Utrecht University

MSc Mathematical Sciences

Master Thesis

Generalized Linear Mixed Models in the competitive non-life insurance market

Author:
Marvin Oeben

Supervisors:
Frank van Berkum, MSc (UvA/PwC)
Prof. Dr. Roberto Fernández (UU)
Dr. Karma Dajani (UU)

December 7, 2015

Abstract

Generalized Linear Models (GLMs) have been the standard tool in non-life insurance pricing for the last few decades. The Giro APT working party [63] states overlooked facts on the use of GLMs in the current highly competitive market for motor insurance in the UK. This thesis describes the models currently used in non-life insurance pricing and looks whether an extension of the GLM, the Generalized Linear Mixed Model (GLMM) can solve some of the problems occurring in this market.
In order to see how a competitive market behaves, we provide a definition of a market with the necessary assumptions on its costs, conversion and consumer behavior. Three markets are simulated to study how parameter choice, market composition and the use of a mixed model effect market share, sales and profit in a market with two insurers and fully economically rational customers.

Acknowledgements

I would like to thank Frank van Berkum for all his guidance, patience and comments during the process of writing this thesis. Moreover, I would like to thank Roberto Fernández for allowing me to roam free in an unknown application of mathematics. This freedom allowed me to find the beautiful world of predictive modeling. I also thank Karma Dajani for agreeing to be the second reader for this thesis.

Everyone at PwC PAIS and Hubert-Jan van der Laar, thank you for introducing me to not only non-life pricing but the diverse actuarial world. The last 9 months have been among the most fun I've had.

Contents

Chapter 1

Introduction

Insurance is the business of sharing risks among groups. In its simplest form it is a group of people agreeing to mutual aid. An example can be found in sharing risk within a village. If one house burns down, the rest of the village agrees to rebuilt it. In modern times, insurance is facilitated by companies, the insurers. Insurers provide coverage against certain risks agreed upon in advance in exchange for a premium.
Insurance can be divided into two types, life and non-life. This thesis focuses on non-life insurance. This can be defined as insurance against accidental or financial risk. Examples of this are theft, fire damages and car insurance. A main topic in insurance is the mathematically determination of the premiums also called 'pricing'. This is done using models which assign relative risks to different groups of customers. This thesis studies these models and investigates whether a relatively new model can solve some of the problems occurring in the competitive market today.

Background

Modeling insurance premiums is often done using a two-step approach. Instead of estimating a full model, claim count and severity are modeled separately. Their model outcome is then combined to form a relative risk. For the last decades this has happened using generalized linear models (GLMs), these models are a generalization of the ordinary linear regression models used in linear regression problems common to statistics.
The last few years however, the non-life insurance market has changed. Due to the rise of the Internet and as a consequence the ease of entering an insurance contract through that medium many contracts are entered through comparison websites. This newly formed type of competition has changed the market and due to the subjective process of choosing a GLM, premiums for different insurers can therefore differ substantially due to these model choices.
An example of this can be seen on comparison websites. On the Dutch market the quoted premiums can easily vary 2-3 fold for some types of motor insurance (source: independer.nl). As the quoted premium relies heavily on the risk of harm of the insured and less on the chosen insurer, these differences are remarkable.

The GIRO APT working party from the Institute and faculty of Actuaries in the UK researched this phenomenon. They noted the big difference in quoted premiums for customers and aim to provide evidence for discussion whether the use of GLMs is still fit for purpose for non-life pricing. In their paper they raised six different problems which can occur with use of the GLM in the highly competitive UK market for motor insurance. In this market, comparison websites take up a big part of all sales. To illustrate this, they state different features of a price comparison market of which we adapted the following. The market has a large number of buyers and sellers, provides homogeneous products. A customer has perfect knowledge of the market and there is no attachment between the buyers and sellers.
This study was the driving force behind this thesis. We will investigate whether and how a newer model, called the Generalized Linear Mixed Model (GLMM) can provide a possible solution to some of the overlook facts when using the GLM. The problems when using GLMs in non-life insurance pricing stated by the GIRO APT working party are [63]:

1. Either zero or full credibility is given to the data and there is no way to do blending.
2. Prediction of a risk depends on data in other completely independent segments.
3. Model predictions depend on the mixture of rating factors in the data.
4. Maximum likelihood estimate of prediction is lower than mean of prediction distribution.
5. Link function could bias the model prediction and significantly change the lower and upper bound of prediction.
6. Model diagnostics is only relevant in the segments where the model is used.

This thesis will be loosely based on problems 1 and 2. The rest of the problems could not be directly linked to the use of a GLMM.

Structure

This thesis will focus on the models used in pricing non-life insurance premiums. Chapter 2 will focus on the models itself, their estimation and practical aspects in application by the notion of 'expert judgement' in their selection and validation. Chapter 3 will provide a quantitative definition of the competitive market with some assumptions. Further, it provides a way to simulate and compare different model choices among insurers. Three different simulations will be shown using two insurers forming a market. We will finish with a short view on how GLMMs can solve some of the shortcomings stated above.
The appendix will provide some mathematical background and addresses some practical issues which may come up when fitting GLMMs in the programming language R as well as some R-code used throughout the text.

Chapter 2

Linear models in pricing

The goal of this chapter is to show several commonly used methods to model non-life insurance data in order to predict future claim counts and severity. We will start with the actuarial use of these models to show
Linear models consist of predictors and responses. In these models, the relationship between the predictors and the response is assumed linear and can therefore be written as a linear equation. The content of this equation differs among the models but for each model we can use a similar setup. In every section we first define or derive the model and possible reasoning and incentives for use of the model are motivated. After this, possible estimation and approximation techniques are either provided or described.
The end of the chapter is dedicated to the application of the models in actuarial science. Model building, testing and comparing are discussed as well as the notion of 'expert judgement' which has a big impact on the final choice of model used in application.
Models will increase in complexity. Classical Linear Models (LM) are introduced in the first section. Next we derive a generalized version in the Generalized Linear Model (GLM) followed by their extensions, the Linear Mixed Model (LMM) and the Generalized Linear Mixed Model (GLMM).

2.1 Actuarial use of linear models

Non-life insurance pricing is the modeling of both claim count and severity. This thesis will focus on claim count estimation using a Poisson distribution and the severity will be modeled using a Gamma distribution. Even though many other distributions can be used, these are the most commonly used in both academic literature as in application across the market.

Claim count

Claim counts count the number of reported insurance claims relative to some exposure. These can be given as claims per year or total claims for a given customer. Counts are often modeled using a Poisson distribution across mathematics (e.g. in queuing theory). A random variable X is Poisson distributed

$X \sim Pois(\lambda)$ if

$$f(k;\lambda) = \mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Claim severity

Each claim that arrives, arrives with a certain claim size also denoted as severity. In this thesis, we assume that the size of a claim is independent of the time it arrives. We focus on claim severity models which are best modeled using the Gamma distribution. A random variable X is Gamma distributed $X \sim Gamma(\alpha, \beta)$ if

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

Calculating the premium

Before we discuss the models, we briefly discuss how they are used in actuarial application. This thesis will consider three different premiums. The first is the relative premium (or relative risk). This is the premium for a given risk class relative to the intercept group. The second is the pure premium, a premium which is the direct output of the combined models. The last is the quoted premium, the premium the insurer brings to the market. The relative and pure premium are calculated as the relative claim count multiplied by the relative severity. This means that the output of the two models is multiplied. The exact calculation using model output is given in 2.6.

2.2 Linear Models

Linear Models are the basic models used in linear regression. They are often used due to their simplicity and effectiveness in many different regression problems. The one dimensional version is defined as

$$y = \beta_0 + \beta_1 x + e. \tag{2.1}$$

In this setting, x is often called the predictor and y the response variable. β_0 and β_1 are the regression parameters where β_0 is also referred to as the intercept. e is a noise/error term which will be further discussed below.
For one predictor variable x, this notation is often sufficient. When more predictors are introduced a matrix based setup is more appropriate as it will ease computational and notational effort. Let p be the number of parameters, $\beta_1, \ldots, \beta_p$ the used parameters and n the number of observations y_i. Then, we can rewrite the expanded version of equation (2.1) for one observation as

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + e_i \tag{2.2}$$

In matrix notation, we can rewrite the vectorized version as

$$Y = X\beta + e.$$

If we assume there is an intercept in X such that $x_{i1} = 1$ the corresponding matrix notation for the elements Y, X, β and e is

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \; e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

Y is the response vector, β the predictor vector and X is in computational application often referred to as the 'model matrix'.

2.2.1 Estimation

In application, the observed values of y do not exactly match the expected modeled value $\mathbb{E}[Y] = X\beta$. As X and Y only share an approximately linear relationship and possible errors can occur in the measurement. Hence, data provides observed values $\hat{y_1} \dots \hat{y_n}$. Therefore, we need to compute an estimated "best" fit for the observations. This requires a measure of fit. For this, we will follow the notation and notion of $\mathcal{L}_p$ spaces from chapter 4 in [56] and chapter 1 in [45]. Different values of p can be used to minimize the difference (often called distance or deviation) between y and $\hat{y}$. A first example is the $\mathcal{L}_1$-norm (also called taxicab or Manhattan norm)

$$S_1(y, \hat{y}) = \sum_{i=1}^{n} |y_i - \hat{y_i}|$$

which minimizes the total summed absolute piecewise difference between the theoretical and fitted values $(y, \hat{y})$ aiming to make the average deviation as small as possible.
Another measure of fit is using the maximum or $\mathcal{L}_\infty$-norm

$$S_\infty(y, \hat{y}) = \max_i |y_i - \hat{y_i}|.$$

This norm aims on minimizing the biggest difference. Hence, eliminating outliers as much as possible.
In general however, the $\mathcal{L}_2$-norm is used as it provides some nice advantages such as differentiability and equivalence with maximum likelihood when data is normally distributed. This norm is widespread in statistics and was introduced by Gauss in the 19th century [57]. Using this norm will eventually lead to the least squares regression method often used for linear regression. The norm requires minimizing

$$S_2(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

and thus to minimize the squared deviation over each of the observations.
In the simple case given by equation 2.1, usage of the least squares method results in the following calculations. The optimal values for β_0 and β_1 for the least squares regression can be found by deriving S_2 with respect to β_0 and β_1. Setting these derivatives equal to zero and solving the equations for β_0 and β_1 will result in an optimal value. The derivatives can be calculated as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.3}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.4}$$

where $\bar{x}$ is the average over x_i, $\frac{1}{n}\sum_{i=1}^{n} x_i$. The full computation of these values can be found in [51] or other basic statistical texts.
In the case of the matrix notation, least square minimization requires minimizing $S_2(Y, \hat{Y})$ which can be rewritten as $||Y - \hat{Y}||^2$. This has the direct solution $\hat{\beta} = (X'X)^{-1}X'Y$ if $X'X$ is non-singular, calculated as the orthogonal projection of Y on the set X. Other possibilities for solving this are also available but beyond the scope of this thesis (see [14] and [43]). In R, we rely on the function `lm(...)` to solve problems of this nature. `lm(...)` uses a QR-decomposition of the matrix X. Documentation on this function is available on CRAN and [6].

Common statistical assumptions

Assuming the model defined in (2.1), we can further investigate the 'noise' parameter e_i. We often assume that the parameters e_i are i.i.d. with $e_i \sim N(0, \sigma)$. Therefore in equation (2.3) it still holds that $\mathbb{E}\beta_i = \beta_i$. Moreover this leads to $y_i \sim N(\mu_i, \sigma^2)$ and $\text{Cov}(y_i, y_j) = 0$ for $i \neq j$.
This can easily be extended to the matrix notation of the model by assuming that $\mathbb{E}e_i = 0$, $\text{Var}(e_i) = \sigma^2$, $Y \sim N(\mu, \Sigma)$ to be multivariate-normal (Σ is in this case the diagonal matrix with value σ^2). And moreover, Y, X and β are defined as above. In this case, Maximum Likelihood Estimation (MLE) is used as described in 2.3.3. For the LM, the MLE coincides with that of the minimized least squares under the above assumption.

Extensions of the linear model

Other examples include partial least squares, generalized least squares and two stage least squares. These examples can be found in many econometrics books such as [60], [32] and [38]. A possible and well used extension is the weighted least squares in which a weighing matrix W is introduced on the variance of Y. This will result in $\mathbb{E}Y = X\beta$ and $\text{Var}(Y) = \sigma^2 W^{-1}$. Which then leads to the least squares regression solution given by $\hat{\beta} = (X'WX)^{-1}X'WY$.

2.3 Generalized Linear Models

Linear Models provide excellent solutions to many applications across all of science, they are however bound to two constraints as two assumptions are made:

1. The components of Y are i.i.d normally distributed random variables with constant variance σ^2.

2. $\mathbb{E}Y = \mu$ and $\mu = X\beta$

As in nature many things are normally distributed due to the central limit theorem described in A.6 and LMs are therefore often sufficient. In actuarial application however, Y is often not normally distributed with non-trivial connection between the predictors and the response. Generalized linear models (GLMs) provide a solution for this. GLMs are a generalization of the linear model. They differ as they do not require the data Y_i to be normally distributed and moreover, do not necessarily impose the rule that $\mathbb{E}Y = X\beta$. Instead, they allow more freedom for the choice of distribution for Y and so-called link function g such that $g(\mathbb{E}Y) = \eta = X\beta$. These generalizations can be summarized as the following three assumptions:

1. The observations Y_i have density according to a member of the exponential dispersion family (2.3.1) and are identically independently distributed.
2. There is a linear predictor η_i such that $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. η is linear in the predictors β_j with $\eta = X\beta$.
3. The link function g uniquely connects $\mathbb{E}Y_i = \mu_i$ and η_i such that $g(\mu_i) = \eta_i$.

This section will focus on theoretical aspects of the GLM, we will start with a discussion of the exponential dispersion family and the link function and finish with the standard method used for estimation.

2.3.1 The exponential dispersion family

This family is defined in mathematical context in A.2. Here, we transform the definition to the standard form used in GLM theory, and look at the Poisson and Gamma distributions and state their properties.

Definition 2.1. The exponential dispersion family is a family of densities which are of the following type:

$$f_y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right)$$

Here, $\phi > 0$ is the the dispersion parameter, $\theta \in \Theta$ is the parameter of the given distribution and $\Theta \subset \mathbb{R}$ is an open set containing θ. We have a function a which allows for use of weights in the distribution. In general however, $a(\phi) = \phi$. $b : \Theta \to \mathbb{R}$ is the cumulant function, and the function c is the normalization factor not depending on θ.

For the Poisson distribution, if $Y \sim Pois(\lambda)$ then

$$\mathbb{P}(Y = y) = e^{(-\lambda)}\frac{\lambda^y}{y!} = \exp(y\log(\lambda) - \lambda - \log(y!))$$

with $b(\theta) = e^\theta$, $a(\phi) = 1$ and $c(y;\phi) = -\log(y!)$
Similarly for the Gamma distribution $Y \sim G(\alpha,\beta)$ with

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

Then after applying a logarithmic transformation

$$-\log\Gamma(\alpha) + \alpha\log\beta + (\alpha-1)\log(y) - \beta y$$

which leads to $a(\phi) = \phi = \frac{1}{\alpha}$, $\theta = -\frac{\beta}{\alpha}$, $b(\theta) = -\log(-\theta)$. Thus $c(y;\phi) = \alpha\log\alpha + (\alpha-1)\log y - \log(\Gamma(\alpha))$.

2.3.2 The link function

The link function is the function which connects the observable $\mathbb{E}Y$ to the linear predictor η. In linear regression models this is always the identity link ($g(\mu_i) = \mu_i = \eta_i$). When using Gamma or Poisson distributions, this may be less useful as these distributions only have values on the positive line $\mathbb{R}_+$. For example, we may require that the mean needs to be strictly positive (as in claim counts and severity). Hence, links that only take positive values may be more appropriate. In general, the following link functions are considered [45]

1. identity:
$$\eta = \mu$$
2. log:
$$\eta = log(\mu)$$
3. logit :
$$\eta = \log(\mu/(1-\mu))$$
4. probit:
$$\eta = \Phi^{-}1(\mu)$$
where $\Phi(\cdot)$ is the normal cumulative distribution function.
5. complementary log-log
$$\eta = \log(\log(1-\mu))$$
6. reciprocal
$$\eta = \mu^{-1}$$

The canonical link

Every member of the exponential family has a canonical link function. This is the link function for which the θ of the exponential family has the property $\theta = g(\mu)$ and thus

$$\theta = X\beta.$$

In a pricing setting, the Poisson distribution has canonical link $\eta = \log(\mu)$ and the Gamma distribution has canonical link $\eta = -\mu^{-1}$. In application, due to the high number of small claims the Gamma distribution is often fitted with the log-link. In general, the canonical link is often preferred in modeling as it leads to a more direct computation of parameters.

The offset

As an addition to normal fixed effects β_i, GLMs can be outfitted with an offset. An offset is defined as an additional model variable with coefficient 1. If we choose to include an offset ξ there is the modified model equation

$$\eta = X\beta + \xi.$$

This offset only scales η_i according to its value in i and can be seen as having fixed value β which is indifferent of calculation [36].
The most common use of offsets is in exposure for Poisson counts with log-link function. This has the corresponding equation

$$\log(\eta_i) = \beta X + log(\xi).$$

Which is equivalent to the distributional notation

$$Y_i \sim Pois(\xi e^{X_i'\beta})$$

Note that we also apply a logarithmic transformation to the offset itself as it lies within the distribution.
Often, instead of claims counts, claim frequency is used. As claim frequency equals the claim count divided by the exposure it is easy to see that claim frequency modeling with weighted exposure is equivalent to claim count modeling with log-linked exposure.

2.3.3 Estimation

In section 2.2 an optimal solution was found using an orthogonal projection on the set $X\beta$. This is a unique property of the normal distribution, where the maximum likelihood coincides with the least squares method. In the generalized case maximum likelihood estimation (MLE) is required. For more details on MLE and its properties see [51]. MLE requires the definition of the likelihood function,

Definition 2.2 (Likelihood functions)**.** Suppose that we have random variables $Y_1, \dots Y_n$ with joint density function $f(Y|\theta)$. Then we can define the likelihood function of θ as

$$L(\theta) = f(Y|\theta)$$

Moreover, if the Y_i are assumed to be i.i.d. then their joint density is the product of the marginal densities and thus

$$L(\theta) = \prod_{i=1}^{n} f(Y_i|\theta).$$

We define the log-likelihood as

$$\log(L(\theta)) = \mathcal{L}(\theta) = \sum_{i=1}^{n} \log(f(Y_i|\theta))$$

Maximizing the likelihood function provides us with the MLE. In the case of the exponential dispersion family, we have the function f_y from 2.1.

$$f_y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right)$$

with likelihood function

$$L(\beta|Y) = \prod_{i=1}^{n} \exp\left(\frac{y_i\theta - b(\theta_i)}{a_i(\phi)} + c_i(y_i;\phi)\right)$$

and corresponding log-likelihood function

$$\mathcal{L}(Y|\beta) = \sum_{i=1}^{n}\left(\frac{y_i\theta - b(\theta_i)}{a_i(\phi)} + c_i(y_i;\phi)\right).$$

The MLE can be found by deriving the log-likelihood with respect to all the β_j. This requires conversion from θ_i and ϕ to β_j. Notice that using the link function and mean μ there is a connection from β_j to θ_i. Moreover, ϕ is constant. And thus there is the relationship

$$\begin{aligned}\mathbb{E}Y_i &= \mu_i = b'(\theta_i)\\ g(b'(\theta_i)) &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.\end{aligned}$$

By using the chain rule, there is the relationship $\frac{\partial}{\partial\beta_j} = \frac{\partial}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}$. This has the following derivative

$$\frac{\partial\mathcal{L}(Y;\beta)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ij}}{a_i(\phi)b''(\theta_i)g'(\mu_i)}$$

Often, $a_i(\phi)$ is defined as the weighed dispersion, in this case it is given as $\frac{\phi}{w_i}$. In our application, w_i is often kept 1 for all i. In counts, we use an offset instead of weighted GLM. In other applications however, weights may be preferred as they can be directly inserted into the likelihood. Using the weights leads to a more pleasant notation no longer directly depending on a and b

$$\frac{\partial\mathcal{L}(Y;\beta)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{w_i(y_i-\mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)}$$

where the variance function $V(\mu_i) = b''(b^{'-1}(\mu_i))$ is the function which represents the relationship between the variance and the mean of a distribution in the exponential family.

The IRLS algorithm

Estimation of GLMs is often done using the Iteratively Reweighted Least Squares estimation algorithm. We only briefly state the result. For a more thorough examination of the IRLS algorithm for GLMs, see [27], [45] or [48]. For more info on the IRLS algorithm itself, [14] or [15] are recommended.

The method used to calculate every iteration is called the Fisher scoring algorithm [39]. This method focuses on solving the following equation in step t, with corresponding vector $\beta^{(t)}$,

$$X'WX\beta^{(t)} = X'Wz$$

In this, the vector z has elements

$$z_i = \sum_{k=1}^{p} x_{ik}\beta_k^{(t-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

where μ_i and $\partial\eta_i/\partial\mu_i$ are evaluated at $\beta^{(t-1)}$. W is a diagonal matrix with elements $w_{ii} = \frac{1}{\mathrm{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$ As this method requires costly matrix manipulation, QR or SVD decompositions are used for the final solving of the next step in the scoring algorithm. These methods however are beyond the scope of this text. For a possible R implementation of the IRLS algorithm for GLM see [3].

2.4 Linear mixed models

Regular (Generalized) Linear Models do not account for possible structure in the fitted data. Pricing data however, is in general highly structured. Policyholders hold contracts over multiple years, are grouped geographically and can own multiple contracts. Mixed models are a method of accounting for structure in our modeling process. The core idea lies in the fact that fixed models either completely pool data together or use no pooling whatsoever. In a regression setting, this will lead to either under- or over fitting of a model holding some sort of structure. It can therefore be desirable to make distributional assumptions on some effects inside the model. One solution is using credibility theory [21]. Another is by introducing a random effect which interacts with the fixed effects and adds a covariance structure to the model. Linear models using both fixed and random effects are called linear mixed models.
These models introduce random effects as a way to handle the variance-covariance structure of Y. Besides the earlier used distributional assumption $Y \sim N(\mu, \sigma^2)$ with fixed effects β and model matrix X we add random effects u with model matrix Z. In general, u is chosen with mean zero and covariance matrix D which is often diagonal and independent of the distribution on Y. This will lead to distributional assumptions

$$\begin{aligned} Y &= X\beta + Zu + \epsilon \\ u &\sim (0, D) \\ \epsilon &\sim (0, \Sigma) \end{aligned}$$

Note that where β is a fixed constant, u is random according to some distribution. Therefore, instead of giving the direct value of the mean and variance, we denote them as conditional on u,

$$\mathbb{E}[Y|u] = X\beta + Zu \tag{2.5}$$
$$\mathrm{Var}[Y|u] = \Sigma \tag{2.6}$$

Moreover, under the above distributional conditions for u and Y, $Y \sim (X\beta, V = ZDZ' + \Sigma)$ as

$$\mathbb{E}Y = \mathbb{E}[X\beta + Zu + \epsilon] = \mathbb{E}[X\beta] + \mathbb{E}[Zu] + \mathbb{E}[\epsilon] = X\beta$$

and

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y|u)] + \mathrm{Var}(\mathbb{E}[Y|u]) \\ &= \Sigma + \mathrm{Var}(X\beta + Zu) \\ &= \Sigma + ZDZ'.\end{aligned}$$

For the calculation of the variance, the law of total variance is used in the first step. The last step uses the assumption that $X\beta$ and Z are constant in probability.
In general, the vector u is considered independently normal with mean 0, independent of Y. Therefore, D is a diagonal matrix and we can define the distribution as the multivariate normal

$$\begin{pmatrix} u \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}\right) \tag{2.7}$$

Which results normality in Y with the model $Y \sim N(X\beta, V)$ and we see that the fixed effects enter only the mean where the matrix Z and the variance of the random effects only effect the variance of Y.

2.4.1 Estimation

In regular (generalized) linear models, only calculation of β was required. In this case, we also need a fitted value for u. Even though LMMs are not the goal of this thesis, their estimation and calculation will give an indication where the problems occur when attempting to use these techniques in estimation of the GLMM in the next section. The following results rely heavily on the normality in the LMM for both the response Y as the random factor u in the MLE calculation. Therefore, the methods are explained to some detail. We start by assuming that the covariance parameter V is known beforehand. This will lead to a closed-form estimation for β and u. Note that we first estimate β and use its estimation to estimate u. Calculation and estimation in the case of an unknown covariance parameter V can be found in [46].

Estimation of β

Estimation of β is the first step to estimating the LMM. Thus we will give a thorough calculation. Suppose that Y is the response vector of size n then the distribution

$$Y \sim N(\mu = X\beta, V)$$

has the density function for the non-degenerate multivariate normal distribution [64]

$$f(y|\mu, V) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp[-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)]$$

Where $|V|$ is the determinant of V. This distribution has log-likelihood estimator

$$\mathcal{L}(u, V) = \frac{1}{2}(y-\mu)'V^{-1}(y-\mu) - \frac{1}{2}\log(|V|) - \frac{1}{2}n\log 2\pi.$$

In order to find the derivative in μ and V, let μ be a function of a vector θ and V be a function of an unrelated vector ϕ. Then

$$\mathcal{L}(\theta, \phi) = \frac{1}{2}(y-\mu(\theta))'V(\phi)^{-1}(y-\mu(\theta)) - \frac{1}{2}\log(|V(\phi)|) - \frac{1}{2}n\log 2\pi$$

As ϕ and θ are assumed unrelated, we see that the former two parts equal 0 in derivation w.r.t. θ. This will give the following calculation

$$\begin{aligned}\frac{\partial}{\partial\theta}\frac{1}{2}(y-\mu(\theta))'V(\phi)^{-1}(y-\mu(\theta)) &= \frac{1}{2}\frac{\partial}{\partial\theta}(y-\mu(\theta))'V^{-1}(y-\mu(\theta)) \\ &= \frac{\partial\mu'}{\partial\theta}V^{-1}(y-\mu) \\ &= \frac{\partial X\beta}{\partial\beta}V^{-1}(y-X\beta)\end{aligned}$$

As in the last step, $\theta = \beta$ because $\mu = X\beta$ and $\frac{\partial X\beta}{\partial\beta} = X$ we can state the MLE for β as

$$\hat{\beta}_{MLE} = (XV^{-1}X)^{-1}XV^{-1}y$$

Estimation of u

For the estimation of u, we resort to the estimated best linear unbiased predictor (BLUP) for u as described in [30] and [46]. This estimator depends on the estimated β in the previous tep. As the random effects are assumed to have structure in variance, we estimate them depending on the observations in y. Thus, we try to estimate u by estimating $\mathbb{E}[u|Y]$. Therefore we need to describe a relationship between Y and u. This can be done by describing their covariance structure. Setting up a multivariate normal distribution in the same manner as above will provide such a structure.
In this case, for $Y \sim N(X\beta, V)$, $u \sim N(0, D)$ there is covariance structure

$$\begin{aligned}\text{Cov}(Y, u) &= \text{Cov}(X\beta + Zu + \epsilon, u) \\ &= \text{Cov}(X\beta, u) + Z\text{Var}(u, u) + \text{Cov}(\epsilon, u) \\ &= ZD\end{aligned}$$

And thus we have a multivariate normal distribution for $(Y, u)'$ defined as

$$\begin{pmatrix} Y \\ u \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} V & ZD \\ DZ' & D \end{pmatrix}\right) \tag{2.8}$$

Which leads to the BLUP [46].

$$\mathbb{E}[u|Y] = DZ'V^{-1}(y - X\beta)$$

Other estimations can be made for u but as our focus lays in the calculations for the GLMM, we refer the interested reader to [46], [29] and [1].

2.5 Generalized linear mixed models

As linear mixed models provide an extension for the linear model. Generalized linear mixed models (GLMMs) provide an extension for the GLM by the addition of random factors. This extension is however not as seamless as the extension from the LM to the LMM by mismatch of distributions. In this section we will look a the extension from the GLM and several estimation algorithms.

Extension from the GLM

The GLM above was defined as a generalization of the LM using three criteria. Observations Y_i have density belonging to the exponential dispersion family, there is a linear observable η_i of the predictors β_j with $\eta = X\beta$. Finally, there is a link function g such that $g(\mu) = \eta$. In the LMM example for identity link and normal distribution in both the error and the random effects a direct definition of the model could be given. As we now try to generalize this model, we need to check if the model still holds under the three conditions defining the GLM.
Note that in equations 2.5 and 2.6 we use the conditional mean and variance for Y. In order to define these for the GLMM, we need to define the conditional distribution for a density from the exponential dispersion family. Suppose that f is the density of y according to the exponential dispersion family, then for i.i.d. y_i

$$\begin{aligned} y_i|u &\sim f_{Y_i|u}(y_i|u) \\ f_{Y_i|u}(y_i|u) &= \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi)\right) \\ u &\sim f_U(u) \end{aligned}$$

The first equation describes the conditional distribution of y_i, the second the density function of this distribution, the last is the distribution of the random factors. This distribution can be assumed non-normal. In many applications however, like in the LMM case, it is assumed to be normally distributed.
Let the mean of $y_i|u$ be defined as $\mathbb{E}[y_i|u] = \mu_i$ then for the exponential dispersion family, the following still holds: $\mathbb{E}[y_i|u] = \mu_i = b'(\theta_i)$ and $\mathrm{Var}(y_i|u) = \phi V(\mu_i)$ for the same variance function V as in the case for the GLM, $V(\mu_i) = b''(b'^{-1}(\mu_i))$. Then the piecewise model equation can be defined as

$$g(\mu_i) = x_i'\beta + z_i'u.$$

This definition follows all criteria imposed by the GLM and thus represents the natural extension form the GLM [46].
The mean, variance and covariance of the marginal distribution y_i can now be calculated. Note that the model equation depends on choice of link-function g and therefore no closed form solution exists.

$$\begin{aligned}\mathbb{E}[y_i] &= \mathbb{E}[\mathbb{E}[y_i|u]] \\ &= \mathbb{E}[\mu_i] \\ &= \mathbb{E}[g^{-1}(x_i'\beta + z_i'u)]\end{aligned}$$

$$\begin{aligned}\text{Var}(y_i) &= \text{Var}(\mathbb{E}[y_i|u]) + \mathbb{E}[\text{Var}(y_i|u)] \\ &= \text{Var}(\mu_i) + \mathbb{E}[\phi V(\mu_i)] \\ &= \text{Var}(g^{-1}[x_i'\beta + z_i'u]) + \mathbb{E}[\phi V(g^{-1}[x_i'\beta + z_i'u])]\end{aligned}$$

$$\begin{aligned}\text{Cov}(y_i, y_j) &= \text{Cov}(\mathbb{E}[y_i|u], \mathbb{E}[y_j|u]) + \mathbb{E}[\text{Cov}(y_i, y_j|u)] \\ &= \text{Cov}(\mu_i, \mu_j) \\ &= \text{Cov}(g^{-1}[x_i'\beta + z_i'u], g^{-1}[x_j'\beta + z_j'u])\end{aligned}$$

These solutions give little to no insight in the behavior of the GLMM. If we add the assumption that $u \sim N(0, \mathbf{I}\sigma_u^2)$, the multivariate normal distribution with zero covariance and variance $\sigma_u > 0$ depending on the vector u with log-link $g(\mu) = \log(\mu)$ (as in most actuarial applications) then

$$\log(\mathbb{E}[y_i]) = x_i\beta + \sigma_u^2/2$$

$$\begin{aligned}\text{Var}(y_i) &= \text{Var}(\exp[x_i'\beta + z_i'u]) + \mathbb{E}[\phi V(\exp[x_i'\beta + z_i'u])] \\ \text{Cov}(y_i, y_j) &= \text{Cov}(\exp[x_i'\beta + z_i'u], \exp[x_j'\beta + z_j'u]) \\ &= \exp[x_i'\beta + x_j'\beta]\text{Cov}(\exp[z_i'u], \exp[z_j'u]) \\ &= \exp[x_i'\beta + x_j'\beta]\left[\exp[\sigma_u^2](\exp[z_i'z_j\sigma_u^2] - 1)\right]\end{aligned}$$

Note that by the moment generating function described in subsection A.1.4, $M_u(z_i) = \exp(\sigma_u^2/2)$ for the normal distributed vector u. For $\text{Var}(y_i)$ no further simplification will lead to more insight without assumptions on the distribution of $y_i|u$. In the covariance, $z_iz_j = 0$ if i and j do not share a random factor and 1 otherwise. In the latter, we can define the correlation as

$$\begin{aligned}\text{Corr}(y_i, y_j) &= \frac{\text{Cov}(y_i, y_j)}{\sigma_{y_i}\sigma_{y_i}} \\ &= \frac{1}{\sqrt{(1 + \xi\exp[-x_i'\beta])((1 + \xi\exp[-x_j'\beta]))}}\end{aligned}$$

for $\xi = 1/(\exp[3\sigma_u^2/2] - \exp[\sigma_u^2/2])$. This correlation can be used find interaction between elements of y which were assumed i.i.d. in previous models. Therefore, structure can now be found and investigated in underlying data.

2.5.1 Estimation

In the case of the LMM we could define a multivariate normal distribution for Y. This resulted a likelihood estimation depending only on $f(y, |\mu, V)$.

In the GLMM case, the distribution for the random effects and that of the response variable often differ. Thus no closed form estimation for β and u can be calculated. To illustrate, a direct computation of the likelihood

$$L = \int f_{Y|u}(y|u) f_U(u) du \tag{2.9}$$

or, individually

$$L = \int \prod_{i=1}^{n} f_{Y_i|u}(y_i|u) f_U(u) du$$

For functions of the exponential family, this leads to a likelihood equation of the form

$$L = \int \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi)\right) \frac{1}{\sigma_u\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma_u^2}\right] du$$

with corresponding log-likelihood

$$\mathcal{L} = \log\left(\int \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi)\right) \frac{1}{\sigma_u\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma_u^2}\right] du\right)$$

For this equation, no analytical solution can be given. Therefore, numerical approximation must be used to estimate the maximum likelihood.
As we require estimation of the likelihood, we need to resort to approximation. We will discuss three different methods. One will focus on approximation of the function itself, the second on the integral and the last uses Bayesian statistics to approximate the likelihood. The Laplace and the Gauss-Hermite quadrature approach will be given in full detail and briefly outline a Bayesian/Monte Carlo approach and its (dis)advantages.

The Laplace method

The first (and default method used in `glmer(...)` for R) is the Laplace method introduced by Laplace in 1774 and translated in [42]. This method is fully outlined in [58]. It uses a Taylor expansion of an exponential term. It approximates integrals

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

in which u is a q-dimensional vector and $h(u)$ is a sufficiently smooth function with local maximum in its domain. Suppose that the function has a local maximum in u_0. We can define the second order Taylor expansion for in h in u_0 as

$$h(u) \approx h(u_0) + \frac{1}{2}(u-u_0)'h''(u_0)(u-u_0) \tag{2.10}$$

Which provides a plug in approximation for the integral as

$$\int \exp[h(u_0) + \frac{1}{2}(u-u_0)'h''(u_0)(u-u_0)] du \tag{2.11}$$

Which, by the Laplace method has approximate function

$$\int e^{h(u)}du \approx \exp[h(u_0)](2\pi)^{q/2}|-h''(u_0)|^{-1/2}$$

In order to approximate the likelihood function, we need to rewrite the likelihood to the form used by the Laplace approximation. Note that

$$\begin{aligned}\mathcal{L} &= \log \int f_{Y|u}(y|u)f_U(u)du \\ &= \log \int \exp[\log f_{Y|u}(y|u) + \log f_U(u)]du\end{aligned}$$

And thus we can set $h(u) = \log f_{Y|u}(y|u) + \log f_U(u)$. In order to find (2.10) we need to compute the second order derivative of $h(u)$.
Firstly, when we assume $u \sim N(0, D)$ then u is multivariate normal and

$$\log(f_U) = -\frac{1}{2}u'D^{-1}u - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|D|$$

This function has first and second derivatives $\frac{\partial \log f_u}{\partial u} = -D^{-1}u$ respectively $\frac{\partial^2 \log f_u}{\partial u \partial u'} = -D^{-1}$
The derivative of $\log f_{Y|u}(y|u)$ can be found by using the chain rule on the exponential family, similar to the derivative found in subsection 2.3.3.

$$\begin{aligned}\frac{\partial \log f_{Y|u}(y|u)}{\partial u} &= \frac{1}{\phi}\sum_i \left(y_i\frac{\partial \theta_i}{\partial u} - \frac{\partial b(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial u}\right) \\ &= \frac{1}{\phi}\sum_i (y_i - \mu_i)\frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)z_i'} \\ &= \frac{1}{\phi}Z'W\Delta(y-\mu)\end{aligned}$$

in which $W = [V(\mu_i)g'(\mu_i)^2]^{-1}$ and $\Delta = g'(\mu_i)$ with functions V and g as defined in subsection 2.5. In order to find u_0 we need to solve

$$\frac{\partial h(u)}{\partial u} = \frac{1}{\phi}Z'W\Delta(y-\mu) - D^{-1}u = 0$$

which is highly non-trivial as all factors involved except for y are functions of u.
Second order derivative calculation follows as

$$\frac{\partial^2 h(u)}{\partial u \partial u'} = \frac{\partial}{\partial u'}\left(\frac{1}{\phi}Z'W\Delta(y-\mu) - D^{-1}u\right) \tag{2.12}$$

$$= \frac{1}{\phi}\left(-Z'W\Delta\frac{\partial \mu}{\partial u'} + Z'\frac{\partial W\Delta}{\partial u'}(y-\mu) - D^{-1}\right) \tag{2.13}$$

For calculative convenience we choose to ignore the second term in the last equation. This as in the case of Poisson distributed Y it is valued 0 and in all

other cases, it has expectation 0 as $\mathbb{E}y = \mathbb{E}\mu$[46]. Thus we can formulate the following approximation.

$$\frac{\partial^2 h(u)}{\partial u \partial u'} = -\frac{1}{\phi}\left(Z'WZD + I\right)D^{-1}$$

Insertion into the log-likelihood gives

$$\mathcal{L} \approx \log f_{Y|u}(y|u_0) - \frac{1}{2}u_0' D^{-1} u_0 - \frac{1}{2}\log|\left(\frac{1}{\phi}Z'WZD + I\right)D^{-1}|$$

Which has the following derivative in β

$$\frac{\partial l}{\partial \beta} = \frac{\partial \log f_{Y|u}(y|u_0)}{\partial \beta} + \frac{\partial}{\partial \beta}\frac{1}{2}\log|Z'WZD/\phi + I| \tag{2.14}$$

$$\approx \frac{1}{\phi}X'W\Delta(y-\mu) \tag{2.15}$$

Here, W is assumed to change negligibly in respect to β. This gives an estimate of β and u by solving the equations

$$\frac{1}{\phi}X'W\Delta(y-\mu) = 0$$
$$\frac{1}{\phi}Z'W\Delta(y-\mu) = D^{-1}u$$

Which is equivalent to solving the quasi-likelihood $\log f_{Y|u}(y|u_0) - \frac{1}{2}u'\frac{\partial \log f_u}{\partial u}$ with a penalty in the second term.

Numerical quadrature

A second way to estimate the likelihood in equation (2.9) is by estimation of the integral itself instead of its inner function. This method works using the adaptive Gauss-Hermite quadrature (AGQ). It uses a weighted sum approximation of the clustered distribution of y. We describe the non-adaptive method followed by the adaptive method and the approximation for the GLMM.

Non-adaptive Gauss-Hermite quadrature

Non-adaptive Gauss-Hermite quadrature is an approximation technique for integrals of the form

$$\int h(z)\exp(-z^2)dz \tag{2.16}$$

where $h(z)$ is a function integrable on $\mathbb{R}$ and sufficiently smooth (at least twice differentiable). The method uses a weighted sum of order Q to estimate the integral in the following manner

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

Here, Q is the order of the approximation. z_i are the zeros of the Q'th order Hermite polynomial

$$H_Q(z) = (-1)^Q \exp[z^2] \frac{d^Q}{dz^Q} \exp[-z^2]$$

with corresponding weights

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

For more information on Hermite polynomials and the Gauss-Hermite quadrature formula see [18] and [33]. This method does not depend on the values in h and is symmetric around 0. Therefore, the approximation is in general relatively poor as h may have its weight elsewhere. Therefore, for the application in the GLMM we will define an adaptive version.

Adaptive Gauss-Hermite quadrature

The adaptive version of the Gauss-Hermite quadrature (AGQ) uses a Gaussian approach, in which a Gaussian function replaces the factor $\exp(-z^2)$ with suitable changes in the weights and approximation points. We follow the outline given by [44] and adapt it to the case of the GLMM.

Let $\phi(t; \mu, \sigma)$ be the probability density function of the normal distribution with mean μ and standard deviation σ. Define a function $g(t)$ such that $g(t) > 0$, is unimodal (i.e. has a unique mode) and is sufficiently smooth. The goal is approximation of $\int g(t)dt$ by transformation of the factors used to solve (2.16). To achieve this, replace the Gauss-Hermite quadrature in (2.16) for the integral

$$\int f(t)\phi(t; \mu, \sigma)dt.$$

This requires a transformation of the sampling nodes z_i to t_i according to the transformation from $\exp[z_i]$ to $\phi(t; \mu, \sigma)$ which equals

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

Moreover as we want to sample the integral in te region of $g(t)$, we can define $\hat{\mu}$ as the mode of $g(t)$ and $\hat{\sigma} = 1/\sqrt{\hat{j}}$ for

$$\hat{j} = -\frac{\partial^2}{\partial t^2} log(g(t))\Big|_{t=\hat{\mu}}$$

Define $h(t) = \frac{g(t)}{\phi(t;\hat{\mu},\hat{\sigma})}$ then we can rewrite the integral for $g(t)$ as

$$\int g(t)dt = \int h(t)\phi(t; \hat{\mu}, \hat{\sigma}))dt$$

Which after applying the transformed Gauss-Hermite quadrature equals

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

for $w_i^* = w_i \exp(z_i^2)$.

In the case of the GLMM, we will show the implementation of a single random effect. This effect can be seen as clustered into different groups. Every cluster i has a random effect which is distributed as $u_i \sim N(0, \sigma^2)$. Thus we need to determine the posterior mode of u_i. This depends on the factors β, ϕ and σ. We replace these by the current estimate (and in the first step a well chosen value) $\hat{\beta}^*$, $\hat{\phi}^*$ and $\hat{\sigma}^*$. Then using these estimates, define $\hat{u}_i$ which maximizes

$$f(y_i|u_i)f(u_i|\hat{\sigma}^*) \propto f(u_i|y_i).$$

Thus we can use $\hat{u}_i$ as the mode for u_i and use the above Gauss-Hermite quadrature to approximate

$$\int f_{Y|U}(y_i|u_i)f_U(u_i)du_i \approx \sum_{l=1}^{Q} w_l^* \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij} z_l^*) \right)$$

in which n_i is the size of the cluster i, y_{ij} is the j-th element of cluster i and we have the adaptive weights $w_l^* = \sqrt{2}\hat{\sigma}_i w_l \exp(z_l^2)\phi(z_l^*; 0, 1)$ for $z_l^* = \hat{\delta}_i + \sqrt{2}\hat{\sigma} z_l$ where $\hat{\delta}_i$ is the approximation for $\sigma^{-1}u_i \sim N(0, 1)$. Moreover, we have linear predictor $x'_{ij} + \sigma z_l^*$ for $f_{Y|U}(y_{ij} z_l^*)$.
Multiplication of this sum leads to the approximated maximum likelihood. This method leads to new current best estimates until convergence.

Bayesian approach

Even though this thesis is written in a frequentist's view, Bayesian algorithms provide a solid way to compute the GLMM. As both the random effects and the response variable have explicit distributions, a Bayesian framework with prior distributions on ϕ and D can lead to good approximations. As the Bayesian framework is less restricted by design it can be of good use in simulation from a posterior distribution. More information on Bayesian statistics and their role in the GLMM can be found in [65] and its references and Chapters 13 and 14 from [30]. For simulating Bayesian statistics in R, a combination of the `glmmbugs` package [7] and the WinBUGS program for windows [11] is used in [30].

Method comparison

Among these methods, Laplace tends to behave poorly [16]. Even though Laplace approximations are the easiest to fit, they use a large amount of approximations on a single value and can therefore in some cases provide a bad fit. AGQ approximations give a significantly better fit but are restricted to only 2-3 nested random effects, where with nested we mean that additional random effects are modeled within another random effect. The use of crossed or high-level nested effects is not possible. Moreover, Laplace and GHQ are only designed for use on normally distributed random effects. Finally, Bayesian methods are relatively slow (due to the high amount of simulations) and are technically challenging due to the Bayesian framework but provide great flexibility in more exotic models.

2.6 Choosing, building and comparing models

Up until now, the actual modeling of data has not been discussed. This section will be dedicated to expert judgement which can be roughly described as 'choices made by an expert in the field'. These choices involve which model is chosen, how the model parameters are selected and what criteria are used to compare different models and setups. We start with the choice between the above given models. Then discuss best actuarial practice when building a model. Different criteria and how they function in different models are discussed next. We finish with a way to compare entirely different models in out of sample testing, a special case of cross-validation.

2.6.1 Model choice

We start this subsection that it is known that a LM is a GLM with identity link and normally distributed Y. Therefore, choice between LM and GLM is the same as the choice between the different distributions within the exponential dispersion family. This choice comes down to the distribution that best fits the observed data in Y in both a statistical and explanatory sense. The same holds for the choice of link function. We often resort to the log-link in non-life modeling as it best handles non-negative sparse data with many small values and few very large ones. Some studies have been done on this subject, the interested reader is referred to [24] and [31].
In our application we use the Poisson distribution for claim counts and Gamma distribution for severity, the choice of model is centered around the choice of parameter.

Model building, parameter inclusion

As any model building, the first step in setting up a GL(M)M model is always data exploration. The researcher should and is often encouraged to thoroughly explore the data. With exploration is in this case meant looking for structure (which will be especially important when fitting GLMMs) and finding correlation between parameters in the data. Knowing the way data is structured and finding possible correlation is crucial to fitting the correct model.
Setting up a GL(M)M requires more than decisions based on only the data. In actuarial and other applications, experience and explanatory value of the parameters are just as important. The reason for this is that the researcher will need to explain why he chose certain parameters in the model, other than just their explanatory value according to some criterion. This is important as he can be held accountable for the calculation of the risks in the portfolio. A good example of probable misspecification is an investigation done by the Dutch Consumentenbond [10]. In this investigation, adding a letter to the address resulted in a higher premium, this means that on the address "street 1" the client payed significantly less than if he would live on "street 1A". In Dutch house numbering, a letter is the result of having multiple houses on the same numerical address. This often happens with apartment complexes in urban areas. Data may suggest that an address with a letter added to it, results in a higher claim amount/severity. Adding a parameter "Street address has a letter behind it" to

the model may result in a higher premium for all houses with an added letter. Therefore, even though this parameter may have good predictive power, it has little explanatory power in the general case. One could use the parameter nested in other geographical data. In that case this relationship only holds in the area where it is actually the case instead of in an entire country. A random effect approach as explained in the next paragraph could also detect structures and dampen unwanted effects like this outside of the date where the effects manifest itself.
Apart from the choice of parameters, the method in which the model is built up can also vary. There are in general two approaches. The first is from the bottom up, the second is a top down approach. The first starts as an empty model, adding factors which are significant and add explanatory value according to the researcher. Often, several approaches are used as correlation between factors may disrupt this nested approach (as we add a factor in each step, we create a new nested model). The second approach starts with a full model, factors are then removed which show little to no predictive power. This is also done several times as the order in which factors are removed may contaminate judgement. An example is that a model with factors A and B which are heavily correlated, then if A and B both reside in the model, their individual predictive power may be not significant. Removing A or B may add to the predictive power for the other. But naively doing solely this step once may remove the factor with the most predictive power.
In general the first method is considered best practice. The reason behind this is that the factors which reside in the final model can all be explained (if not they would not have passed the above 'test'). Therefore, in applications GL(M)Ms should be build from the ground up.
If we look at the common practice, we see that actuaries often start with a set of parameters for which experience and expert judgement say should be in the model and build the model from there. This can be called a hybrid but is technically a bottom up approach as these parameters are rarely removed.

Fixed or random factors

After adding factors according to expert judgement and predictive power, one needs to decide whether these should be considered fixed or random factors.
Choice of fixed or random effects for the LMM and GLMM are under a lot debate. The most cited text [46] describes the following process:
"In endeavoring to decide whether a set of effects is fixed or random, the context of the data, the manner in which they were gathered and the environment from which they came are determining factors. In considering these points the important question is: are the levels of the factor going to be considered a random sample from a population of values which have a distribution? If 'yes' then the effects are to be considered as random effects; if 'no' then, in contrast to randomness, we think of the effects as fixed constants and so the effects are considered as fixed effects. "
This choice is supported by several econometric texts such as [28].
Other econometric text take a different approach, [32] states: *"Again, the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not. "*

And thus, they need not be drawn but show heavy correlation between subsets (often called clusters) in the data.
The latest actuarial source [30] describes a more natural choice. They introduce the notion of a multilevel model with clusters, sometimes called hierarchical as data often has a hierarchical structure (e.g. aging cars or policyholders). However, multilevel or mixed model is the preferred name as data needs not be nested in different levels. Geographic data is an example of a structure which is non-nested but structured (often in postal/zip codes). And thus in this thesis, instead of choosing whether data comes from a random sample, known structure of data will decide the use of randomness for a factor.
The most recent and complete text [23] supports the last definition. This text is the most complete text on choice of effect to date. Thus we will follow this text and [30] in the modeling of our data for the GLMM. We will select random effects on basis of structures in the data. Even though there is no solid mathematical background for the choice of effect, there is however a method which tests whether a effect should not be randomized. This test is called the Hausman test.

Hausman test

As choosing between a fixed or random effect is a combination between expert judgement and theoretical considerations, the Hausman test can give some help in detecting whether estimates in the fixed effects are similar to estimates in the random effects. Suppose we define a fixed effect as $\hat{\beta}$ and its random counterpart as $\hat{u}$. Then the Hausman test H is defined as

$$H = (\hat{u} - \hat{\beta})' \left[\mathrm{Var}(\hat{\beta}) - \mathrm{Var}(\hat{u}) \right]^{-1} (\hat{u} - \hat{\beta})$$

H is assumed to be χ^2 distributed with degrees of freedom equal to the number of regressors in the model. If $p < 0.05$ we should reject the random effects model in favor of the fixed effects model [37].
In the case that $p > 0.05$ no conclusion can be given and expert judgement will have to decide to implement the random element or not. A possible R-implementation has been given in Appendix B.1.

2.6.2 Scoring criteria

As building and fitting models is often more of an art than an exact science, comparing models is up to the researcher who does the fitting. In general, three questions come to mind. The first is: How well does my model fit? The second question is: Does my model behave like I expect it? The last holds importance especially in actuarial science: How well does my model predict the future? The first one can be answered using various tests which will be covered in this subsection. We view each of these test and look at their applicability for the different models. Five criteria will be covered, the first is the Root mean squared error (RMSE) followed by the log-likelihood. After this the related AIC and BIC are discussed. The deviance and anova tests are examined last. Many more tests are used in different applications but we will only consider these 5.
The second and third question require expert judgement but are in applications just as important as the first. An experienced researcher often knows how data

should behave and does not stare blindly on the data in front of him as data can be misleading. Answering the third question often comes down to doing out of sample tests. These are discussed in a later subsection.
Before we can discuss criteria on which we can compare our models, we need define the notion of 'nested' in modeling. We say that a model A is nested in model B if model A consists of all factors of model A plus an extra factor.

RMSE

The Root Mean Squared Error is an absolute test for error which works for both nested and non-nested models. It is given by the root of the mean squared error and equals

$$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y)^2}{n}}$$

This gives the mean error of prediction and can be of great help in the modeling process. The advantage is that it can quickly compare many different fits, indifferent of model choice, nestedness and choice of parameters.

log-likelihood

As we are in the process of maximizing the log-likelihood, one can make the argument that bigger is always better. Hence, a bigger log-likelihood will might show a better fit. Using the linear models described above, one needs to remember that in calculation for the LM and GLM, one calculates an estimated true likelihood where LMM and GLMM often provide asymptotic likelihoods and therefore a likelihood test is not suited for testing between different models. If two similar models are compared based on log-likelihood, the model with more factors will most likely provide a better score. This may interfere with comparison as more predictors do not necessarily provide a better fit. This problem is partially solved in the next two tests.

AIC and BIC

To address the problem offered by comparison on log-likelihood, there is the Akaika information criterion. This criterion gives a penalty on the number of fitted parameters. And is therefore defined as

$$AIC = 2[-\mathcal{L}(Y;\hat{\theta}) + r]$$

where r is the number of fitted parameters and l the fitted likelihood.
The Bayesian information criterion adds the size of the model to the test. Thus the BIC equals

$$BIC = 2[-\mathcal{L}(Y;\hat{\theta}) + r\log(n)].$$

By many researchers, the BIC is preferred over the AIC as the size of the dataset should be incorporated in the assessment of the fit.
In the case of mixed models, the use of AIC and BIC is not as clear. Usage up until recently was based on [34] and [59]. For the LMM they define different versions of the AIC. When the model contains random effects, the AIC cannot

be defined directly as the value of r is unclear. Questions arise as to which likelihood to use, to see whether random effects are parameters and how to count the degrees of freedom. Therefore, [59] focuses on this question and determines that it is dependent on the focus of the research. They distinguish between two types of research: inference concerning the population and that of clusters. Note that these 2 different types of research coincide with the problem of adding random effects as given in the paragraph on choice of fixed or random factors. In the case of the LMM, conditional and marginal AIC criteria have been set up. These are however not directly viable for use with the GLMM. Therefore we focus on a newer article [52]. It is not the goal of this thesis to give the full layout of this criterion. We will however state the definition of the AIC and a small passage on the BIC.
In the case of the AIC we can define the conditional AIC (cAIC) for Poisson distributed response $y_i|u \sim Pois(\lambda_i)$ as

$$\text{cAIC} = -2\log(f(y|\hat{\beta}, \hat{\mu})) + 2\Psi$$

with

$$\Psi = \sum_{i=1}^{n} y_i \left(\hat{\theta}_i(y) - \hat{\theta}_i(y_{-i}, y_i - 1)\right)$$

where y_{-i} is the vector of observed responses without the i-th observation y_i and y_i is the i-th observation with $y_i\hat{\theta}_i(y_{-i,y_i-1}) = 0$ if $y_i = 0$ by convention. It is available in the package `cAIC4` in R [5].
On BIC, the article states: *"The behaviour of other information based criteria like the Bayesian information criterion, BIC, for the selection of random effects in GLMMs needs further investigation."* Therefore we currently do not advice GLMM selection based on BIC.
Questions then arise whether we can compare values for the AIC in GLM with the AIC found for a fit in the GLMM. As show above, this is not advised. A good overview can be found in [35] which discusses different flavors of the AIC and other more exotic versions such as the DIC and the FIC and their possible uses.

Deviance and Anova

The Deviance is defined as twice the difference between the log-likelihood of the fitted model and the 'saturated' model. The saturated model is the model which perfectly fits the data. The test statistic used in this case is the χ^2 test. The `glm(...)` and `glmer(...)` functions in R output values for the deviance. The `glm(...)` outputs 2, the null deviance which is the residual deviance for a model with only a constant term and the residual deviance itself. This provides an upper bound on the deviance. This deviance can be used to compare two nested models and is equivalent to a log-likelihood statistic. As the deviance is only used in nested models it, as a test can only say whether a certain factor should be added to the model or not. It can provide no useful information on non-nested models. In the case of the GLMM, the deviance is almost never used as a test statistic, this as it provides problems in the definition of the 'saturated' model. Comparison between models is therefor not useful and a comparison between fixed and mixed models can not be given based on the

Deviance.
The analysis of variance (Anova) test is a test to compare two nested models, it shows the residual degrees of freedom and deviance for each model. It can also show statistical tests which compare the reduction in deviance with the change in degrees of freedom. This test is particularly useful to quickly see whether adding an extra parameter significantly increases the model fit.

Other methods

Likelihood ratio tests are common in nested GLMM models. In the light of coefficients of determination R^2 tests which are common for LM and GLM, these tend to be extremely hard in the case of GLMM. There is a package available to build these, based on [47] in the `MuMIn` package for R [8].

Out of sample comparison of models

The best way to compare distinctively different models is by using cross-validation or out of sample testing. These tests test the predictive capability of a model when it is introduced to new data. Therefore, the test does not depend on the used model itself. In general, two tests are used to test the predictive power of a model, the RMSE as defined above and the Mean Absolute Deviation, which is the test on the $\mathcal{L}_1$-norm defined earlier [16] [22]. Competitions such as Kaggle [2] use more sophisticated methods as these are less easy to fool. Examples are the normalized Gini coefficient [54] and the AUC (Area Under Curve) metric [61]. These are however not always applicable to regression data.

2.7 Fitting GLMMs, practical aspects and troubleshooting

In order to find the convergence behavior for the GLMM in the case for Laplace and AGQ GLMMs we will start with a theoretical approach and then result to examples and solutions for failing approximations in `glmer`. The cause for this can often be found in the optimizers used, and assumptions made by these optimizers. In statistical texts however, there is only little attention for optimizers. Therefore we start with a view on optimizers in general and why they may fail. Almost all statistical calculations contain some form of optimization. For example, maximum likelihood estimation involves fitting the most plausible parameters to a given dataset. Most of these solutions can be found in a closed-form expression as in the calculation of the LM and GLM. In this case, software will produce a good and stable estimation of the solution. For the GLMM, no closed form solution can be given. Thus we need to result to numerical estimations like the AGQ algorithm. These algorithms can fail for many reasons. The iteration might reach a singular gradient and be forced to stop or it reaches the maximum number of iterations without any convergence. These problems are often a result of failed constraints. These constraints can be statistical (positive variance, positive definite covariance matrices etc.) or arise from the data used. In statistics it holds that if the dataset is large enough and the model is well-specified the maximum likelihood estimation is close to the true optimal parameters. Thus finding optima should not be that hard. Optimizers however are

software and can easily get stuck in singularities, local optima or plateaus.
A lot of theory has been written on optimization. A good way to see why and how optimizers get 'stuck' and how to trick them into leaving these positions can be found in heuristics. A good start on this subject is [55].
In R, a good and often used package for optimization is `optimx` [9]. This package has many different optimization tools and can be incorporated in `glmer` for the calculation of the Laplace or AGQ approximation. As optimization requires us to look at both our data and our model we will now look at how and why Laplace and AGQ approximation can fail.

Laplace approximation failure

We now give a quick view on possible points of failure for the Laplace approximation, these points are then shortly analyzed. It is beyond the scope of this thesis to explore the full structure of these approximations.

1. The assumption that $u \sim N(0, D)$ may not hold for our chosen random effect.
2. The approximation for the first order derivative may fail.
3. For the second order derivative, The assumption that the second term can be ignored in (2.13) may be false in some cases.
4. The approximation of the second order derivative may fail.
5. The assumption that W varies negligibly with respect to β in (2.15) may not be true in some cases. And thus, ignoring the factor may be wrong.
6. The approximation for the likelihood in the last step may fail.

The first assumption is important. Even though the GLMM can in theory handle non-normally distributed random effects, the Laplace function can only model normally distributed random effects. And thus, the model will not be specified correctly. The second, fourth and sixth possible points of failure can be traced back to the above discussion on optimizers.
In the case of the third point, the second term in (2.13) can be ignored for Poisson distributed GLMMs and with expected value 0 it is assumed safe to ignore the term in other cases. If the link function is chosen properly and the variance function is relatively small, this point should in general not cause many problems.
The same reasoning holds for the fifth point, as β has little influence on the (co)variance structure of Y, it will not be of great influence on W and can therefor safely be ignored in the approximation.
As we will see in subsection 2.7.1 points 1,2,4 and 6 will cause problems in the GLMM fitting.

Adaptive Hermite-Gaussian Quadrature failure

For the AGQ we look at the algorithm in a similar fashion as for the Laplace, we look at possible points which can cause trouble and shortly analyze them.

1. As with the Laplace approximation, AGQ uses a normal transformation. The effect however may not be normally distributed.

2. The maximization of the mode may fail in some occasions.

3. Solutions may depend heavily on starting values for $\hat{\beta}^*$, $\hat{\phi}^*$ and $\hat{\sigma}^*$.

The first point is the same as for the Laplace approximation. The assumption of normality may not be true and thus our model may not be correctly specified.
For the second point, as we assume a unimodal function g, in theory there are no local optima. As we optimize however, due to the iteration and choice for $\hat{\beta}^*$, $\hat{\phi}^*$ and $\hat{\sigma}^*$, solutions for the mode of g may stay away from the true mode in some optimizers.
The same goes for solution 3 and 4. The algorithm relies heavily on the optimization method.

2.7.1 Errors and failures in the glmer function

This subsection provides an overview of possible errors generated by the `glmer` function, their origin and provides some solutions. This subsection is descriptive, general example code can be found in Appendix B.2 or the corresponding reference. Possible solutions will be given for each problem and if encountered in Chapter 3 for our example dataset.
We consider two examples. The first is a scaling problem, the second is a convergence problem.

Example 2.3. The first error is an error referring to scaling of variables. This error occurs rather often, but can be fixed with relative ease without changing the outcome of the model in a severe way.

```
Warning messages:
1: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
Model failed to converge with max|grad| = 0.00133114 (tol = 0.001, component 1)
2: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
Model is nearly unidentifiable: very large eigenvalue
- Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
- Rescale variables?
```

Several approaches can be taken to remove the error. As it suggests rescaling fixed variables, we can do this in multiple ways. The first is shrinking the fixed effects x_k to be closer to zero

$$\tilde{x_k} = x_k - \min(x).$$

Another approach is to get rid of this error is to center the value as e.g. in [30]. Here, the fixed effect x_k becomes

$$\tilde{x_k} = x_k - \frac{1}{n}\sum_{i=1}^{n} x_i.$$

A last approach is rescaling the effects using the `scale` function in R. This method is preferred as it does not change the interacting between different

effects. This method is similar to the first, but the `scale` function adjusts the values to the standard deviation and is therefor equal to

$$\tilde{x_k} = \frac{x_k - \sum_{i=1}^{n} x_i}{\mathrm{sd}(x)}.$$

Standardizing the results using the `standardize` function as outlined in [35] from the `arm` package [4] it can be shown that all three methods will give similar fit. We will sketch a proof of this more theoretically in subsection 2.7.2.

These scaling examples can be fixed with relative ease, convergence problems are in general harder to deal with. We start with an example and then work towards a way to tackle these problems in a more general framework.

Example 2.4. Convergence errors are errors often giving output similar to

```
Warning messages:
1: In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
Model failed to converge with max|grad| = 0.00385383 (tol = 0.001, component 1)
```

Several things can be done to improve the convergence. The rescaling options mentioned above can help in order to try and solve these problems, in general however it is advised to switch between several different optimizers to find one that does meet the conditions on convergence. In appendix B.2.2 the method to do this using the `glmer` function is fully outlined in R. This code includes the sources for the code provided by Ben Bolker, author of the `glmer` function. It should be noted that all optimizers find similar optima, but some seem to converge quicker or have slightly different conditions for this convergence such that they do not fail.
First we will look at the different options to view results given by `glmer` with respect the convergence. These commands can give insight in convergence and errors in the calculation.
Suppose we have fitted the model according to some data/distribution and have the error above, the first step is to check for singularities in the random-effects parameter estimates θ for which the lower bounds on the random parameters is 0. If the estimate is very close to zero ($< 10^{-4}$) then this singularity can lead to convergence issues and more often false positives.
The next step is checking the derivative calculus, the reason for this is that `glmer` may in some cases fail to derive proper Jacobian and Hessian Matrices. In this case, the package `numDeriv` may prove to give better approximations of the derivatives. In some cases, this will give better approximations for the Jacobian and Hessian matrices which may lead to better convergence but may also worsen it.
Sometimes these subtle changes to the model do not work, in that case next to trying different optimizers, we are left with one more option which is increasing the number of iterations. This can be done in two ways. The first is allowing `glmer` more iterations through the `control` command, the second is to use the latest values as starting values and hoping that more iterations will lead to convergence.

2.7.2 Consistency of fit when rescaling and centering fixed variables

In the previous examples, fixed effects were scaled and centered to assist convergence in the numerical approximation. As our data contains many zeros, large eigenvalues are common and can therefore provide scaling errors.
Even though scaling seems necessary, it is important to show that scaling fixed effects does no change the fitted values. An example on the fit for scaled variables in the GLMM is given in appendix B.3. This example shows very similar fits for the data. Next we look at the calculation for the fit using Laplace and AGQ and see whether a change in X leads to a change in the fit $\hat{y}$. Suppose that we have an n by p matrix X of predictors and rescale one fixed effect in X. This is the same as transforming one column in X. Let $\hat{\mu}_j$ be the mean and $\hat{\sigma}_j$ the standard deviation of X'_j, a column of X. Define the scaling of elements x_i of X'_J as $\tilde{x_i} = \frac{x_i - \hat{\mu}_j}{\hat{\sigma}_j}$. Suppose w.l.o.g. that $j = 1$ then we have the matrix

$$\tilde{X} = \begin{pmatrix} \frac{x_{1,1}-\hat{\mu}}{\hat{\sigma}} & x_{1,2} & \cdots & x_{1,p} \\ \frac{x_{2,1}-\hat{\mu}}{\hat{\sigma}} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n,1}-\hat{\mu}}{\hat{\sigma}} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

which itself is also a predictor matrix. Therefore, we only have to see whether the fit $\hat{y}$ differs when optimizing β and u in the Laplace and AGQ equations when changing from X to X'_j. Note that the w.l.o.g. holds as we can change any column of X and reorder them at will. Moreover, as any rescaling of a column of X gives a new valid matrix $\tilde{X}$ by transitivity only the above case is required. Another important notion is that there is no change in the values of y. Therefore, no distributional changes happen when rescaling from X to $\tilde{X}$. Therefore, no change in distribution occurs.

The Laplace case

As we saw in 2.5.1, the Laplace method resolves to solving the two equations in β and u

$$\frac{1}{\phi} X' W \Delta (y - \mu) = 0 \tag{2.17}$$

$$\frac{1}{\phi} Z' W \Delta (y - \mu) = D^{-1} u \tag{2.18}$$

with X the matrix of fixed effects, Z the matrix of random effects, $W = [V(\mu_i) g'(\mu_i)^2]^{-1}$ and $\Delta = g'(\mu_i)$ for link function g and variance function $V(\mu_i) = b''(b'^{-1}(\mu_i))$. Note that $\mu_i = g^{-1}(x'_i\beta + z'_i u)$ and μ the vector $(\mu_1, \ldots, \mu_n)$. Therefore, X is an integral part of the calculation.
We therefore formulate the following proposition.

Proposition 2.1. *Suppose we have matrices X and $\tilde{X}$ described above. Then the fitted values $\hat{y}$ and $\hat{\tilde{y}}$ differ only in approximation.*

For an exact proof of this, full detail on the above equations in Laplace approximation should be given for all link functions and members of the exponential family. In this case, however we restrict ourself to the linearity in X and β in a sketch of a possible proof. Note that X is used in two different ways in the equations to be solved. In the first, it is directly used in calculation, in the second it is only used as input for μ_i.
Suppose we have an approximately optimal solution to (2.17) and (2.18). Then we have optimal values $\hat{y}, \hat{\beta},$ and $\hat{u}$ and therefore, an optimal value μ as function of $\hat{\beta}$ and $\hat{u}$. The optimality of this solution will be the key to this proof. As, we have near optimality, 2.17 and 2.18 will be approximately true.
Let X be replaced with $\tilde{X}$ in our optimal solution. Assume we lose optimality. In equation (2.18), the change $X \to \tilde{X}$ is only trough $\mu_i = g^{-1}(x_i'\beta + z_i'u)$. Therefore, a transformation $\hat{\beta} \to \tilde{\hat{\beta}}$ is required to regain the original optimal $\hat{\mu_i}$. This transformation leads to previous optimal $\hat{\mu_i}$ in equation (2.17). Moreover, the linear change in the first column of X (and therefore the first row of X') has no effect as the equation equals 0 to the right. Thus the transformation $X \to \tilde{X}$ has no difference except in approximation and thus our sketch is complete.

Possible changes in actual fitted values will probably be caused by the optimizer. As the errors suggest, optimizers tend to prefer scaled variables which will not lead to large eigenvalues or big difference in fixed effect values. Thus, it may even be true that as an optimizer performs better on a scaled effect, the fitted values may be closer to the true optimal value y.

The AGQ case

In AGQ case, there is no straight forward calculation. Above, we showed the following approximation

$$\int f_{Y|U}(y_i|u_i) f_U(u_i) du_i \approx \sum_{l=1}^{Q} w_l^* \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij} z_l^*) \right).$$

This was approximated by a linear predictor $x_{ij}' + \sigma z_l^*$ for $f_{Y|U}(y_{ij} z_l^*)$. The same reasoning as for the Laplace method applies. As the prediction is done in a linear manner, a linear shift in one of the columns of X will result in a linear adjustment for β to still achieve the same fit. And therefore, scaling and centering has no result on the fit of the model in the AGQ case.

2.7.3 Actuarial application, from model output to premium

In the beginning of this chapter the calculation of a premium was shortly outlined. We will now give the complete method which is used in the next chapter. In this thesis we use a pure premium, which is the combined pure premium of the count and severity model. For a risk class we multiply the expected count by the expected severity. The calculation of the fixed and mixed models slightly differs, therefore both are treated separately.

Premium calculation for a fixed model

Linear models are estimated by estimating the best β in the equation

$$Y = X\beta.$$

Which for a single observation comes down to estimating

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

The estimated model provides the estimated optimal values for all β_i. These values can be used to predict estimated counts/severity for customers both inside and outside the dataset (as in an out-of-sample test).
In application, all factors are categorical. Therefore, each category is assigned its own value β_i. This provides for direct building of risk classes.
Calculating the premium is then straightforward. As every observation y_i in a dataset falls in a risk class within the model, a relative count/severity can be calculated in the following manner. For N the number of counts, the use of link function g and observation i in a certain risk class we have

$$N = g^{-1}(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \ldots + \hat{\beta}_p x_{ip}).$$

Here, $x_i j$ for fixed i is a binary vector linked to the risk class for observation i. Similarly, for S the severity, h the used link function and observation i in a certain risk class we have

$$S = h^{-1}(\hat{\beta}_1 + \hat{\beta}_2 \tilde{x}_{i2} + \hat{\beta}_3 \tilde{x}_{i3} + \ldots + \hat{\beta}_p \tilde{x}_{ip}).$$

Note that the models for claim count and severity may involve different risk classes. As the chosen variables for both may change, therefore x and $\tilde{x}$ can differ for certain model choices. This will not pose a problem in the final calculation for the premium which will be $N \cdot S$ for every observation in a given dataset. In later simulations however, we often choose the same parameters for an illustrative example.

Chapter 3

The competitive market

The highly competitive market for non-life insurance is a rather new phenomenon. Even though comparison websites have existed for more than a decade (independer.nl was founded in 1999, source: independer.nl), their influence has only recently begun to grow. The exact size of this influence in the Netherlands is currently unknown. In other markets, such as the UK motor insurance market, comparison websites contribute to a large part of the market [63]. Thus, investigating how competition may influence the market is worth investigating. In the previous chapter different models were described which can be used to price non-life insurance products using historical data and models were compared on theoretical aspects.
This chapter will focus on simulating the competitive market using a dataset and to illustrate the effect different pricing models have on policies that are sold using a simulation exercise.
First we provide a quantitative definition of the insurance market for a single product. This definition will be expanded by assumptions on the structure of risk classes and contracts. Next, possible ways to simulate customer decision making are discussed. A theoretical splitting and a credibility approach to using the competitor's quoted premium are provided. The chapter ends with a sample simulation and several splitting and model choices.

3.1 The market

To define the competitive market, several assumptions are required. The market itself however requires none. Throughout this chapter, focus lies on a single product type (e.g. motor insurance, moped insurance)

Definition 3.1 (The non-life insurance market for a product class)**.** For a type of insurance product we define the market S as a set of risk classes and L as the list of insurers acting on the market. The size of S is denoted by M, the size of L by N. An element $i \in S$ is a risk class. Every risk class i is a unique combination of risk factors in the market. The list L is a list of insurer names with L_j denoting the j'th insurer.

This definition allows for the introduction of pricing vectors next. These vectors display the quoted premiums for every insurer acting on the market.

Definition 3.2. Consider the market S from definition 3.1. Define the pricing vector P_i as the vector of quoted premiums for risk class i. An element of P_i is denoted by p_{ij}, the premium of the j'th insurer for the i'th risk class. If the insurer does not list a premium, the price p_{ij} is denoted as ∞.

A possible example market with corresponding price vector can be motor insurance in the Netherlands.

Example 3.3. Consider the product class "motor insurance" in the Netherlands for both third party and full physical damage. Let the input consists of vehicle age, vehicle (sub)model and bonus-malus years. The set S then consists of all possible combinations between these factors, excluding combinations which are not possible (e.g. driver aged 18 with 20 bonus-malus years) and thus often, $M < |S|$. Table 3.1 is an example of an ordered market S.

Riskclass	Driver Age	Vehicle Age	Vehicle model	Bonus-malus years
1	18	0	AAAA	0
2	18	0	AAAB	0
⋮	⋮	⋮	⋮	⋮
i	19	0	AAAA	0
⋮	⋮	⋮	⋮	⋮
M	...	...	...	...

Table 3.1: Structure of the market S

Any risk class in S is a combination of the factors in table 3.1. Therefore, for every class given in table 3.1, we can define an example pricing vector

$$P_i = (p_{i1}, p_{i2}, p_{i3}, \ldots, p_{iN}) = (100, \infty, 250, \ldots, 350)$$

Assumption 3.4 (Insurance policy content)**.** Let the market be defined as in Definition 3.1. Assume that for any class $i \in S$ the insurance policies $p_{ij} \in P_i$ are similar in content. With similar in content we mean that contracts in P_i cover the same damages and have equivalent terms and conditions.

Assumption 3.4 lets the price of the product be the only quantitatively discrimination factor between policies. Therefore, only the pricing vector P_i is needed to discriminate between insurers. To justify the direct mapping from quoted premium to the risk classes one more assumption is required.

Assumption 3.5 (Equal costs)**.** In this chapter, we assume that all the quoted premiums in a set P_i have the same relative costs. Hence, the conversion from risk to quoted premium is equivalent. Therefore the risk of harm is the only driver for the price.

In order to make a proper comparison we require risk to be the only driver for the quoted premium. If insurers have different conversion rates, the prices cannot be compared as if they are relative risks and we can less easily incorporate them in our pricing method.

3.1.1 The customer

An insurance market S is observed by customers. These customers are sometimes referred to as agents in agent based modeling. In this thesis we will refer to them as customers. For more information on agent based modeling and its applications see [17]. To help us define and simulate customers, we make some assumptions. First, they are assumed almost perfectly economically rational in their choice of insurer. This states that a customer will always choose one of the cheapest quoted premiums in P_i. In order to quantify customer choice, we want to assign a probability distribution to the choice of premium by the customer. This distribution has to respect the following assumption.

Assumption 3.6 (Customers prefer cheaper policies)**.** We assume that customers prefer cheaper (not necessarily the cheapest) quoted premiums. Suppose there exists a random variable X with probability density P_X. Then P_X is assumed to be decreasing in probability along the domain. To strengthen this, we assume that the more expensive premiums are almost never converted i.e. the distribution P_X is light tailed.

Thus, customers prefer cheaper premiums over the more expensive ones. Even though this last assumption is not as strict as one may want, expert judgement will decide on whether a chosen distribution suits the behavior of the customer in the observed market. In order to assign a probability distribution to the choice of policy we assume that the insurer has no historic data on this choice. Therefore, assumptions need to be made on the distribution of the choice of insurer per risk class or in general.

Possible distributions for policy choice of customers

Using the previous assumptions, we can suggest several distributions. We will provide a few suggestions in this paragraph, the final choice will be a decision based on the appropriateness of the fit, the ease of implementation and professional judgement. Derivation of means and variance can be found in appendix C. Using assumptions 3.4 and 3.6 we can start by defining the first possible distribution:

Definition 3.7 (Cut-off Uniform)**.** Consider the risk class $i \in S$ with quoted premium vector P_i. We define the random variable X on P_i as follows

$$\mathbb{P}(X = x) = \begin{cases} \frac{1}{k} & \text{if } x \text{ is amongst the } k \text{ cheapest premiums} \\ 0 & \text{else} \end{cases}$$

Here, we can choose k anywhere between 1 and the number of insurers offering an insurance ($\#P_{ij} < \infty$). Therefore we have a uniform distribution on the lowest k quoted premiums in P_i. The mean of this distribution is $\mathbb{E}X = \frac{k+1}{2}$ and variance $\text{Var}(X) = \frac{1}{12}(k^2 - 1)$.

We can choose to apply one scalar k to our entire portfolio S (in [63] they decide that k ranges between 5 and 10) or we can make it risk class specific. Hence we get a vector $k = (k_1, \ldots, k_M)$ in which we choose a k_i for every $i \in S$. This can also be done in an automated way, one can for example take k_i as

the lowest 5 or 10 percent of the set P_i. If data is available, this scalar can be chosen according to the distribution of sales from a comparison website.
Another possible distribution makes use of a finite geometric sum. The advantage of this distribution is that it can scale according to the given rate factor r.

Definition 3.8. Suppose we have class $i \in S$ with quoted premium vector P_i of size $n = m_i$ then we can define a discrete random variable X on $\tilde{P}_i$, the ordered quoted premium vector as

$$\mathbb{P}(X = k; r, n) = \frac{1-r}{1-r^n} \cdot r^{k-1}, \text{ for } k = 1, \ldots, n$$

Assuming again that $m_i = n$, we have the following mean $\mathbb{E}X = \frac{nr^{n+1}-(n+1)r^n+1}{(1-r)(1-r^n)}$ and variance $\text{Var}(x) = \frac{r}{(r-1)^2} - \frac{n^2 r^n}{(r^n-1)^2}$.

Discussion

Before we look at the implementation of the above market, we discuss some of the definitions and assumptions thus far. Many of these are debatable. Here are some views on them and an explanation of the choices made.

- Market definition: One can argue that Definition 3.1 is too general. In my opinion however it is left to the user to specify this in more detail. This framework should provide sufficient flexibility even for a non-insurance setting.

- Assumption 1 is rather strong. In general, not all insurance policies are of similar content. We can decide to specify all possible input such that any two policies in P_i would have matching content. This rigorous classification would however lead to an increase in size for the set S of distinct classes and a large decrease in size for many of the sets P_i. And we would then be left without sufficient policies to compare. Note however that not all insurers have to provide policies in every risk class. Thus we decide to choose a specification of the input which would lead to almost distinct classes i which are large enough such that the elements p_{ij} can still be compared.

- The definitions and assumptions on the side of the customer are highly debatable. It is in general true that price is of big influence on the decision in policy, there are however other (less easily quantifiable) influences that play a big role in the decision. Examples of this are experience, exposure (marketing etc.) and an undefined gut feeling. These are however difficult to measure and generalize and are therefore beyond the scope of this thesis. It may be an idea to introduce these into the model as noise, but this is left for further research.

- We assume that no data on the market is available but there are however alternatives. Full data on customer choice can be available but will almost never appear. As only the owner of a comparison website and an insurer can set this up together (Independer for example is owned by Achmea).

Hence, it is a theoretically viable case. Another is a market where customer choice is estimated by surveys, in this case there is a distribution per risk class available.

3.2 Simulating the competitive market

After defining the market and assigning a probability distribution to the choice of insurer by customers we need a solid and reproducible method to simulate the market. This simulation can be using an existing or using a simulated dataset. We will describe both methods. The former can be done using data from an insurer or scientific source. The latter can be done using mathematical models (e.g. Poisson for arrivals and Gamma for severity) and publicly available data on the population (e.g. data from centraal bureau voor de statistiek (cbs.nl), sociaal cultureel plan bureau (scp.nl) or verbond van verzekeraars (verzekeraars.nl) in the Netherlands).
Regardless of the way we acquire the dataset it will consist out of customers on the rows (every row is a customer) and risk factors and claim data as columns. In this simulation it is desirable that the amount of customers exceeds the number of risk classes resulting from the modeling choice. This is required for proper model fit as well as assigning premiums to all the different pricing vectors. It may be useful to remove factors in order to decrease the size of S and increase the relative size of P_i. Hence, we will have a dataset similar to the one in table 3.2.
Here, c_i represents the total claim amount of customer i

Customer	Factor 1	...	Claim Number	Claim Amount
1	–	...	z_1	c_1
2	–	...	z_2	c_2
⋮	⋮	⋱	⋮	⋮
C	...	...	z_C	c_C

Table 3.2: Structure of the dataset representing the market

Splitting and calculation

As there exists clear structure in the used dataset, it can be split into subsets. When splitting, one can make two choices. Whether all insurers are the same size or some may differ and the choice whether they all serve the same market segments. The first needs no explanation, with the seconde it is meant that an insurer focuses on a certain market segment. An example is the dutch insurer Promovendum who is aiming for the higher educated segment in the market. No matter which method we choose, the overall mapping will be similar. Every insurer will have a consumer portfolio in the form of table 3.1. This portfolio can also be transformed to the risk-class setup defined in definition 3.1. In this thesis we will only look at the size of the insurer, no distinction in segmentation along the market is used.

Calculating the quoted premiums and the sales

On each subset assigned to an insurer, a predictive model can be fitted using models in the previous chapter. This model will provide the market with relative premiums per risk class and thus per customer. According to our assumptions, these relative premiums can be directly converted to the quoted premium for class P_{ij}.
Conversion can then be measured by using one of the probability distributions mentioned above. In the case of a readjustment of the model, this process can be repeated for different models, providing different quoted premiums P_{ij}. This can be the case in the following approach.

Incorporating the competitors price, a credibility approach

Due to the lack of data on the entire market and the risk of underpricing and overpricing premiums an insurer may want to include quoted premiums by competitors in its own model. A way to include these market prices is a credibility approach. This approach is centered around using the quoted premiums observed in the market in reweighing the quoted premiums of the own portfolio. In application it will be rescaling according to a formula equal to a pricing equation with own price p, some market price p_{market}, a parameter α and the new price $\hat{p}$ given by

$$\hat{p} = \alpha p + (1 - \alpha) p_{market}$$

The choice of p_{market} is non-trivial and for a great deal expert judgement. In some cases, one may want to use the mean price, in others the median, mode or a set place in the market. Mostly, this place is a strategic decision, expert judgement assigns a market price p_{market}. As an example, we can use a weighted average of the premiums for p_{market}. This weighted average is calculated in the following form

$$p_{market} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\tilde{p_i}}$$

in which $\tilde{p}$ is the ordered pricing vector. This choice gives greater weight to the lower priced premiums (where actual sales occur) and lower weight to higher priced premiums which are rarely converted into contracts. Whether this choice of market price is desired from a marketing point of view is left out of consideration.
Choice of α is what drives the credibility approach. The original use is due to the Buhlmann Straub model [20]. That model however, is originally meant to be used inside a porfolio, leading to GLMM-like modeling of claim data. In that case, α can be chosen as relying on variance within the model. In general this method can not be applied to external premiums. Therefore expert judgement is required in this scenario. Whether or not there exists a value α which optimizes this equation in this scenario is currently unknown and beyond the scope of this thesis. In other scenarios, where information on market share and customer behavior are available some information on the choice of α can probably be derived.

Chapter 4

A simulation example, a market with two insurers

This chapter provides an example on a competitive market with two insurers. We will discuss three different simulations. In the first simulation, one insurer uses an almost full model (includes all parameters in the model), where the other uses only two parameters. This simulation is done to investigate how possible under and over-fitting will influence the relative risks for both insurers.
In the next example, both insurers use the same model but the market is split up in a 70/30 ratio. The purpose of this is to investigate how availability of data influences the relative risks.
In the last simulation, the first insurer sticks with this model while the other uses several GLMM approaches. We investigate how different GLMM approaches affect the relative risks and compare sales conversion for a fully random model.

The Data

We will use a dataset which is large enough, contains sufficient claims and parameters, is freely available and used in [49]. This dataset is from the former Swedish insurance company 'Wasa', and concerns partial casco insurance for motorcycles in the years 1994-1998. An overview of the set and the cuts given in the book are given in Table 4.1. In this table, the rating factor and class show how the data can be factorized (such that only categorical variables remain). A short description and the size of each cut is also given. In our use of the dataset contracts with zero exposure are removed and zone 7 is merged with zone 6 due to the small exposure to claims which may lead to problems when splitting the market. An overview of the data is given in Figure 4.1. In these figures, the relative size of the class is given in black, the relative claim size is given in red to give a quick view on where claims occur.
Quick data exploration shows that newer vehicles have relatively high damages and vehicles older than 20 years take up a small portion of both the dataset and the relative claims. The exposure spikes every half year which is probably the moment most customers switch contracts. With respect to the ages, we see maxima around 25 and 50 where the spike on age 50 which corresponds to a single outlier. Also, men report relatively more claims than women and zone 4 (small towns and countryside) is the only zone with relatively more claims than

Rating factor	Class	Class description	Size
Geographic zone	1	Central and semi-central parts of Sweden's three largest cities	7.678
	2	Suburbs plus middle-sized cities	4.227
	3	Lesser towns, except those in 5 or 7	1.336
	4	Small towns and countryside, except 5–7	1.000
	5	Northern towns	1.734
	6	Northern countryside	1.402
	7	Gotland (Sweden's largest island)	1.402
Engine Power	1	EV ratio -5	0.625
	2	EV ratio 6–8	0.769
	3	EV ratio 9–12	1.000
	4	EV ratio 13–15	1.406
	5	EV ratio 16–19	1.875
	6	EV ratio 20–24	4.062
	7	EV ratio 25–	6.873
Vehicle age	1	0–1 years	2.000
	2	2–4 years	1.200
	3	5– years	1.000
Bonus Malus	1	1–2	1.250
	2	3–4	1.125
	3	5–7	1.000

Table 4.1: Structure and class size used in the motorcycle insurance dataset from [49]

customers.
In this thesis, bonus malus years will not be considered in the simulations. Bonus Malus scales provide a direct bonus for drivers making no damages but provides a punishment for those who make damages. This provides a new dynamic to modeling and is therefore beyond the scope of this thesis. For info on bonus malus modeling see [40].

The simulation setup

In every simulation, the dataset is split up in two subsets each representing an insurer using random selection on the rows. In every simulation this process is repeated multiple times to make the result less dependent on the sampling. The first and second simulation will use 10 different market cuts, the last will use 5 as the fitted GLMMs show some convergence issues. Every market cut simulation has the following setup:

1. The dataset is split up into two insurers using random sampling on the rows for the first insurer and assigning the rest of the rows to the second insurer.

2. Both insurers calculate the relative risks with a GL(M)M and provide relative risks for all the customers in the market.

3. Customers view the relative risks in the market and choose the cheapest.

In the last step, customers choose according to the Cut-off uniform distribution for two insurers with $k = 1$ given in 3.7.
Each simulation run provides us with relative risks and customer choices. To find the average risk and customer choice, the average relative risk for both insurers is considered. The customer choice is then reevaluated according to these averages. This allows us to investigate both the relative risks and the customer choice to see how the chosen market setup influences the relative risks and sales for both insurers.

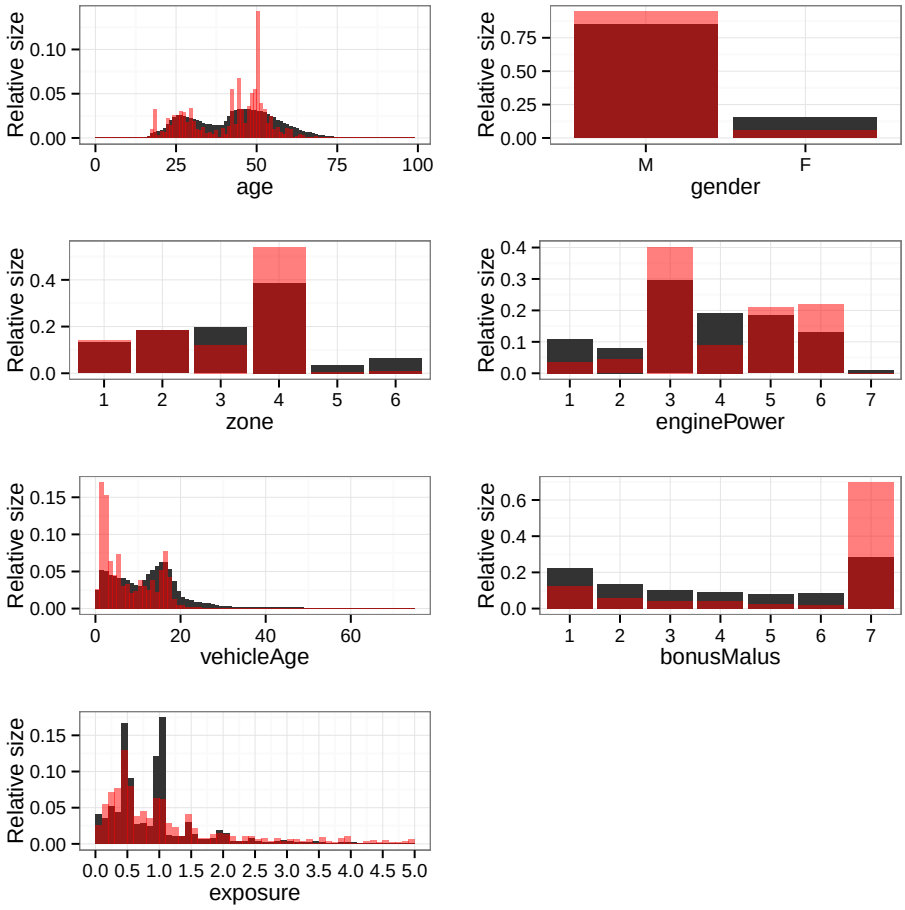

Figure 4.1: Black: The relative size of the classes from Table 4.1. Red: The relative claim size per class calculated as the portion of the total claim size per class.

Testing profitability

For every simulation we check whether an insurer is profitable. With this check there is the choice between two options. Use the reported claims in the data or use a modeled approach on the entire dataset. As the existence of exposures spanning multiple years makes the first approach inadequate for this data we choose the second. We modeled a GLM best suited for this dataset. The relative risks from this model are used as benchmark to see how well an insurer performs and whether a market will be winning or loosing.

4.1 Over and under-fitting in an equal market

In the first simulation we investigate the effect of over and under-fitting in a GLM. To illustrate this, the first insurer uses an almost full model. Both the

count and the severity model use the parameters gender, zone, engine power, age and vehicle age. The second insurer uses only zone and vehicle age in the model.
A scatter plot shows the relative risks for both insurers in the way they are presented to the customers. We can color them per parameter. These plots are shown in Figure 4.2. On the bottom right corner the relative size of the benchmark is represented. It can immediately be seen that the over-fitted model used for the first insurer will acquire most of the sales as in the upper left diagonal, the value for the second insurer is always higher than that for the first. As the second insurer uses only two parameters it only brings 18 different risk classes to the market. Moreover, there is a clear segregation in values for zone as it is included in both models. As the factors appear diagonally it can be seen that the factors of zone act the same in both models as expected. For vehicle age as similar pattern emerges, both insurers value the older vehicles as lower in risk than the newer vehicles. In the rest of the figures it can be seen that the second insurer does not discriminate for these parameters and therefore the first fully dominates the position of the colorings.
When looking at the coloring for the benchmark it shows that the benchmark risks the contracts in a similar manner as would be expected of similar models. Figure 4.3 provides a more detailed comparison of the benchmarked risk factors and those for both insurers. The over-fitted model used by the first insurer severely underestimates the risks. The second does in general however also underestimate the risk where one may expect that under-fitting will lead to higher relative risks. This underestimation may be due to three reasons. The first is that the poorly chosen model underestimates the risks in general in an under-fitting setting. The second may be that as the insurer estimates its model on only half the dataset it estimates lower relative premiums when exposed to less data. A third may lie in the averaging done in the simulation. A quick view of the results however shows that the average risk for each run of the simulation only slightly varies.

Testing profitability

Figure 4.3 suggests that these insurers will both reside in a loosing situation. As 99% of the customers prefer the first over the second insurer we look at the average revenue, cost and profit per policyholder. These values are shown in Figure 4.4. The first insurer reports huge losses due to the underestimated risks. The second insurer also reports losses. Possible reasons for this loss have been stated in the above paragraph. As customer choice is based on the cheapest premium there is an another reason. If the first insurer greatly underestimates almost all of the risks then in the cases where the second insurer is cheaper it will almost certain also be cheaper than the benchmark. This situation therefore shows that over-fitting in this market leads to greatly underestimating risks compared to the benchmark leading to not only a loss for the insurer causing the underestimated risks but also for its competition.

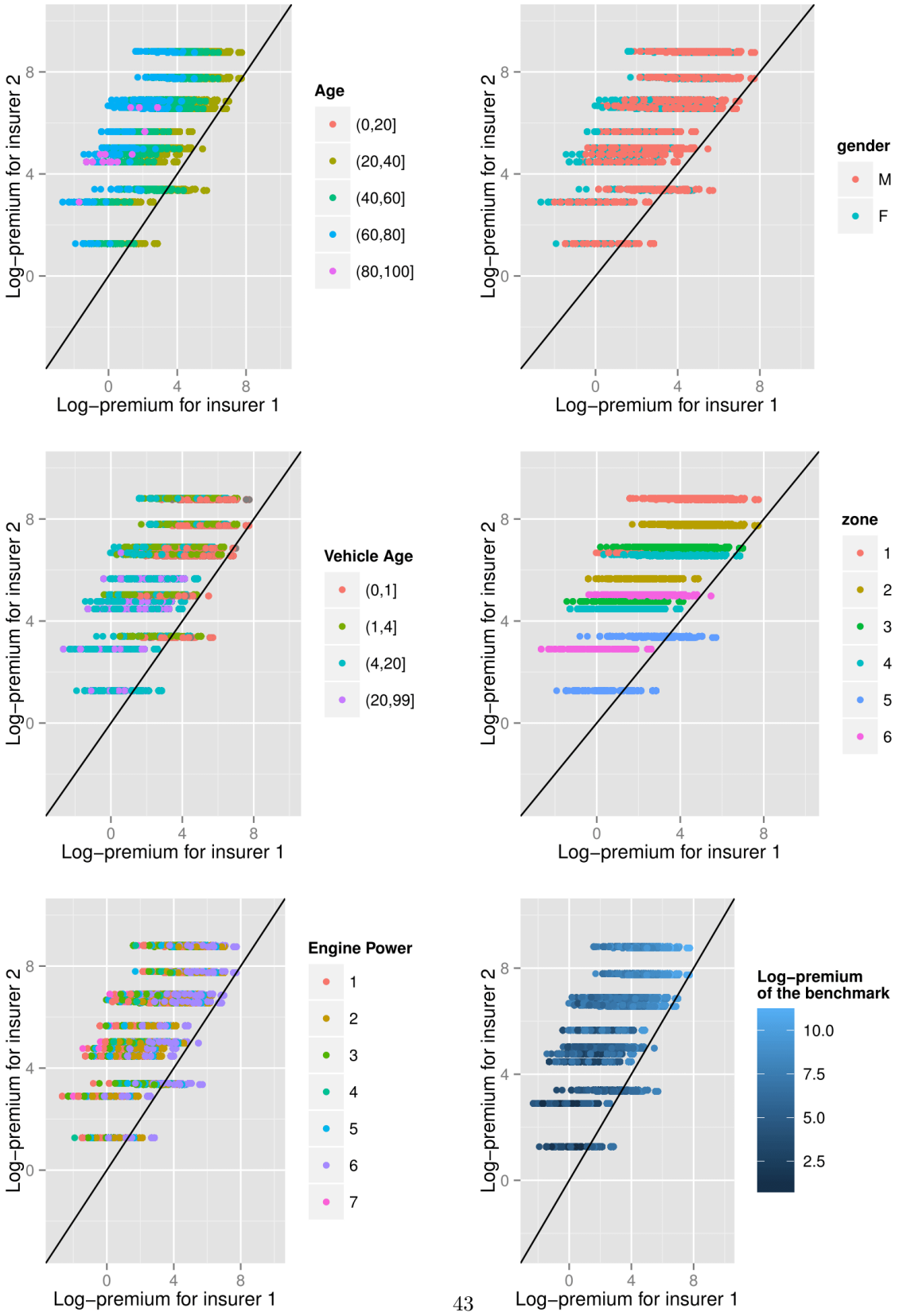

Figure 4.2: Scatter plots showing the relative risks colored per parameter in the model combined with the relative risk for the benchmark.

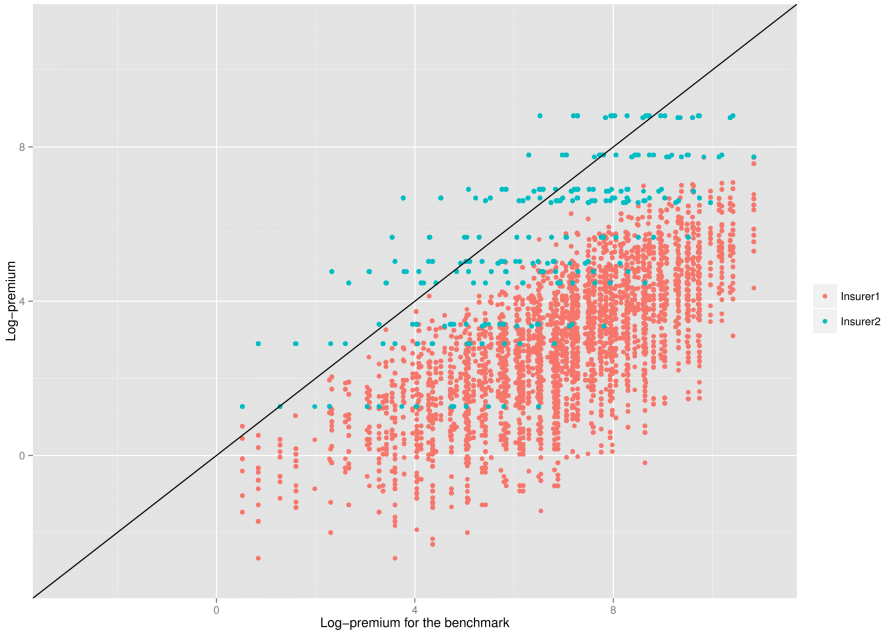

Figure 4.3: Relative risks for both insurers compared to the benchmark.

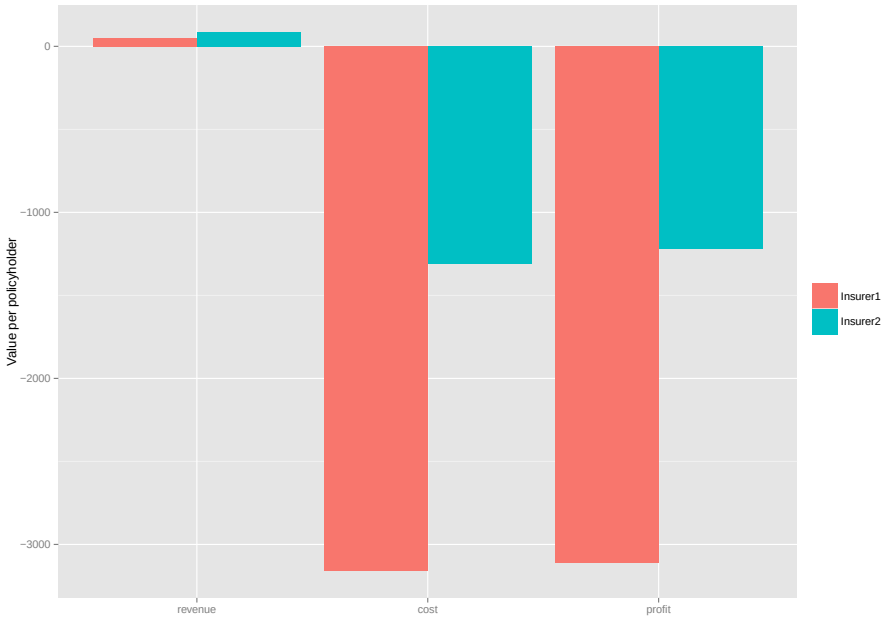

Figure 4.4: Revenue, costs and profit per policyholder for both insurers when compared to the benchmark.

4.2 An unequal market

The second simulation tests how market share effects the relative risks. The previous simulation suggested that less market share may cause for lower relative risks. This simulation is set up to test this hypothesis. The first insurer has 70% market share, the second has 30%. Both estimate their relative risks according to the model setup identical to the benchmark. This provides us with a comparison of a 30%/70%/100% subset of the market.
As with the previous simulation we look at the scatter plot with colored parameters and have the lower right plot reserved for the benchmark coloring. The plots are shown in Figure 4.5. These plots show that both models provide similar relative risks. The second insurer (having only a 30% market share) shows lower relative risks for the higher priced policies where the first insurer shows lower relative risk for the lower priced policies.
The location of the parameters show that vehicle age and zone are probably driving the model as their relative risks are more closely grouped than engine power, gender (which isn't included in the model) and age. Moreover, for customers in zone 5 and vehicle age group 20 to 99, the first insurer provides a lower relative risk. Looking at the benchmark coloring, we see that it is consistent with the two subsets as expected.
Figure 4.6 was made to see how both insurers stack up to the benchmark. As in the previous simulation we see that the insurers provide often lower relative risks than the benchmark. We see however (as with the comparison between the insurers) that for the lower priced premiums, the benchmark is often cheaper.
We can therefore conclude that for this data and this model choice, that an increase in availability of data leads to higher relative risks communicated to the market. The exact reason for this is currently unknown.

Testing profitability

As in the previous example, Figure 4.6 suggests that both insurers will loose against the market. Figure 4.7 clearly supports this theory. In this simulation the second insurer is in general cheaper and therefore it takes up 95% of all sales in the portfolio. The first insurer is only cheaper in the lower risk region as shown in Figure 4.5 therefore its relative revenue and cost are both lower.

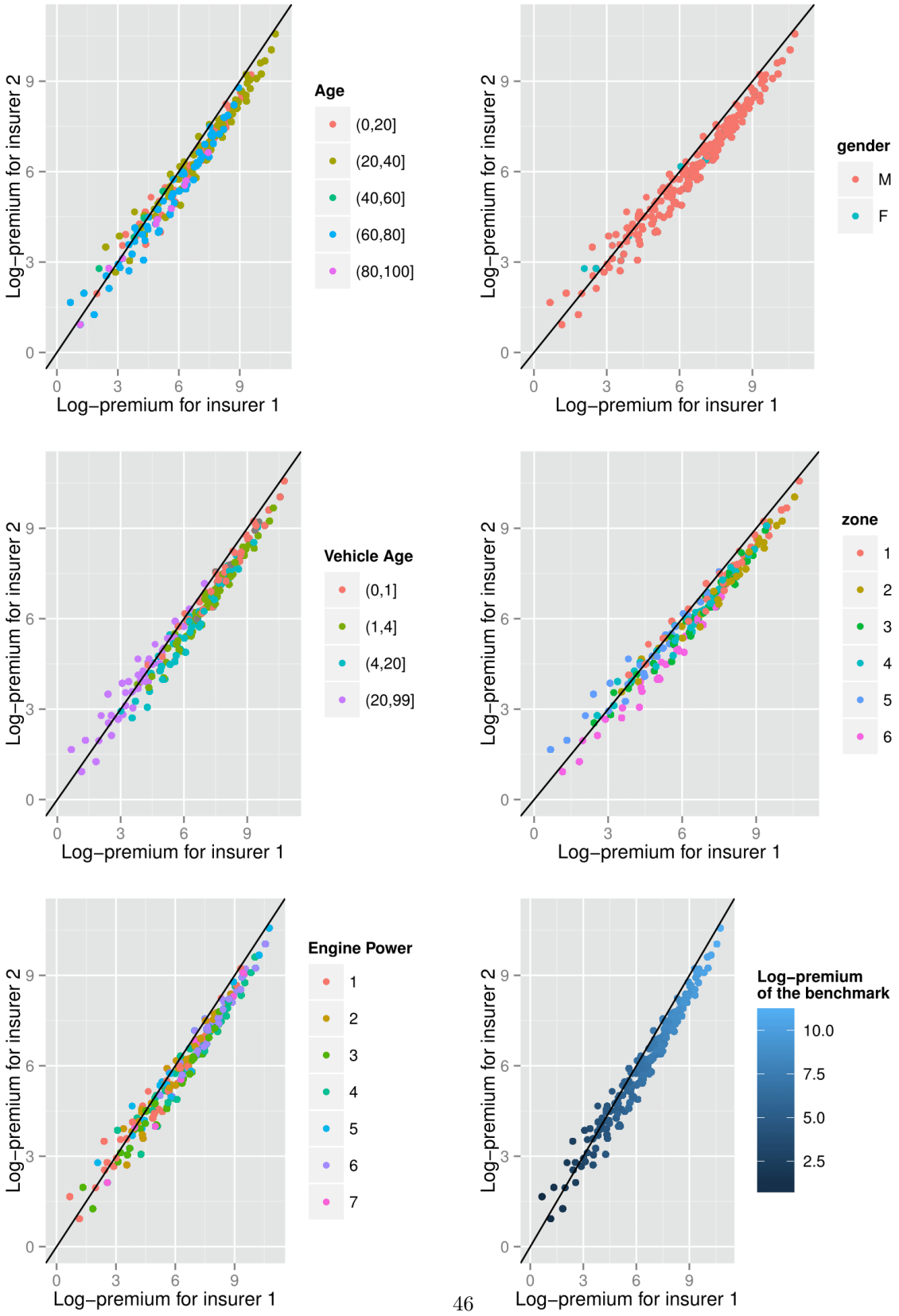

Figure 4.5: Scatter plots showing the relative risks colored per parameter in the model combined with the relative risk for the benchmark.

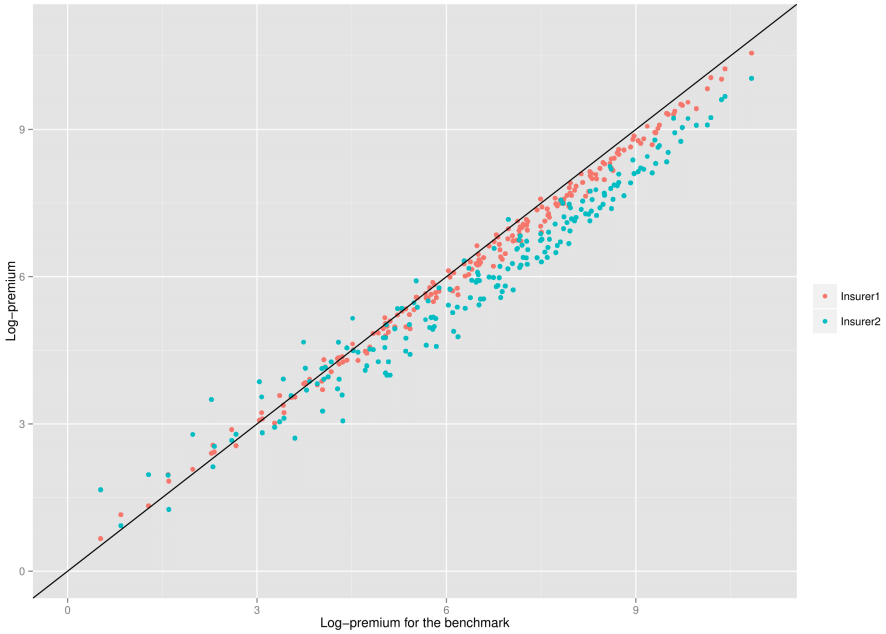

Figure 4.6: Relative risks for both insurers compared to the benchmark.

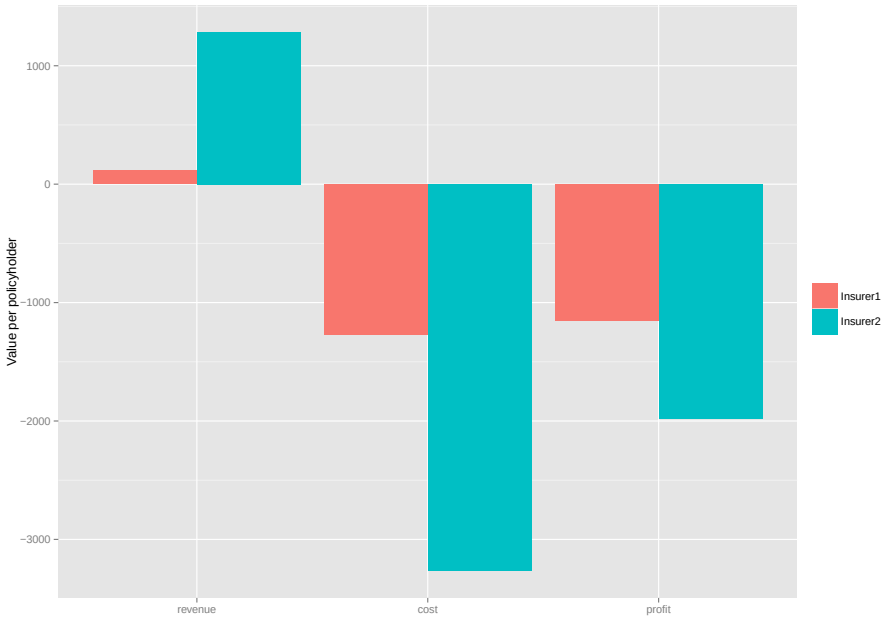

Figure 4.7: Revenue, costs and profit per policyholder for both insurers when compared to the benchmark.

4.3 GLM vs GLMM

The previous examples show how choice of number of parameters and insurer size influence the relative risks for insurers in a competitive market. Considering these effects on this dataset and model choice we let the second insurer switch to a GLMM approach and see how this effects the relative risks and sales. As GLMMs are known to have convergence issues as discussed in section 2.7 and Appendix B.2.2 we decide to only do 5 runs per simulation. The market is split up in two subsets of equal size. The first insurer chooses a simplified version of the model in the previous simulation. This model includes zone, engine power and vehicle age. The second insurer starts with an identical model to the first and then randomizes several effects. We consider four different model, one with random zone, one with random engine power, one with random vehicle age and a full random model considering all these factors as random effects. As we use a different model approach, we first view the relative risks per parameter before looking at the difference in relative risk presented to the customers. The factors for the models are shown in Figure 4.8. In this figures, the differently shaped

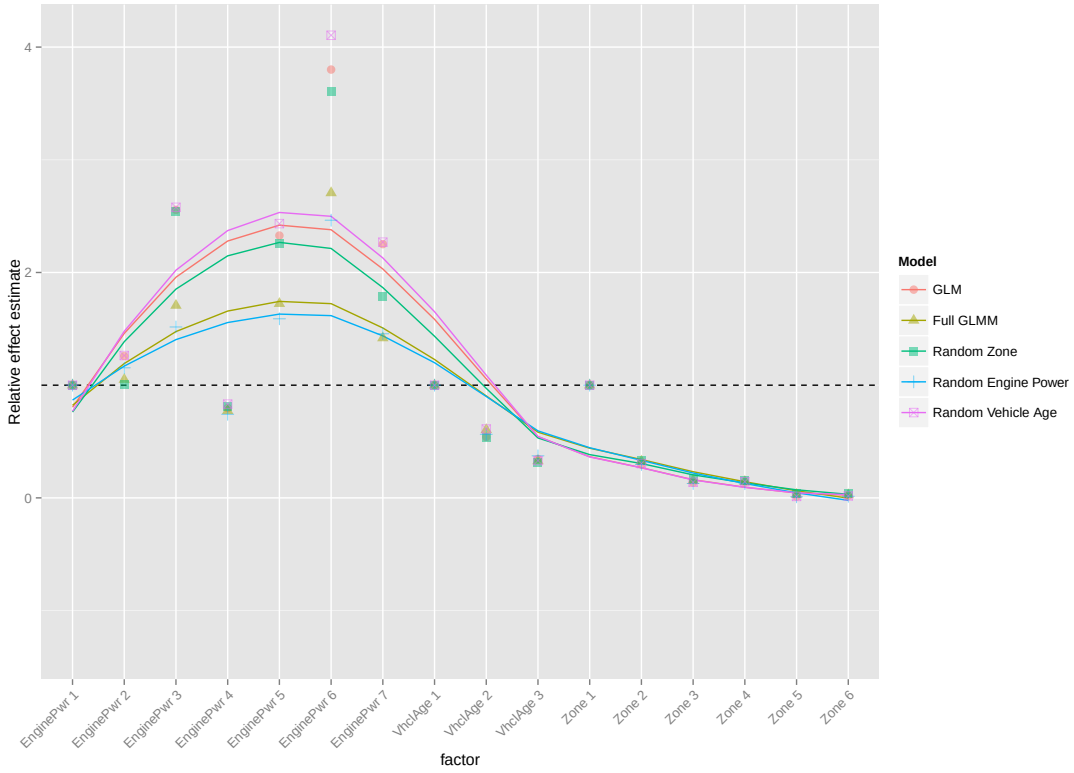

Figure 4.8: Estimated relative effects for the different GL(M)Ms for the second insurer.

and colored dots represent the different models, the lines show the trend in the relative effects. For each model, the intercept has relative effect 1, therefore it is plotted on the the dotted line. Note that as the GLMM assumes a distribution

on the random effect, the expected value for the random effect is taken as the effect in the figure. We see that for vehicle age and zone, the random model and the GLM coincide in their model estimates. Therefore, adding random effects for these parameters will not influence the model outcome. For engine power however, the models vary a lot. This data seems to be structured in some sense. Therefore we see that the random engine power model takes values closest to the relative effect 1. The full random model is the next closest, followed by the other models who provide more extreme estimations.
Knowing how the different models act, we decide to compare the full GLMM and the GLM as we expect these to vary the most. As in Figure 4.9 we create a scatter plot for the first insurer using the GLM and the second insurer using a full GLMM approach. The GLM approach for the second insurer perfectly coincides with that for the first as their model and data split are similar, therefore it will not be displayed.
Comparing this figure with Figure 4.8, we see that as the GLM and the GLMM coincide for relative effects smaller than 1 represented by the lower priced relative risks. Therefore, the first insurer using a GLM approach will have a small advantage on sales in that section. For the higher priced relative risks however, the second insurer using a GLMM approach will have smaller relative risks by the dampened effect of the GLMM model estimation.
For the factors, only zone shows a clear trend in grouping different effects. Zone 5 and 6 prefer the fixed model whereas the rest prefer the random model. The rest of the factors seem oblivious of the random choice.
Figure 4.10 shows the relative risks compared to the benchmark, here the same trend is seen but the second insurer also provides relatively high premiums compared to the benchmark in the lower regions.

Testing profitability

Lastly we also test the profitability, when comparing the GLM to the full GLMM we see that the second insurer acquires an 89% market share. Its revenue per customer is higher than it is for the first insurer but so are its costs. As in the previous three examples, the overall market is loosing. Keeping in mind that the insurer using the GLMM has sales in the higher priced part of the market, it is not strange that its revenue, cost and loss are also larger.

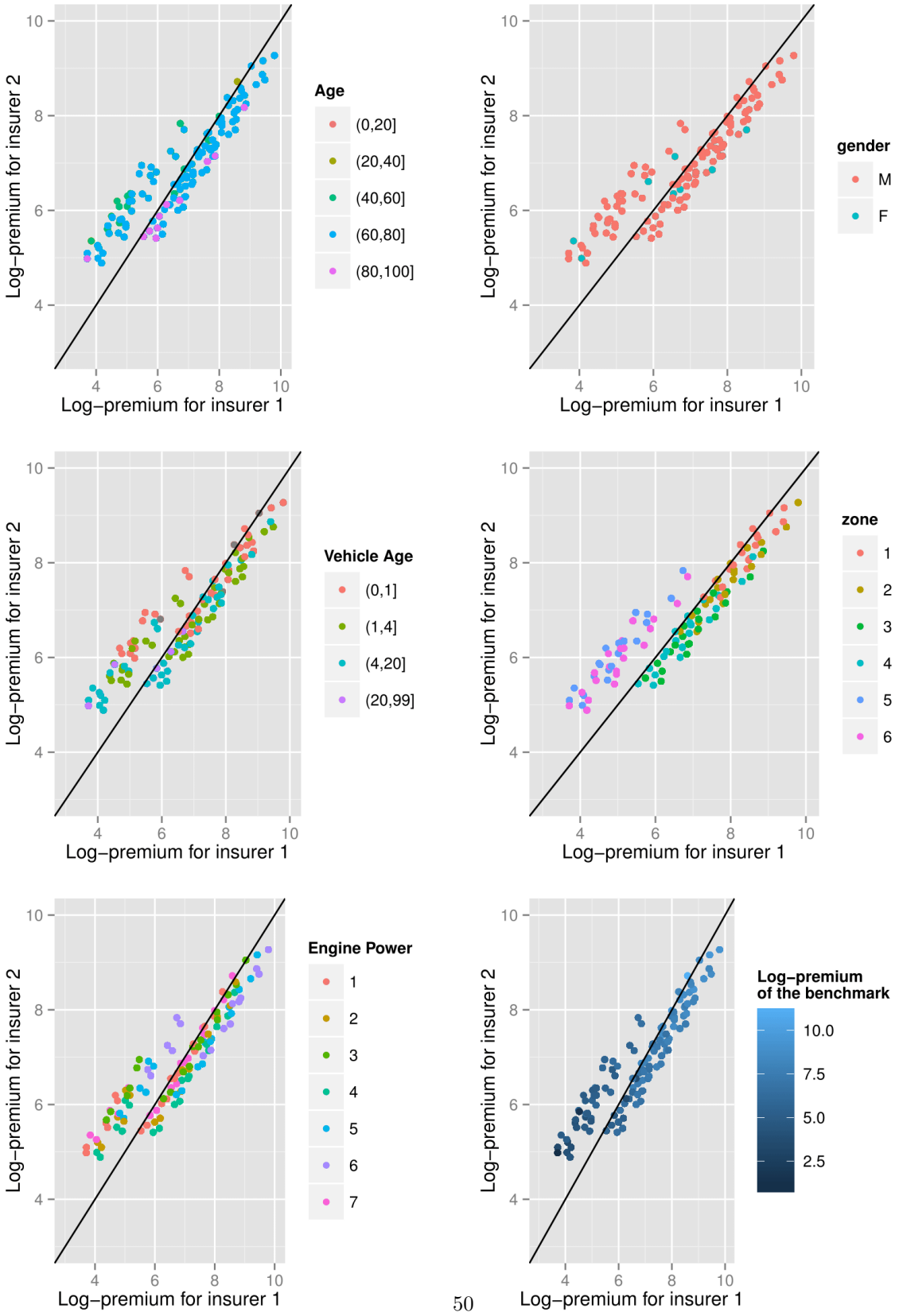

Figure 4.9: Scatter plots of the relative risks colored per parameter in the model combined with the relative size of the benchmark.

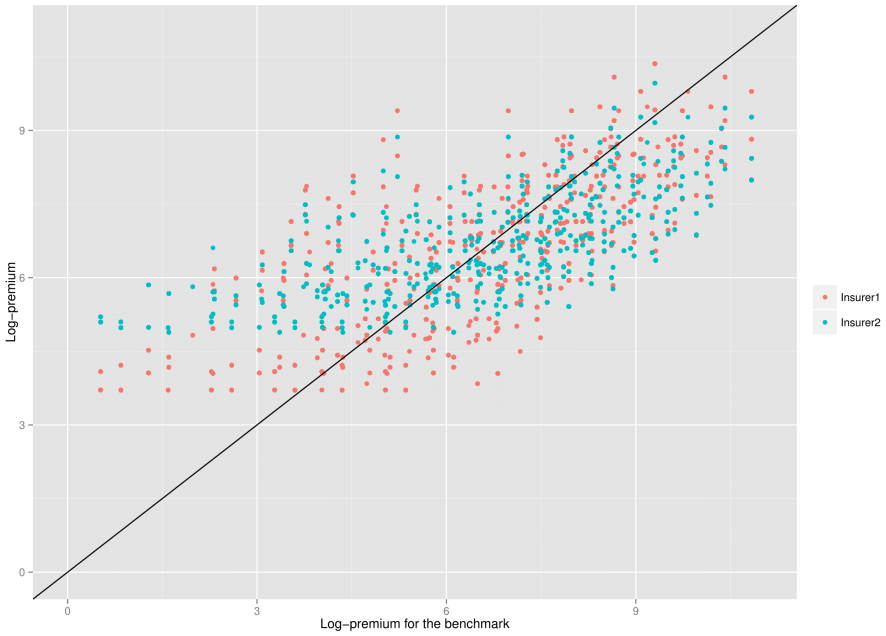

Figure 4.10: Relative risks for both insurers compared to the benchmark.

Figure 4.11: Revenue, costs and profit per policyholder for both insurers.

Chapter 5

Conclusion and further research

Pricing non-life insurance premiums is as much an art as it is statistical practice. Experience and explainability are for most practitioners as important as the statistical significance when building a model. Therefore, pricing these products most of all requires experience. This does not only hold for the model itself but also for the choice between random or fixed models.
Generalized Linear Mixed Models provide a good alternative for Generalized Linear Models. As seen in the theory and the third simulation, estimated factors are closer to the intercept leading to less variance in relative risks. Less variance in the effects with a minimal cost in model accuracy is certainly desirable. There are however many practical issues with calculation and understanding of the model output. From a viewpoint that a model should be as simple as possible but not simpler, GLMMs may fail to meet this criterion when comparing their understandability and performance with the GLM. Any practitioner should ask themselves whether the increase in complexity makes up for the (often) minimal increase in model performance.
In a competitive market the choice of model can heavily effect the profitability of both insurers acting on that market in this setup. Over-fitting can lead to a loosing market for both insurers. This can be due to choosing relatively many factors in a GLM approach or less availability of data. Surprisingly we saw that on the used dataset, when using the same model an insurer with much less data will quote lower premiums leading to a dominant place on the market. Even though more data should provide a better fit, less data can lead to more sales. A GLMM approach seems to outperform the GLM approach currently employed by market due to less extreme quoted premiums. Therefore, in the region with higher relative risks more sales are made.
With respect to the questions raised by Giro APT [63], a GLMM approach will directly lead to credibility and blending as supposed by the working party itself. A GLMM approach can help decrease the fact that the prediction of a risk depends on data in other completely independent segments as the GLMM accounts for structure in the data. The rest of the questions raised do not seem to be directly related to a GLMM based solution but require either different model setups or machine learning solutions where relative risks are updated according

to market movements and future expected mixture of the market.

Further research

With respect to further research, a Bayesian or Monte Carlo Markov Chain approximation may lead to more stability, better convergence and more flexibility in the model setup for GLMMs. Bayesian frameworks can however be technically challenging and require a different approach than the usual frequentist's view used by most statisticians. Another price to pay will be the longer running time of the process.
Above used simulation can be done on another dataset to see if the relationship between the increase in available data and relative risk hold for different situations. If so, a theoretical explanation may be found within either the GLM or its use on pricing data.
Apart from the GLMM, other regression and classification models often used in statistical learning may be tested in this context. It may be that they outperform a GL(M)M with less computation time or better convergence.
With respect to the competitive market, more complicated simulations can be done. Multiple insurers, dynamic customers and marketing strategies can all be build on top of this framework. Seeing how customers or insurers act on a more realistic market could provide different insights in the dynamics of the market.

Appendix A

Non-Life insurance mathematics

A.1 Fundamental probability and statistics

Throughout this appendix we assume the reader has knowledge of probability theory and some knowledge of measure theory. For an introductory book into probability theory see Dekking [26], for books introducing measure theory see either Shilling [53], Spreij [56] or Billingsby [13]. Throughout this text, we will follow the notation used by Spreij.
As we only have to deal with probability spaces in this context, we restrict ourself to use of probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, we only state definitions and results needed by models and estimation and approximation techniques used in this thesis.

A.1.1 Probability and measures

Suppose we have the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which consists of a non-empty set Ω, with $\sigma-$algebra $\mathcal{F}$ and probability measure $\mathbb{P}$. We can define the Borel sets of $\mathbb{R}$ as follows:

Definition A.1. The Borel sets $\mathcal{B} = \mathcal{B}(\mathbb{R})$ are the smallest $\sigma-$algebra generated by all the open sets $\mathcal{O}$ of $\mathbb{R}$.

Random variables on these sets can be defined as follows

Definition A.2. A mapping $X : \Omega \to \mathbb{R}$ is called a random variable if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$

As we do not want to go into the measure theoretical details too much, it is safe to assume that these definitions work in a similar fashion for products of measure spaces. In this text only finite such products are considered therefore we can almost directly translate these results to so called random vectors.
Having constructed these foundations, we can start by defining distribution functions. Define a measure $\mu : \mathcal{B} \to [0, 1]$ as $\mu(B) = \mathbb{P}(X^{-1}[B])$. Now as these sets B are generated by open subsets of $\mathbb{R}$ we see that $\mu((-\infty, x]) = \mathbb{P}(X \leq x)$. And thus a distribution function F can be defined.

Definition A.3. A distribution function F is a function $F : \mathbb{R} \to [0, 1]$ defined as $F(x) = \mu((-\infty, x]) = \mathbb{P}(X \leq x)$

A.1.2 Integration of random variables

The Lebesgue integral is a way of integrating random variables and evaluating the distribution function F. We will however, not make any distinction between discrete or continuous distributions as it limits our handling of cases which are neither discrete nor continuous.
Later on we will define the Stieltjes and Riemann integrals and show how and when these coincide with our Lebesgue integral.

Definition A.4. Let f be a function, then we can define the Lebesgue integral of f with respect to the measure μ as

$$\int f d\mu.$$

We need this definition to show that the expectation of a random variable is a Lebesgue integral. And moreover, can be given as such without making assumptions on any properties of the underlying distribution. Hence, if we change the definition of A.4 to the space $(\Omega, \mathcal{F}, \mathbb{P})$ then a random variable X,

$$\mathbb{E}(X) = \int_\Omega X d\mathbb{P},$$

which is well defined if $\mathbb{P}(|X|) < \infty$. This definition is nicer than the one given by either the sum or the Riemann integral. It is however, quite easy to show that they are the same.

Lemma A.1. *Suppose we have a countable set $\Omega = \{\omega_1, \omega_2, \ldots\}$ and we have a sequence of real numbers $p_j \in [0, 1]$ for $j \in \mathbb{N}$ such that their summation equals 1. Now, define $\mathcal{F} = \sigma(\Omega)$, then we can define the function $P : \Omega \to [0, 1]$ as follows*

$$P(A) = \sum_{j:\omega_j \in A} p_j = \sum_{j \in \mathbb{N}} p_j 1_{w_j}(A)$$

Hence, this function is a probability measure on the space (Ω, cf). Now we can define a random variable X such that $X = \sum_0^\infty x_i 1_{X_i}$. And hence, it has Lebesgue integral

$$\int_\Omega X d\mathbb{P} = \int_\Omega \sum_0^\infty x_i 1_{X_i} d\mathbb{P}$$

Which in this case equals $\sum_j x_j \mathbb{P}(X = x_j)$ which is the familiar expression of our expectation. Hence, the Lebesgue integral expectation is the same as the expectation of a discrete random variable.

Now forward to the continuous case. We state Example 4.28 from [56]. Here, $\lambda(\cdot)$ is the Lebesgue measure and we have a Borel measureable function $h : \mathbb{R} \to \mathbb{R}$ with $h \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

Lemma A.2. *Suppose there exists $f \geq 0$, Borel-measurable such that for all $B \in \mathcal{B}$ one has $\mathbb{P}(X \in B) = \lambda(1_B f)$, in which case it is said that X has a density f. Then, provided that the expectation is well defined, we see that we have*

$$\mathbb{E}h(X) = \int_R h(x)f(x)dx,$$

which is the familiar formula for the expectation of h(X).

Moments of X can be constructed as

$$\mathbb{E}X^k = \int_{\mathbb{R}} x^k d\mathbb{P}$$

Therefore we can define the usual moments of X.

Definition A.5. By usage of the above moment, we can define the following important functions for a random variable X.

- We can define the mean or first moment of X by

$$\mu_X = \mathbb{E}X$$

- The variance of X can be given by

$$\sigma_X^2 = Var(X) = \mathbb{E}(X - \mathbb{E}X^2)^2 = (\mathbb{E}X)^2 - \mathbb{E}X^2$$

- The standard deviation of X is given by

$$\sigma_X = Var(X)^{-1/2}$$

- The coefficient of variation of X is given by

$$Vco(X) = \frac{\sigma_X}{\mathbb{E}X} \text{ for } \mathbb{E}X > 0$$

- We can define the skewness of X by

$$\varsigma_X = \frac{\mathbb{E}(X - \mathbb{E}X)^3}{\sigma_X^3}$$

Stieltjes and Riemann Integrals

Riemann integrals are the go-to way in applied mathematics, we can derive these from the Lebesgue version. From that, we can also define the Riemann-Stieltjes and Lebesgue-Stieltjes integrals.

Lemma A.3. *Suppose that f is Borel measurable and it is Lebesgue integrable on a finite interval $[a, b]$. Then the Lebesgue integral must coincide with the Riemann integral if it exists. Moreover, if f is continuous then it is Riemann integrable on the interval $[a, b]$.*

This difference seems quite small, but consider the following example. Suppose we have a function defined as 0 on $\mathbb{Q}\cap[a,b]$ and 1 on $\mathbb{R}\cap[a,b]$. This function has Lebesgue integral 1, but is not Riemann integrable. Hence, Lebesgue integrals allow us to deal with more general cases.
Next up are the Riemann-Stieltjes integrals, which are Lebesgue measures with respect to the measure bestowed by the distribution functions F. For this we need the Fundamental theorem of calculus given by

Theorem A.4 (Generalized fundamental theorem of calculus)**.** *Let $f:\mathbb{R}\to\mathbb{R}$ be measurable such that $\int_K |f|d\lambda < \infty$ for every compact $K\subset\mathbb{R}$. Define for any $a\in\mathbb{R}$ the function $F:[a,\infty)\to\mathbb{R}$ such that $F(x)=\int_{(a,x]} f d\lambda$. Then outside of a null-set N (A set of Lebesgue measure zero), F is differentiable and $F'(x)=f(x)$, for all $x\notin N$.*

This theorem may seem somewhat technical, however it leads to the well know fact that the distribution function F can in some cases be seen as the primitive function of f while integrating. And thus in these cases

$$\int_a^b f(y)dy = F(b)-F(a)$$

Moreover, using theorem A.4, for distribution functions F such that $\mu((a,b]) = F(b)-F(a)$ we can define the Reimann-Stieltjes integral as

Definition A.6. Suppose that f is bounded and Lebesgue measurable on $A=(a,b]$, then we can define the Riemann-Stieltjes integral as

$$\int_A f(x)dF(x) = \int_A f d\mu$$

Hence, the Riemann-Stieltjes integral is a version of the Lebesgue integral over the real line.

A.1.3 Useful theorems and inequalities

Theorem A.5 (Strong law of large numbers [56])**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ and finite variance σ^2. Then*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i = \mu \qquad a.s.$$

Theorem A.6 (Central Limit Theorem[56])**.** *Suppose that we have $X_1,\dots,X_n$ i.i.d. random variables with mean μ, positive, finite variance σ^2 Then the classical Central Limit Theorem states*

$$\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^{n}(X_j-\mu) \xrightarrow{W} \mathcal{N}(0,1)$$

Lemma A.7 (Hölder's ineqaulity [56])**.** *Let $p,q\in[0,\infty]$, $f\in\mathcal{L}^p(s,\Sigma,\mu)$ and $g\in\mathcal{L}^q(s,\Sigma,\mu)$. If $\frac{1}{p}+\frac{1}{q}=1$, then $fg\in\mathcal{L}^1(s,\Sigma,\mu)$ and $||fg||_1 \le ||f||_p||g||_q$*

Lemma A.8 (Markov's inequality [56])**.** *Suppose we have a real valued random variable X and an increasing function $g:\mathbb{R}\to[0,\infty]$ Then*

$$\mathbb{E}g(X) \ge g(c)\mathbb{P}(X\ge c)$$

A.1.4 Moment generating functions

Having defined the usual probabilistic tools, we can define a tool which will appear to be quite useful when identifying the properties of certain random variables. We will cover moment generating functions as used in [62].
A different way of finding the properties of a random variable X is through moment generating functions. They have the upside of being easier to handle and evaluate but have the big downside of requiring an exponentially fast decaying tail of the distribution.
It is however a nice tool with some really nice properties. We can define it as follows

Definition A.7. Let μ be a probability measure on $(\mathbb{R}, \mathcal{B})$. Its moment generating function $M(s) : A \to \mathbb{R}$ is defined by

$$M(s) = \int_{\mathbb{R}} e^{sx} \mu(dx)$$

Now, we define a subset A to be the set on which $M(s)$ exists. To be more precise, it is the interval $(-s_0, s_0)$, such that $M(s)$ exists for all $s \in A$. It turns out that $M(s)$ is finite, if it has finite bounds. Hence, we can proof using the Markov inequality given by A.8 that for bounded tails, the $M(s)$ is finite and thus exists. Moreover, we can rewrite our function to the usual sense $m(s) = \mathbb{E}e^{sX}$.
We need to now need to show two properties of this interval, first that if the existence holds on the boundary of the interval, it holds on its interior. Second, we need to show that we can pick the boundary such that finite tails lead to a finite value for $m(s)$.

Lemma A.9. *Suppose that we have a value s_0 such that $m(s_0) < \infty$. Then $m(s_0) < \infty$ for all $s \in [-s_0, s_0]$.*

Proof. We will use the convexity of the exponential function combined with monotonicity of the integral. As we look at values $s \in (-s_0, s_0)$, we can say that $s = (1 - 2\lambda)s_0$ for $\lambda \in [0, 1]$.
Thus we can write

$$\begin{aligned} e^{sX} &= e^{(1-2\lambda)s_0} X \\ &\leq (1 - 2\lambda)e^{s_0 X} < \infty \end{aligned}$$

Hence, we have finite values of $m(s)$ for $s \in [-s_0, s_0]$ □

Knowing this, we can move to the final theorem.

Theorem A.10. *Assume we have a probability measure μ on $(\mathbb{R}, \mathcal{B})$ with random variable X then the moment generating function $M(s)$ is finite and exists if and only if*

$$\mathbb{P}(X > x) \leq ce^{-\gamma x}$$

for some $c, \gamma > 0$

Proof. ($\Rightarrow$) Suppose $m(\gamma) < \infty$ for some $\gamma > 0$. Then by Markov's inequality we have that

$$\mathbb{P}(X > x) = \mathbb{P}(e^{tX} > e^{tx}) \leq e^{-tx}\mathbb{E}e^{t}X = m(t)e^{-tx}$$

Now, this is allowed as Markov's inequality works for any increasing function $g(t)$ by A.8. Now define $c = m(\gamma)$ and we are done.
($\Leftarrow$) Assume that there exist $c, \gamma > 0$ such that $\mathbb{P}(X > x) \leq ce^{-\gamma x}$. Then according to ex. 5.11 of [56] we can write that for any nonnegative random variable Y it holds that for $\alpha > 0$.

$$\mathbb{E}Y^{\alpha} = \alpha \int_0^{\infty} y^{\alpha-1}(1 - F(y))dy = \alpha \int_0^{\infty} y^{\alpha-1}\mathbb{P}(Y > y)dy$$

Now as our expression $e^{sX} > 0$ we can put $Y = e^{s}X$ and with $\alpha = 1$ obtain

$$\mathbb{E}e^{sX} = \int_0^{\infty} \mathbb{P}(e^{sX} > y)dy$$

Using this formula, we can obtain that using a property of the exponential function that

$$\begin{aligned}
\mathbb{E}e^{sX} &= \int_0^{\infty} \mathbb{P}(e^{sX} > y)dy \\
&\leq 1 + \int_1^{\infty} \mathbb{P}(e^{sX} > y)dy \\
&\leq 1 + \int_1^{\infty} cy^{-\gamma/s}dy
\end{aligned}$$

Which is finite for any $0 < s < \gamma$ by A.9. □

As we have established existence for the mgf, we can now look at some of its properties. The first lemma follows from the existence of all the moment generating functions in the interval $(-s_0, s_0)$.

Lemma A.11. *Assume that the moment generating function M(s) of a random variable X exists on the interval (s_o, s_0) then for $s \in (s_o, s_0)$ $M(s)$ has a power expansion of the form*

$$M(s) = \sum_{k \geq 0} \frac{s^k}{k!}\mathbb{E}X^k$$

Proof. Our proof is mostly based on the above lemma. Lemma A.9 implies that the $M(s)$ is finite, and thus $\mathbb{E}e^{sX}$ is finite. This also implies finiteness of $\mathbb{E}|X|^k$ for all $k \in \mathbb{N}$ as $|x^k| < e^{|sx|}$ for large enough x. Hence, by this bound, we see that by dominated convergence in the expansion we have the result. □

As an analytic result (see for example *Gevolg 4.35,[12]*) we see that the moment generating function is infinitely differentiable on its domain. Moreover, these derivatives at the origin for $k \in \mathbb{N}\backslash\{0\}$ are given by

$$\frac{d^k}{ds^k}M(s)|_{s=0} = \mathbb{E}X^k$$

Hence, the moment generating function gives us exactly the moments we need. And can thus help us tremendously in determining variance and skweness of many (light-tailed) distributions. We now state one more lemma for which the proofs can be found in section 30 of Billingsley[13]. It mainly states that the distribution F of X is uniquely determined by its moments (and thus moment generating function).

Lemma A.12. *Consider a random variable X, with finite $M(s)$ on some interval $(-s_0, s_0)$.*

(a) The distribution function F of X is completely determined by its moment generating function $M(s)$

(b) If $X \geq 0$ a.e. Then the distribution function F of X is completely determined by $M(s)$ independent of the finite constraint.

Moreover, we see that the moment generating function carries some similar properties as the characteristic functions known to measure theory.

Lemma A.13. *Assume that for a random variables X, Y the mgf $M(s)$ exists for $s \in (-s_0, s_0)$ then, we have the following properties.*

1. *If X and Y are independent $M_{aX+bY}(s) = M_X(as) \cdot M_Y(bs)$*
2. *Let X and Y be independent random variables. Define Z to be the random variable which equals X with probability p and Y with probability 1-p. Then we have that the mgf of Z is given by*

$$M_Z(s) = pM(s) + (1-p)M(s)$$

A.2 The exponential family

The exponential family plays a key role in the modeling of non-life mathematics. It is an essential part of the GL(M)M and requires a proper definition. Therefore, we will define it in a measure theoretic fashion. We will use [19] for the definition and the results with an update to modern notation. Original credits are given to [50], [25] and [41].

Definition A.8 (Standard exponential family)**.** Let μ be a σ-finite measrue on the Borel subsets of $\mathbb{R}^k$. Let

$$N = N_\mu = \{\theta : \int e^{\theta x} \mu(dx) < \infty\}$$

And define a function $\lambda(\theta)$ as

$$\lambda(\theta) = \int e^{\theta x} \mu(dx)$$

With $\lambda(\theta) = \infty$ if the integral is infinite. Next if we let

$$\Psi(\theta) = \log \lambda(\theta)$$

And define the function $p_\theta(x) = \exp(\theta x - \Psi(\theta))$. Then the family of probability densities given by $\{p_\theta : \theta \in \Theta\}$ for $\Theta \subset N$ is called the k-dimensional standard exponential family.

The distributions $p_\theta(A) = \int_A p_\theta(x)\mu(dx)$ is called the standard exponential family. N is the natural parameter space, Ψ is the log Laplace transform (conform [64]) and θ is the canonical parameter.
Convexity of the parameter space N and function Ψ is needed to allow use the mgf on this family. And we can use A.11 to build distributions in the exponential family.

Theorem A.14. *N is a convex set and Ψ is a convex function on N.*

Proof. Convexity of N follows from the definition, as it is an open set. Let $\theta_1, \theta_2 \in N$ then by Hölder's inequality A.7 we have that for some $\alpha \in (0, 1)$:

$$\begin{aligned}\exp(\psi(\alpha\theta_1 + (1-\alpha)\theta_2)) &= \int \exp\left(\psi(\alpha\theta_1 + (1-\alpha)\theta_2)x\right)\mu(dx) \\ &= \int \exp\left(\theta_1 x\right)^{\alpha} \cdot \exp(\theta_2 x)^{1-\alpha}\mu(dx) \\ &\leq \exp(\alpha\Psi(\theta_1) + (1-\alpha)\Psi(\theta_2))\end{aligned}$$

□

For the convex set n we can use A.11 to show that the interval $(-s_0, s_0)$ exists in all k dimensions (by convexity). Hence, if we choose μ as the Lebesgue-measure restricted to Ω, we can define a probability measure $\mathbb{P}_\theta(dx) = p_\Psi(x;\theta)\mu(dx)$. And hence, our function $p_\theta(x;\theta)$ ia a probability distribution, following the definition $p(x;\theta) = \exp\left(\theta x - \Psi(\theta) + k(x)\right)$ where $k(x)$ is the carrier measure following from the chosen probability measure $\mathbb{P}_\theta$. Thus we can define the moment generating function

$$m_\theta(x) = \exp(\Psi(\theta + x) - \Psi(\theta))$$

From which, we can now generate with an extra parameter ϕ

$$p(x;\theta,\phi) = \exp((\theta x - \psi(\theta))/\phi + k(x,\phi)) \tag{A.1}$$

Which leads to the known notation for exponential families.

Using the differentiability and the moments of Ψ, we see that $\mathbb{E}x = \psi'(\theta)$ and $\mathrm{Var}(y) = \phi\psi''(\theta)$. The full proof is shown in [19] Chapter 2.

Appendix B

R-code

B.1 Hausman Test

```
## Code from: http://stackoverflow.com/questions/23630214/hausmans-specification-test-for-
phtest_glmer <- function (glmerMod, glmMod, ...)  {  ## changed function call
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)  ## added as.matrix()
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))  ## changed
class(res) <- "htest"
return(res)
}
```

B.2 Errors and failures in the glmer function

B.2.1 Example 2.3

Code for shrinking, centering and rescaling fixed effects. Here the data is denoted by `data` and a predictor is denoted by `fixed_effect`.

```
# Shrinking the fixed effect
```

```
data$fixed_effect_small <- data$fixed_effect-min(data$fixed_effect)
# Centering the fixed effect manually:
data$fixed_effect_centered <- data$fixed_effect-mean(data$fixed_effect)
# Centering the fixed effect automatically:
data$fixed_effect_centered2 <- scale(data$fixed_effect, center=TRUE, scale = FALSE)
# Centering and scaling the fixed effect manually:
data$fixed_effect_scaled <- (data$fixed_effect-mean(data$fixed_effect))/sd(countData$fixed
# Centering and schaling the fixed effect automatically:
data$fixed_effect_centered2 <- scale(data$fixed_effect, center=TRUE, scale = TRUE)
```

B.2.2 Example 2.4

We now present first the code given in example 2.4 followed by a way to use a different optimizer and a way to try them all (beware of the long running time). We define the model as follows. Suppose we have some `data` with effects `effect_1`,...,`effect_N` and a response variable denoted by `response` which is assumed to be Poisson distributed with offset `exposure` then we can define our model to be of the form

```
glmmfit <- glmer(response ~ effect_i + (1 | effect_j) + offset(log(exposure)), data=data,
```

which is a standard Laplace approximation of our data. Then we can find singularities and new derivatives with the following R-code

```
## Code to check for singularities and poor derivatives
library(numDeriv)
# Find singularity
glmmfit_theta <- getME(glmmfit, "theta")
glmmfit_lower <- getME(glmmfit, "lower")
min(glmmfit_theta[glmmfit_lower == 0])

# Recalculate the derivatives
derivs_1 <- glmmfit@optinfo$derivs  # extract derivatives
# solve the equation for the Hessian and gradient
glmmfit_derivs_new <- with(derivs_1, solve(Hessian, gradient))
# compares the minima of the derivatives.
# In essence reproduces the value given by the error
max(pmin(abs(glmmfit_derivs_new), abs(derivs_1$gradient)))

# Repeat the calculations with numDeriv
# Only calculates deviance function hwen updating
glmmfit_devFunOnly<- update(glmmfit, devFunOnly=TRUE)
# Substract effect parameters
glmmfit_pars <- unlist(getME(glmmfit, c("theta", "fixef")))
derivs_2 <- list(gradient = grad(glmmfit_devFunOnly, glmmfit_pars),
Hessian = hessian(glmmfit_devFunOnly, glmmfit_pars)) # Retrieve new gradient and Hessian
# solve the equation for the Hessian and gradient
glmmfit_numDeriv <- with(derivs_2, solve(Hessian, gradient))
max(pmin(abs(glmmfit_numDeriv), abs(derivs_2$gradient))) # test new value

## Create new starting point for the model with more iterations
```

```
new_start <- getME(glmmfit, c("theta", "fixef"))
glmmfit_new <- update(glmmfit, start=new_start,
control=glmerControl(optCtrl=list(maxfun=2e4)))
```

Choice of optimizer functions

In the spirit of: "If all else fails, change optimizer" we here state a script which switches optimizers for `glmer`. This code is due to Ben Bolker and given on Stackoverflow. Notable to mention is the explained version given by Rstudio.

```
## Code to switch optimizers when using the glmer function.
# Credits go to http://stackoverflow.com/questions/23478792/
# warning-messages-when-trying-to-run-glmer-in-r
# and https://rstudio-pubs-static.s3.amazonaws.com/
# 33653_57fc7b8e5d484c909b615d8633c01d51.html
# Contrary to these sources, we load the function allFit
# from the afex package in R
# Loading the needed packages
library(afex) # Contains the allFit Function
library(optimx) # Contains some optimizers
library(nloptr) # Contains some more optimizers

# Start at the old solution

new_start <- getME(glmmfit, c("theta", "fixef"))

# Manual calculation with the "obyqa" optimizer

glmmfit_2 <- update(glmmfit, start = new_start,
control=glmerControl(optimizer="bobyqa",
optCtrl = list(maxfun=2e4)))

# Rescaling, choice of scaled parameters has to be done
# manually

data_scaled <- data
# cols_scaled needs to be chosen as set of columns
data_scaled[, cols_scaled] <- scale(data_scaled[, cols_scaled])
glmmfit_scaled <- update(glmmfit,data=data_scaled)

# Automated version using allFit

glmmfit_all <- allFit(glmmfit_scaled)
check_ok <- sapply(glmmfit_all, is, "merMod")
glmmfit_all_ok <- glmmfit_all[check_ok]

# Pull out the warnings

lapply(glmmfit_all_ok,function(x) x@optinfo$conv$lme4$messages)
```

```
# Check log-likelihood and AIC

summary(sapply(glmmfit_all_ok,logLik),digits=6)
summary(sapply(glmmfit_all_ok,AIC),digits=6)

# Optional: Make a ggplot of all the optimizers

library(ggplot2)
library(reshape2)
library(plyr)

glmmfit_fixef <- t(sapply(glmmfit_all_ok,fixef))
glmmfit_fixef_melt <- melt(glmmfit_fixef)
glmmfit_models <- levels(glmmfit_fixef_melt$Var1)

(gplot1 <- ggplot(glmmfit_fixef_melt,aes(x=value,y=Var1,colour=Var1))+
geom_point()+
facet_wrap(~Var2,scale="free")+
scale_y_discrete(breaks=models,
labels=abbreviate(glmmfit_models,6)))

# Calculate coefficients of variation of the fixed-effect
# parameter estimates:
summary(unlist(daply(glmmfit_fixef_melt,"Var2",summarise,
sd(value)/abs(mean(value)))))
```

B.3 Rescaling fixed effects from 2.7.2

Here, we give example code, in which we use the dataset and fitted values from [30] chapter 16.

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

library(lme4)
wc <- read.table("http://instruction.bus.wisc.edu/jfrees/
jfreesbooks/PredictiveModelingVol1/files/chapter-16/CountsWorkers.txt",header=T)

glmmfitnAGQ <- glmer(count ~ year
+ (1|riskclass)
+ offset(log(payroll)),family = poisson(link = "log"),
data = wc, nAGQ=25)

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

wcScaled <- cbind(wc, yearScaled=scale(wc$year))
wcCentrd <- cbind(wc, yearCentrd=scale(wc$year, scale=FALSE))

# Rerun the glmm/glm fits
glmmfitnAGQSc <- glmer(count ~ year + (1|riskclass)
```

```
+ offset(log(payroll)),family = poisson(link = "log"),
data = wcScaled, nAGQ=25)
glmmfitnAGQCt <- glmer(count ~ yearCentrd
+ (1|riskclass)
+ offset(log(payroll)),family = poisson(link = "log"),
data = wcCentrd,nAGQ = 25)

# Build the predictions

predictionTableGlmmnAGQ <- cbind(glmmfitnAGQ=predict(glmmfit, type="response"),
glmmfitnAGQSc=predict(glmmfitSc, type="response"),
glmmfitnAGQCt=predict(glmmfitCt, type="response")
)
## Output:
head(predictionTableGlmmnAGQ)
# glmmfitnAGQ glmmfitnAGQSc glmmfitnAGQCt
# 1    2.361078      2.361080      2.361078
# 2    2.499743      2.499745      2.499743
# 3    3.264530      3.264532      3.264530
# 4    3.545302      3.545303      3.545302
# 5    2.669812      2.669812      2.669813
# 6    2.560298      2.560297      2.560298
# Calculate the total squared error

avgDiffGlmmnAGQ <- c(sum((predictionTableGlmmnAGQ[, 2]-predictionTableGlmmnAGQ[, 1])^2),
sum((predictionTableGlmmnAGQ[, 3]-predictionTableGlmmnAGQ[, 1])^2))
## Output:
print(avgDiffGlmmnAGQ)
# [1] 9.925802e-08 1.206280e-08
```

Appendix C

Appendix to chapter 3

Customer choice distributions, mean and variance

The variance derivation for the uniform case:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 \\ &= \frac{1}{k}\sum_{i=1}^{k} i^2 - \frac{(k+1)^2}{4} \\ &= \frac{k^2}{3} + \frac{k}{2} + \frac{1}{6} - \frac{(k+1)^2}{4} \\ &= \frac{1}{12}(k^2-1)\end{aligned}$$

The mean and variance derivation for the finite geometric case:

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=1}^{n} k\mathbb{P}(X=k) \\
&= \sum_{k=1}^{n} k\frac{1-r}{1-r^n}\cdot r^{k-1} \\
&= \frac{1-r}{r-r^{n+1}}\sum_{k=1}^{n} kr^k \\
&= \frac{1-r}{1-r^n}\frac{1-(n+1)r^n+nr^{n+1}}{(1-r)^2} \\
&= \frac{nr^{n+1}-(n+1)r^n+1}{(1-r)(1-r^n)} \\
\mathrm{Var}(X) &= \mathbb{E}X^2-(\mathbb{E}X)^2 \\
&= \sum_{k=1}^{n} k^2\mathbb{P}(X=k)-(\mathbb{E}X)^2 \\
&= \frac{1-r}{r-r^{n+1}}\sum_{k=1}^{n} k^2r^k-(\mathbb{E}X)^2 \\
&= \frac{1-r}{1-r^n}\frac{1+r-(n+1)^2r^n+(2n^2+2n-1)r^{n+1}-n^2z^{n+2}}{(1-r)^3}-(\mathbb{E}X)^2 \\
&= \frac{r}{(r-1)^2}-\frac{n^2r^n}{(r^n-1)^2}
\end{aligned}$$

Bibliography

[1] Linear mixed models. https://www.statistics.ma.tum.de/fileadmin/w00bdb/www/czado/lec10.pdf, 2004.

[2] Claim prediction challenge (allstate). https://www.kaggle.com/c/ClaimPredictionChallenge/data, 2011.

[3] Generalized linear models, abridged. https://github.com/bwlewis/GLM, 2011.

[4] arm: Data analysis using regression and multilevel/hierarchical models. https://cran.r-project.org/web/packages/arm/, 2015.

[5] caic4: Conditional akaike information criterion for lme4. https://cran.r-project.org/web/packages/cAIC4/, 2015.

[6] Fitting linear models. https://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html, 2015.

[7] glmmbugs: Generalised linear mixed models and spatial models with winbugs, brugs, or openbugs. https://cran.r-project.org/web/packages/glmmBUGS/index.html, 2015.

[8] Mumin: Multi-model inference. https://cran.r-project.org/web/packages/MuMIn/, 2015.

[9] optimx: A replacement and extension of the optim() function. https://cran.r-project.org/web/packages/optimx/, 2015.

[10] Premies verzekeringen verschillen tot op huisnummer. http://www.consumentenbond.nl/actueel/nieuws/2015/verzekeringspremies-verschillen-tot-op-huisnummer/, 2015.

[11] winbugs. http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/, 2015.

[12] Erik P Ban. Dictaat functies en reeksen. 2014.

[13] Patrick Billingsley. Probability and measure. wiley series in probability and mathematical statistics. 1995.

[14] Ake Björck. *Numerical methods for least squares problems*. Siam, 1996.

[15] Åke Björck. *Numerical methods in matrix computations*. Springer, 2015.

[16] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135, 2009.

[17] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.

[18] I Bronshtein and K Semendyayev. *Handbook of mathematics*. Springer, 2013.

[19] Lawrence D Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages i–279, 1986.

[20] Hans Bühlmann. Experience rating and credibility. *Astin Bulletin*, 4(03):199–207, 1967.

[21] Hans Bühlmann and Alois Gisler. *A course in credibility theory and its applications*. Springer Science & Business Media, 2006.

[22] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

[23] Tom S Clark and Drew A Linzer. Should i use fixed or random effects? *Political Science Research and Methods*, 3(02):399–408, 2015.

[24] Claudia Czado and Adrian E Raftery. Choosing the link function and accounting for link uncertainty in generalized linear models using bayes factors. *Statistical Papers*, 47(3):419–442, 2006.

[25] Georges Darmois. Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, 260:1265–1266, 1935.

[26] Frederik Michel Dekking. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005.

[27] Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2008.

[28] Christopher Dougherty. *Introduction to econometrics*. Oxford University Press, 2011.

[29] Edward W Frees. *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press, 2004.

[30] Edward W Frees, Richard A Derrig, and Glenn Meyers. *Predictive Modeling Applications in Actuarial Science*, volume 1. Cambridge University Press, 2014.

[31] José Garrido and Jun Zhou. Credibililty theory for generalized linear and mixed models. 2006.

[32] William H Greene. *Econometric Analysis*. Prentice Hall, 2011.

[33] Robert E Greenwood and JJ Miller. Zeros of the hermite polynomials and weights for gauss' mechanical quadrature formula. *Bulletin of the American Mathematical Society*, 54(8):765–769, 1948.

[34] Sonja Greven and Thomas Kneib. On the behaviour of marginal and conditional akaike information criteria in linear mixed models. 2009.

[35] CE Grueber, S Nakagawa, RJ Laws, and IG Jamieson. Multimodel inference in ecology and evolution: challenges and solutions. *Journal of evolutionary biology*, 24(4):699–711, 2011.

[36] James William Hardin, Joseph M Hilbe, and Joseph Hilbe. *Generalized linear models and extensions*. Stata Press, 2007.

[37] Jerry A Hausman. Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pages 1251–1271, 1978.

[38] Christiaan Heij, Paul De Boer, Philip Hans Franses, Teun Kloek, Herman K Van Dijk, et al. *Econometric methods with applications in business and economics*. OUP Oxford, 2004.

[39] Robert I Jennrich and PF Sampson. Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18(1):11–17, 1976.

[40] Rob Kaas, Marc Goovaerts, Jan Dhaene, and Michel Denuit. *Modern actuarial risk theory: using R*, volume 128. Springer Science & Business Media, 2008.

[41] Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3):399–409, 1936.

[42] Pierre Simon Laplace. Memoir on the probability of the causes of events. *Statistical Science*, pages 364–378, 1986.

[43] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 161. SIAM, 1974.

[44] Qing Liu and Donald A Pierce. A note on gauss—hermite quadrature. *Biometrika*, 81(3):624–629, 1994.

[45] Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[46] Charles E McCulloch and John M Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

[47] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013.

[48] John A Nelder and RJ Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[49] Esbjörn Ohlsson and Björn Johansson. *Non-life insurance pricing with generalized linear models*. Springer Science & Business Media, 2010.

[50] Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 567–579. Cambridge Univ Press, 1936.

[51] John Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[52] Benjamin Saefken, Thomas Kneib, Clara-Sophie van Waveren, Sonja Greven, et al. A unifying approach to the estimation of the conditional akaike information in generalized linear mixed models. *Electronic Journal of Statistics*, 8(1):201–225, 2014.

[53] René L Schilling. *Measures, integrals and martingales*, volume 13. Cambridge University Press, 2005.

[54] Amartya Sen. *On economic inequality*. Oxford University Press, 1973.

[55] Edward Allen Silver. An overview of heuristic solution methods. *Journal of the operational research society*, 55(9):936–956, 2004.

[56] Peter JC Spreij. Measure theoretic probability. *UvA Course Notes*, 2012.

[57] Stephen M Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, pages 465–474, 1981.

[58] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

[59] Florin Vaida and Suzette Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.

[60] Marno Verbeek. *A guide to modern econometrics*. John Wiley & Sons, 2008.

[61] Shaomin Wu and Peter Flach. A scored auc metric for classifier evaluation and selection. In *Second Workshop on ROC Analysis in ML, Bonn, Germany*, 2005.

[62] Mario V Wuthrich. Non-life insurance: Mathematics & statistics. *Available at SSRN 2319328*, 2014.

[63] Ji Yeo. Generalized linear models for non-life pricing - overlooked facts and implications. *Institute and Faculty of Actuaries*.

[64] Harry Zanten. An introduction to stochastic processes in continuous time. *Lecture Notes*, 2004.

[65] Yihua Zhao, John Staudenmayer, Brent A Coull, and Matthew P Wand. General design bayesian generalized linear mixed models. *Statistical Science*, pages 35–51, 2006.