. The left-hand panel shows the plot for the observed losses divided by the number of claims and the right-hand panel for the observed losses.

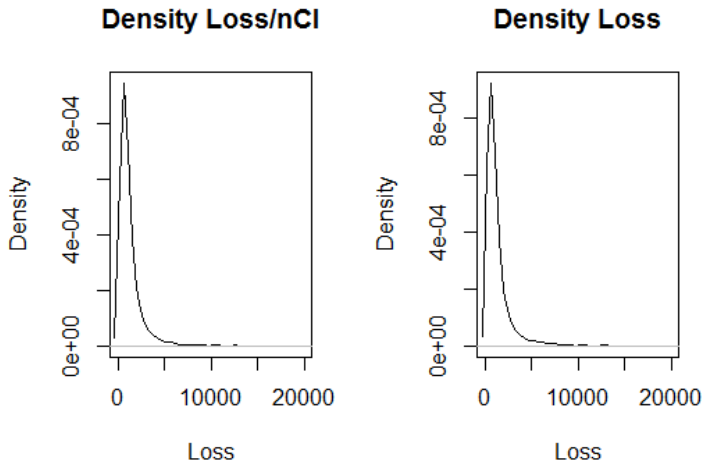


Figure 6.2: The density of the observed losses. The left-hand panel shows the plot for the observed losses divided by the number of claims and the right-hand panel for the observed losses.

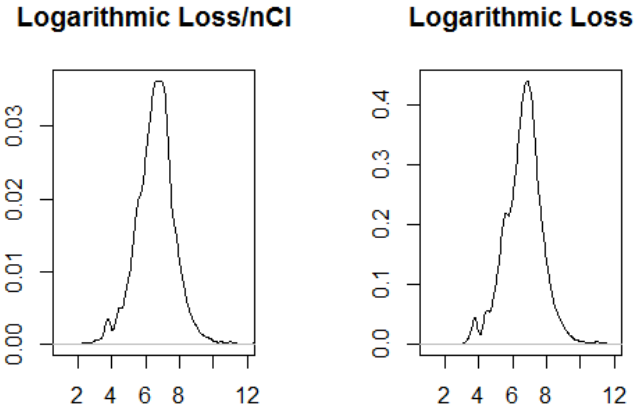


Figure 6.3: The density of the observed losses on a logarithmic scale. The left-hand panel shows the plot for the observed losses divided by the number of claims and the right-hand panel for the observed losses.

From figures 6.1 and 6.2 a general indication of the underlying distribution of

the observed losses (and observed losses divided by the number of claims since the graphs are very similar) can be made. A symmetric distribution, such as the Normal distribution, is probably not a good fit, since the right tail of the density is much larger than the Normal distribution will model. Also, the large difference between the median, 824.59, and the mean, 1743.71, of the losses indicate a non-symmetric distribution. On the other hand, the Pareto is probably too right-skewed, with too small a left tail. The Gamma or Lognormal distribution can probably fit the observed loss distribution better, this can also be observed in figure 6.3 (figure 4.1 indicates indeed a similar behaviour). However, these observations are not enough to be sure about the loss distribution. Therefore, based on the observed losses, some other figures are made.

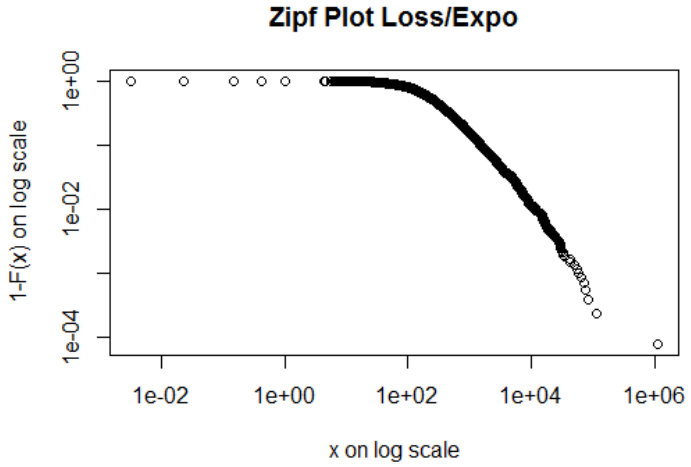


Figure 6.4: Zipf plot of the observed losses divided by the number of claims.

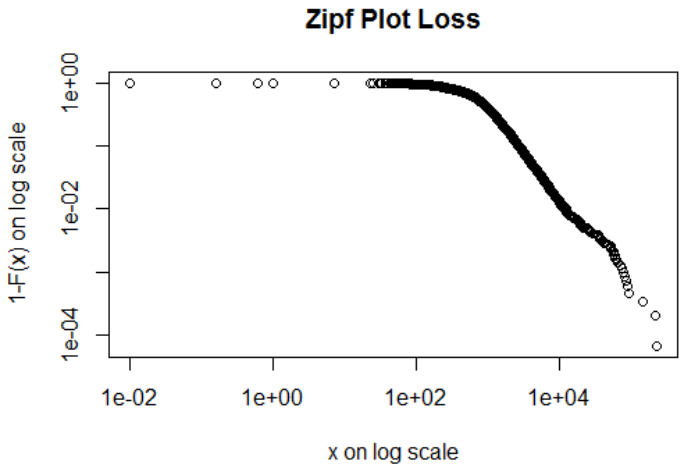


Figure 6.5: Zipf plot of the observed losses.

When figures 6.4 and 6.5 are given for the losses only between 1000 and 6000, one might think that a Pareto distribution is a good choice for the underlying distribution of the observed losses (and observed losses divided by the number of claims since the figures are very similar) since a linear trend is observed. However, as indicated in [22], it can be dangerous to use solely the Zipf plot when investigating the theoretical distribution that could be a good fit for the observations. Also, figure 6.4 does not show a linear trend in general. However, now discriminating, based on this plot, between the Normal and Lognormal (or Gamma $\alpha$>1) distributions and between the Exponential and Lognormal distributions can be difficult.

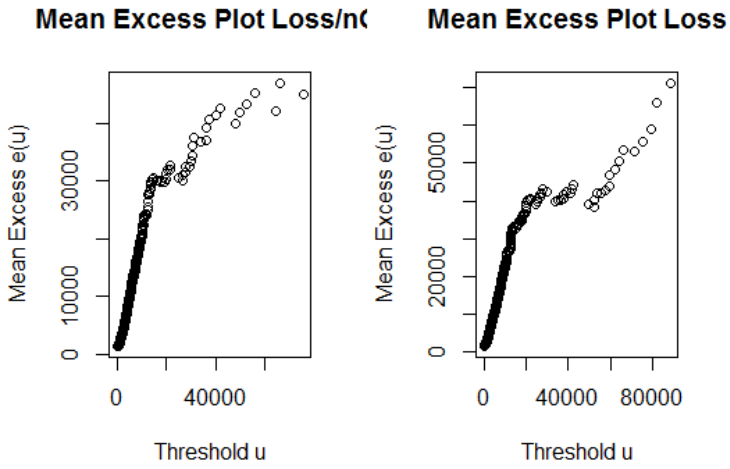


Figure 6.6: Mean excess function plot of the observed losses. The left-hand panel shows the plot for the observed losses divided by the number of claims and the right-hand panel for the observed losses.

Based on the first part of figure 6.6 (for example take the figure for threshold $u$<18000), there is a clear linear trend. Also, now the Pareto distribution might seem a better fit than the Lognormal or Gamma distribution. However, in general a very large sample is needed, more than 10000 observations, to observe a real concave trend [22]. In the used data set, less than 7500 losses can be found. Also, the linear trend does not continue for threshold $u$>20000. Hence, the linear trend at the beginning of the graph cannot be solely used.

In figure 6.7 the maximum to sum plot can be found for the first four moments. A convergence towards zero can be observed for all moments which suggests that the corresponding moments are finite. Hence, it not only indicates that the losses are not Pareto distributed, it also indicates that the necessary moments for the discriminant moment-ratio plot exist.

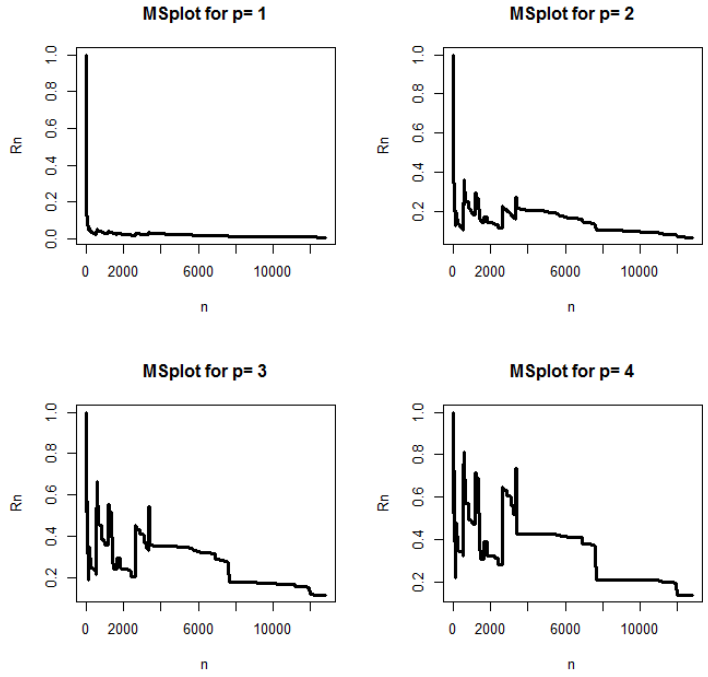


Figure 6.7: Maximum to sum plot for the observed losses (larger than 0) for the first four moments ($p = 1, ..., 4$).

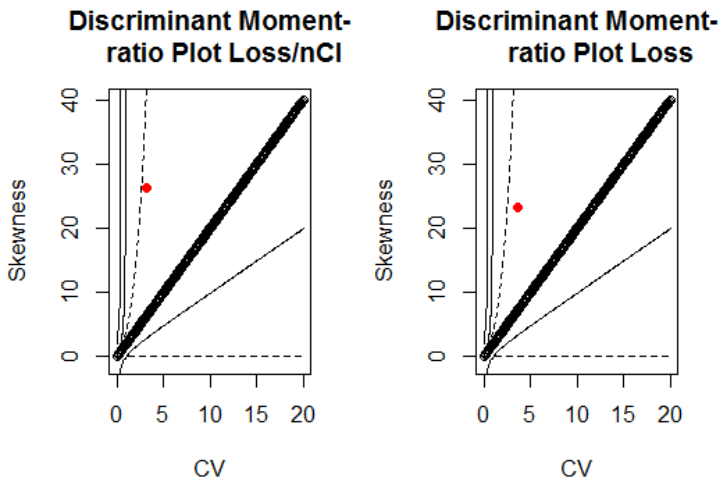


Figure 6.8: Discriminant moment-ratio plot where the red dot is based on the observed losses and the lines indicate the theoretical distributions. The left-hand panel shows the plot for the observed losses divided by the number of claims and the right-hand panel for the observed losses.

The red dot in figure 6.8 based on equations 4.3 and 4.4 lies in the Lognormal-like zone. Also, the point lies closer to the Lognormal line than to the Gamma line. Therefore, this figure discriminates in favour of the Lognormal distribution.

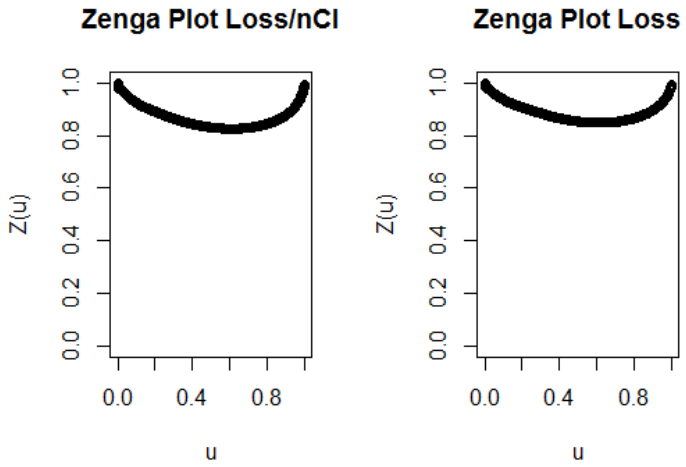


Figure 6.9: Zenga plot of the observed losses. The left-hand panel shows the plot for the observed losses divided by the number of claims and the right-hand panel for the observed losses.

Figure 6.9 indicate a Lognormal distribution with large $\sigma$ since the graph is almost constant.

The observations mentioned in this section lead to the conclusion that the Lognormal or Gamma distribution fit the observed losses (and the observed losses divided by the number of claims) probably pretty well in comparison to for example the Pareto, Normal or Exponential distribution. To discriminate between the Gamma and Lognormal distribution the discriminate moment-ratio plot has been used. The Lognormal seems to be a slightly better fit than the Gamma distribution. However, the Gamma distribution also seems to be a good choice.

## 6.3 Frequency Distribution

For the claim frequencies, often used distributions are the Poisson, and the Negative Binomial distribution. However, time-dependence models are also suggested [27].
It is expected that the number of claims would increase over time since the number of policies increase in the given data. And indeed, the number increases over the years as can be seen in table 6.1. However, it is not necessarily expected that the number of claims grows over the years if only the policies are considered that started before a certain year, and only the claims of these policies after a

certain year. Indeed, these observations can be made from table 6.1, that is, there is no clear trend for the increase of the number of claim over the years.

Table 6.1: The number of claims per year. The second row considers all data, the third row only considers the data with the policies that started before 2010 and the corresponding claims after 2010 and the fourth row only considers the data with the policies that started before 2012 and the corresponding claims after 2012. Notice that all policies were in force until 2016.

| year | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|
| All data | 8 | 122 | 189 | 309 | 322 | 447 | 734 | 1119 | 1737 |
| Data 2010 | | | | 276 | 213 | 213 | 222 | 230 | 265 |
| Data 2012 | | | | | | 332 | 366 | 342 | 444 |

Time dependency can also occur when the months of a year are considered. It can be expected that in months where there is more rain or ice, the claim frequencies increase. Also, depending on what exactly is insured, the claim frequency could be different in months where there are a lot of persons on vacation. From figures 6.10 and 6.11 the observations can be made that in August generally the least number of claims are made and in December the most.

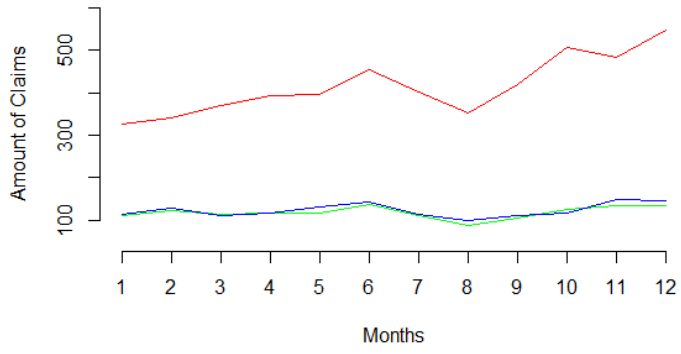


Figure 6.10: The number of claims per month. The red line is based on all data, the blue line on the data with the policies that started before 2012 and the corresponding claims after 2012 and the green line on the data with the policies that started before 2010 and the corresponding claims after 2010. Notice that all policies were in force until 2016.

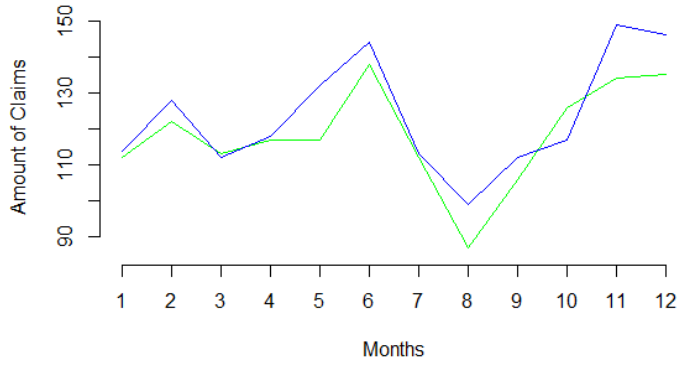


Figure 6.11: The number of claims per month. The blue line is based on the data with the policies that started before 2012 and the corresponding claims after 2012 and the green line on the data with the policies that started before 2010 and the corresponding claims after 2010. Notice that all policies were in force until 2016.

Another, important factor that can change the frequency model, more specifically the distribution that fits the observations well, is the dispersion factor. For the data the dispersion factor is 1.22, which means that there is overdispersion. For the Poisson distribution, the mean and variance are equal, the overdispersion implies that this is not the case. Using for example a Negative Binomial distribution might be a better option in this case.

# Chapter 7

# Risk Characteristics

## 7.1 Overview of Chapter 7

The risk characteristics that can be found in table 5.1 will be investigated in this chapter. That is, important values will be calculated to determine the risk profile of the risk levels of the risk characteristics. Also an explanation of the risk characteristics will be given. Furthermore, it will be evaluated if the amount of claims is sufficient, to determine if the fitted value will be accurate enough. Finally, decisions will be made on how the risk characteristics will be taken into account in the model building process. That is, which possibilities will be evaluated in the implementation of the models.

The information of this chapter can be considered as confidential information. Therefore, the outcomes are not presented.

# Chapter 8

# Generalized Linear Model Analysis

## 8.1 Overview of Chapter 8

In this chapter generalized linear models are build and analyzed. Before starting the generalized linear model analysis, note that the models are chosen using a bottom-up approach. Hence, the analysis is started with the null model and risk characteristics are added in order of how much they improve the model. In this way the intuition is clear, only the risk characteristic that improves the model most significantly is added. In each step the risk characteristic which improves the model most significantly should be evaluated due to correlations. More information about the tests and values used can be found in chapter 4. For the Akaike Information Criterion table 13.4 is used. Furthermore, the option of allowing interaction between risk characteristics is evaluated.

The information of this chapter can be considered as confidential information. Therefore, the outcomes are not presented.

# Chapter 9

# Other Models

## 9.1 Overview of Chapter 9

In chapter 8 a frequency-severity approach is fitted on the data. The two regression models are generalized linear models assuming a Poisson distribution with a log-link and a Gamma distribution with a log-link for the frequency and severity model respectively. The models have been fitted allowing interaction or not allowing interaction. In chapter 9 other models will be considered. Other generalized linear models will be considered, such as zero inflated models and a Tweedie regression model. Furthermore, hierarchical generalized linear models are considered. All these models are compared with the other models that are used to evaluate which model can be considered as the preferred model.

The information of this chapter can be considered as confidential information. Therefore, the outcomes are not presented.

# Chapter 10

# Expectations

Some expectations about outcomes have been presented. The purpose of this chapter is not to see the exact outcomes of the performed study. However, this chapter is based on outcomes that follow from the performed study. The purpose is to give a general idea of the outcomes regarding the difference of the claim frequencies, average losses and premium per general risk level when compared to the claim frequencies, average losses and premium when all data is taken into account, and compare them with expected outcomes that have been presented.

The information of this chapter can be considered as confidential information. Therefore, the outcomes are not presented.

# Chapter 11

# Conclusions

## 11.1 Recommendations

In this chapter, some recommendations will be made based on the outcomes of the study, how the premium is now determined and how to keep the company solvable. Note that some recommendations can be considered as confidential information. Therefore, these recommendations are not presented.

- The premium should be based on a sound line of reasoning for every part of the premium. One part should be based on the ability to pay the future claims and another part should be based on costs. Note that the part 'costs' can also be subdivided into two parts. One part is based on the costs of the company (processes). Another part is based on capital requirements.
- Start by applying statistical research on the products of the insurance company to develop a premium that is based on managing future liabilities.
- Apply statistical research on the products of the insurance company periodically to incorporate changes in expected future liabilities.
- Use a multiplicative formula to calculate the premium instead of an additive formula. The premium for the policyholders with a heavier car is more likely to be a percentage of the premium for policyholders with a lighter car when all other risk characteristics are the same.
- Change a parameter for a risk characteristic or risk level instead of changing one factor by which the whole formula is multiplied. This approach will lead to a far better diversification of the premium which will lead to less adverse selection and healthy profits for all groups, when based on statistical research.

- Implement the weight of the car and mileage as a continuous trend.
- Verify whether the claim frequency of insured part level 2 is indeed larger than the claim frequency of insured part level 1.
- Allow for interaction between risk groups. For example, if no interaction is allowed between the age of a driver, then the younger and the older driver will have a fixed percentage of additional claims when all other risk characteristics are the same. When interaction is allowed between the age of the person and the region, a younger driver will have a different fixed percentage of additional claims compared to an older driver in region 2 than in region 3. Therefore, allowing for interaction is preferred, when the data allows for this approach.
- Correlations should be taken into account. Basing the premium on a one-dimensional data analysis can, for example, lead to punishing or rewarding a policyholder twice, although this may not coincide with the expected future liability of this policyholder.
- A market based study should be based on a far-reaching diversification of risk characteristics to compare the obtained premiums from the market based study with the premiums obtained from the statistical study.
- Keep an eye on the actuarial developments regarding pricing. New techniques and data could lead to more accurate predictions of future liabilities.
- Make decisions on how to incorporate costs into the premiums. Base these decisions on the expected claim numbers and the Solvency (II) Capital Requirements.
- Make data more easily accessible and less polluted. Also, keep an eye on whether other data can be added to the already existing data.
- Regarding the pricing system that is now used for the third-party car insurance products, the following can quickly be changed. The premium for older policyholders should be higher than the premium for middle aged policyholders. Policyholders with a small number of Bonus-Malus number of years should generally pay more premium and policyholders with a larger number of Bonus-Malus number of years should generally pay less premium. The premium should increasingly decrease when the Bonus-Malus number of years increases. The same holds for mileage but the other way around (increasing instead of decreasing). The risk profile of the different insurance parts should be investigated again.

## 11.2 Summary

Here, the research questions will very briefly be answered.

Q: Which statistical models can be used to base the premiums on expected future liabilities?
A: Generalized linear models in general can be used for this purpose. Keep an eye on which variations can be made to improve the predictions. When too little data is available the Bühlmann-Straub model can be used.

Q: Are there enough data available to predict future liabilities accurately enough?
A: Based on the data of the third-party car products, generalized linear models can be implemented where quite a lot of risk characteristics are taken into account. However, more data will lead to better predictions.

Q: How can the 'best' model be chosen?
A: One can use for example the analysis-of-deviance, Akaike Information Criterion, coefficient of variation and $t$ and $z$ values correctly to make a trade-off between the complexity, accuracy and goodness-of-fit of the models. Also, comparing the outcomes of a model with what is expected based on the data is important.

Q: How can the models be implemented?
A: Statistical programming languages such as `R` and `SAS` are convenient to use for this purpose, as functions and algorithms that have been developed for this purpose are incorporated in these languages.

Q: What are the results when using these models for the third-party car products?
A: Results of the data and model analysis can be found in previous sections. For some models that have been implemented an order of preference has been made.
Models for the claim numbers in order of preference.

- Generalized linear model assuming the Negative Binomial distribution and a log link function with as risk factors the Bonus-Malus number of years (dynamically until 15), the weight of the car dynamically, mileage dynamically, age of the person as a variate and insured part as a factor. In addition, interaction between age and the Bonus-Malus number of years should be allowed.
- Generalized linear model assuming a Negative Binomial distribution and a log link function with as risk factors the Bonus-Malus number of years (dynamically until 15), the weight of the car dynamically, mileage dynamically, age of the person as a variate and insured part as a factor.
- A generalized linear model assuming the Poisson distribution instead of

the Negative Binomial distribution.

- A zero inflated model variant.
- A generalized linear mixed model variant.

Models for the losses in order of preference.

- Generalized linear model assuming the Gamma distribution and a log link model with Bonus-Malus number of years as a variate.
- Generalized linear model assuming the Gamma distribution and a log link without the number of claims as weights.

Furthermore, the Tweedie regression model is less preferred than the frequency-severity approach.

# Part III

# Appendix

# Chapter 12

# A, Statistical Background Information

## 12.1 Introduction

This thesis will, in the most general description, propose several models to predict future events based on historical data. The models will estimate parameters to make these predictions. Theory about which data should be or can be used and how this data behaves is important, but also which estimators will be used and how to compare these. The purpose of this chapter is to provide background information about models, statistics and estimators if they are unknown.

## 12.2 Estimator Criteria

### 12.2.1 Consistency

**Definition 12.2.1.** Let $\theta_n^*$ be an estimator of a parameter $\theta$ based on a sample of size $n$. Then $\theta_n^*$ is said to be *consistent* in probability if $\theta_n^*$ converges in probability to $\theta$ as $n$ approaches infinity; that is, for any $\epsilon>0$,

$$\mathbb{P}\{\mid \theta_n^* - \theta \mid > \epsilon\} \to 0, \text{ as } n \to \infty. \tag{12.1}$$

Nearly always it is a desirable property for a statistical estimator to be consistent. It also intuitively implies that the amount of non superfluous data for the estimator is important.

### 12.2.2 Bias

**Definition 12.2.2.** Let $\theta_n^*$ be an estimator of a parameter $\theta$ based on a sample of size $n$. Then $\theta_n^*$ is said to be *unbiased* if the bias, that is the difference of the expected value of the estimator and the true parameter value, of an estimator is zero; that is

$$E[\theta_n^*] - \theta = 0. \tag{12.2}$$

Intuitively one would say that it is a desirable property for an estimator to be unbiased, which is true in most cases. However, in some cases, like the Stein's paradox, there are estimators that are not unbiased but are better than the unbiased estimators.

### 12.2.3 Variance

The variance of an estimator is a measure of how much the estimator can differ from the expected value of the estimator. Intuitively a small variance is preferred, if the estimator is unbiased. However, often there is a trade-off between the bias and variance of an estimator. Also, the decision of which estimator should be used is difficult and not straight-forward. The sample size is an important factor in the decision. A large sample size often means that the unbiased estimator is preferred because the variance is probably quite small, but with a small sample size, one may decide (possibly with good reason) to choose an estimate with a small bias.
Under very general conditions, there is a lower bound on the variance for unbiased estimators. This is the so called Cramer-Rao Lower Bound. Hence, it is known that if the conditions are satisfied, no unbiased estimator can have a smaller variance than the Cramer-Rao Lower Bound. Hence, if an estimator has a lower variance than the Cramer-Rao Lower Bound, under these conditions, then it has to have a bias not equal to zero, and this estimator has a lower variance than all possible estimators that have no bias.

### 12.2.4 t- and p-value

Judging which parameters have a significant statistical influence can be done by using $t$ and $p$-values. Suppose that the null hypothesis is that a parameter does not have strong statistical significance.
The $t$-value measures the size of the difference relative to the variation in the sample data. That is, the $t$-value is simply the calculated difference represented in units of standard error. The greater the magnitude of the $t$-value (it can be either positive or negative), the greater the evidence against the null hypothesis that there is no significant difference.
In contrary, the $p$-value is the probability that the statistical summary (such as the sample mean difference between two compared groups) would be the same

as or more extreme than the actual observed results. The conclusions based on the $p$-value are opposite, the greater the magnitude of the $p$-value the greater the evidence to accept the null hypothesis.
The two values are strongly linked. For example, for large values of the $t$-value, the $p$-value is expected to be close to zero.
This section concludes with a mathematical definition. The scaled deviance, under a broad array of distributions of the data (such as the Poisson, Normal and Gamma distributions), will have a $\chi^2$ distribution with $k$ degrees of freedom for large sample sizes. The $p$-value is then given by

$$P = \mathbb{P}(\chi^2_k > \frac{D}{\phi}). \tag{12.3}$$

## 12.3 Statistics

A *statistic* is such that it only depends on the data, for example a particular function of the data that does not involve unknown parameters.

**Definition 12.3.1.** The statistic $T = T(X)$ is *sufficient* for $\theta$ if the distribution of $X$, conditional on $T(X) = t$, is independent of $\theta$.

However, there can still be information that is superfluous for the estimate. Therefore, the statistic that contains as little as possible information but is still sufficient is preferred. This is the *minimal sufficient statistic*.

**Definition 12.3.2.** A sufficient statistic $T(X)$ is *complete* if for any real function $g$,

$$E_\theta[g(T)] = 0 \text{ for all } \theta, \tag{12.4}$$

implies

$$\mathbb{P}_\theta\{g(T) = 0\} = 1 \text{ for all } \theta. \tag{12.5}$$

The Lehman-Scheffe theorem states that if there is a statistic which is sufficient and also complete then the statistic is also mimimal. If in this case some function of the statistic is an unbiased estimator, then it is also unique. The question that arises for the best unbiased estimator, is that of which unbiased estimator has the minimum variance. Given a convex loss function, there is a best unbiased estimator which is a function of the sufficient statistic, and that, if the sufficient statistic is also complete, this estimator is unique. If the loss function is the squared error loss, then this is equivalent to the Rao-Blackwell theorem on the existence of a minimum variance unbiased estimator [1]. However, there can be biased estimators that have less variance than the unbiased estimator with least possible variance. Unbiased estimators are therefore not always the best estimates.

## 12.4 Bayesian Modelling

The key conceptual point of Bayesian approach of inference is the way that the *prior* distribution $\pi(\theta)$ on the unknown parameter $\theta$, which is chosen a priory, is updated, on observing the realised value of the data $x$, to the *posterior* distribution, via Bayes' law, by

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x;\theta)}{\int_\Theta \pi(\theta') f(x;\theta') d\theta'}, \tag{12.6}$$

where $\Theta$ denotes the parameter space of $\theta$. Note that the function $f(x;\theta)$ is treated as a function of $\theta$ for fixed $x$ and is called the likelihood function. Inference about $\theta$ is then extracted from this posterior. The question about which prior distribution to choose arises immediately.
In the Bayesian statistics the observed data $X$ as well as the parameter $\theta$ (which you want to estimate) are both regarded as random variables, which implies that they both have their own distribution with their own density. For example, $X \mid \theta \sim N(\theta, 1)$ and $\theta \mid \tau^2 \sim N(0, \tau^2)$ is a Bayesian approach where the distribution of $\theta$ has been chosen and the parameter $\tau^2$ is, as in non-Bayesian statistics is customary, treated as a fixed parameter. If this is not the case then $\tau^2$ will have a distribution which first has to be evaluated, then the model is a so-called *hierarchical model*.
Furthermore, briefly stated is that *empirical Bayes* analysis is characterized by the estimation of prior parameter values from marginal distributions of data. Having estimated the prior parameter values, then proceed as if these values had been fixed at the beginning.
In a lot of cases, such as some special cases of hierarchical generalized linear models, there is in some sense, for example asymptotically, equivalence between empirical Bayes estimates and maximum likelihood estimates.

### 12.4.1 Hierarchical Modelling

In general hierarchical modelling is used if the assumption that the observations are independent of each other given the predictor variables may not be true. This assumption is made for generalized linear models and can be omitted if hierarchical modelling is used.
Point estimation with for example maximum likelihood estimation or restricted maximum likelihood estimation and Bayesian inference by choosing prior distributions of the parameters and compute the posterior distribution(s), can both be used for calculating values for the parameters in a hierarchical model.

### 12.4.2 Prior Distribution

When the Bayesian approach is chosen, the prior distribution has to be defined. Both fixed and random effects can have a different prior. Sometimes the prior

distribution belongs to a different distribution family than the posterior distribution. However, sometimes only the parameters have different values, in this case the common parametric form is called the *conjugate prior family*. Having a conjugate prior family can have several advantages such as a closed analytic form of the likelihood and more easy and accurate algorithmic approaches. For the prior their can be a density that does not integrate to 1, if integrated over the whole parameter space. If this is the case then the prior is *improper*. Furthermore, the prior can be *informative* and *uninformative*. Typically there are many nuance variations between the two. Strictly speaking, an informative prior gives specific, definite information about a variable (such as the distribution and its parameter values), and an uninformative prior expresses only vague or general information about a variable (such as equal probabilities to all possible outcomes). The question arises if the information given to the prior are based on facts or on subjective judgment.

## 12.5 Compound Poisson Distribution

The compound Poisson distribution can be used in cases where there is an interested for the claim sizes rather then the number of claims. Let $N$ be the number of claims payable by the insurer generated by a portfolio of insurance policies in a fixed time period and let $N$ follow a Poisson distribution with a constant parameter. Furthermore, assume that the claim size, denoted as $X_i, i = 1, ..., N$, are identically distributed and each $X_i$ is independent of $N$. If $Y$ denotes the aggregated size of the claims, that is $Y = X_1 + ... + X_N$, then $Y$ follows a compound Poisson distribution. This can also be applied to a certain group within the portfolio, characterized by risk factors.

**Lemma 12.5.1.** *Suppose that $Y_1, ..., Y_k$ are independent random variables such that $Y_i \sim compoundPoisson(\lambda_i)$, $i = 1, ..., k$. Let $Y = \sum_{i=1}^{k} Y_i$ then $Y \sim compoundPoisson(\sum_{i=1}^{k} \lambda_i)$.*

*Proof.* The moment generating function that belongs to $Y_i$ is given by $M_{Y_i}(t) = e^{\lambda_i(M_i(t)-1)}$. Then the moment generating function of the independent sum $Y = Y_1 + ... + Y_k$ is

$$M_Y(t) = \Pi_{i=0}^{k} e^{\lambda_i(M_i(t)-1)} = e^{\lambda[\sum_{i=0}^{k} \frac{\lambda_i}{\lambda} M_i(t)-1]}. \tag{12.7}$$

The moment generating function of the independent sum $Y$ has now the form of a $compoundPoisson(\sum_{i=1}^{k} \lambda_i)$. □

## 12.6 Tweedie Family

An exponential dispersion family is called a Tweedie Family if the domain of its variance function $V$ is $(0, \infty)$ with

$$V(\mu) = \mu^p, \tag{12.8}$$

for some $p \in \mathbb{R}$. The parameter $p$ is called the shape parameter.
The Tweedie family contains many distributions, characterized by the value $p$. For example, for $p = 0$ a Normal distribution is obtained, for $p = 1$ a Poisson distribution is obtained, for $p \in (1, 2)$ a Compound Poisson-Gamma distribution is obtained, for $p = 2$ a Gamma distribution is obtained and for $p = 3$ an inverse Gamma distribution is obtained [31].

The Compound Poisson-Gamma distribution is further considered since this distribution can be useful to obtain the pure premium when used in a generalized linear model.
Let $N$, $Y$ and $X_i$ be as defined in section 12.5. Assume $N$ to be $Poisson(\lambda)$ distributed and $X_i$ to be identically $Gamma(\alpha, \beta)$ distributed. Let $Y$ follow a Compound Poisson distribution with parameter $\lambda$. Suppose that the parameters satisfy $\lambda = \frac{\mu^{2-p}}{\phi(2-p)}$, $\alpha = \frac{2-p}{p-1}$ and $\frac{1}{\beta} = \phi(p-1)\mu^{p-1}$. The density function for $Tweedie(\lambda, \alpha, \beta)$ is given by

$$f_Y(y) = \sum_{n=1}^{\infty} \frac{\beta^{n\alpha} y^{n\alpha-1} e^{-\beta y}}{\Gamma(n\alpha)} \frac{\lambda^n e^{-\lambda}}{n!}. \tag{12.9}$$

It can be shown that this density function satisfies the definition of the density function of the exponential dispersion family [2].

## 12.7 Model for the Number of Claims

Suppose that the number of claims are modelled with a Poisson distribution in a generalized linear model setting with a log-link. Furthermore, let $N_i$ be the number of reported claims for policyholder $i$. Then the Poisson distribution comes down to

$$\mathbb{P}\{N_i = n_i\} = \frac{\lambda_{i,t}^{n_i}}{n_i!}, \tag{12.10}$$

where $\lambda_{i,t} = \exp(\sum_j x_{ij}\beta_j)$.
When the time dependence is included then, when assuming the following functional form of the intensity function

$$\lambda_i(t) = \delta t^{\delta-1} \exp(\sum_j x_{ij}\beta_j), \tag{12.11}$$

the following equation holds

$$\Lambda_i(t) = t^{\delta}\exp(\sum_j x_{ij}\beta_j), \tag{12.12}$$

since the cumulative intensity, at time $t$, is such that

$$\Lambda_i(t) = \int_0^t \lambda_i(u)du. \tag{12.13}$$

Suppose that the claim numbers are modelled with a Negative Binomial distribution in a generalized linear model setting with a log-link. Then the following equation holds

$$\mathbb{P}\{N_i = n_i\} = \frac{\Gamma(n_i + \alpha^{-1})}{\Gamma(n_i + 1)\Gamma(\alpha^{-1})}\Big(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\Big)^{n_i} \text{ x } \Big(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\Big)^{\alpha^{-1}}, \tag{12.14}$$

where $\lambda_i = \exp(\sum_j x_{ij}\beta_j)$ and the function $\Gamma$ is the gamma function [27].

## 12.8 Model for the Losses

Suppose that the Lognormal distribution seems to be the better fit and a generalized linear model will be implemented. Since the Normal distribution does belong to the exponential family and the Lognormal distribution does not, the logarithmic function is used. The model becomes a linear model on log-scale with an identity link. Then the expected value of the losses comes down to

$$\exp\big(\mu + \frac{\sigma^2}{2}\big), \tag{12.15}$$

where $\mu$ and $\sigma^2$ are the mean and variance of the Normal distribution. Hence, the estimated mean comes down to

$$\mu = \exp(\sum_j x_{ij}\beta_j + \frac{\sigma^2}{2}). \tag{12.16}$$

## 12.9 Laws of Large Numbers

First some definitions are given.

**Definition 12.9.1.** A sequence of random variables $\{Y_1, Y_2, ...\}$ is said to *converge in probability* to $a \in \mathbb{R}$ if, given $\epsilon$>0 and $\delta$>0, there exists a $n_0 \equiv n_0(\delta, \epsilon)$ such that, for all $n$>$n_0$,

$$\mathbb{P}\{\mid Y_n - a \mid > \epsilon\} < \delta. \tag{12.17}$$

Often $Y_n \xrightarrow{p} a$ is notated.

**Definition 12.9.2.** A sequence of random variables $\{Y_1, Y_2, ...\}$ is said to *converge almost surely* to $a \in \mathbb{R}$ if, given $\epsilon$>0 and $\delta$>0, there exists a $n_0 \equiv n_0(\delta, \epsilon)$ such that

$$\mathbb{P}\{\mid Y_n - a \mid > \epsilon \text{ for some } n > n_0\} < \delta. \tag{12.18}$$

Often $Y_n \overset{a.s.}{\to} a$ is notated.

**Definition 12.9.3.** A sequence of random variables $\{Y_1, Y_2, ...\}$ is said to *converge in distribution* if there exists a distribution function $F$ such that,

$$\lim_{n\to\infty} \mathbb{P}\{Y_n \leq y\} = F(y) \tag{12.19}$$

for all $y$ that are continuity points of the limiting distribution $F$. If $F$ is the distribution function of the random variable $Y$, then $Y_n \overset{d}{\to} Y$ is often notated.

Two law of large numbers, and afterwards the well-known Central Limit Theorem, will be stated. Proofs can be found in [3].

**Property 12.9.1.** *Let $X_1, X_2, ..., X_n$ be independent, identically distributed random variables with finite mean $\mu$. The strong law of large numbers says that the sequence of random variables $Y_n = n^{-1}(X_1 + ... + X_n)$ converges almost surely to $\mu$, if and only if $\mathbb{E}[\mid X_i \mid]$ is finite.*

**Property 12.9.2.** *Let $X_1, X_2, ..., X_n$ be independent, identically distributed random variables with finite mean $\mu$ and assume that $X_i$ has finite variance for all $i \in \{1, 2, ..., n\}$. The weak law of large numbers says that the sequence of random variables $Y_n = n^{-1}(X_1 + ... + X_n)$ converges in probability to $\mu$.*

**Theorem 12.9.3.** *Let $X_1, X_2, ..., X_n$ be independent, identically distributed random variables with finite mean $\mu$ and assume that $X_i$ has finite variance $\sigma^2$ for all $i \in \{1, 2, ..., n\}$. Furthermore, write $Y_n = n^{-1}(X_1 + ... + X_n)$. The Central Limit Theorem says that $Z_n = \sqrt{n}(Y_n - \mu)/\sigma$ converges in distribution to a random variable $Z$ having the standard Normal distribution $N(0, 1)$.*

# Chapter 13

# B, Tables

Table 13.1: Commonly used distributions and link functions with their, Bayesian conjugate, which can be fit by `hglm()`.

| Model name | $y \mid u$ distribution | Link $g(\mu)$ | $u$ distribution | Link $v(\mu)$ |
|---|---|---|---|---|
| Linear mixed model | Gaussian | identity | Gaussian | identity |
| Binomial conjugate | Binomial | logit | Beta | logit |
| Binomial GLMM | Binomial | logit | Gaussian | identity |
| Binomial frailty | Binomial | comp-log-log | Gamma | log |
| Poisson GLMM | Poisson | log | Gaussian | identity |
| Poisson conjugate | Poisson | log | Gamma | log |
| Gamma GLMM | Gamma | log | Gaussian | identity |
| Gamma conjugate | Gamma | inverse | Inverse-Gamma | inverse |
| Gamma-Gamma | Gamma | log | Gamma | log |

Table 13.2: Interpretation of the Bayes factor, the null-hypothesis is that model $M_1$ is more supported than model $M_2$.

| Bayes factor | Interpretation |
|---|---|
| $B>1$ | evidence supports $H_0$ |
| $1>B>10^{-\frac{1}{2}}$ | slight evidence against $H_0$ |
| $10^{-\frac{1}{2}}>B>10^{-1}$ | substantial evidence against $H_0$ |
| $10^{-1}>B>10^{-\frac{3}{2}}$ | strong evidence against $H_0$ |
| $10^{-\frac{3}{2}}>B>10^{-2}$ | very strong evidence against $H_0$ |
| $10^{-2}>B$ | decisive evidence against $H_0$ |

Table 13.3: The main classes of distributions in the generalized linear model exponential dispersion family, with the customary parameters as well as the $(\mu, \phi)$ and $(\theta, \phi)$ reparameterizations, and more properties

| Distribution | Density<br>Domain | $(\mu, \phi)$ reparameterization; canonical link<br>Canonical link $\theta(\mu)$<br>Variance function $V(\mu)$ | Cumulant function $b(\theta)$<br>$E[Y;\theta] = (\theta) = b'(\theta)$ |
|---|---|---|---|
| $N(\mu, \sigma^2)$ | $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ | $\phi = \sigma^2$<br>$\theta(\mu) = \mu$<br>$V(\mu) = 1$ | $\frac{\sigma^2}{2}$<br>$\theta$ |
| $Poisson(\mu)$ | $e^{-\mu}\frac{\mu^y}{y!}$<br>$y = 0, 1, 2, ...$ | $\phi = 1$<br>$\sigma(\mu) = log(\mu)$<br>$V(\mu) = \mu$ | $e^{\theta}$<br>$e^{\sigma}$ |
| $Poisson(\mu, \phi)$ | $e^{\frac{-\mu}{\phi}} \frac{(\frac{\mu}{\phi})^{\frac{y}{\phi}}}{(\frac{y}{\phi})!}$<br>$y = 0, \phi, 2\phi, ...$ | $\theta(\mu) = log(\mu)$<br>$V(\mu) = \mu$ | $e^{\theta}$<br>$e^{\theta}$ |
| $Binomial(m, p)$<br>($m \in \mathbb{N}$ fixed) | $\binom{m}{y} p^y (1-p)^{m-y}$<br>$y = 0, ..., m$ | $\mu = mp$; $\phi = 1$<br>$\theta(\mu) = log(\frac{\mu}{m-\mu})$<br>$V(\mu) = \mu(1 - \frac{\mu}{m})$ | $mlog(1 + e^{\theta})$<br>$\frac{me^{\theta}}{1+e^{\theta}}$ |
| $Negbin(r, p)$<br>($r$>0 fixed) | $\binom{r+y-1}{y} p^r (1-p)^y$<br>$y = 0, 1, ...$ | $\mu = \frac{r(1-p)}{p}$; $\phi = 1$<br>$\theta(\mu) = log(\frac{\mu}{r+\mu})$<br>$V(\mu) = \mu(1 + \frac{\mu}{r})$ | $-rlog(1 - e^{\theta}$<br>$\frac{re^{\theta}}{1-e^{\theta}}$ |
| $Gamma(\alpha, \beta)$ | $\frac{1}{\Gamma(\alpha)} \beta^{\alpha} y^{\alpha-1} e^{-\beta y}$<br>$y$>0 | $\mu = \frac{\alpha}{\beta}$; $\phi = \frac{1}{\alpha}$<br>$\theta(\mu) = -\frac{1}{\mu}$<br>$V(\mu) = \mu^2$ | $-log(-\theta)$<br>$-\frac{1}{\theta}$ |
| $IG(\alpha, \beta)$ | $\frac{\alpha y^{-\frac{3}{2}})}{\sqrt{2\pi\beta}} exp(\frac{-(\alpha-\beta y)^2}{2\beta y}$<br>$y$>0 | $\mu = \frac{\alpha}{\beta}$; $\phi = \frac{\beta}{\alpha^2}$<br>$\theta(\mu) = -\frac{1}{2\mu^2}$<br>$V(\mu) = \mu^3$ | $-\sqrt{-2\theta}$<br>$\frac{1}{\sqrt{-2\theta}}$ |
| $Tweedie(\lambda, \alpha, \beta)$<br>($\alpha$ fixed;<br>$p = \frac{\alpha+2}{\alpha+1}$) | $\sum_{n=1}^{\infty} \frac{\beta^{n\alpha} y^{n\alpha-1} e^{-\beta y}}{\Gamma(n\alpha)} \frac{\lambda^n e^{-\lambda}}{n!}$<br>for $y$>0;<br>$e^{-\lambda}$ for $y = 0$ | $\mu = \frac{\lambda\alpha}{\beta}$; $\phi = \frac{\alpha+1}{\beta}\mu^{1-p}$<br>$\theta(\mu) = \frac{\mu^{1-p}}{p-1}$<br>$V(\mu) = \mu^p$ | $\frac{[(1-p)\theta]^{(2-p)/(1-p)}}{2-p}$<br>$[(1-p)\theta]^{1/(1-p)}$ |

Table 13.4: Interpretation of $\Delta AIC$, $i$ is the reduced/restricted model.

| $AIC_i - AIC_j$ | Relative likelihood $(j : i)$ | Interpretation |
|---|---|---|
| between 0 and 2 | between 1 and 2.7 | substantial support for model $i$ |
| between 4 and 7 | between 7.4 and 33.1 | considerably less support for model $i$ |
| greater then 10 | greater then 148 | substantial support for model $j$ |

Table 13.5: Explanation of the risk characteristic insured part (VO). Note that * can be replaced by 1 until 4 and indicate the region. Also note that all insurances have a third-party insurance.

| Level | Code | Explanation |
|---|---|---|
| 1 | 48*1 | Royal car third-party insurance without casco cover. |
| 2 | 48*2 | Royal car third-party insurance with limited casco cover. |
| 3 | 48*3 | Royal car third-party insurance with full casco cover. |
| 4 | 48*4 | Royal car third-party insurance with limited casco and window cover. |
| 5 | 48*5 | Royal car third-party insurance with full casco and window cover. |

# Chapter 14

# C, Algorithms used by R

## 14.1 Frequentist Approach

### 14.1.1 Fitting Generalized Linear Models

The R function `glm()` can be applied for generalized linear models. When used appropriately it will try to compute maximum likelihood estimates.

#### Calculation of Maximum Likelihood Estimates

It uses the *iteratively reweighted least squares* method for this purpose. This method is an iteratively method in which each step involves solving a weighted least squares problem of the form

$$\beta^{(t+1)} = argmin_\beta \sum_{i=1}^{n} w_i(\beta^{(t)}) \mid y_i - f_i(\beta) \mid^2 . \tag{14.1}$$

This method however can only guarantee to converge to a local maximum. When finding a local maximum, it is possible that the method does not continue in searching for another maximum when for a small enough neighbourhood there is not a better value. Note that if the function is concave, then, for this function, there is an unique global maximum. Furthermore, since the function is concave the method will always find a better value in the neighbourhood of where it is at that particular point, unless it is at the global maximum. Therefore, for concave functions the method guarantees that it will give the global maximum. For the maximum likelihood estimates a value for which the likelihood function is maximum has to be found. Hence, if the likelihood function is concave, then this method will find the maximum likelihood estimate, which is also unique. Applying the canonical link function will guarantee a concave

likelihood function. A proof of this is as follows.
Since only generalized linear models are used, assume that the observations are independent random variables with a density in the exponential dispersion family. The density of the exponential family is given by

$$p(x \mid \theta) = h(x)exp(\theta^T \phi(x) - A(\theta)), \tag{14.2}$$

where $A(\theta)$ is the log-partition function. The following properties are known, which can be generalized to higher dimensions, $\frac{dA}{d\theta} = E[\phi(x)]$, $\frac{d^2A}{d\theta^2} = E[\phi^2(x)] - E[\phi(x)]^2 = var(\phi(x))$, $\Delta^2 A(\theta) = cov(\phi(x))$, $cov(\phi(x))$ is positive semidefinite, and $A(\theta)$ is convex. The likelihood function is given by

$$p(D \mid \theta) = [\Pi_{i=1}^N h(x_i)]exp(\theta^T [\sum_{i=1}^N \phi(x_i)] - NA(\theta)). \tag{14.3}$$

Thus the log-likelihood function is given by

$$log(p(D \mid \theta)) = \theta^T [\sum_{i=1}^N \phi(x_i) - NA(\theta) = \theta^T [\phi(D)] - NA(\theta). \tag{14.4}$$

Note that $\theta^T[\phi(D)]$ is linear in $\theta$. It can be concluded that $A(\theta)$ is concave. Hence, any link function $\eta(\theta)$ for which the whole log-likelihood is concave in $\theta$ (such as the canonical link function) will guarantee concavity for $A(\eta(\theta))$. A very important result is that when the canonical link function is used, the R function `glm()` will give as result the unique maximum likelihood estimates. Note that it was assumed that the maximum is not infinitely large. However, this can happen if for example complete separation occurs. If the maximum is infinitely large, then a warning will be showed that the fitted probabilities are numerically 0 or 1. The algorithm then did not converge.

#### Helpful Functions Applied to `glm()`

One function which helps to compare different (nested) models is `anova()`. This function can be applied to a generalized linear model and gives an analysis of deviance table. Also the function `summary()` can be applied to a generalized linear model. This function will give more information about errors, deviance residuals and degrees of freedom.

### 14.1.2 Fitting Generalized Linear Mixed Models

The `R` package `lme4` allows the use of the function `glmer()`, which can be applied for generalized linear mixed models. Under rather general conditions, the `glm()` function finds maximum likelihood estimates for generalized linear

models. Unfortunately, the calculation, or rather approximation, of maximum likelihood estimates for generalized linear mixed models is more complicated. The `glmer()` function uses adaptive Hermite-Gaussian quadrature and Laplace approximations. However, there are many reasons for these methods to fail.

### Laplace Method

The method used, unless specified differently, by `glmer()` is the *Laplace method*. The Laplace method approximates integrals of the form

$$\int e^{h(u)} du, \tag{14.5}$$

where $u$ is a $q$-dimensional vector and $h(u)$ is a sufficiently smooth function with a local maximum $u_0$ in its domain. The method uses a second order Taylor expansion of the exponential term in $u_0$

$$h(u) \approx h(u_0) + \frac{1}{2}(u - u_0)^T h''(u_0)(u - u_0). \tag{14.6}$$

Substituting this approximation for the exponent and approximating $u$ with a multivariate Gaussian distribution $N(u_0, (-h''(u_0))^{-1})$ leads to

$$\int e^{h(u)} du \approx (2\pi)^{\frac{q}{2}} \mid -h''(u_0) \mid^{-\frac{1}{2}} e^{h(u_0)}. \tag{14.7}$$

State

$$h(u) = log[f_{Y|U}(y \mid u)] + log[f_U(u)] \tag{14.8}$$

since the log-likelihood can be written as

$$\begin{aligned} l &= log[\int f_{Y|U}(y \mid u) f_U(u) du] \\ &= log[\int e^{log[f_{Y|U}(y|u)] + log[f_U(u)]} du]. \end{aligned} \tag{14.9}$$

Given the second order Taylor expansion, there is the need of calculating the second order derivative of the new defined $h(u)$.
Assume $u$ to be multivariate normal distributed, $u \sim N(0, D)$. Then the following equality holds

$$log(f_U) = -\frac{1}{2} u^T D^{-1} u - \frac{q}{2} log(2\pi) - \frac{1}{2} log(\mid D \mid). \tag{14.10}$$

Then the derivatives are given by $\frac{\partial log(f_U)}{\partial u} = -D^{-1}u$ and $\frac{\partial^2 log(f_U)}{\partial u \partial u^T} = -D^{-1}$.
Using the chain rule leads to the derivative

$$\begin{aligned} \frac{\partial log f_{Y|U}(y \mid u)}{\partial u} &= \frac{1}{\phi} \sum_i (y_i - u_i) \frac{1}{V(\mu_i) g'(\mu_i) z_i^T} \\ &= \frac{1}{\phi} Z^T W \Delta (y - u), \end{aligned} \tag{14.11}$$

where $W = [V(\mu_i)g'(\mu_i)^2]^{-1}$ and $\Delta = g'(\mu_i)$.
In order to find $u_0$ there is the need to solve

$$\frac{\partial h(u)}{u} = \frac{1}{\phi} Z^T W \Delta (y - u) - D^{-1} u = 0, \tag{14.12}$$

which is highly non-trivial as all factors involved except for $y$ are functions of $u$.
For the second order derivative the following expression holds

$$\frac{\partial^2 h(u)}{\partial u \partial u'} = \frac{1}{\phi}(-Z^T W \Delta \frac{\partial \mu}{\partial u^T} + Z^T \frac{\partial W}{\partial u^T} \Delta (y - u) - D^{-1}). \tag{14.13}$$

The second term of this last expression is ignored. Note that for a Poisson distribution of the observations it is 0 and in all other cases it has expectation 0. This expression is now substituted in the log-likelihood function

$$l \approx log(f_{Y|U}(y \mid u_0)) - \frac{1}{2} u_0^T D^{-1} u_0 - \frac{1}{2} log(\mid (\frac{1}{\phi} Z^T W Z D + I) D^{-1} \mid). \tag{14.14}$$

Then, the derivative with respect to $\beta$ is given by

$$\frac{\partial l}{\partial \beta} \approx \frac{1}{\phi} X^T W \Delta (y - u), \tag{14.15}$$

where $W$ is assumed to change negligibly with respect to $\beta$. The estimates of $\beta$ and $u$ are obtained by solving the equations

$$\begin{aligned} \frac{1}{\phi} X^T W \Delta (y - \mu) &= 0, \\ \frac{1}{\phi} Z^T W \Delta (y - \mu) &= D^{-1} u, \end{aligned} \tag{14.16}$$

which also arises if a *penalized quasi-likelihood* is used, where the quasi-likelihood term is augmented with a penalty term [18], that is,

$$log[f_{Y|U}(y \mid u)] - \frac{1}{2} u^T D^{-1} u. \tag{14.17}$$

There are now some possible characteristics given that could lead to failure of this method.

- The assumption $u \sim N(0, D)$ may not hold for the random effect.
- The assumption that the second term, for the second order derivative, can be ignored may be false.
- The assumption that $W$ varies negligibly with respect to $\beta$ may not be true.
- The approximation for the first order derivative may fail.

- The approximation for the second order derivative may fail.
- The approximation in the last step may fail.

In some cases approximations can be made better by using optimizer theory. Furthermore, the first assumption is very important, because random effects do not have to follow a Normal distribution. The other two assumptions, when the link function is chosen properly and the variance function has small values, should not be of too much problem.

### Non-adaptive Gauss-Hermite Quadrature

The *non-adaptive Gauss-Hermite quadrature* is an approximation technique for integrals that have the form

$$\int h(z)e^{-z^2}dz, \tag{14.18}$$

where $h(z)$ is an integrable function on $\mathbb{R}$ and sufficiently smooth, that is, at least twice differentiable. The non-adaptive Gauss-Hermite quadrature uses a weighted sum of order $Q$ for the approximation

$$\int h(z)e^{-z^2}dz \approx \sum_{i=1}^{Q} w_i h(z_i). \tag{14.19}$$

The Hermite polynomial of order $Q$, and $z_i$ the zero's corresponding to this order Hermite polynomial, is defined as

$$H_Q(z) = (-1)^Q e^{z^2}\frac{d^Q}{dz^Q}e^{-z^2}, \tag{14.20}$$

with corresponding weights

$$w_i = \frac{2^{Q-1}Q!\sqrt{\pi}}{Q^2[H_{Q-1}(z_i)]^2}. \tag{14.21}$$

Note that the method does not depend on the values of $h$ and is symmetric around 0. Since $h$ may have its weight elsewhere, the approximation is often very poor. Therefore, the adaptive Gauss-Hermite quadrature is proposed and used in `glmer()`.

### Adaptive Gauss-Hermite Quadrature

In this method the factor given by $e^{-z^2}$ is replaced by a Gaussian functions with suitable changes in the weights and approximation points.
Suppose $\phi(t;\mu,\sigma)$ to be the probability density function of the Normal distribution with mean $\mu$ and standard deviation $\sigma$. The integral $\int g(t)dt$ is going to

be approximated, where $g(t)$ is such that $g(t)>0$, sufficiently smooth and with a unique mode. Now $\int h(z)e^{-z^2}dz$ is replaced by

$$\int f(t)\phi(t;\mu,\sigma)dt. \tag{14.22}$$

The sampling nodes $z_i$ and $t_i$ are transformed according to the transformation from $e^{z_i}$ to $\phi(t;\mu,\sigma)$ which is given by

$$t_i = \mu + \sqrt{2}\sigma z_i. \tag{14.23}$$

Let $\mu^*$ be the mode of $g(t)$ and $\sigma^* = \frac{1}{\sqrt{j^*}}$ where $j^* = -\frac{\partial^2}{\partial t^2} log(g(t)) \mid_{t=\mu^*}$. Also, define $h(t) = \frac{g(t)}{\phi(t;\mu^*,\sigma^*)}$ then

$$\int g(t)dt = \int h(t)\phi(t;\mu^*,\sigma^*)dt. \tag{14.24}$$

Using the Gauss-Hermite quadrature gives

$$\int g(t)dt = \sqrt{2}\sigma^* \sum_{i=1}^{Q} w_i^* g(\mu^* + \sqrt{2}\sigma^* z_i), \tag{14.25}$$

where $w_i^* = w_i e^{z_i^2}$.

For generalized linear mixed model given a single random effects, the method is given below. Note that the adaptive Gauss-Hermite quadrature indeed does not allow to have many random effects. The effect can be considered as clustered over different groups. For every cluster $i$ there is a random effect $u_i$ that follows a Normal distribution with mean 0 and variance $\sigma^2$. The posterior mode of $u_i$ has to be determined, which depends on the factors $\beta$, $\phi$ and $\sigma$. Let $\beta^*$, $\phi^*$ and $\sigma^*$ be current estimate according to the definitions given before. Let $\mu_i^*$ be such that the posterior

$$f(y_i \mid u_i)f(u_i \mid \sigma^*) \propto f(u_i \mid y_i) \tag{14.26}$$

is maximized. Then $\mu_i^*$ can be used as the mode for $u_i$. The Gauss-Hermite quadrature gives the following approximation

$$\int f_{Y|U}(y_i \mid u_i)f_U(u_i)du_i \approx \sum_{l=1}^{Q} w_l^* [\Pi_{j=1}^{n_i} f_{Y|U}(y_{ij} z_l^*)], \tag{14.27}$$

where $n_i$ is the size of cluster $i$, $y_{ij}$ the $j$th element of cluster $i$ and $w_l^*$ the adaptive weights given by $w_l^* = \sqrt{2}\sigma_i^* w_l e^{z_l^2}\phi(z_l^*;0,1)$. Also, $x_{ij}^T\beta + \sigma z_l^*$ is the linear predictor $f_{Y|U}(y_{ij}z_l^*)$.

There are now some possible characteristics given that could lead to failure of this method.

- The assumption that the random effect follows a Normal distribution may not be true.
- The maximization of the mode may fail.
- Solutions may depend heavily on starting values for $\beta^*$, $\phi^*$ and $\sigma^*$.

In this case also, the optimization method is of big influence on the algorithm.

In general, assuming a generalized linear mixed model, with a small number of random effects, the adaptive Gauss-Hermite quadrature approach is recommended because of its accuracy [18]. In other cases, the Laplace method could be used which can lead to bad fits because of the large amount of approximations that have to be done.

### Solving Failures of `glmer()`

Unfortunately the `glmer()` function sometimes fails. One of the errors this function gives, is an error referring to scaling of variables. This error can be fixed relative easily, and more importantly, the outcome of the model will not change in a severe way when used properly on fixed effects. There is a numerical example provided in section 15.4 and a sketch of a proof outlined below.

Suppose that $X$ is the matrix of predictors and $X'$ is the matrix after rescaling one fixed effect, which is equivalent to changing one column in $X$. Furthermore, assume that $\mu$ and $\sigma$ is respectively the mean and standard deviation of this particular column. Then the column is transformed for a fixed column $j$ to $\frac{x_{ij}-\mu}{\sigma}$ for all $i \in \{1, ..., n\}$. Let $y^*$ be the fit where $X$ is used and $y^{*\prime}$ where $X'$ is used. It is sufficient to show that $y'$ does not differ substantially from $y^*$ since any rescaling of a column of $X$ gives a new valid matrix $X'$. Also, note that there are no distributional changes when rescaling from $X$ to $X'$ since there is no change in the value of $y$.

Assume that the estimates are obtained using the previously described Laplace method, where the specific denoted system of equations

$$\begin{aligned} \frac{1}{\phi} X^T W \Delta (y - \mu) &= 0 \\ \frac{1}{\phi} Z^T W \Delta (y - \mu) &= D^{-1} u \end{aligned} \tag{14.28}$$

is solved. Then the optimal values of $y^*$, $\beta^*$ and $u^*$ have been found and therefore, an optimal $\mu$ as function of $\beta^*$ and $u^*$. For a sketch of a proof with a restriction to linearity in $X$ and $\beta$, suppose that by transforming $X$ to $X'$, as described above, the optimality is lost. In the last equation the change of the matrix of predictors is only through $\mu_i = g^{-1}(x_i^T \beta + z_i^T u)$. Therefore, $\beta^*$ is changed to $\beta^{*\prime}$ if the original optimality of $\mu_i^*$ is regained, but this change

leads to previous optimality $\mu_i^*$ in the first equation. Moreover, the linear change in the particular column of $X$ has no effect as the equation equals 0 to the right.

Now assume that the estimates are obtained using the previously described adaptive Gauss-Hermite quadrature. To recall, the following approximation is made

$$\int f_{Y|U}(y_i \mid u_i) f_U(u_i) du_i \approx \sum_{l=1}^{Q} w_l^* [\Pi_{j=1}^{n_i} f_{Y|U}(y_{ij} z_l^*)]. \tag{14.29}$$

Note that $x_{ij}^T \beta + \sigma z_l^*$ is the linear predictor $f_{Y|U}(y_{ij} z_l^*)$. Using the same reasoning and the knowledge that a linear shift in one of the columns of $X$ will result in a linear adjustment of $\beta$, leading to the same fit since the prediction is done in a linear manner, there can be concluded that the scaling and centering has no result on the fit of the model.

Another error that can arise is due to convergence problems. These can sometimes be solved by rescaling. However, the use of another solution can be necessary. The use of different optimizers can be helpful and if this does not work, the number of iterations can be increased. For different optimizers and the choice of optimizers a code is provided in section 15.4. This code tries to find the possible reason of failure. For example, the singularities of the random effects parameter estimates are checked. If these are very close to zero it could lead to convergence issues. On the other hand, the derivative approximations may turn out to be bad and these can be replaced by approximations that may be better.

## 14.2 Bayesian Approach

### 14.2.1 Introduction

More information about Bayesian statistics can be found in section 12.4. The Bayesian approach leads often to difficulties in computation, especially in high dimensional cases. The reason for this is often the difficult normalising constant term that is needed to make the posterior density a proper density function. Approximating the integral is often needed, the Markov Chain Monte Carlo methods are widely used for this purpose.

### 14.2.2 Monte Carlo Methods

The *Monte Carlo* methods are methods in which random numbers are drawn to simulate a sample from the posterior distributions. These methods use computational algorithms known as *pseudo-random number generators* to obtain streams

of numbers, which look like independent, identically distributed uniform random numbers over (0,1). Afterwards, a variety of transformation techniques are used to convert these uniform numbers to any desired distribution.

### 14.2.3 Markov Chain

A *Markov Chain* is a stochastic process which satisfies the *Markov Property*. The Markov property refers to the memoryless property of a stochastic process. A stochastic process has the Markov property if the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the sequence of events that preceded it.

### 14.2.4 Markov Chain Monte Carlo

The *Markov Chain Monte Carlo* methods are well-known for (hierarchical) generalized linear models where the Bayesian approach is used. The parameters are now not estimated by using point estimation, such as maximum likelihood estimation but by choosing a prior distribution and adjust this to the data. A posterior distribution is obtained. The Monte Carlo part of the method denotes that the random variable is for many times simulated. Afterwards, probabilities from these simulated random numbers are calculated. This is done by simply looking at the proportion of numbers that is in a particular interval. Hence, it is very important to generate enough data to obtain a good estimation of the posterior distribution, but even then it is still inherently random. This last disadvantage is omitted by using Markov Chains. If the same limiting distribution is applied with the memoryless property of a Markov Chain, then, again when enough data have been generated, we have probably managed to take independent samples as if they are from the posterior distribution that we wanted to know. What is needed to make this method possible is, a test that decides if a generated number is or is not added to the list of generated number that estimates the distribution. This can depend on the expected distribution for example.
When a generalized linear model is fitted with a Bayesian approach, then the `R` function `bayesglm()` will assume a $t$-distribution for the prior, if another prior is used then this should be specified in this function. If a generalized linear mixed model or a hierarchical generalized linear model is implemented where the Bayesian conjugate distribution is used, then the `MCMCglmm` package in `R` and the `MCMCglmm()` function [9]. Then the deviance is calculated for the lowest level of the hierarchy, furthermore the Deviance Information Criterion is calculated. Also the software `WinBUGS` uses the Markov Chain Monte Carlo method.
Since `WinBUGS` is very flexible, the use of this software will be recommended. The models chosen can be constructed without the limitations of the assumption that for example `bayesglm()` makes for the prior distribution.

The following steps have to be made when using WinBUGS.

- State the Bayesian model.
- Develop a Markov Chain that has an expected distribution of the joint posterior distribution of interest.
- Run the chain until output converges in distribution to draws from the target distribution.
- Base inference regarding unidentified parameters in the simulated outcome of the model with successive iterations of the chain.

### 14.2.5 Markov Chain Monte Carlo methods

In this section, two Markov Chain Monte Carlo methods are explained, that WinBUGS uses. One of the two methods is the *Metropolis-Hastings algorithm*, the other is the *Gibbs sampler* which is a special case of the Metropolis-Hastings algorithm.

#### Metropolis-Hastings algorithm for a continuous state space

Suppose that the state space $\chi$ is continuous, often a subspace of $\mathbb{R}^d$ for some $d \in \mathbb{Z}_{>0}$. Assume that the true density of $X$ is $f(x)$, where $X$ is often written as $\theta$. Let $q(x, y)$ be a trial density such that $q(x, y) \geq 0$ and $\int_\chi q(x, y)dy = 1$, for all $x$. Also, an irreducibility condition is assumed, that is, from any starting point $X^{(0)} = x$, it should be possible to get arbitrarily close to any other point $x$, for which $f(y)>0$, in a finite number of steps, with positive probability. The steps below summarises the algorithm in the case of a continuous state space.

- Step 1: Start from an arbitrary $X^{(0)}$.
- Step 2: Given $X^{(n)} = x$, generate a trial value $Y = y$ from the probability denisty $q(x, y)$.
- Step 3: Define $\alpha = \min(\frac{f(y)q(x,y)}{f(x)q(x,y)}, 1)$. If $\alpha = 1$ then set $X^{(n+1)} = Y$. If $0<\alpha<1$ perform an auxiliary randomisation to accept $Y$ with probability $\alpha$. If $Y$ is accepted then $X^{(n+1)} = Y$; else $X^{(n+1)} = X^{(n)}$.
- Step 4: Replace $n$ by $n + 1$ and return to step 2.

If the state space is discrete then the steps above are similar. A more specified explanation of the steps in the and a proof of convergence of the Metropolis-Hastings algorithm in the discrete case can be found in [1].

**Gibbs sampler**

Suppose $\theta = (\theta_1, ..., \theta_d) \in \Theta \subset \mathbb{R}^d$. Define an arbitrary initial vector $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$. The following procedure describes one iteration of the Gibbs sampler.

- Step 1: Holding $\theta_2^{(0)}, ..., \theta_d^{(0)}$ fixed, generate a new value of $\theta_1$ conditional on $\theta_2 = \theta_0^{(0)}, ..., \theta_d = \theta_d^{(0)}$.
- Step 2: Generate a new value $\theta_2 = \theta_2^{(1)}$ from the conditional distribution given $\theta_1 = \theta_1^{(1)}, \theta_3 = \theta_3^{(0)}, ..., \theta_d = \theta_d^{(0)}, X = x$.
- Step 3 until d-1: The procedure is followed analogue to the first and second step.
- Step d: Generate a new value $\theta_d = \theta_d^{(1)}$ from the conditional distribution given $\theta_1 = \theta_1^{(1)}, ..., \theta_{d-1} = \theta_{d-1}^{(1)}, X = x$.

### 14.2.6 Convergence

For Markov Chain Monte Carlo processes, testing the convergence is very important. The convergence determines when it is secure to halt sampling and utilize the outcome to evaluate and estimate characteristics of the distribution of interest. There are several ways of testing the convergence. Some diagnostic procedures are *visual inspection of history plots*, *autocorrelation*, *Brook-Gelman-Rubin test* and *Monte Carlo Error Estimation*.

# Chapter 15

# D, R Codes

## 15.1 Credibility GLM

```
#Data example
RADATA <- read.table("H:\\vDesktop\\R Code & txt\\RATEST.txt",header=TRUE)

#suppose that Expo is the number of claims.
RADATA
#    Expo R A
#1    286 1 1
#2    333 2 1
#3    413 3 1
#4      4 4 1
#5   4100 1 2
#6   4848 2 2
#7   6920 3 2
#8    106 4 2
#9   1380 1 3
#10 1877 2 3
#11 2942 3 3
#12   34 4 3

attach(RADATA)

Claims.A <- rep(0,max(A))

###!Note: take corresponding columns in other data
for(i in 1:max(A)){
  Claims.A[i] <- sum(RADATA[which(RADATA[,3]==i),1])
}
```

```
Min.level.A <- which.min(Claims.A)
Max.level.A <- which.max(Claims.A)

Claims.R <- rep(0,max(R))

for(i in 1:max(R)){
  Claims.R[i] <- sum(RADATA[which(RADATA[,2]==i),1])
}

Min.level.R <- which.min(Claims.R)
Max.level.R <- which.max(Claims.R)

j.A <- rep(0,(max(A)-2))
for(i in 1:(max(A)-2)){
  j <- ifelse(Min.level.A != i & Max.level.A != i, i,
              ifelse(Min.level.A<max(A) & Max.level.A<max(A),
                     max(A), max(A)-1))
  j.A[i] <- j
}

j.R <- rep(0,(max(R)-2))
for(i in 1:(max(R)-2)){
  j <- ifelse(Min.level.R != i & Max.level.R != i, i,
              ifelse(Min.level.R<max(R) & Max.level.R<max(R),
                     max(R), max(R)-1))
  j.R[i] <- j
}


#filling Q
###!Note: this code is valid only when there are less
###then 5 factor levels for the risk classes
Length.Q <- ncol(RADATA)+max(A)-2+max(R)-2
Q <- rep(0,Length.Q)
Q[1] <- sum(RADATA[,1])
Q[2] <- min(Claims.A)
i <- 0 #if risk class has 1 or 2 levels
if(max(A)>2){ #if riskclass has 1 or 2 levels
  for(i in 1:(max(A)-2)){
    Q[2+i] <- min(Claims.A)+max(Claims.A)+ifelse(max(A) == 3, 0,
                  ifelse(max(A) == 4, Claims.A[j.A[i]], NA))
  }
}
Continue <- 2+i+1
Q[Continue] <- min(Claims.R)
```

```
i <- 0
if(max(R)>1){
  for(i in 1:(max(R)-2)){
    Q[Continue+i] <- min(Claims.R)+max(Claims.R)+ifelse(max(R) == 3, 0,
                          ifelse(max(R) == 4, Claims.R[j.R[i]], NA))
  }
}
#If another riskfactor included:
#Continue2 <- Continue+i+1
#Q[Continue2] <- min(..)
#i <- 0
#if en for loop
#writing again a for loop

Q
#[1] 23243  1036 17010   144 16185 17477

#upper bound of variance of the worst risk class combination
#Note: Lowest value of Q has most influence on how large u is
u_data <- sum(1/Q)

c <- 0.1
z_p <- 1.96
#max variance to be sufficiently precise
u_max <- (log(1-c)^2)/(z_p^2)

#testing how much difference there is
Dif <- u_data-u_max

Dif
#[1] 0.005240872
#little difference, the variance is a bit larger then for the
#accurate level that is chosen. This could be an accurate enough GLM.

#calculating how many more data is needed if variance is bigger
#(hence, accuracy is not sufficient)
if(Dif > 0){
  f <- (z_p^2*u_data)/(log(1-c)^2)
}

f
#[1] 2.813677
#Sample size needed (for all levels uniformly) is 2.813677
```

## 15.2 Hausman test

```
#Hausman test

phtest_glmer <- function (glmerMod, glmMod, ...) {
  coef.wi <- coef(glmMod)
  coef.re <- fixef(glmerMod) ## changed coef() to fixef() for glmer
  vcov.wi <- vcov(glmMod)
  vcov.re <- vcov(glmerMod)
  names.wi <- names(coef.wi)
  names.re <- names(coef.re)
  coef.h <- names.re[names.re %in% names.wi]
  dbeta <- coef.wi[coef.h] - coef.re[coef.h]
  df <- length(dbeta)
  dvcov <- vcov.re[coef.h, coef.h] - vcov.wi[coef.h, coef.h]
  stat <- abs(t(dbeta) %*% as.matrix(solve(dvcov)) %*% dbeta)
  pval <- pchisq(stat, df = df, lower.tail = FALSE)
  names(stat) <- "chisq"
  parameter <- df
  names(parameter) <- "df"
  alternative <- "one model is inconsistent"
  res <- list(statistic = stat, p.value = pval, parameter = parameter,
              method = "Hausman Test", alternative = alternative,
              data.name=deparse(getCall(glmerMod)$data))
  class(res) <- "htest"
  return(res)
}
```

## 15.3 Bühlmann-Straub model

The codes of the Bühlmann-Straub model are small adjustments of codes that can be found in [2].

```
#Credibility premium with B hlmann-Straub model
#comparing estimates also for theoretical (exact,
#since chosen) a, s2 and m and estimates

######
#first creating a theoretical data set
#####

# exposure for 5 years (claims over 5 years,
# contracts over 5 years)
```

```
# J is number of contracts
J <- 10

# K is exposure
K <- 5

# calculating exposure
j <- rep(1:J, each=K)

#exposure as a factor
j <- as.factor(j)

#m, a and s2 normally have to be estimated from the data
#see below for estimating from real data using (8.44)
m <- 100; a <- 100; s2 <- 64

#to generate everytime the same dataset
set.seed(6345789)

#natural weights, such as total premium
#if weights =1 then (homogeneous) B hlmann model
#now 'random' generated for each observations
w <- 0.50 + runif(J*K)

#generate the 'observations' such as number of claims
X <- m + rep(rnorm(J,0,sqrt(a)),each=K) +
            rnorm(J*K,0,sqrt(s2/w))

#Testing for homogeneous or heterogeneous portfolio
anova(lm(X~j,weight=w))
#If homogeneous then don't have to continue

#Output:
#Analysis of Variance Table
#
#Response: X
#Df Sum Sq Mean Sq F value   Pr(>F)
#j          9   5935  659.45  14.836 3.36e-10 ***
#  Residuals 40   1778   44.45
#---
#  Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1

#Conclusion:
#The probability of obtaining a larger F-value than the one
#we observed here is 3.36e-10, so the null-hypothesis that
```

```
#the group means are all equal is rejected at 5% level
#Hence: heterogeneous portfolio.

#calculate the parameters can be found (8.32)
w.js <- tapply(w,j,sum)
w.ss <- sum(w.js)
z.j <- 1/(1+s2/(a*w.js))
z.s <- sum(z.j)
X.jw <- tapply(X*w,j,sum)/w.js
X.ww <- sum(X.jw*w.js)/w.ss
X.zw <- sum(X.jw*z.j)/z.s

#calculating the premium (8.34)
pr.j <- z.j*X.jw+(1-z.j)*X.zw

###
#calculating estimates (real data) with (you need above)
#using (8.42), (8.43) and (8.44)
##

#calculate (8.42)
SSW <- sum(w*(X-X.jw[j])^2)

#calculate (8.43)
SSB <- sum(w.js*(X.jw-X.ww)^2)

#calculate (8.44)
m.tilde <- X.ww
s2.tilde <- SSW/J/(K-1)
a.tilde <- (SSB-(J-1)*s2.tilde)/(w.ss-sum(w.js^2)/w.ss)

#calculate from (8.32) some needed parameters
z.j.tilde <- 1/(1+s2.tilde/(a.tilde*w.js))
z.s.tilde <- sum(z.j.tilde)
X.zw.tilde <- sum(X.jw*z.j.tilde)/z.s.tilde

#calculating the new credibility premiums (vector,
#for each cell) from (8.34)
pr.j.tilde <- z.j.tilde*X.jw+(1-z.j.tilde)*X.zw.tilde

#pr.j and pr.j.tilde almost the same
#> pr.j
#1         2         3         4         5         6
#85.79192 107.13432 105.32094 109.84200  98.60685  90.32069
#7         8         9        10
#101.86342 110.93065  95.70341  82.54719
```

```
#> pr.j.tilde
#1         2         3         4         5         6
#85.18367 107.48218 105.69892 110.43350  98.59832  89.90267
#7         8         9        10
#102.01843 111.57450  95.58069  81.82776
```

## 15.4 Testing Scaling

```
####### File for testing the effect of scaling and centering on gl(m)m fits

library(lme4)

RADATA <- read.table("H:\\vDesktop\\R Code & txt\\RATEST.txt",header=T)

attach(RADATA)

#making factor levels (each level arbitrary)
RF <- as.factor(RADATA$R)
AF <- as.factor(RADATA$A)

glmmfitnAGQ <- glmer(Expo ~ R
                     + (1|A),family = poisson(link = "log"),
                     data = RADATA, nAGQ=25)

# Compare these fits to the scaled and centered version for 'year'

RAScaled <- cbind(RADATA, RegionScaled=scale(RADATA$R))
RACentrd <- cbind(RADATA, RegionCentrd=scale(RADATA$R, scale=FALSE))

# Rerun the glmm/glm fits
glmmfitnAGQSc <- glmer(Expo ~ RegionScaled
                       + (1|AF),family = poisson(link = "log"),
                       data = RAScaled, nAGQ=25)

glmmfitnAGQCt <- glmer(Expo ~ RegionCentrd #RegionCentrd
                       + (1|AF),family = poisson(link = "log"),
                       data = RACentrd,nAGQ = 25)

# Build the predictions
predictionTableGlmmnAGQ <- cbind(glmmfitnAGQ=predict(glmmfitnAGQ,
                                                     type="response"),
                                 glmmfitnAGQSc=predict(glmmfitnAGQSc,
                                                       type="response"),
                                 glmmfitnAGQCt=predict(glmmfitnAGQCt,
```

```
                                                    type="response"))

head(predictionTableGlmmnAGQ)
#  glmmfitnAGQ glmmfitnAGQSc glmmfitnAGQCt
#1    358.0366      358.0366      358.0366
#2    282.0187      282.0187      282.0187
#3    222.1408      222.1408      222.1408
#4    174.9761      174.9762      174.9762
#5   5513.9691     5513.9689     5513.9689
#6   4343.2498     4343.2497     4343.2497

#calculate the total squared error
avgDiffGlmmnAGQ <- c(sum((predictionTableGlmmnAGQ[, 2]
                          -predictionTableGlmmnAGQ[, 1])^2),
                     sum((predictionTableGlmmnAGQ[, 3]
                          -predictionTableGlmmnAGQ[, 1])^2))


print(avgDiffGlmmnAGQ)
#[1] 6.041523e-08 6.856661e-08
#the fitted values are almost equal.
```

## 15.5 Plots

These codes for the plots are small adjustments of codes that can be found in [22].

```
##Zipf Plot
zipfplot=function (data,type=    plot    ,title=T) {
  # type should be equal to     points     if you want to add the
  # Zipf Plot to an existing graph
  # With other strings or no string a new graph is created.
  # If title is set to be F, the title of the plot is not given.
  # This can be useful when embedding the Zipf plot into other
  # plots.
  data <- sort(as.numeric(data)) #sorting data
  y <- 1 - ppoints(data) #computing 1-F(x)
  if (type==    points    ){
    points(data, y, xlog=T, ylog=T, xlab = "x on log scale",
           ylab = "1-F(x) on log scale")}
  else{
    if (title==F) {plot(data, y, log="xy", xlab = "x on log scale",
                        ylab = "1-F(x) on log scale")}
```

```
    else {plot(data, y, log="xy", xlab = "x on log scale",
               ylab = "1-F(x) on log scale", main= Zipf  Plot )}
  }
}

# In cut you can specify the number of maxima you want to exclude.
# The standard value is 5
meplot=function(data,cut=5) {
  data=sort(as.numeric(data));
  n=length(data);
  mex=c();
  for (i in 1:n) {
    mex[i]=mean(data[data>data[i]])-data[i];
  }
  data_out=data[1:(n-cut)];
  mex_out=mex[1:(n-cut)];
  plot(data_out,mex_out,xlab="Threshold u", ylab="Mean Excess e(u)",
       main="Mean Excess Plot Loss")
}

##Zenga plot
zengaplot=function(data){
  # Since the code relies on the Lorenz curve
  # as computed by the "ineq" library,
  # we upload it
  library(ineq)
  # Empirical Lorenz
  est=Lc(data)
  # Zenga curve
  Zu=(est$p-est$L)/(est$p*(1-est$L))
  # We rescale the first and the last point for
  # graphical reasons
  Zu[1]=Zu[2]; Zu[length(Zu)]=Zu[(length(Zu)-1)]
  # Heres the plot
  plot(est$p,Zu,xlab="u",ylab="Z(u)",ylim=c(0,1), main="Zenga Plot Loss",lty=1)
}


##Moment plot
moment_plot=function(data,i){
  # "data" is a vector containing the sample data
  ################################################
  ################################################
  # CV and Skewness functions
  coefvar=function(data){
    CV=sd(data)/mean(data)
```

```
  CV}
skewness=function(data) {
  m_3 <- mean((data-mean(data))^3)
  skew <- m_3/(sd(data)^3)
  skew}
##################################################
##################################################
# Computation of CV and Skewness
# CV
CV=coefvar(data);
# Skewness
skew=skewness(data)
# Rule of Thumb
if (CV<0 || skew <0.15){print("Possibly neither Pareto
                                nor lognormal. Thin tails."); stop}
##################################################
# Preparation of the plot
##################################################
# Paretian Area
# The upper limit - Pareto I
if(i==1){
p=seq(3.001,400,length.out=250)
g2brup=1/(sqrt(p*(p-2)))
g3brup=(1+p)/(p-3)*2/(sqrt(1-2/p))
# The lower limit, corresponding to the Inverted Gamma
g2ibup=seq(0.001,0.999,length.out=250)
g3ibup=4*g2ibup/(1-g2ibup^2)
##################################################
# Lognormal area
# Upper limit: Lognormal
w=seq(1.01,20,length.out=250)
g2log=sqrt(w-1)
g3log=(w+2)*sqrt(w-1)
# Lower limit - Gamma
g2iblow=seq(0,20,length.out=250)
g3iblow=2*g2iblow
##################################################
#Exponential Area
# The upper limit corresponds to the lower limit of the
# lognormal area
# The lower limit - Bernoulli
g2below=seq(0,20,length.out=250)
g3below=g2below-1/g2below
##################################################
# The Gray area is obtained for free from
# the previous lines of code.
```

```
#####################################################
# Normal / Symmetric distribution
g2nor=seq(0,20,length.out=250)
g3nor=rep(0,250)
#####################################################
# PLOT
# Limits
plot(g2iblow,g3iblow,xlab="CV",ylab="Skewness",main="Discriminant Moment-
     ratio Plot Loss",xlim=c(0,20),ylim=c(-1,40)) #,    l
lines(g2ibup,g3ibup)#,    l
lines(g2brup,g3brup)#,    l
lines(g2below,g3below)#,    l
lines(g2log,g3log,lty=2) # Lognormal
lines(g2nor,g3nor,lty=2) # Normal
# Strictly Paretian Area
polygon(c(g2ibup,g2brup),c(g3ibup,g3brup))#,col=    green
points(0,2,pch=1,cex=0.8) # Pareto limit point
}
# Hints for interpretation
'/
text(-0.2,20,cex=0.8,srt=90,"Pareto I")
text(1.2,20,cex=0.8,srt=90,"Inverted Gamma")
text(2.5,12,cex=0.8,srt=70,"Lognormal")
text(12,21,cex=0.8,srt=23,"Gamma")
text(14,11,cex=0.8,srt=10,"Bernoulli")
text(15,1.5,cex=0.8,"Normal or Symmetric")
/'
if(i==1){
points(CV,skew,pch=16,col="red")#
}
if(i==2){
  points(CV,skew,pch=16,col="black")
}
if(i==3){
  points(CV,skew,pch=16,col="purple")
}
if(i==4){
  points(CV,skew,pch=16,col="brown")
}
if(i==5){
  points(CV,skew,pch=16,col="green")
}
if(i==6){
  points(CV,skew,pch=16,col="blue")
}
  return(c(CV,skew))
```

```
}
```

## 15.6 Codes for Chapter 7

### 15.6.1 Figures and Tables

```
#Function to obtain Number of Contracts, Total Exposure, Number of Claims,
#Claim Frequency, Average Loss, Minimum Loss, First Quantile, Median,
#Third Quantile and Maximum Loss
TABLES <- function(Loss, LengthContracts, TotalExposure){
  Qs <- quantile(Loss)
  Ml <- mean(Loss)
  Ncl <- length(Loss)
  Ncontracts <- LengthContracts
  TExpo <- TotalExposure
  Mcl <- length(Loss)/TotalExposure*100
  return(rbind2(c("Number of Contracts","Total Exposure","Number of Claims",
  "Claim Frequency (%)","Average Loss","Minimum Loss","First Quantile",
  "Median","Third Quantile","Maximum Loss"),
          c(Ncontracts,TExpo,Ncl,Mcl,Ml,
          Qs[[1]],Qs[[2]],Qs[[3]],Qs[[4]],Qs[[5]])))
}

FrVGEW <- rep(0,6)
x <- rep(0,6)
y <- rep(0,6)
z <- rep(0,6)
q <- rep(0,6)

#GEW6
#subset for the sixth weight level
GEW6 <- subset(WAMdf,WAMdf$GEW>=1500)

#obtaining all losses for every claim separately
SeverityData1 <- subset(GEW6,GEW6$BETAALD1>0)
SeverityData2 <- subset(GEW6,GEW6$BETAALD2>0)
SeverityData3 <- subset(GEW6,GEW6$BETAALD3>0)
SeverityData4 <- subset(GEW6,GEW6$BETAALD4>0)
SeverityData5 <- subset(GEW6,GEW6$BETAALD5>0)

#all claims in a vector
CL <- c(SeverityData1$BETAALD1,SeverityData2$BETAALD2,SeverityData3$BETAALD3,
        SeverityData4$BETAALD4,SeverityData5$BETAALD5)
```

```
#generating outcomes
TABLES(Loss=CL, LengthContracts = length(GEW6$DEKKING.ID),
    TotalExposure = sum(GEW6$EXPO))

#vectors to make the graphs where the trends can be seen
x[6] <- TABLES(Loss=CL, LengthContracts = length(GEW6$DEKKING.ID),
    TotalExposure = sum(GEW6$EXPO))[2,5]
y[6] <- TABLES(Loss=CL, LengthContracts = length(GEW6$DEKKING.ID),
    TotalExposure = sum(GEW6$EXPO))[2,8]
z[6] <- TABLES(Loss=CL, LengthContracts = length(GEW6$DEKKING.ID),
    TotalExposure = sum(GEW6$EXPO))[2,7]
q[6] <- TABLES(Loss=CL, LengthContracts = length(GEW6$DEKKING.ID),
    TotalExposure = sum(GEW6$EXPO))[2,9]

#trend of average, median and first and third quantile for the losses
plot(log(as.numeric(x)), type="l", xlab='Risk Levels',
    ylab='Logarithmic Loss', col="green", ylim=c(5,8))
lines(log(as.numeric(y)), col="blue")
lines(log(as.numeric(z)), col="red")
lines(log(as.numeric(q)), col="black")

#vector for the frequency
FrVGEW[6] <- length(CL)/sum(GEW6$EXPO)*100

#two plots next to each other
par(old.par)
old.par <- par(mfrow=c(1, 2))

#plot for the claim frequencies
plot(FrVGEW, type="l", xlab='Risk Levels', ylab='Claim Frequencies',
    col="green", ylim=c(2,4))
lines(FrVBJ, col="blue") #age of the car
lines(FrVVERM, col="red") #capacity of the car
lines(FrVLFD, col="black") #age of the person
lines(FrVKM, col="darkorange") #mileage
lines(FrVASS, col="hotpink") #assertivity of the car
#new plot for the Bonus-Malus risk characteristics
plot(FrVTR.BM, type="l", xlab='Risk Levels', ylab='Claim Frequencies',
    col="purple", ylim=c(0,13))
lines(FrVSVJ, col="brown")

#adding sixth point for discriminant moment-ratio plot
moment_plot(data=CL, i=6)

#generating the loss ratio and income
```

```
#note that the premium can be found in another dataset
X <- GebPR$dek.id %in% GEW6$DEKKING.ID
X <- which(X == TRUE)
sum(GEW6$TOTSEV)/sum(GebPR$geboekte.premie[X])*100
sum(GebPR$geboekte.premie[X])

#claim frequencies and the mean of the weight
m6 <- sum(length(CL))/sum(GEW6$EXPO)*100
L6 <- log(mean(GEW6$GEW))

#vector with claim frequency and mean for all weight levels
L <- c(L1,L2,L3,L4,L5,L6)
m <- c(m1,m2,m3,m4,m5,m6)

#obtaining the plot with the median against the weight of the car
#and the claim frequency against the weight of the car, all on log scale
plot(L,log(m), ylab="Claim frequency on log scale",
     xlab="Weight of the car on log scale",type="o")
plot(L,log(as.numeric(y)),ylab="Median on log scale",
     xlab="Weight of the car on log scale",type="o")
```

### 15.6.2 Risk Characteristics

```
#Capacity of the car, variate and factor approach
VERM.F <- function(){
  BRSTL <- rep(0,length(WAMdf$DEKKING.ID))
  for(i in 1:length(WAMdf$DEKKING.ID)){
    BRSTL[i] <- if(WAMdf$VERM[i]<65){1}else{if(WAMdf$VERM[i]>65 &
                    WAMdf$VERM[i]<95){2}else{3}
    }
  }
  return(BRSTL)
}
WAMdf$VERML <- VERM.F()
WAMdf$VERML <- as.factor(WAMdf$VERML)
WAMdf$VERMLV <- as.numeric(WAMdf$VERML)

#Capacity of the car, dynamic approach
VERM1 <- subset(WAMdf,WAMdf$VERM<52)
VERM2 <- subset(WAMdf,WAMdf$VERM>=52 & WAMdf$VERM<65)
VERM3 <- subset(WAMdf,WAMdf$VERM>=65 & WAMdf$VERM<80)
VERM4 <- subset(WAMdf,WAMdf$VERM>=80)
VERMD.F <- function(){
  BRSTL <- rep(0,length(WAMdf$DEKKING.ID))
  for(i in 1:length(WAMdf$DEKKING.ID)){
    BRSTL[i] <- if(WAMdf$VERM[i]<52){1}else{if(WAMdf$VERM[i]>52 &
```

```
                    WAMdf$VERM[i]<65){2}else{
                    if(WAMdf$VERM[i]>65 & WAMdf$VERM[i]<80){3} else{4}}}
  }
  return(BRSTL)
}
WAMdf$VERMD <- VERMD.F()
WAMdf$VERMD <- as.factor(WAMdf$VERMD)
ActualVERM <- c(mean(VERM1$VERM),mean(VERM2$VERM),
                mean(VERM3$VERM),mean(VERM4$VERM))
#1=+-52.29653  2=+-72.05052  3=+-91.30751  4=+-133.18355
WAMdf$VERMD <- log(ActualVERM/mean(VERM1$VERM))[WAMdf$VERMD]

#Bonus-Malus number of years dynamic approach until 9
SVJD.F <- function(){
  BRSTL <- rep(0,length(WAMdf$DEKKING.ID))
  for(i in 1:length(WAMdf$DEKKING.ID)){
    BRSTL[i] <- if(WAMdf$SVJ.BM[i] >= 9 ){15} else{WAMdf$SVJ.BM[i]+6}
  }
  return(BRSTL)
}
WAMdf$SVJD <- SVJD.F()
WAMdf$SVJD <- as.numeric(WAMdf$SVJD)
WAMdf$SVJDminus1 <- WAMdf$SVJD-1
WAMdf$Bis15 <- as.numeric(WAMdf$SVJD==15)
WAMdf$SVJD <- as.factor(WAMdf$SVJD)
SVJD <- as.factor(WAMdf$SVJD)
```

## 15.7 Codes for Chapter 9

```
#dispersion test
library(AER)
dispersiontest(gSVJD2.GEWD.KMD.LFDLVCF.VOL)

#fitting Negative Binomial
library(MASS)
gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb <- glm.nb(nCl ~ SVJD2minus1 + Bis21 +
    W + MIL + LFDLVS + VOL + offset(log(EXPO)), link=log, data=WAMdf)

#1 divided by theta and log-likelihood difference
1/gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb$theta
-2*(logLik(gSVJD2.GEWD.KMD.LFDLVCF.VOL)-logLik(gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb))

#obtaining the estimated dispersion parameter with Pearson X^2
phi <- sum(pr^2)/df.residual(gSVJD2.GEWD.KMD.LFDLVCF.VOL)
round(c(phi,sqrt(phi)),4)
```

```
#obtaining largest difference in |CV| with adjusted factor
max(summary(gSVJD2.GEWD.KMD.LFDLVCF.VOL)$coefficients[,2]*1.0237/
abs(summary(gSVJD2.GEWD.KMD.LFDLVCF.VOL)$coefficients[,1]) - summary(gSVJD2.GEWD.
abs(summary(gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb)$coefficients[,1]))

#obtaining mean-variance relation plot
xb <- predict(gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb)

g <- cut(xb, breaks=quantile(xb,seq(0,100,5)/100))
m <- tapply(WAMdf$nCl, g, mean)
v <- tapply(WAMdf$nCl, g, var)

png("c4afig1.png",width=500,height=400)

plot(m, v, xlab="Mean", ylab="Variance",
       main="Mean-Variance Relationship")

x <- seq(0.01,0.6,0.01)

lines(x, x*phi, lty="dashed")

lines(x, x*(1+x/gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb$theta))

legend("topleft", lty=c("dashed","solid"),
     legend=c("Q. Poisson","Neg. Binom."), inset=0.05)

lines(x, x, lty="dashed")

#estimate proportion of no claims based on data, Poisson model and
#zero inflated Poisson model
zobs <- WAMdf$nCl==0
mean(zobs)
zpoi <- exp(-exp(predict(gSVJD2.GEWD.KMD.LFDLVCF.VOL)))
mean(zpoi)
library(pscl)
mzip <- zeroinfl(nCl ~ SVJD2minus1 + Bis21 + W + MIL + LFDLVS + VOL + offset(log
pr <- predict(mzip,type="zero")
mu <- predict(mzip,type="count")
zip <- pr + (1-pr)*exp(-mu)
mean(zip)

#obtaining the shape parameter by profile likelihood
library(tweedie)
library(statmod)
out=tweedie.profile(TOTSEV ~ 1,
```

```
                    data=WAMdf,p.vec=seq(1.1,1.9,length=9),
                    method="interpolation",do.ci=TRUE,
                    do.smooth=TRUE,do.plot=TRUE)
out$p.max;out$phi.max #out$p.max is het estimate of the shape parameter

#Example of checking outcomes, checking AIC outcomes
aic <- -2*logLik(gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb) +
            2*gSVJD2.GEWD.KMD.LFDLVCF.VOL.nb$rank
```

# Bibliography


[1] Young, G.A. and Smith, R.L., *Essentials of Statistical Inference*, Cambridge University Press, 2005.

[2] Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M, *Modern Actuarial Risk Theory*, Springer, Second Edition.

[3] Rice, J.A., *Mathematical Statistics and Data Analysis*, Brooks/Cole, Third Edition.

[4] Levy, R., *Probabilistic Models in the Study of Language*, draft, November 6, 2012.

[5] Fienberg, S.E. and Rinaldo, A., *Maximum Likelihood Estimation in Log-Linear Models*, The Annals of Statistics, No. 2, 996-1023, 2012.

[6] Molas, M., Lesaffre, E., *Hierarchical Generalized Linear Models: The R Package HGLMMM*, Journal of Statistical Software, Volume 39, Issue 13, March 2011.

[7] Ji Yeo, *Generalized Linear Models for Non-Life Pricing- overlooked facts and implications*, Institute and Faculty of Actuaries.

[8] Clark, T.S., Linzer, D.A., *Should I Use Fixed or Random Effects?*, Political Science Research and Methods, 3(02):399-408, March 24, 2012.

[9] Hadfield, J., *MCMC Methods for Multi-response Generalized Linear Mixed Models: The MCMCglmm R package*, University of Edinburgh.

[10] AG-werkgroep, *Zuivere schadevrije jaren- een uniforme terugvaltabel*, Utrecht, 5 June 2013.

[11] Verbond van Verzekeraars, *Roy-data*, Utrecht, 5 June 2013.

[12] The insurance company, *Tarievenboek*, 23 februari 2017.

[13] The insurance company, *Tarief Royaal autoverzekeringen*, 7 December 2016.

[14] Brzeźniak, Z. and Zastawniak, T., *Basic Stochastic Processes*, Springer-Verlag London, 7th printing, 2005.

[15] Ohlsson, E., Johansson, B., *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag Berlin, 2010.

[16] Schmitter, H., *The sample size needed for the calculation of a GLM tariff*, ASTIN Bulletin 34(1), 249-262.

[17] Garrido, J. and Zhou, J., *Credibility Theory for Generalized Linear and Mixed Models*, Concordia Univeristy, August, 2006.

[18] Frees, E.W., Derrig, R.A., Meyers, Glenn, *Predictive Modeling Applications in Actuarial Science*, Cambridge Univeristy Press, 2014.

[19] Rönnegård, L., Shen, X. and Alam, M., *hglm: A Package for Fitting Hierchical Generalized Linear Models*, The R Journal Vol. 2/2, December 2010.

[20] Murtaugh, P.A., *In defense of P values*, Ecology Society of America, 95(3), 2014, pp. 617-621.

[21] Czado, C., Raftery, A., *Choosing the Link Function and Accounting for Link Uncertainty in Generalized Linear Models using Bayes Factors*, Sonderforschungsbereich 386, Paper 262 (2001).

[22] Cirillo, P., *Are your data really Pareto distributed?*, Physica A: Statistical Mechanics and its Applications, Vol. 392, No. 23, 2013, p. 5947-5962.

[23] Burnecki, K., Härdle, W. and Weron, R., *An Introduction to Simulation of Risk Processes*, Hugo Steinhaus Center of Stochastic Methods, Research Report HSC/03/4.

[24] Bonetti, M, Cirillo, P., Tanzi, P.M., Trinchero, E., *An Analysis of the Number of Medical Malpractice Claims and Their Amounts*, PLOS ONE — DOI:10.1371/journal.pone.0153362, April 2016.

[25] Koyluoglu, H.U., Hickman, A., *A GENERALIZED FRAMEWORK FOR CREDIT RISK PORTFOLIO MODELS*, Risk magazine, October 1999.

[26] Hirz, J., Schmock, U., Shevchenko, P.V., *Crunching Mortality and Life Insurance Portfolio with extended CreditRisk+*, SSRN Electronic Journal, January 2016.

[27] Boucher, J-P., Denuit, M., and Guillén, M., *Models of Insurance Claim Counts with Time Dependence Based on Generalization of Poisson and Negative Binomial Distributions*, Casualty Actuarial Society, Volume 2/Issue 1.

[28] Cameron, A., and Trivedi, P., *Regression-based tests for overdispersion in the Poisson model*, Journal of Econometrics, vol. 46, issue 3, 347-364, 1990.

[29] Ruoyan, M., *Estimation of Dispersion Parameters in GLMs with and without Random Effects*, Mathematical Statistics Stockholm University, Examensarbete 2004:5, January 2004.

[30] Rodríguez, G., *Models for Count Data With Overdispersion*, http://data.princeton.edu/wws509/notes/, 2007.

[31] Quijano, O., and Garrido, J., *Generalized linear models for aggregate claims; to Tweedie or not?*, Concordia University, Montreal, Canada, May 2013.

[32] Vandekerckhove, J., Matzke, D., Wagemakers,E. *The Oxford Handbook of Computational and Mathematical Psychology*, Chapter 14, (pp. 300-318).

[33] Lawless, J. F., *Regression Methods for Poisson Process Data*, Journal of the American Statistical Association, Vol. 82, No. 399 (Sep., 1987), pp. 808-815.

[34] Antonio, K., Zhang, Y., *Mixed models for predictive modeling in actuarial science*, KU Leuven, AFI Research Report, wol. AFI1376, September 17, 2012.

[35] Cui, J., Pitt, D., and Qian, G., *Model Selection and Claim Frequency for Workers' Compensation Insurance*, Peeters online Journals, Vol. 40, issue 2, 779-777, 2010.