Stockholms universitet

Bayesian credibility methods for pricing non-life insurance on individual claims history

Daniel Eliasson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2015:1
Försäkringsmatematik
April 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2015:1**
http://www.math.su.se

Bayesian credibility methods for pricing non-life insurance on individual claims history

Daniel Eliasson*

April 2015

Abstract

Claim frequencies in non-life insurance are typically modelled using generalised linear models (GLMs), making the assumption that all insurance policies with the same covariates have homogeneous risk. It is known in the insurance industry that there remains a relatively large heterogeneity between policies in the same tariff cell. Credibility theory is the study of how best to combine the collective prediction for a tariff cell with the experienced claim frequency of an individual policy, in order to obtain a more accurate policy-level prediction. In this thesis, we consider a credibility model in the form of a generalised linear mixed model (GLMM) which includes a random intercept for each policy, allowing us to model the correlation between repeated observations of the same policies. We compare this GLMM with a corresponding GLM which lacks the per-policy random intercepts. The claim frequency models are evaluated in the setting of third party liability (TPL) motor insurance, using a representative data set from the Swedish insurance company Trygg Hansa. The models are estimated under the Bayesian paradigm, using Markov Chain Monte Carlo methods. The main aim of the thesis is to determine whether the predictive performance of the GLMM model is better than that of the GLM model. Due to the use of Bayesian inference, the predictions obtained are not point predictions, but full posterior predictive distributions. This allows the use of proper scoring rules to evaluate and compare the predictive performance of the models. Using a panel of comparison metrics, we find that the GLMM model with per-policy random intercepts outperforms the reference GLM model, making it an attractive option for use in non-life insurance pricing. The thesis also contains a discussion on the computational difficulties encountered, and a brief overview of possible future extensions of the GLMM model.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: daniel@danieleliasson.com. Supervisor: Michael Höhle.

Acknowledgements

I would like to thank my advisor Michael Höhle for all his advice and encouragement, his availability and disconcertingly rapid responses to emails sent at all times and all days of the week. I am also indebted to my senior colleague and advisor at Trygg Hansa, Peter Byström, who taught me all I know about being a pricing actuary. Finally, I wish to thank my employer Trygg Hansa for lending me both data and time while I wrote my thesis.

Contents

Chapter 1

Introduction

A non-life insurance contract involves the insured party paying a regular *premium* to the insurer, in return for the insurer covering the costs involved in any *insurance claims*. The claims are random events of generally unpredictable nature, such as traffic accidents in motor insurance, or storm damage in home insurance. The aim of this thesis is to quantify the impact on third party liability motor insurance pricing when using the claim history of individual policyholders to adjust their premiums. To do so, we will compare a generalised linear model (GLM) tariff of the kind which is commonly used by insurance companies to a generalised linear mixed model (GLMM) credibility model. The models will be adjusted to and tested against data from Trygg Hansa's car insurance portfolio.

We will model *claim frequency*, i.e. the number of claims that arise from an insurance policy during a year of insurance. Our aim is to predict with good accuracy the number of claims a policy will generate during its next year of insurance, using the historical data on this individual policy as well as the other policies in the portfolio. The disposition of the thesis is as follows:

- In the rest of the introduction, we will outline what third party liability insurance is, and discuss the basics of insurance tariffs and the adverse selection problem that motivates a focus on improving risk premium modelling. We will also introduce and describe the set of insurance data from Trygg Hansa that will be used in the comparison of our models.

- Chapter 2 contains a brief overview of Bayesian modelling, with a discussion of a few metrics for model assessment and comparison.

- Chapter 3 goes into further detail on insurance pricing models. We introduce the terminology and notation that will be used in this thesis, and discuss using generalised linear models (GLMs) to model insurance claim frequencies.

- Chapter 4 introduces credibility theory, the branch of insurance mathematics that deals with the question of how to combine individual and collective data to model insurance claims and premiums. We describe the generalised linear mixed models (GLMMs) that we will compare to the GLM models described in Chapter 3.

- Since the focus of this thesis is on the comparison of the GLM and GLMM models, Chapter 5 describes a few metrics used for comparing insurance tariffs: Lorenz curves and Gini scores, the so-called quotient test, and a brief look at proper scoring rules.

- In Chapter 6, we define the two models that we will compare and describe the procedure for fitting them. We apply the goodness-of-fit measures and model comparison metrics described in Chapter 2, as well as the more insurance-specific tariff comparisons from Chapter 5. The data analysis is carried out on a set of third party liability motor insurance data from the insurance company Trygg Hansa.

- Finally, Chapter 7 contains a discussion of the results from Chapter 6, and our concluding remarks.

The theory chapters 2–5 serve to establish a context to the work, and provide a reference for the notation and concepts used. They are not intended, nor do they suffice, as thorough reviews of the theory of Bayesian models, non-life insurance pricing or credibility theory. Each chapter contains references to more thorough treatments of the topics. A reader familiar with the respective areas may safely skim these chapters, or visit them when referred to in Chapter 6.

1.1 The problem—using policy-level claims experience to calculate premiums

In motor insurance, *third party liability* insurance, abbreviated TPL, covers the costs a driver becomes liable for in case of causing an accident. This includes both damage to property, such as other vehicles collided with, and personal injury damage and related invalidity, such as income replacement for people who become disabled in a traffic accident. TPL insurance is mandatory in many countries, including Sweden. The premium charged for all policies in an insurance portfolio needs to be sufficiently large to cover the cost of all claims, the need for reserves, and the running costs of the insurance company. In this thesis, however, we will restrict ourselves to only the premium needed to cover the expected costs of claims. This is called the *risk premium*.

Before we move on to discuss insurance pricing further in Chapter 3, we will introduce some key terminology and concepts in this chapter, and look at the analysis data set we will use to evaluate our models.

Terminology of non-life insurance

To simplify for the reader, we will introduce a few key insurance terms that are used throughout this thesis. These definitions can be found in many entry-level texts about insurance pricing, e.g. Ohlsson and Johansson (2010). A non-life insurance contracts involves the insured party paying a regular *premium* to the insurer, in return for the insurer covering the costs involved in any *insurance claims*. The claims are random events of generally unpredictable nature, such as traffic accidents in motor insurance, or storm damage in home insurance.

The *duration* or *exposure* of a policy is the amount of time it is in force, and for a typical insurance policy this is one year. The measurement unit is called *risk years* or *policy years*. The exposure of a portfolio of insurances is the sum of the duration of each included insurance, and is typically taken over a calendar year. The number of claims that a policy incurs while it is in force is referred to as the *claim frequency*. When dealing with a portfolio, the claim frequency is defined as the number of claims incurred, divided by the exposure of the portfolio. Thus, the claim frequency is measured in claims per risk year. Dividing the costs for all claims on a portfolio by the number of claims in the portfolio gives an average cost per claim, a number referred to as the *claim severity*.

If instead the costs for all claims on the portfolio are divided by the portfolio exposure, the result is an average cost per risk year, a number referred to as the *pure premium* or *burning cost*. The pure premium is equal to the product of claim frequency and claim severity. The amount of premium income for an insurance portfolio during a period is called the *earned premium*. Premiums are commonly paid per year up front, and the premium is considered to be earned *pro rata temporis*, so that 1/12 of the premium is earned after 1 month, etc. Therefore, the earned premium is the annual premium times the duration.The *loss ratio* is the total claim costs divided by the earned premium.

Adverse selection

The need for a good risk premium model stems from competition between insurance companies. Imagine an insurance market with two companies, A and B, and 100 000 drivers. Each year, the 100 000 drivers generate claims costing a total of SEK 80 million. Of the drivers, 20 000 are poor drivers, while 80 000 are good drivers. The poor drivers generate SEK 60 million in claim costs, while the good drivers generate only SEK 20 million. A simple

risk premium for the market would be to divide the total claim costs per year with the number of drivers, giving a risk premium of SEK 800 per policy. If both A and B charge this premium, they might each get half of the market, and so have a portfolio of 10 000 poor drivers and 40 000 good drivers each. A and B each charge SEK 40 million in risk premium from their respective portfolios.

Now assume that company A changes their pricing, offering a premium of SEK 3000 to the poor drivers, and SEK 600 to the good drivers. A good driver will now get a lower premium at company A. Assuming that price is the only factor in the choice of insurance company, the good drivers will move to company A. Poor drivers get lower premiums at company B, so the poor drivers will flock there. The situation now is that A has 80 000 good drivers in their portfolio, generating a premium of SEK 48 million and claim costs of SEK 20 million. B instead has a portfolio of 20 000 poor drivers, generating a premium of SEK 16 million, but claim costs of SEK 60 million. B will be forced to increase their premiums to SEK 3000 per policy simply to break even, while A is generating a large profit.

Company B has suffered from *adverse selection*. This example shows the importance of charging each customer a correct and fair premium, that is in line with the risk the customer presents. In other words, a good risk premium model is essential for a good insurance tariff.

Risk premium modelling

The aim of a risk premium model is to predict the cost of an insurance policy. To create a good risk premium model for TPL insurance, companies gather a lot of data about the cars and drivers they insure, and use this together with the history of claims in their portfolio to divide policies into groups based on their risk, and predict the cost of a single policy in such a group. To further improve the price, it is possible to use each individual policy's claim history to give it an individual price. The question of how to best balance the estimated cost of the collective and the estimated cost of the individual is the domain of what actuaries call *credibility theory*. This thesis will consider one type of credibility model, a Poisson GLMM with per-policy random intercepts, and compare it to a GLM model that only uses information on a collective level when setting a risk premium.

Claim costs arise in a portfolio of insurance policies when different events take place: vehicles collide, burn, are stolen, suffer engine failure, etc. The same total cost could arise from a large number of small claims, such as many parking lot bumps and scratches, or through a small number of large claims, such as high-speed collisions, resulting in written-off cars and personal injury costs. Recall from Section 1.1 that

$$\text{pure premium} = \text{claim frequency} \cdot \text{claim severity}. \tag{1.1}$$

Even though the pure premium is not technically a premium, the name is motivated by the pure premium being the risk premium we would charge, if we had knowledge of the future. Alas, we are instead forced to use a risk premium which is merely a prediction of our pure premium, e.g. the expected value of the pure premium.

When creating a tariff, there are two broad approaches to the structure of the model. Either the pure premium, the claim costs per risk year, is modelled directly, or the model is split into two parts: a model for the claim frequency, and a separate model for claim severity. A discussion on the reasons for splitting the premium when modelling can be found in Brockman and Wright (1992) and Murphy, Brockman, and Lee (2000). A prominent reason is that the distribution of claim frequency is more stable than that of claim severity, and having separate models allows a better ability to see if trends in claim costs are driven by changes in severity or in the underlying rate of events.

In this thesis, we will concern ourselves only with models for claim frequency. This choice is made in order to simplify the models, and keep focus on the comparison between models, rather than the details of frequency and severity modelling. We consider a portfolio of N insurance policies, indexed $i = 1, \ldots, N$. A given policy i may have been renewed multiple times, leading us to have J_i repeated observations of this policy, indexed $j = 1, \ldots, J_i$. The time period for each observation is measured in risk years, w_{ij}. For each observation indexed by a pair (i, j), we have the number of claims incurred, Z_{ij}, and the claim frequency $Y_{ij} = Z_{ij}/w_{ij}$ (measured in claims per year). As previously discussed, we are interested in predicting the number of claims that policy i will incur during the next year, i.e. Z_{i,J_i+1}, given the claims experience of the policies in our data set, i.e. given all Z_{ij} and w_{ij} for $i = 1, \ldots, N$ and $j = 1, \ldots, J_i$.

1.2 The data—TPL motor insurance

To compare the GLM and GLMM models for claim frequency, we will apply them on a real data set from the Swedish insurance company Trygg Hansa. The analysis data set consists of *third party liability* (TPL) insurance data for personal cars. The data set contains information about 12 000 policies with a total of 44 186 risk years of exposure and 2 683 claims. This is a subset of Trygg Hansa's portfolio, obtained by taking a simple random sample of all policies that were in force at some point between 1 January 2010 and 31 May 2014. No weighting was performed by exposure, so some of the policies have little exposure in the data set, while some have been in force for the full 4.5 years under consideration. The data is censored on both edges, so that some policies will have been in force before 1 January 2010 or after 31 May 2014, or both. Due to the confidentiality of the data, this thesis does

not contain a full descriptive analysis of the data, but we will describe as much as is possible.

Due to the computationally heavy nature of the MCMC methods we will use to estimate the formulated Bayesian models, the data was split by random division of the policies into 6 subsets of 2 000 policies each. Each subset was split into a training and an evaluation data set chronologically, by taking the latest 1/3 of the observations for each policy aside for evaluation, e.g. for a policy with 3 observations of a risk year each, the training data set would contain the first two, while the evaluation data receives the third. Claims are relatively scarce events, with a claim frequency in the analysis data set of 0.061 claims per risk year. The distribution of claims over the policies and observations is shown in Table 1.1. We see that no single observation has more than 3 claims, and no policy has more than 5.

Claims	0	1	2	3	4	5
Policies	9 813	1 792	314	65	12	4
Observations	47 366	1 585	73	3	0	0

Table 1.1: Distribution of claims over policies and observations in the Trygg Hansa analysis data set.

The first seven rows of the data are shown in Table 1.2, and an explanation of the five tariff rating factors included in the data set are shown in Table 1.3. An observation in the data set has the identifying variables of policy number and j, which indexes the repeated observations for a single policy. Each period has an associated exposure, measured in risk years, which is the period of time spanned by the observation. An insurance contract normally covers a one-year period, but an observation in the data set does not necessarily cover a full contract period. There are several events

obs. no.	policy no.	j	expo-sure	no. claims	for-dar	fve-hic	kkarb	ko-rstr	ztrkof
1	526	0	1.000	1	24	25	61	10	83
2	526	1	1.000	1	26	26	61	10	83
3	526	2	0.419	0	27	27	61	10	83
4	1416	0	0.833	0	9	10	53	5	27
5	1416	1	1.000	0	10	11	54	5	27
6	1416	2	1.000	0	11	12	55	5	27
7	1416	3	0.167	0	12	13	56	5	27

Table 1.2: The first seven rows of the analysis training data set, showing repeated observations of two policies.

that causes an observation to be less than a full year long. The policy can end earlier than contracted due to the vehicle being sold, scrapped or temporarily de-registered. An event may also occur which causes the rating factors to change and a new premium to be calculated, such as the policy holder moving to an area with different geographical risk factors.

Factor	*Description*	*Type*
fordar	Vehicle age in years, 0–90. The highest class contains all vehicles 90 years and older.	Numerical, 0–90
fvehic	A vehicle related parameter. 52 classes from 1–52, where class 52 contains all higher values.	Ordinal, 52 classes
kkarb	Number of years the customer has possessed a driving licence (class B). The highest class contains those that have had their licence for 61 years or more.	Numerical, 0–61
korstr	Distance travelled per year, in intervals of 5 000 km. E.g. class 1 represents a distance of 0–5 000 km, class 2 a distance of 5 001–10 000 km. Truncated at class 6, representing 25 001 km and upwards.	Ordinal, intervals of 5 000 km per year
ztrkof	Geographical risk zone for collision events. The classes go from 1, with the lowest risk, over 50 with average risk up to 100 with highest risk.	Ordinal, 100 classes

Table 1.3: Rating factors included in the analysis data.

The first policy in Table 1.2, policy number 526, is renewed twice. The first two observations are full years, while the last is only 0.419 years long. The first two years have one claim each, the last has none. The second policy, number 1416, begins with an observation of duration 0.833 risk years, due to left-censoring, and is then renewed twice. The last observation is also shorter than a year. No claims are observed for this policy. Note that several of the rating factors change at the time of renewal. This is due to a year having passed by the time the policy is renewed, so that the vehicle is one year older and the customer has had their driving licence for another year.

To get an idea of how the claim frequency depends on the rating factors, we include Figures 1.1–1.5 on pages 8–10. Each figure consists of two panels. The lower panel is a histogram which shows how the exposure in the data set is distributed over the different values of the rating factors. The upper panel shows how the claim frequency varies over the different values of the rating factor, including a loess smoothing line. To make it easier to get a feel for how much the claim frequency varies for a rating factor, and to facilitate comparison between rating factors, the claim frequency line is normalised so that the value of the rating factor with the most exposure is set to 1, and

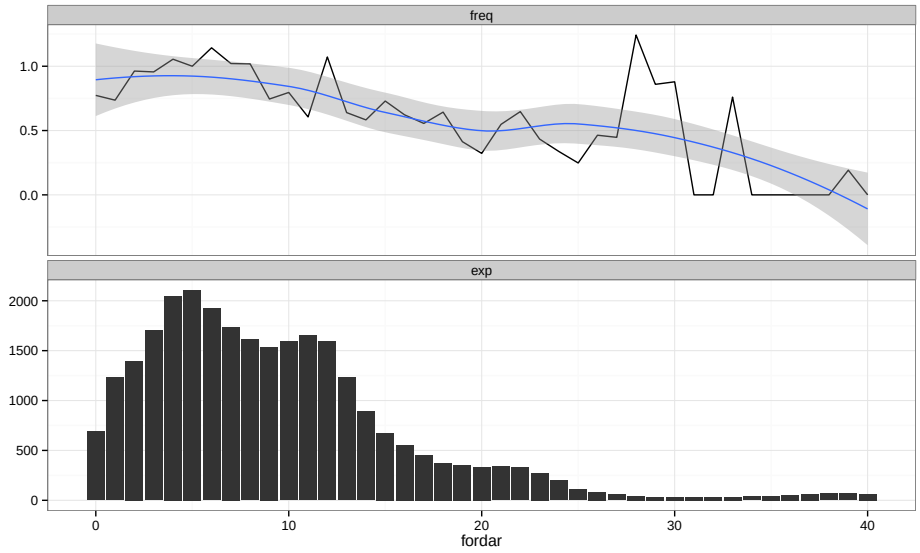

Figure 1.1: Lower panel: histogram of the distribution of the rating factor fordar in the data. Upper panel: relative claim frequency as a function of fordar, normalised against the value with the most exposure. A loess smoothing line has also been added.

the other values are scaled accordingly. As an example, in Figure 1.2 (p. 9), the value fvehic = 1 has the most exposure, so the relative claim frequency there is set to 1. When fvehic = 5, the relative claim frequency is 0.8, i.e. 20 % lower than when fvehic = 1.

1.3 Notes on software used

The statistical analysis in this thesis was performed in the statistical software R (R Core Team, 2014), using the RStudio integrated development environment (RStudio Team, 2012). The models were estimated using JAGS v. 3.4.0 (Plummer, 2003), and the rjags package (Plummer, 2014). The graphs were created using the excellent ggplot2 package (Wickham, 2009), and much of the data manipulation was performed with the plyr package (Wickham, 2011). I am deeply grateful for the enormous amount of work that has been put into making these excellent software programs and packages available to use for free.

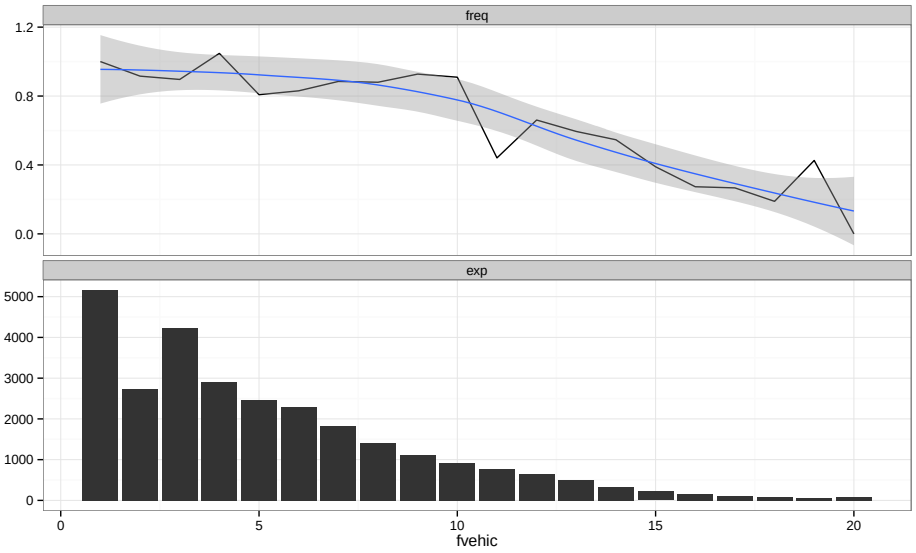

Figure 1.2: Exposure and relative claim frequency plot of the rating factor fvehic.

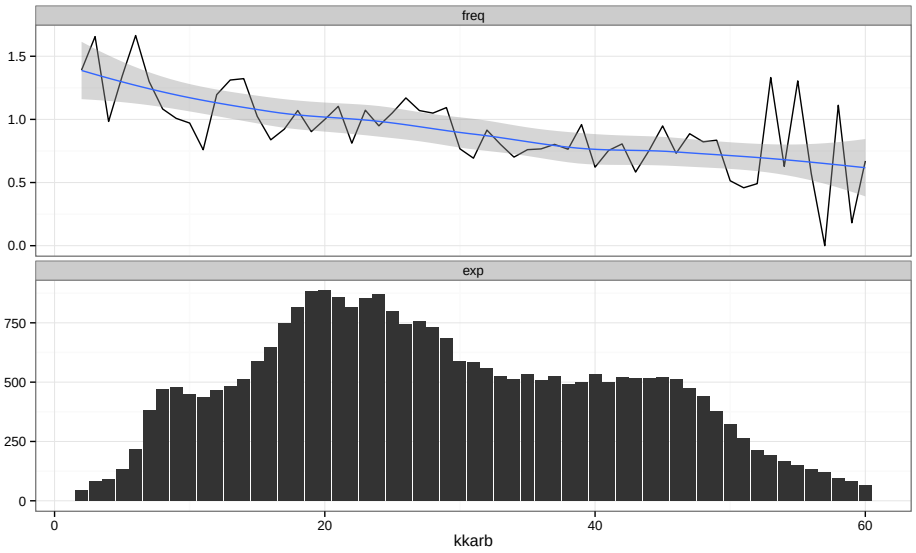

Figure 1.3: Exposure and relative claim frequency plot of the rating factor kkarb.

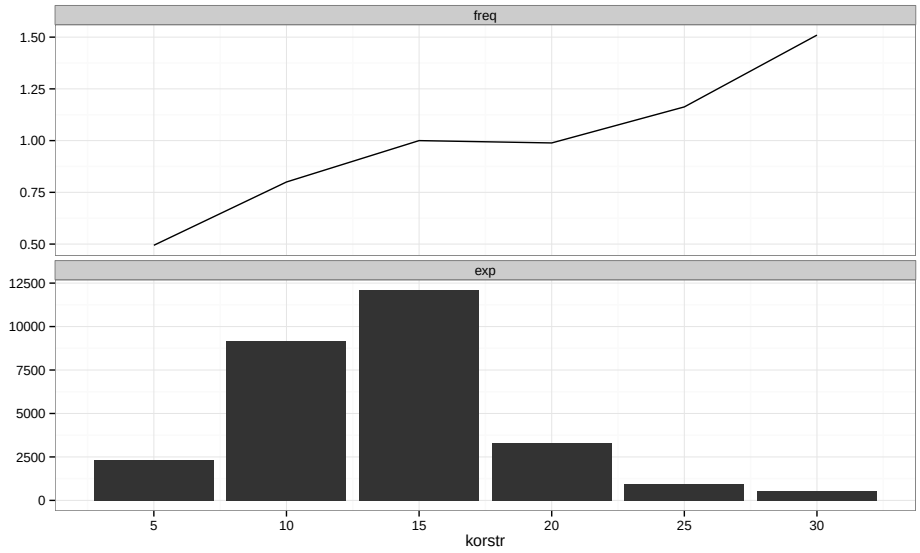

Figure 1.4: Exposure and relative claim frequency plot of the rating factor korstr.

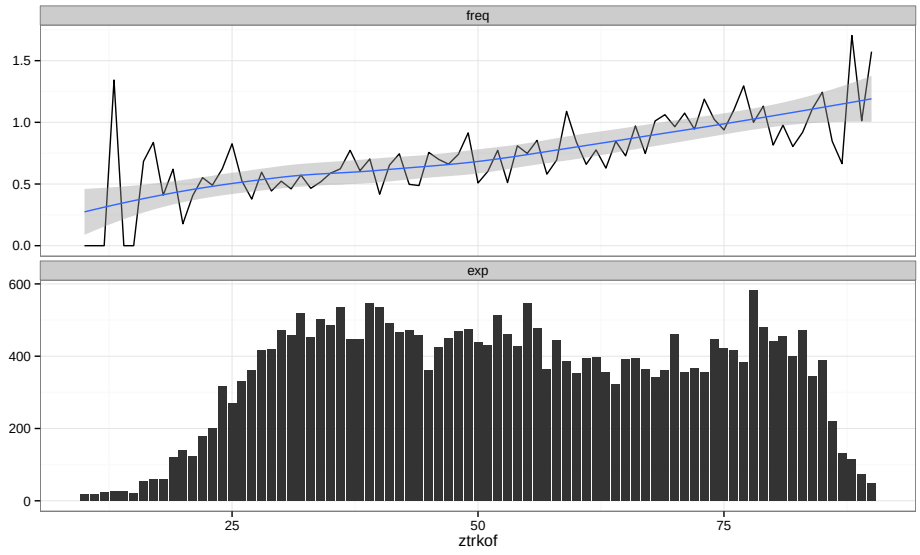

Figure 1.5: Exposure and relative claim frequency plot of the rating factor ztrkof.

Chapter 2

Bayesian models

The GLM and GLMM models for claim frequency prediction that will be implemented and compared in this thesis are considered within the Bayesian statistical framework. This chapter begins with a cursory introduction to Bayesian models and a short discussion of model fitting. This is followed by an overview of metrics used to assess the goodness-of-fit of these models, and metrics for comparing models. The chapter is intended mostly to provide a reference for the notation and concepts that will be used in Chapter 6. A more thorough treatment of Bayesian statistics is available in Carlin and Louis (2009), or in any number of textbooks on the topic.

2.1 Bayesian statistical modelling

Assume that we have some observed vector of data $\mathbf{y} = (y_1, \ldots, y_n)'$ that we wish to model. We begin with a sampling model depending on some parameters $\boldsymbol{\theta}$, typically given in the form of a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$. The probability distribution function is perceived as a function of the data $\mathbf{y}$ with a parameter vector $\boldsymbol{\theta}$. It can also be considered a function of the parameters $\boldsymbol{\theta}$. Under this interpretation, it is often written $L(\boldsymbol{\theta}; \mathbf{y})$ and called the *likelihood function*.

In the frequentist framework, we consider $\boldsymbol{\theta}$ to be a fixed but unknown quantity, that we typically estimate by the value $\widehat{\boldsymbol{\theta}}$ which maximises $L(\boldsymbol{\theta}; \mathbf{y})$. This is known as the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$.

In the Bayesian framework, we instead consider $\boldsymbol{\theta}$ to be a random variable as well. To do so, we specify a *prior distribution* for $\boldsymbol{\theta}$ which incorporates any information about it that we have *independently* of the data $\mathbf{y}$, e.g. prior to conducting the experiment where $\mathbf{y}$ is gathered. This prior distribution will itself depend on some vector of hyperparameters $\boldsymbol{\eta}$, which we can either assume to be known, or to in turn be random with its own prior distribution. The latter case leads to the class of *hierarchical models*.

Assuming for now that $\boldsymbol{\eta}$ are known, we write for the prior distribu-

tion $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\boldsymbol{\eta})$. Inferences about $\boldsymbol{\theta}$ can now be made by calculating the *posterior distribution*, which is obtained by weighing together the prior distribution and the information about $\boldsymbol{\theta}$ that is carried in the data $\mathbf{y}$ via the likelihood function $f(\mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y})$.

This is performed using *Bayes' theorem*, and the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{\int p(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{2.1}$$

The posterior distribution of $\boldsymbol{\theta}$ is the product of the prior and the likelihood of the data, renormalised to be a proper distribution function.

The prior distribution contains our prior information about the distribution of the parameters $\boldsymbol{\theta}$. This can be subjective information, such as expert opinions or results from data assumed to be similar to that under study. There is also the possibility of using so-called *non-informative priors*. These are prior distributions that do not in themselves favour any specific value of $\boldsymbol{\theta}$, which leads to the posterior distribution depending only on the data.

Non-informative priors are easy to find in some situations, such as for discrete and finite probability spaces, where a uniform distribution on all values of $\boldsymbol{\theta}$ is considered to be non-informative. In continuous and unbounded cases, however, a uniform distribution would be of the form $p(\theta) = c, c > 0$ for all $\theta \in \mathbb{R}$, but such a function p is not a probability density function, as it has an unbounded integral. Bayesian inference is however still possible as long as the integral of the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is finite. This is not always the case, and so care must be taken when using improper priors.

In this thesis we are concerned with the problem of predicting a claim frequency based on historical data, so the question of how to make a prediction based on a Bayesian model is of interest. Once a posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ has been obtained, it is possible to predict a future observation y_{n+1} by the *posterior predictive* distribution:

$$\begin{aligned} p(y_{n+1}|\mathbf{y}) &= \int p(y_{n+1}, \boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} && (2.2) \\ &= \int p(y_{n+1}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} && (2.3) \\ &= \int p(y_{n+1}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta}. && (2.4) \end{aligned}$$

The last equality holds because of the assumption that y_{n+1} and $\mathbf{y}$ are independent conditional on the parameters $\boldsymbol{\theta}$.

A good feature of a Bayesian model is that we get not just point predictions of y_{n+1}, but a predictive distribution from which we can calculate any quantity of interest regarding y_{n+1}, e.g. the mean, median or quantiles.

2.2 Estimating Bayesian models

The multiple integral in equation (2.1) above is simple enough in theory, but depending on the dimensionality of the parameter vector $\boldsymbol{\theta}$ and the form of the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ and prior distribution $\pi(\boldsymbol{\theta})$, it is typically too complicated to calculate without resorting to numerical methods. This is the case for the GLM and GLMM models we use in this thesis.

There are several different numerical methods for calculating the posterior distribution in complex Bayesian models. The most common approach are Markov chain Monte Carlo methods (MCMC), provided for example by the software package JAGS (Plummer, 2003). These work by sampling from a Markov chain which has been designed to have a stationary distribution that is precisely the joint posterior distribution of the parameters $\boldsymbol{\theta}$. A common implementation is the Gibbs sampler together with the Metropolis-Hastings algorithm, see Carlin and Louis (2009).

MCMC methods have the advantage of being applicable for a very wide variety of Bayesian models, but there are drawbacks. Most importantly, they are computationally intensive, as there is a need to first simulate and discard enough samples for the Markov chain to converge on its stationary distribution ("burn in"), and then to draw a large number of samples in order to well approximate the posterior distributions of the parameters.

Another approach to Bayesian model estimation is via integrated nested Laplace approximation (INLA), introduced in Rue, Martino, and Chopin (2009). The INLA method avoids simulation by instead using deterministic numerical methods. The posterior marginal distributions of the parameters $\boldsymbol{\theta}$ are evaluated in a number of points by means of numerical integration and Laplace approximations. We will not here delve into the rather extensive details, but instead refer the reader to Rue et al. (2009) for a full treatment. Owing to the use of deterministic numerical integration instead of simulation, the INLA method is less general than MCMC, but has major computational advantages, being frequently orders of magnitude faster than MCMC. INLA is applicable to a class of additive regression models called *latent Gaussian models* (LGMs) by Rue et al. (2009). Both the GLM and the GLMM are latent Gaussian models, and can therefore be estimated using INLA.

Once the posterior distributions of a model have been obtained by some method, it is important to verify that the model is a good fit of the data, which is the topic of the next section.

2.3 Model assessment

A generally common approach to model validation is to split the data vector $\mathbf{y} = (y_1, \ldots, y_n)'$ into a *training* data set $\mathbf{z} = (z_1, \ldots, z_k)'$ and a *validation*

data set $\mathbf{u} = (u_1, \ldots, u_m)'$, $\mathbf{y} = (\mathbf{z}, \mathbf{u})$, and then to compare the model's predictions of the validation data with the observed data.

Carlin and Louis (2009) describe in Chapter 2 a number of metrics for assessment of model fit and comparison of models. We will here give a brief overview of those that we will apply to our models in Chapter 6, in order to introduce the concepts and notation. We begin with considering *Bayesian residuals*, defined as

$$r_i = u_i - \mathbb{E}\left[U_i|\mathbf{z}\right], \quad i = 1, \ldots, m. \tag{2.5}$$

These residuals may be plotted against the fitted values to identify if assumptions such as normally distributed errors or homogeneous variance fail to hold. We can remove the effect of the scale of the data on our residuals by standardising them. The standardised residuals are given by

$$d_i = \frac{u_i - \mathbb{E}\left[U_i|\mathbf{z}\right]}{\sqrt{\mathrm{Var}\left(U_i|\mathbf{z}\right)}}, \quad i = 1, \ldots, m. \tag{2.6}$$

By summing the absolute values and squares of the standardised residuals, we get metrics that can be used to assess the goodness-of-fit of the model. Carlin and Louis (2009) suggest that an observation can be considered an outlier if its standardised residual has an absolute value greater than 1.5.

Another useful goodness-of-fit metric is the *Bayesian p-value*, which hinges on the discrepancy measure

$$D(\mathbf{u}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \frac{\left[u_i - \mathbb{E}\left[U_i|\boldsymbol{\theta}\right]\right]^2}{\mathrm{Var}\left(U_i|\boldsymbol{\theta}\right)}. \tag{2.7}$$

To calculate the Bayesian p-value, p_B, we compare the distribution of $D(\mathbf{u}, \boldsymbol{\theta})$ for the observed validation data $\mathbf{u}$, with the distribution of $D(\mathbf{u}^\star, \boldsymbol{\theta})$ for a future observation vector $\mathbf{u}^\star$. Specifically, we consider the probability that new data would have a higher discrepancy measure D than the observed:

$$p_B = \mathbb{P}\left[D(\mathbf{u}^\star, \boldsymbol{\theta}) > D(\mathbf{u}, \boldsymbol{\theta})|\mathbf{z}\right] = \int \mathbb{P}\left[D(\mathbf{u}^\star, \boldsymbol{\theta}) > D(\mathbf{u}, \boldsymbol{\theta})\right] p(\boldsymbol{\theta}|\mathbf{z}) \mathrm{d}\boldsymbol{\theta}. \tag{2.8}$$

A small p_B-value then indicates a lack of fit, since there is a low probability that new data would be less well fit than the training data. In other words, the model fits the training data poorly enough that new data is likely to be better fit by the model.

2.4 Model comparison

Given two competing models, M_1 and M_2 for some data $\mathbf{y}$, we would like a way to determine which model better fits the data. A favoured tool for this

is the *Bayes factor* and its more robust version, the *partial Bayes factor*. Another popular metric is the deviance information criterion (DIC). As in Section 2.3, we give here only a brief overview to introduce the metrics and notation. For a fuller treatment, refer to Carlin and Louis (2009).

Consider the two competing models M_1 and M_2, and let $\pi_i(\boldsymbol{\theta}_i), i = 1, 2$ be the prior distributions of their respective parameter vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Integrating out the parameters give the marginal distributions of $\mathbf{y}$ under the two models:

$$p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_\mathrm{i}, \quad i = 1, 2. \tag{2.9}$$

We can now apply Bayes' theorem to calculate the posterior probabilities of the models, $P(\mathrm{M}_1|\mathbf{y})$ and $P(\mathrm{M}_2|\mathbf{y}) = 1 - P(\mathrm{M}_1|\mathbf{y})$. The Bayes factor, BF, is defined as the ratio of the posterior odds of M_1 to the prior odds of M_1':

$$\begin{aligned}
\mathrm{BF} &= \frac{P(\mathrm{M}_1|\mathbf{y})/P(\mathrm{M}_2|\mathbf{y})}{P(\mathrm{M}_1)/P(\mathrm{M}_2)} && (2.10)\\
&= \frac{\left[\frac{p(\mathbf{y}|\mathrm{M}_1)P(\mathrm{M}_1)}{p(\mathbf{y}}\right] / \left[\frac{p(\mathbf{y}|\mathrm{M}_2)P(\mathrm{M}_2)}{p(\mathbf{y}}\right]}{P(\mathrm{M}_1)/P(\mathrm{M}_2)} && (2.11)\\
&= \frac{p(\mathbf{y}|\mathrm{M}_1)}{p(\mathbf{y}|\mathrm{M}_2)}, && (2.12)
\end{aligned}$$

which is the ratio of the observed marginal densities of the two models, see (2.9) above. If the Bayes factor is greater than 1, then the evidence is in favour of model M_1, while if it less than 1, then it is in favour of M_2. Kass and Raftery (1995) give a table for the strength of the Bayes factor, suggesting that $BF > 3.2$ is substantial evidence in favour of M_1, $BF > 10$ is strong evidence in favour, and a $BF > 100$ is decisive.

The Bayes factor does have its drawbacks. It is sensitive to the prior distributions $\pi_i(\boldsymbol{\theta}_i)$, and worse, if an improper prior is used, as is common for non-informative priors, then the Bayes factor is not well-defined. For a more in-depth treatment of Bayes factors, see Lavine and Schervish (1999).

One way to deal with the problem of improper priors in the Bayes factor is by using the partial Bayes factor. This is done by splitting the data $\mathbf{y}$ into two parts, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and using the first portion to obtain posterior densities $p(\boldsymbol{\theta}_i|\mathbf{y}_1), i = 1, 2$ and then using these as priors in the equation for the Bayes factor, while using $\mathbf{y}_2$ as the data there. This is the partial Bayes factor

$$\mathrm{BF}(\mathbf{y}_2|\mathbf{y}_1) = \frac{p(\mathbf{y}_2|\mathbf{y}_1, \mathrm{M}_1)}{p(\mathbf{y}_2|\mathbf{y}_1, \mathrm{M}_2)}. \tag{2.13}$$

The partial Bayes factor is easily calculated in cases where the models have been trained against part of the available data, while some data has been held off for validation purposes, such as in in Chapter 6.

The deviance information criterion (DIC), introduced in Spiegelhalter, Best, Carlin, and van der Linde (2002) and treated in Carlin and Louis (2009), is a generalisation of the familiar Akaike information criterion (AIC). The DIC is based on the *deviance* statistic, defined as

$$D(\boldsymbol{\theta}) = -2\log f(\mathbf{y}|\boldsymbol{\theta}) + 2\log h(\mathbf{y}), \tag{2.14}$$

where $h(\mathbf{y})$ is some standardising function of the data alone. Since $h(\mathbf{y})$ does not depend on the model, it has no impact on model selection, where we consider differences in DIC, and so can be assumed to be zero for our purposes. The DIC has a part rewarding good fit of the model to the data, and another part which penalises model complexity. The fit is measured by the posterior expectation of the deviance,

$$\overline{D} = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}\left[D\right], \tag{2.15}$$

and the complexity is measured by the effective number of parameters, p_D. This metric is typically less than the total number of parameters in the model, and is defined as the expected deviance minus the deviance evaluated at the posterior expectation of $\boldsymbol{\theta}$:

$$p_D = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}\left[D\right] - D(\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}\left[\boldsymbol{\theta}\right]) = \overline{D} - D(\overline{\boldsymbol{\theta}}). \tag{2.16}$$

The deviance information criterion itself is then defined as

$$\text{DIC} = \overline{D} + p_D = 2\overline{D} - D(\overline{\boldsymbol{\theta}}). \tag{2.17}$$

Smaller values of the DIC indicate a better fitting model, and to compare two models M_1 and M_2, we consider the difference in DIC between them. Carlin and Louis (2009) suggests that the difference should be at least 3 to 5 in order to be of interest. The DIC is easily obtained when a model is estimated using MCMC simulation, by sampling the deviance $D(\boldsymbol{\theta})$ together with $\boldsymbol{\theta}$ itself.

This concludes our discussion of Bayesian models. We will now move on to a brief overview of non-life insurance pricing and tariff models.

Chapter 3

The basics of non-life insurance pricing

3.1 Risk premiums and tariffs

In non-life insurance, the price of an insurance policy is dependent on a number of *rating factors*, which determine the risk level of the policy, and thus the appropriate price to charge for it. The pricing begins with a statistical model to calculate an expected cost of the policy in terms of the insurance claims it is expected to bring. This is the *risk premium*, and is the premium that we will be concerned with in this thesis.

Beyond the direct claim costs, the premiums for a portfolio of insurance policies also need to cover the running costs of the insurance company (claims handling costs, customer acquisition costs, salaries and office rents, etc.) as well as a cost of capital or profit margin. On top of this, the insurance company may adjust the price for individual customers for various reasons, such as rebates given to acquire new business, or extra high margins for price-insensitive customers. At the end of this process, the *market premium* is what is eventually presented to a prospective customer.

However, as mentioned, we shall limit our discussion to the risk premium, and will discuss the form of tariffs that are standard in most non-life insurance companies. For a more thorough discussion of pricing, refer to Ohlsson and Johansson (2010).

The model that determines which risk premium is assigned to an insurance policy is called a tariff. The tariff classifies policies based on a number of *rating factors*, which can be either numerical, such as the engine power of a vehicle; ordinal, such as a geographic risk classification; or categorical, such as the fuel type used in a vehicle.

For practical reasons, not infrequently related to the capabilities of the software systems used, it is very common to divide numerical rating factors into a number of classes. As an example, consider a simple tariff for motor

hull insurance. As rating factors, we use the car's fuel type (petrol or diesel), the engine power (0–100, 101–200, 201–400, 400+ hp), the vehicle's age (0–5, 6–10, 11–20, 21–30, 30+ years) and the age of the driver (18–25, 26–35, 36–45, 46–55, 56–65, 65+ years).

The outcome of this classification is that the policy is placed in a *tariff cell*, which is a unique combination of all the factors, e.g. a diesel car of 101–200 hp, 6–10 years old and with a driver aged 36–45 years. The tariff cell k then has an expected claims cost, or risk premium, μ_k.

3.1.1 Multiplicative tariffs

A desirable feature of an insurance tariff is that it be *multiplicative*. That is to say that if diesel cars have a higher expected claims cost than petrol cars, then this should be a fixed percentage higher, rather than a fixed amount higher.

As an example, if an *additive* model was used, it could be that a diesel car premium is SEK 200 higher than that for a petrol car with all other rating factors the same. But if one petrol car has a premium of SEK 400, then the diesel version of this car is 50 % more expensive, while for a petrol car with a premium of SEK 4000, the corresponding diesel is only 5 % more expensive. In practice, it has been found that a multiplicative tariff better fits the observations. Further discussion of the reasons for multiplicative tariffs can be found in Brockman and Wright (1992).

Let us consider a multiplicative tariff with 4 rating factors. We can identify a tariff cell k by the quadruple (l, m, n, p), where l is the fuel type class (numbered 1 for petrol, 2 for diesel), m the engine power class (numbered 1 for 0–100 hp, 2 for 101–200 hp, etc.), and so on. The risk premium μ_k of tariff cell $k = (l, m, n, p)$ under the multiplicative tariff is given by

$$\mu_k = \gamma_0 \gamma_{1l} \gamma_{2m} \gamma_{3n} \gamma_{4p}, \tag{3.1}$$

where γ_0 is the *base premium*, the risk premium of a defined base cell, and γ_{qr} is the relative risk of the q-th rating factor for class r, which is the class of this factor in tariff cell k. The γ-s are called *relativities*, since they give us the risk of a rating factor class relative to the risk of the base cell's class.

For simplicity, assume that the base cell is $k = 1$ and has classes $(1, 1, 1, 1)$. We let $\gamma_{11} = \gamma_{21} = \gamma_{31} = \gamma_{41} = 1$, so that $\mu_1 = \gamma_0$ in this cell.

An example will be helpful in making this all clearer. Table 3.1 shows the relativities of our example tariff. If we also know the risk premium of the base cell, say $\gamma_0 = \text{SEK}\,500$, then we can calculate the risk premium of any policy. For instance, let the policy fall in tariff cell $(1, 3, 3, 2)$. Then, the premium is

$$\mu_i = \gamma_0 \gamma_{11} \gamma_{23} \gamma_{33} \gamma_{42} = 500 \cdot 1.0 \cdot 1.6 \cdot 0.6 \cdot 0.7 = \text{SEK}\,336. \tag{3.2}$$

Rating factor	Class	Class number	Relativity
Fuel type	Petrol	1	1.0
	Diesel	2	1.2
Engine power	0–100 hp	1	1.0
	101–200 hp	2	1.2
	201–400 hp	3	1.6
	400+ hp	4	2.0
Vehicle age	0–5 years	1	1.0
	5–10 years	2	0.8
	11–20 years	3	0.6
	21–30 years	4	0.5
	30+ years	5	0.3
Driver age	18–25 years	1	1.0
	26–35 years	2	0.7
	36–45 years	3	0.5
	46–55 years	4	0.4
	56–65 years	5	0.4
	65+ years	6	0.3

Table 3.1: Relativities of a simple motor hull insurance tariff.

Tariff cell	Fuel type	Engine power	Vehicle age	Driver age	Relativity	Risk premium
1	1	1	1	1	1.00	500
2	1	1	1	2	0.70	350
3	1	1	1	3	0.50	250
4	1	1	1	4	0.40	200
5	1	1	1	5	0.40	200
6	1	1	1	6	0.30	150
7	1	1	2	1	0.80	400
8	1	1	2	2	0.56	280
⋮	⋮	⋮	⋮	⋮	⋮	⋮
239	2	4	5	5	0.29	144
240	2	4	5	6	0.22	108

Table 3.2: Simple motor hull insurance tariff on list form.

A handy way to show a tariff in a table is what Ohlsson and Johansson (2010) call *list form*, which is simply the way the data would be represented in a database. We enumerate all the combinations (l, m, n, p) and assign them tariff cell numbers $k = 1, \ldots, K$, and for each tariff cell calculate the relativity from the base class, as well as the risk premium. An abridged list form of the example tariff is shown in Table 3.2. In this format, it is easily seen that there are a large number of tariff cells, in this case $K = 240$, despite the tariff only including four different rating factors, and each with only a few classes. A full tariff of the kind used in industry may include a couple of dozen rating factors, some with dozens of levels.

3.2 Pricing with generalised linear models

We have seen a brief overview of the structure of a non-life insurance tariff. We now turn our focus to the method by which such tariffs are created. The industry standard way is to use generalised linear models (GLMs) to accomplish this.

Generalised linear models (GLMs) are a class of regression models that extend linear regression to situations where the response follows any distribution from the *exponential dispersion models* family of probability distributions, and where the mean does not need to be linear in the covariates, but can instead be a function of a linear combination of the covariates.

We will provide here a brief overview of GLMs, specifically the cases of interest in non-life insurance pricing, structured on the exposition in Ohlsson and Johansson (2010).

3.2.1 Exponential dispersion models (EDMs)

Exponential dispersion models (EDMs) are models where the distribution of the random variables belong to the *exponential family* of probability distributions, and are the distributions used in generalised linear models.

Consider a vector of key ratios Y_i, observations y_i and corresponding weights w_i, where $i = 1, \ldots, N$. The key ratio could be either the claim frequency or claim severity. In the former case, the weights w_i are exposure, while in the latter, they are the number of claims.

The key ratios Y_i belong to the *exponential dispersion model* if their probability density functions can be written on the form

$$f_{Y_i}(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right\}. \tag{3.3}$$

In this formula, θ_i and ϕ are parameters, and b, c are known functions. θ_i is a location parameter which can depend on i, while the scale parameter $\phi > 0$ is taken to be constant for all i. The cumulant function b is twice continuously differentiable, with invertible first derivative. The function c is a normalising function.

The mean and variance of an EDM are characterised by the parameters θ_i, ϕ, weight w_i and cumulant function b:

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i), \tag{3.4}$$

$$\mathrm{Var}(Y_i) = \sigma_i^2 = b''(\theta_i)\frac{\phi}{w_i}. \tag{3.5}$$

It is convenient to rewrite the variance as a function of the mean μ_i via the *variance function*, v:

$$\theta_i = b'^{-1}(\mu_i) \tag{3.6}$$

$$\Rightarrow \mathrm{Var}(Y_i) = b''(b'^{-1}(\mu_i))\frac{\phi}{w_i} \tag{3.7}$$

$$\Rightarrow \mathrm{Var}(Y_i) = v(\mu_i)\frac{\phi}{w_i}, \tag{3.8}$$

where the variance function is defined as

$$v(\mu_i) = b''(b'^{-1}(\mu_i)). \tag{3.9}$$

Thus, the variance of an EDM is a function of the mean multiplied by a scaling and weighting factor ϕ/w_i.

EDMs encompass many common probability distributions, but not all are useful in insurance pricing. For a distribution to be suitable for insurance pricing, it needs to be *scale invariant*, in other words it must hold that if Y_i has some distribution, then for any constant p, pY_i has the same distribution,

with different parameters. If this were not true, then the choice of measurement unit would affect the distribution. Clearly this is unacceptable, as the choice of currency unit should not affect the distribution of claim severity in any other sense than scale, nor should the distribution of claim frequency be affected if we measure it in percent or per mille.

The subclass of exponential dispersion models that is scale invariant is called the class of *Tweedie* models. These are characterised by their variance function being a power of the mean:

$$v(\mu_i) = \mu_i^p, \quad p \leq 0 \text{ or } p \geq 1. \tag{3.10}$$

Some well known examples of Tweedie distributions are the Normal distribution ($p = 0, v(\mu_i) = 1$), Poisson distribution ($p = 1, v(\mu_i) = \mu_i$) and Gamma distribution ($p = 2, v(\mu_i) = \mu_i^2$).

Since this thesis will investigate the effect of using individual policy data to improve the prediction of claim frequencies, we now go on to look at the specific EDM which is commonly used to model claim frequency in non-life insurance: the Poisson distribution. For a more thorough discussion of EDMs, refer to Jørgensen (1997), and for a discussion of their application in non-life insurance pricing, see Ohlsson and Johansson (2010).

3.2.2 Claim frequency modelling: Poisson distribution

Third party liability motor insurance claims are rare events that occur randomly, and are therefore frequently modelled as Poisson processes, so that the number of claims for a policy or portfolio of policies can be modelled by a Poisson distributed random variable. Other distributions can also be considered, notably the negative binomial distribution, but the Poisson distribution remains a popular choice for insurance tariff modelling, see Ohlsson and Johansson (2010), Jørgensen and De Souza (1994) and Denuit, Maréchal, Pitrebois, and Walhin (2007).

Consider a tariff cell i over a period of time in which the exposure was w_i risk years. Let the number of claims that occurred during this time be Z_i, and let μ_i be the expected number of claims when $w_i = 1$. Then, Z_i follows a Poisson distribution:

$$f_{Z_i}(z_i; \mu_i) = e^{-w_i\mu_i} \frac{(w_i\mu_i)^{z_i}}{z_i!}, \quad z_i = 0, 1, 2, \ldots \tag{3.11}$$

We typically seek to study the claim frequency $Y_i = Z_i/w_i$ rather than the number of claims Z_i. This follows a distribution which Ohlsson and Johansson (2010) call a *relative Poisson distribution*. This can be written

on exponential family form as

$$\begin{aligned} f_{Y_i}(y_i;\mu_i) &= \mathbb{P}\left[Y_i = y_i\right] = \mathbb{P}\left[Z_i = w_i y_i\right] = e^{-w_i\mu_i}\frac{(w_i\mu_i)^{w_i y_i}}{(w_i y_i)!} \\ &= \exp\left\{w_i\left[y_i \log\mu_i - \mu_i\right] + c(y_i, w_i)\right\} \\ &= \exp\left\{w_i\left[y_i\theta_i - e^{\theta_i}\right] + c(y_i, w_i)\right\}, \end{aligned} \tag{3.12}$$

where we have introduced $\theta_i = \log\mu_i$ and $\phi = 1$, $b(\theta_i) = e^{\theta_i}$.

3.2.3 Generalised linear models (GLM)

We are now ready to characterise generalised linear models as used in non-life insurance pricing.

Let Y_i, $i = 1, \ldots, N$ be the dependent variables, with observations y_i, weights w_i. Further, let x_{ik}, $k = 0, \ldots, m$ be the value of independent variable k for observation i.

In our case, Y_i is the claim frequency for tariff cell i, and the independent variables x_{ik} are either numerical variables or *dummy variables* coding for the different rating factors and their classes. For the dummy variables, a given x_{ik} is either 1 or 0, telling us if tariff cell i belongs to the rating factor class coded for by k.

We assume that $Y_i \sim \mathrm{EDM}(\theta_i, \phi)$ for some EDM, and that the mean μ_i depends on the covariates x_{ik} by the formula

$$g(\mu_i) = \eta_i = \sum_{k=1}^{m} x_{ik}\beta_k. \tag{3.13}$$

The values η_i are called the *linear predictors*, and the monotonous, differentiable function g is the *link function*.

In ordinary linear regression, the response distribution is the normal distribution, and the link function is the identity function, so that $\mu_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. In insurance pricing, we typically wish for a multiplicative model, leading to the choice of $g(\mu_i) = \log\mu_i = \eta_i$. This gives

$$\mu_i = \exp\{\eta_i\} = \exp\left\{\mathbf{x}_i'\right\} = \prod_{k=1}^{m} e^{x_{ik}\beta_k}, \tag{3.14}$$

which is on the multiplicative form of Section 3.1.1. The relativities are obtained as $\gamma_k = e^{\beta_k}$. The base class is coded for by letting $x_{i0} = 1$ for all i, so that β_0 is a model intercept which captures the base cell expectation $\mu_0 = e^{\eta_0} = e^{\beta_0}$. In our notation from earlier, we would say $\beta_0 = \log\gamma_0$.

The regression coefficients, β_k, are estimated using maximum likelihood estimation. A description can again be found in Ohlsson and Johansson (2010). In practice, these models are fitted to data using statistical software.

In the industry, this is frequently SAS, or specialised pricing software, such as Towers Watson's Emblem. The use of R is increasing, and this thesis used R and JAGS, see Section 1.3.

Chapter 4

Credibility theory

Insurance tariffs of the kind described in Chapter 3 assume that all policies in the same tariff cell have the same expected claim frequency. In practice, this risk homogeneity assumption is not valid, and significant within-group heterogeneity still remains due to the rating factors failing to capture some relevant information. As an example, certain intersections are more dangerous than others, but it is infeasible to require the customer to supply the insurer with information on which routes they drive on a daily basis.

Each policy belongs to a tariff cell, which holds a collective of policies with similar risk characteristics, but each policy is also different from the others. The history of claims for an individual policy is valuable in determining the risk, but there is likely too little information from a single policy to rely entirely on it. The problem is especially clear when dealing with a new policy, where there is no historical data available at all. On the other hand, the data for all risks in the collective is statistically reliable to determine the overall risk level of the collective, but is not specific enough to an individual policy to correctly assess its risk.

This leads to actuaries needing to tackle the question of how to optimally assess a policy's risk using both the collective and individual data available. This is the topic of the field of *credibility theory*. A good overview with a focus on applications is available in Bühlmann and Gisler (2005).

Credibility theory has a rich history in the insurance industry, with roots going back to the early 20th century, and active development of the ideas from the 1940s on. The seminal Bühlmann-Straub method was published in Bühlmann and Straub (1970), and its familiarity and ease of computation has led to great popularity with actuarial practitioners. The method is still in common use for pricing in the insurance industry today, see Bühlmann and Gisler (2005) and Ohlsson and Johansson (2010).

We will not perform a thorough review of credibility theory, but rather consider an illuminating example before we move on to look at the specific credibility model that we are using in this thesis, namely the generalised

linear mixed model, or GLMM.

4.1 An example—fleet TPL insurance

To illustrate the need for and use of credibility theory, we give a somewhat contrived example inspired by Bühlmann and Gisler (2005).

We consider an artificial portfolio of nine insurance contracts, each one covering third party liability insurance for a company's fleet of vehicles. We have data for the last 5 years, during which the premium for contracts 1, 4, and 7 was SEK 30 000, the premium for contracts 2, 5, and 8 was SEK 300 000, and the premium for contracts 3, 6, and 9 was SEK 3 000 000. Figure 4.1 shows the loss ratios for the nine contracts. The average loss ratio over the 5 years is 60 % for contracts 1–3, 100 % for contracts 4–6 and 140 % for contracts 7–9. Note that each group of three contracts with the same average loss ratio contains one contract with a low premium, one with a medium premium, and one with a high premium.

We now wish to adjust the premiums for next year. A well adjusted risk premium should lead to a loss ratio which averages 100 %. If we take the entirely collective point of view, we could assume that all differences between the contracts are due to random chance, and that since the entire group averages to a loss ratio of 100 %, no adjustment is needed. If we instead take the individual policy view, we find that policies 1–3 deserve a premium decrease, while policies 7–9 should receive a premium increase. Credibility theory studies how to balance between these two extremes.

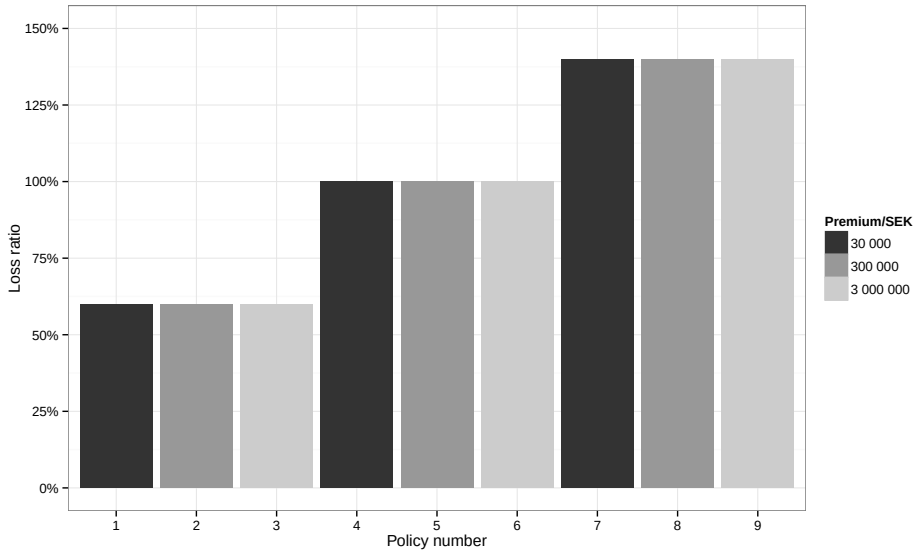

Figure 4.1: Loss ratios of nine fleet insurance contracts averaged over 5 years.

Now let us consider the policies by grouping them by their premium. The contracts with low premiums are likely fleets with fewer vehicles, while the higher premiums cover more vehicles. Therefore, we may expect to see a higher variance in observed yearly loss ratios in the group with a premium of SEK 30 000 than in the group with a premium of SEK 3 000 000.

Figure 4.2 shows the loss ratios for each of the 5 years in our data set for policies 1, 4 and 7, which all have a premium of SEK 30 000. The dark line shows the 5-year average loss ratio. The variation in loss ratio from year to year is very high, so that we may rightfully question whether the difference in average loss ratios here is more than just random chance. Figure 4.3 shows the loss ratios for the policies with a premium of SEK 300 000. Here the variation is much lower, leading us to believe more in the average loss ratio, but there is still significant variation between years. Finally, Figure 4.4 shows the loss ratios for the policies with a premium of SEK 3 000 000. The yearly variation here is small, so that we may put a lot of faith in these average loss ratios.

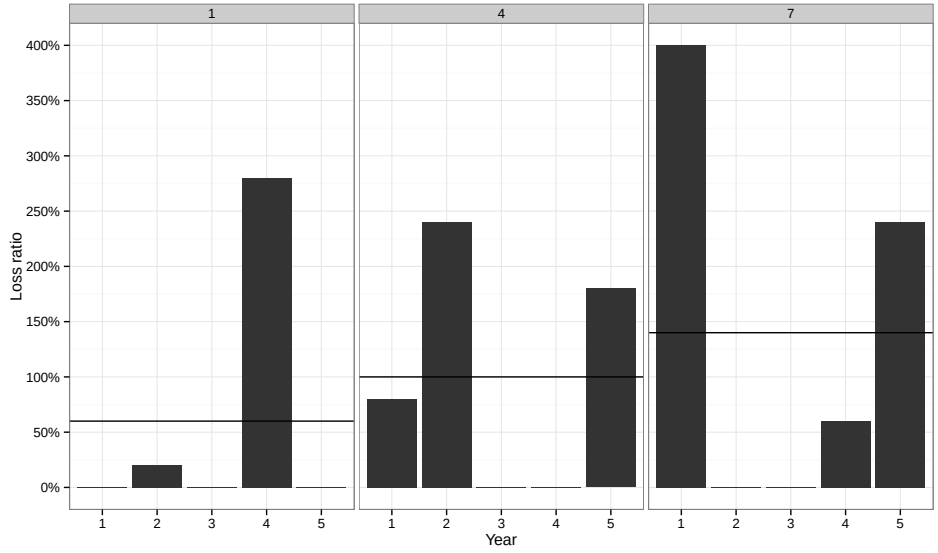

Figure 4.2: Loss ratios of policies with premium SEK 30 000: policies 1, 4 and 7.

If we again look at Figure 4.1 and consider policies 7–9, all three have an average loss ratio of 140 %, but our analysis suggests that we don't give them all the same premium adjustment. For policy 7, we believe the data tells us little about the risk, and we use the average over all nine contracts, 100 %, as our forecast for next year's performance. For policy 9, we believe strongly in our data, and so we forecast that next year's loss ratio will also be 140 %, and increase the premium accordingly. For policy 8, we believe that

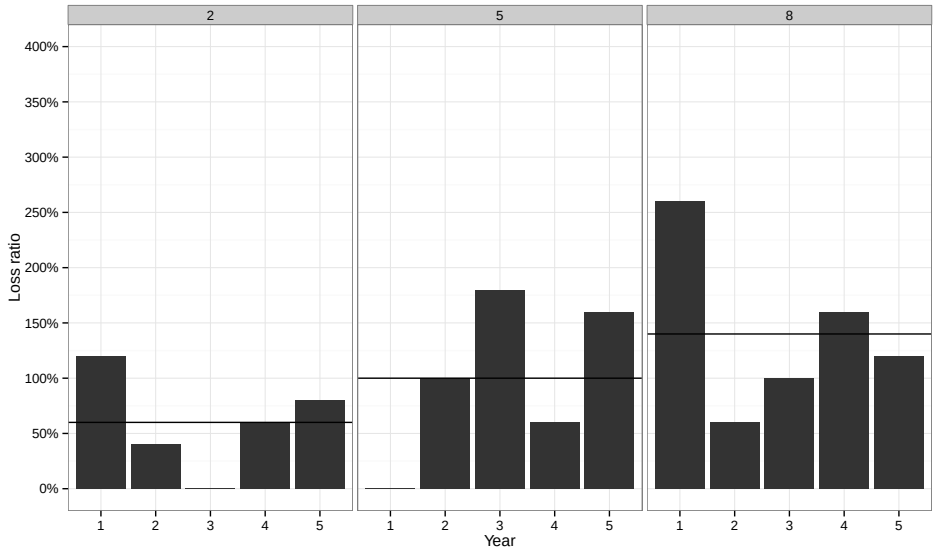

Figure 4.3: Loss ratios of policies with premium SEK 300 000: policies 2, 5 and 8.

we are in an in-between situation. We make a forecast that lies somewhere between 100 % and 140 %, and increase the premium, but by less than for policy 9.

This example shows why, intuitively, we wish to combine both collective and individual experience in insurance ratemaking. Credibility theory gives a solid mathematical foundation for how to do this, and in the section that follows, we shall look at a regression based way to handle it.

4.2 Generalised linear mixed models (GLMM)

Recall from Section 3.2.3 that in the GLM model, the predicted claim frequency for a policy i is determined by its tariff cell, in other words by its vector of rating factors. All policies that fall under the same tariff cell will receive the same predicted claim frequency, regardless of their claim history.

From a credibility perspective, we would like to take into account the specific experience we have of each policy in order to adjust this prediction. Assume that we have 4 years of history on two policies in the same tariff cell, where the predicted claim frequency is one claim per year. The first policy has had 6 claims during these 4 years, while the other has had only 2 claims. We would like to give the first policy a higher predicted claim frequency than the GLM model indicates, while the second policy should receive a lower prediction. The generalised linear mixed model (GLMM) allows us to do precisely this.

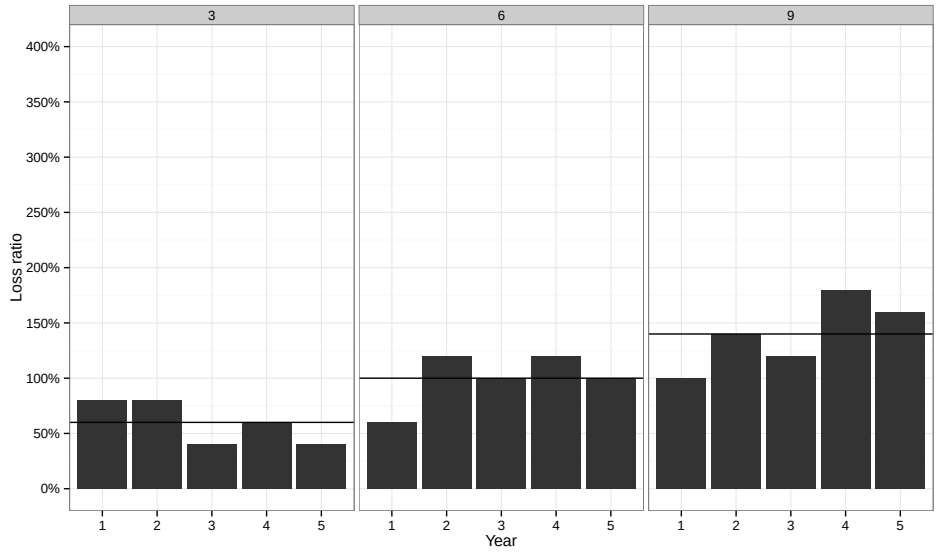

Figure 4.4: Loss ratios of policies with premium SEK 3 000 000: policies 3, 6 and 9.

Generalised linear mixed models extend the GLM by allowing inclusion of random effects in the linear predictor. This allows estimation of the heterogeneity between policies within the same tariff cell. The use of GLMMs in actuarial pricing is discussed in Antonio and Beirlant (2007), and here we follow their exposition. A more extensive overview of GLMMs can be found in, for instance, McCulloch and Searle (2001).

In the GLM model, all observations that have the same rating factors are part of the same tariff cell and are considered independent of each other. In reality, we have repeated observations of the same policies, in other words: we have so called *longitudinal* data. The GLMM model allows us to add a policy-level random effect that helps model the association in the data that occurs due to the same policy being observed multiple times.

Let $i = 1, \ldots, N$ index the individual policies, and $j = 1, \ldots, J_i$ index the subsequent observations for the policy i. We introduce the vector $\mathbf{u}_i$ of random effects for policy i. Given $\mathbf{u}_i$, the repeated measurements $Y_{i1}, \ldots, Y_{iJ_i}$ are assumed independent, and distributed according to an exponential family distribution with density function

$$f\left(y_{ij}|\mathbf{u}_i, \beta, w_{ij}, \phi\right) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi/w_{ij}} + c(y_{ij}, w_{ij}, \phi)\right\}, \quad j = 1, \ldots, J_i. \tag{4.1}$$

Similar to the GLM case described in Section 3.2.3, it holds that

$$\mu_{ij} = \mathbb{E}\left[Y_{ij}|\mathbf{u}_i\right] = b'(\theta_{ij}), \tag{4.2}$$

$$\text{Var}\left(Y_{ij}|\mathbf{u}_i\right) = b''(\theta_{ij})\frac{\phi}{w_{ij}}, \tag{4.3}$$

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i. \tag{4.4}$$

As in the GLM case, $\boldsymbol{\beta}$ is a vector of parameters for the so called fixed effects, determining the effect of each tariff cell, and $\mathbf{x}_{ij}$ is the vector of covariate information for subject i, observation j. New in the GLMM is $\mathbf{z}_{ij}$, a vector of random effects covariate information, as well as the random effects vector $\mathbf{u}_i$ itself.

We assume the random effects $\mathbf{u}_i, i = 1, \ldots, N$ to be mutually independent and identically distributed with density $f\left(\mathbf{u}_i|\boldsymbol{\alpha}\right)$, where $\boldsymbol{\alpha}$ is a vector of unknown parameters in this density. Typically, the random effects are taken to be (multivariate) normally distributed with mean zero and covariance matrix determined by $\boldsymbol{\alpha}$. The motivation for assuming a zero mean is that the overall mean is included in the model as an intercept, so that the random effects are policy-level deviations from the group means supplied by the intercept and other fixed effects.

In the specific GLMM model that we will use, we have a single random intercept u_i for each policy, letting $z_{ij} = 1$, so that we have

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i. \tag{4.5}$$

Note that the fixed effects parameters $\boldsymbol{\beta}$ are the same for all policies, while the random effects are specific to each policy i. However, the parameters $\boldsymbol{\beta}$ cannot generally, as in the GLM case, be seen as the effect of the covariates on the population average. Instead, they represent the effect of the covariates on the response, *conditional on the random effects* $\mathbf{u}_i$. In the GLM,

$$\mathbb{E}\left[Y_{ij}\right] = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta}), \tag{4.6}$$

but in the GLMM, it generally instead holds that

$$\mathbb{E}\left[Y_{ij}\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{ij}|\mathbf{u}_i\right]\right] = \mathbb{E}\left[g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i)\right] \neq g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta}). \tag{4.7}$$

However, there are special cases where the marginal interpretation is valid. In the linear mixed model, where g is the identity link function $g(x) = x$, it does hold

$$\mathbb{E}\left[g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i)\right] = \mathbb{E}\left[\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i\right] = \mathbf{x}'_{ij}\boldsymbol{\beta}, \tag{4.8}$$

since $\mathbb{E}\left[\mathbf{u_i}\right] = 0$.

Predicting the claim frequency Y_{i,J_i+1} of an insurance policy over the next year of insurance is simple using this GLMM model, as the predictive

distribution of Y_{i,J_i+1} can be obtained from the same MCMC sampler that is used to estimate the model. The question arises of how to predict the claim frequency of a new policy, for which there is no historical data to estimate the random intercept u_i. However, since $\mathbb{E}[u_i] = 0$, the fixed effects serve as group-level means, and so a predictive distribution will still be possible to obtain, effectively using $u_i = 0$ for this future observation.

4.3 Ratemaking with the GLMM model

From the perspective of a pricing actuary, the GLMM model described in Section 4.2 has both advantages and disadvantages compared to the GLM model of Section 3.2.

The policy level random intercept u_i in the GLMM model is a way to capture those differences in risk profile between policies that are not captured by the rating factors. There are many such factors, for different reasons. Some are difficult to quantify, such as aggressive driving style; others are difficult to measure with high data quality, such as abstinence from alcohol. Yet others simply incur too high an administrative load to measure, such as the myriad of different safety equipments a car may be fitted with (adaptive cruise control, automatic lane keeping, sleep detectors, automatic braking systems, etc.).

The GLMM model thus has the opportunity to improve prediction of claim frequencies for individual policies by using the data available on each policy. As more data becomes available over time, the predictions should improve. On the other hand, to receive a prediction from the GLMM model for a policy i, the random effect u_i needs to be known. Insurance companies have large portfolios of policies which may begin on any given day of the year. When a policy nears the end of its term, the insurance company will calculate a new premium and send the customer an offer to renew the policy for another year. In order to correctly calculate this new premium, the GLMM model needs to be re-estimated so that the u_i's are updated. This leads to a situation where the tariff needs to be re-estimated on at least a monthly basis, which incurs an administrative cost for the insurance company.

Another drawback for a practitioner is that while GLM tariffs are industry standard and supported by any statistical software that an insurance company may use, GLMM models are less well supported by software, and so may be more difficult to integrate into company processes.

Chapter 5

Tariff comparison metrics

To determine which of two alternative insurance tariffs is the better, one needs good comparison metrics that capture the desirable properties sought in a tariff.

Determining the *level*, e.g. the total number of claims to expect during a period for a portfolio of insurances, or the total costs of the claims, is not the main goal of the insurance tariff itself. The level is estimated based on historical data, and is adjusted by reserving and catastrophe models. Instead, we seek metrics that allows us to quantify the tariff's ability to differentiate between the high and low risk policies. As mentioned in Section 2.3, it is common to split the data into a *training* data set and a *validation* data set. Given models that have been estimated using the training data set, they are not necessarily well calibrated against the validation data set. Thus, the first step when evaluating tariffs using an evaluation data set is to adjust the predictions so that the total sum of predicted claims is equal to the observed number of claims in the validation data set. This allows all tariffs to compete on an even footing.

Note that we will discuss predicted number of claims and claim frequencies, as this is what we are modelling, but that all these methods work equally well when modelling pure premium.

We will consider three different metrics to compare our tariffs: Lorenz curves and corresponding Gini scores, the *quotient test*, and proper scoring rules. Both the Lorenz curves and the quotient test hinge on ordering policies based on a ratio between the two competing models' predictions. To calculate this ratio, the predictive distribution needs to be reduced to a point prediction of some sort, in our case the expected value. This reduction of the full predictive distribution to a point prediction means that a wealth of information is discarded, and means that we are missing out on one large advantage of using Bayesian models. The proper scoring rules have the advantage of considering the full predictive distributions instead of point predictions only.

5.1 Lorenz curves and Gini scores

Lorenz curves are known from welfare economics as a way to illustrate inequality in distribution of income over a population. However, they can be easily extended to other situations where comparisons of distributions are useful. Frees, Meyers, and Cummings (2011) introduce "ordered" Lorenz curves and apply them to the problem of comparing insurance tariffs. This implies comparing premiums, but the methodology can just as easily be applied to any predictive quantity, in our case the claim frequency.

Consider an evaluation data set of K policies for which we know the observed claim frequencies $L_k, k = 1, \ldots, K$. Let $P_{k,ref}$ be the reference model's point prediction of the claim frequency L_k, and $P_{k,alt}$ the alternative model's point prediction. We define R_k as the ratio between these,

$$R_k = \frac{P_{k,alt}}{P_{k,ref}}. \tag{5.1}$$

We now wish to order the policies by R_k, such that $R_{\pi_1} \leq R_{\pi_2} \leq \cdots \leq R_{\pi_k}$ for some permutation of the indices $k = 1, \ldots, K$, i.e. π_1 is the value in $\{1, \ldots, K\}$ such that $R_{\pi_1} = \min_k R_k$; π_2 is the index with the second smallest R_k, and so forth. The Lorenz curve is defined by the pairs (x_k, y_k), $k = 0, \ldots, K$, where $(x_0, y_0) = (0, 0)$ and

$$x_k = \frac{\sum_{l=1}^{k} P_{\pi_l,ref}}{\sum_{l=1}^{K} P_{\pi_l,ref}}, \tag{5.2}$$

$$y_k = \frac{\sum_{l=1}^{k} L_{\pi_l}}{\sum_{l=1}^{K} L_{\pi_l}}. \tag{5.3}$$

The interpretation of this is that the policies are ordered from the one where the alternative prediction is the smallest compared to the reference prediction, to the one where the alternative prediction is the largest compared to the reference prediction. On the x-axis of the Lorenz curve we show the proportion of total *predicted* claims that has been summed up, with the point prediction used being that of the reference model, i.e. $P_{k,ref}$. The y-axis shows the cumulative proportion of total *observed* claims. Figure 5.1 illustrates this; the figure is based on the simulations described in Chapter 6. The dashed 45° line indicates a break-even line, where equal amounts of predicted and observed claims have been accumulated for each point.

At the beginning of the curve, we have the policies where the alternative model believes that the reference model has overestimated the number of claims. Therefore, if the alternative model is correct, the curve should fall below the 45° line, as the reference predictions are higher than the observed claims. Conversely, if the reference model is more accurate, then the curve will go above the 45° line, as the alternative model mistakenly places higher-risk policies to the left. In Figure 5.1, the alternative model is slightly worse

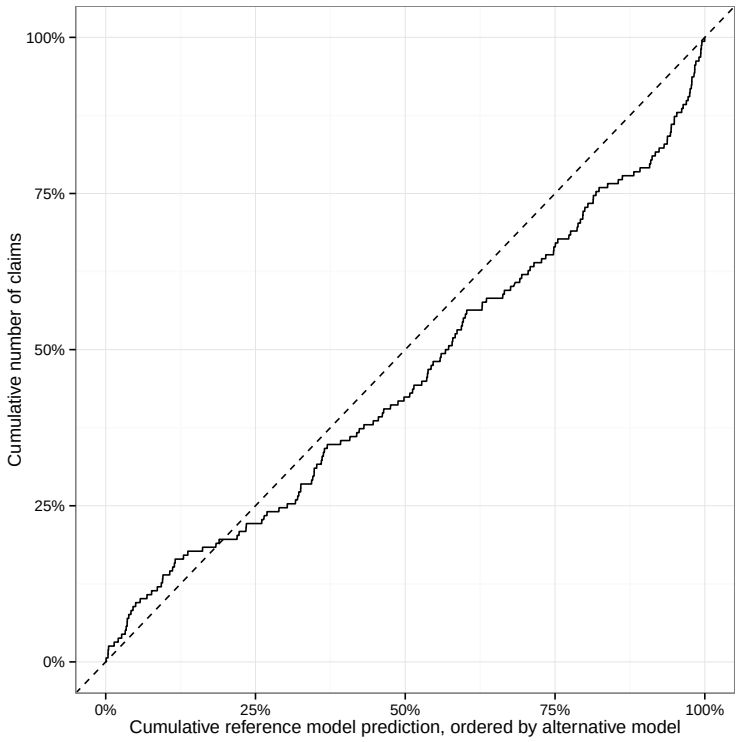

Figure 5.1: Lorenz diagram of alternative vs reference model.

in detecting the lowest-risk policies, but is an improvement on the reference model on the most part.

The Gini score is a summary statistic of the Lorenz curve, and is defined as two times the area between the Lorenz curve and the 45° line, with a Lorenz curve below the line giving a positive Gini score, and a Lorenz curve above the line giving a negative score. The Gini score is thus calculated as

$$1 - \sum_{l=1}^{K} \left(P_{\pi_l, ref} - P_{\pi_{l-1}, ref} \right) \left(L_{\pi_l} + L_{\pi_{l-1}} \right). \tag{5.4}$$

As an example, the Gini score of the Lorenz curve in Figure 5.1 is 0.093, indicating that the alternative model is an improvement on the reference model.

5.2 Quotient test

The quotient test is based on the same ratio (5.1) between the alternative and reference model predictions as is used in the Lorenz curves. Again, we consider an evaluation data set of K policies for which we know the observed claim frequencies $L_k, k = 1, \ldots, K$ and can compare the ratios R_k.

In the quotient test, the policies are divided into two groups: those for which $R_k > 1$, i.e. those where the alternative model implies an increase in predicted number of claims, and those for which $R_k \leq 1$, i.e. where the alternative model decreases the prediction compared to the reference model. For each group, the quotient between observed claims and predicted number of claims in the evaluation data is calculated for both models, and these are then tabulated for comparison. The formulae are $L_k/P_{k,ref}$ and $L_k/P_{k,alt}$.

The two groups can be seen as those for which the alternative model prescribes an increased prediction relative to the reference model, and those for which it prescribes a decreased prediction. If the alternative model is an improvement on the reference model, it should then have predictions which are closer to the observed outcomes for these groups than the reference model does. In other words, the quotients should be closer to 1.

Table 5.1 shows an example taken from the simulations in Chapter 6. For the policies where the alternative model has a lower prediction than the reference, the reference model overestimates the number of claims, with ratio of observed to predicted claims of 0.94. The alternative model has a better prediction, with a ratio of observed to predicted claims of 1.04, which is closer in magnitude to 1. Similarly, for the policies where the alternative model has a higher prediction than the reference, the reference model underestimates the number of claims, having a ratio of observed to predicted of 1.16. The alternative tariff instead has a ratio of 0.93, which again is closer to 1. Switching from the reference model to the alternative model here then improves the predicted number of claims in both these groups. Note that in both cases, the alternative model has "overshot the target" in the sense that it has gone from a group where claims are underestimated to one where they are overestimated, but by a smaller margin, and vice versa.

Change in prediction from reference to alternative	Ratio observed/predicted Alternative	Reference	Winner
alternative pred. lower	1.04	0.94	alt
alternative pred. higher	0.93	1.16	alt

Table 5.1: Quotient test example, with alternative model outperforming the reference.

As an aside, the quotient test has another feature of interest to a practitioner: it can give some insight into the effect on sales of changing tariffs. The group which is given a lower premium by the alternative tariff are likely to be more inclined to purchase policies after the tariff change, while the group which receives increased premiums are less likely to. By comparing the amount of exposure in the two groups, the effect on the portfolio can be estimated.

5.3 Proper scoring rules

The previous two methods for model comparison have been based on point predictions. However, our Bayesian models yield full posterior predictive distributions, and a good tool for comparing those are *proper scoring rules*. Scoring rules allow comparison of predictive distributions by assigning a score to a prediction based on the prediction itself and the observed outcome. The scores are taken to be positive rewards in the sense that a higher score indicates a better prediction. The *calibration* of a predictive distribution is a measure of the statistical consistency between the predictive distribution and the observed outcomes, which means that calibration is a joint property of prediction and outcomes. The *sharpness* of a predictive distribution is a measure of how concentrated the predictive distribution is over the possible outcomes, and thus is a property of the prediction only. Gneiting and Raftery (2007) suggest that the goal of probabilistic forecasting is to maximise the sharpness subject to calibration. Proper scoring rules assign scores that increase for a given outcome if the predicted probability of the outcome increases. This rewards both sharpness and calibration in the sense that a more concentrated probability mass will lead to a higher score for observations, *if* the predictive distribution is well calibrated to the outcomes.

Gneiting and Raftery (2007) give a thorough introduction to scoring rules for continuous prediction, interval and quantile prediction and categorical predictions, and we will follow their exposition here. Technically, we are modelling claim count data, which means that the predictive distributions are defined on $\mathbb{N} = \{0, 1, 2, \ldots\}$. Czado, Gneiting, and Held (2009) study proper scoring rules for such count data. However, in practice the low incidence of claims means that the observations, and thus (hopefully) also the predictive distributions, are overwhelmingly skewed towards zero and the low integer counts, such that they can be considered to have the support $\{0, 1, 2, 3, 4, 5\}$. We will therefore use scoring rules for such finite discrete distributions to evaluate our models. Note that our data is not categorical, but ordinal, so that we may use scoring rules that depend in some way on this, an example being the ranked probability score.

Let P be a predictive distribution that we wish to score, and let x denote the event that materialises. The score of P when x materialises is denoted by $S(P, x)$, where the score function S takes values in $\mathbb{R}$, or possibly the extended real line $\overline{\mathbb{R}} = [-\infty, \infty]$, e.g. the logarithmic score, which may assume the value $-\infty$. When predicting a categorical or ordinal variable with $m + 1$ possible outcomes, a predictive distribution can be represented by a probability vector $\mathbf{p} = (p_0, \ldots, p_m)'$ where $p_i \geq 0,\ i = 0, \ldots, m$ and $p_0 + \cdots + p_m = 1$. A scoring rule for such a predictive distribution can then be represented by $m + 1$ functions $S(\cdot, i),\ i = 0, \ldots, m$. Thus, if the prediction $\mathbf{p}$ is given, and outcome i occurs, the score is $S(\mathbf{p}, i)$.

Let $\mathbf{q} = (q_0, \ldots, q_m)'$ be another probability distribution. We denote the expected value of the score of $\mathbf{p}$ under this distribution by

$$S(\mathbf{p}, \mathbf{q}) = \mathbb{E}_{\mathbf{q}}\left[S(\mathbf{p}, \cdot)\right] = \sum_{i=0}^{m} p_i S(\mathbf{p}, i). \tag{5.5}$$

Assuming that our best prediction is $\mathbf{q}$, we wish for a scoring rule to incentivise us to quote this best judgement, i.e. we want it to hold that

$$S(\mathbf{q}, \mathbf{q}) \geq S(\mathbf{p}, \mathbf{q}), \tag{5.6}$$

for any predictive distribution $\mathbf{p}$. If S fulfils this criteria, then it is called a *proper* scoring rule. If the equality in (5.6) holds only for $\mathbf{p} = \mathbf{q}$, then S is *strictly proper*.

Czado et al. (2009) note that there is no simple or automatic way to decide what scoring rule to use to evaluate predictive distributions, but suggest to use multiple ones to take advantage of their respective emphases and strengths. We will therefore consider four strictly proper scoring rules: the ranked probability score from Czado et al. (2009), and the quadratic, spherical and logarithmic scores from Gneiting and Raftery (2007). The *ranked probability score* is used for ordinal data, and is defined as

$$S(\mathbf{p}, i) = -\sum_{j=0}^{m} \left[\left(\sum_{k=0}^{j} p_k \right) - \mathbf{1}_{\{i \leq j\}} \right]^2, \tag{5.7}$$

where $\mathbf{1}$ is the indicator function that is equal to 1 if the condition is fulfilled, else 0. The ranked probability score thus is the negative of the sum of the squared difference between the cumulative distribution function of the prediction and the empirical cumulative distribution of the outcome, which is simply the step function that goes from 0 to 1 at the observed value. The *quadratic score* is given by

$$S(\mathbf{p}, i) = -\sum_{j=0}^{m} (\delta_{ij} - p_j)^2 = 2p_i - \sum_{j=0}^{m} p_j^2 - 1, \tag{5.8}$$

where the Kronecker delta, δ_{ij}, is equal to 1 when $i = j$, else 0. The *spherical score* is defined as

$$S(\mathbf{p}, i) = \frac{p_i}{\left(\sum_{j=0}^{m} p_j^2\right)^{1/2}}. \tag{5.9}$$

Finally, the *logarithmic score* is given by

$$S(\mathbf{p}, i) = \log p_i. \tag{5.10}$$

The logarithmic score is the only proper scoring rule that only depends on the predictive distribution through the probability p_i at the observed

outcome (Czado et al., 2009). A drawback of the logarithmic score is that it becomes $-\infty$ if $p_i = 0$ and i is the observed outcome. In the case of prediction of rare events, such as the claim events modelled in this thesis, it is possible that a sample from an MCMC algorithm will lead to an empirical distribution where some $p_i = 0$, and for this event to occur. In such cases, the logarithmic score becomes difficult to interpret.

As a simple illustration of the scores we will use, we consider two predictive distributions resulting from the estimation of the GLMM model in Chapter 6. The observed value of the first example was 2 claims, while the observed value of the second example was 0 claims. Figure 5.2 shows the two predictive distributions, while Figure 5.3 show the scores as function of the observed value, with one panel for each scoring rule, and one line per example predictive distribution. We note that the first example predictive distribution has a higher predicted number of claims than the second. In Figure 5.3 we see the results of this. If the observed value is 0, the second prediction receives a higher score, but it quickly drops off for higher observed claim counts, as it has zero probability assigned to claim counts higher than 1. We observe here also that the log score becomes $-\infty$ where the predicted probability of the outcome is zero.

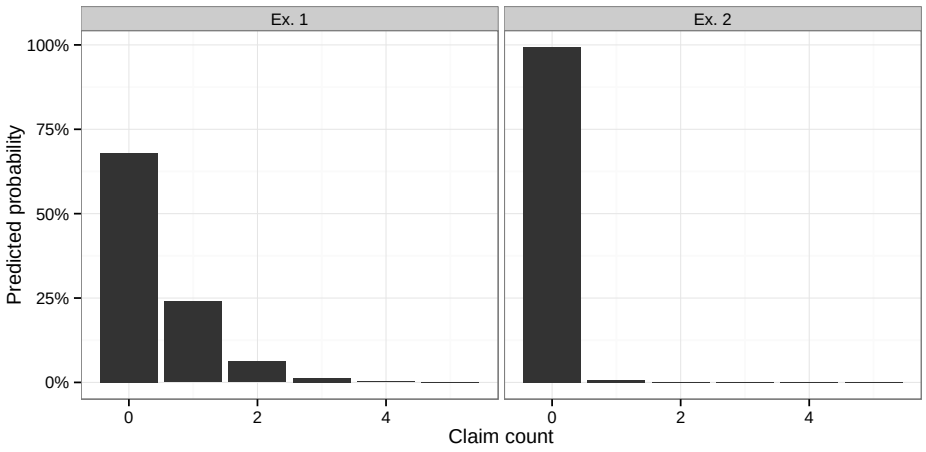

Figure 5.2: Example predictive distributions for two observations from the evaluation data, generated by the GLMM model from Chapter 6.

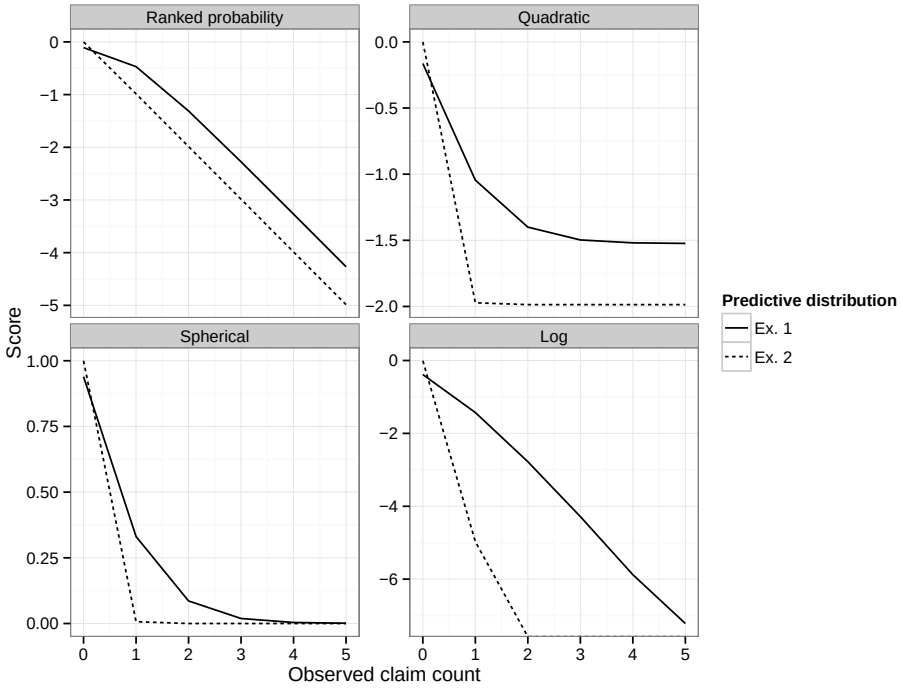

Figure 5.3: Scores as a function of the observed outcome for the two example predictive distributions, one panel per proper scoring rule.

Chapter 6

Data analysis

To determine the effectiveness of using individual policy level data for predicting claim frequencies, we consider two models: a GLM model with fixed effects only, as a reference model, and a GLMM model with a random intercept for each policy. Both models will be estimated using the Trygg Hansa TPL insurance data set described in Section 1.2, and then evaluated against each other using the metrics described in Section 2.4 and Chapter 5.

6.1 Model specification

Let the individual policies be indexed by $i = 1, \ldots, 12000$, and the repeated observations for each policy be indexed by $j = 1, \ldots, J_i$. We denote by Z_{ij} the number of claims for the j-th observation of policy i, and by w_{ij} the exposure for the period. We will model the claim counts Z_{ij}, from which the claim frequency is calculated as $Y_{ij} = Z_{ij}/w_{ij}$.

Both models will use as fixed effects the five rating factors mentioned in Table 1.3 in Section 1.2, namely *fordar, fvehic, kkarb, korstr*, and *ztrkof*. All five rating factors are ordinal, and can be seen as discretised and right-truncated continuous variables, e.g. *fordar* is the age of the vehicle measured in whole years, and truncated at 90. As can be seen in Figures 1.1–1.5 on pages 8–10, the claim frequency's dependence on these rating factors is well approximated by a linear function, at least over the classes with high exposure. We therefore include these rating factors linearly in the models. More advanced treatments, such as generalised additive models (GAMs) (Ohlsson and Johansson, 2010) are also possible, but are outside the scope of this thesis.

The reference claim frequency model we will use is a standard Poisson GLM with exposure, cf. Section 3.2.3:

$$\begin{aligned} Z_{ij} &\sim \mathrm{Po}(\mu_{ij}), \\ \log(\mu_{ij}) &= \log w_{ij} + \mathbf{x}'_{ij}\boldsymbol{\beta}. \end{aligned} \tag{6.1}$$

Here $\mu_{ij} = \mathbb{E}\left[Z_{ij}|w_{ij}, \mathbf{x}_{ij}\right]$ is the expected number of claims of observation j for policy i, $\mathbf{x}_{ij}$ is the vector of covariates, and $\boldsymbol{\beta}$ is the vector of fixed effect coefficients to be estimated, and is constant for all i, j. The parameter β_0 is the intercept, while $\beta_{\text{fordar}}, \ldots, \beta_{\text{ztrkof}}$ correspond to the 5 rating factors. Writing the linear predictor out explicitly, we have

$$\begin{aligned}\log(\mu_{ij}) = \log w_{ij} &+ \beta_0 \\ &+ x_{\text{fordar}}\beta_{\text{fordar}} \\ &+ x_{\text{fvehic}}\beta_{\text{fvehic}} \\ &+ x_{\text{kkarb}}\beta_{\text{kkarb}} \\ &+ x_{\text{korstr}}\beta_{\text{korstr}} \\ &+ x_{\text{ztrkof}}\beta_{\text{ztrkof}}.\end{aligned} \tag{6.2}$$

Note that since the GLM model does not include any policy-level effects, it would have been possible to use a single index k to enumerate all observations, instead of the dual indices i, j. However, the present form is easier to compare to the GLMM model, which does require the dual indices.

To incorporate the individual policy claims history in the tariff, we use a GLMM model of the kind described in Section 4.2, which extends the reference GLM model by adding a random intercept u_i for each policy i, in addition to the fixed effects $\boldsymbol{\beta}$. The model is specified as

$$\begin{aligned}u_i &\sim \mathrm{N}(0, \sigma_u^2), \\ Z_{ij}|u_i &\sim \mathrm{Po}(\mu_{ij}), \\ \log(\mu_{ij}) &= \log w_{ij} + u_i + \mathbf{x}_{ij}'\boldsymbol{\beta},\end{aligned} \tag{6.3}$$

where $\mu_{ij} = \mathbb{E}\left[Z_{ij}|u_i, w_{ij}, \mathbf{x}_{ij}\right]$ is now the expected claim frequency of the j-th observation of policy i, *conditional on* the random intercept u_i. Written explicitly, the model for the linear predictor is

$$\begin{aligned}\log(\mu_{ij}) = \log w_{ij} + u_i &+ \beta_0 \\ &+ x_{\text{fordar}}\beta_{\text{fordar}} \\ &+ x_{\text{fvehic}}\beta_{\text{fvehic}} \\ &+ x_{\text{kkarb}}\beta_{\text{kkarb}} \\ &+ x_{\text{korstr}}\beta_{\text{korstr}} \\ &+ x_{\text{ztrkof}}\beta_{\text{ztrkof}}.\end{aligned} \tag{6.4}$$

With the models specified, we will now discuss the estimation of the parameters $\boldsymbol{\beta}$ and $u_i, i = 1, \ldots, 12000$.

6.2 Model inference

Both the GLM reference model and the GLMM alternative model were estimated as Bayesian models using the software JAGS (Plummer, 2003), which

implements MCMC methods using Gibbs sampling. Due to the computationally heavy and memory-intensive nature of MCMC methods, it was not possible to estimate the models on the entire data set consisting of 12 000 policies using JAGS. Therefore, as described in Section 1.2, the policies were divided by simple random sampling into six subsets of 2 000 policies each. For each such subset, the observations were divided into a training data set consisting of 2/3 of the observations, and an evaluation data set consisting of 1/3 of the observations, refer to Section 1.2 for the details. The models were then estimated separately for all six subsets of the data, and our model comparisons are performed per subset, as well as over the merged result sets.

As prior distributions for the fixed effects vector $\boldsymbol{\beta}$, the GLM model was estimated using standard maximum-likelihood methods on the entire training data set. The resulting parameter estimates and variances were used as the mean and variance in normal distribution priors for the parameters, see Table 6.1 for these values. Thus, each of the six subsets of the data were estimated using the same prior distributions for all parameters. Since we are using the data to obtain MLE estimates of the parameters, and then using these estimates in our prior distributions, we are using a form of *empirical Bayes* method, in a sense using the data twice, both in the prior and in the MCMC estimations. This means that care needs to be taken with the inference, e.g. a naive credible interval for a parameter will be narrower than is warranted due to this double use of the data. As we are concerned only with the predictive performance of our models, this does not lead to issues for us, and the increased speed of convergence of our MCMC simulations is welcome. Carlin and Louis (2009) contains a chapter on empirical Bayes methods, which discusses the adjustments necessary to Bayesian inference in this setting.

The random effects u_i are distributed as $\mathrm{N}(0, \sigma_u^2)$, where $1/\sigma_u^2$ was assigned the frequently used vague prior of $\mathrm{Ga}(0.01, 0.01)$. Gelman (2006) argues that such a prior distribution is not uninformative, especially if the number of random intercepts N is small, or when σ_u is close to zero, and recommends instead a uniform prior $\sigma_u \sim \mathrm{U}(0, A)$ where A is some large number. In our case, the number of random intercepts is large, and the standard deviation σ_u is not expected to be very close to zero. Furthermore, computational efficiency is an important concern for us, and $\mathrm{Ga}(0.01, 0.01)$ is a conjugate prior, leading to faster sampling. Therefore, we choose to proceed with the $\mathrm{Ga}(0.01, 0.01)$ prior.

The GLM and GLMM models were both estimated using each of the six training data subsets. For each combination of model and data subset, two MCMC chains were used, with the first 500 iterations per chain discarded as a burn-in period, and subsequently 37 500 iterations per chain were used for estimation. The parameter estimates are thus based on 75 000 samples each. To obtain samples from the predictive distributions for the evaluation data, each observation in the evaluation data subset was included in the

β	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.2521	0.1362	-23.87	0.0000
fordar	-0.0229	0.0042	-5.45	0.0000
fvehic	-0.0168	0.0075	-2.24	0.0253
kkarb	-0.0079	0.0019	-4.16	0.0000
korstr	0.0225	0.0045	4.98	0.0000
ztrkof	0.0115	0.0013	8.86	0.0000

Table 6.1: The parameter estimates, standard errors and z-test from a maximum likelihood estimation of the Poisson GLM model using the entire training dataset. These values are used as the mean and standard deviations of normal priors in the Poisson GLM and GLMM models for each of the six data subsets.

JAGS model, but with the outcome deleted, thus generating 75 000 samples from the posterior predictive distributions of each of the evaluation data observations. Generating the samples for both the GLM and GLMM models over all six data subsets took approximately 70 hours.

The outputs from each run of the estimation procedure are samples drawn from the posterior distributions of each of the fixed effect parameters $\beta_0, \beta_{\text{fordar}}, \ldots, \beta_{\text{ztrkof}}$ and the random intercept variance σ_u^2, samples drawn from the posterior distribution of each of the 2 000 random intercepts included in the data subset, and samples drawn from the posterior predictive distribution for each observation in the evaluation data subset. Since there are six different data subsets, the GLM and GLMM models each generate samples from six different posterior distributions for $\boldsymbol{\beta}$, one per subset, while each random intercept $u_i, i = 1, \ldots, 12000$ is only included in a single data subset, and therefore all samples for a given u_i are drawn from a single posterior distribution.

To assess the convergence, we will now study some plots over the outcomes of the MCMC simulations. Figure 6.1 shows plots of the samples drawn from the two chains in the GLM model for data subset 1, with one panel per element in $\boldsymbol{\beta}$. The figure indicates that the convergence and mixing are good. The corresponding plots for data subsets 2–6 look essentially the same, and are not shown here. Figure 6.2 shows the corresponding plot for the GLMM model, where again the convergence and mixing is good. Figure 6.3 shows plots of the samples of σ_u drawn from the two chains over each of the six data subsets. Here we see that σ_u does not appear to converge as well as the other parameters do, and there is some variation between the data subsets in how good the mixing is. It can be noted that while the values of the parameters $\boldsymbol{\beta}$ are estimated on the "first level", σ_u is the standard deviation of the parameters u_i, and thus is a second-level estimate. This means that the convergence of σ_u is not very important to the validity of

the posterior predictive distributions that are our main interest.

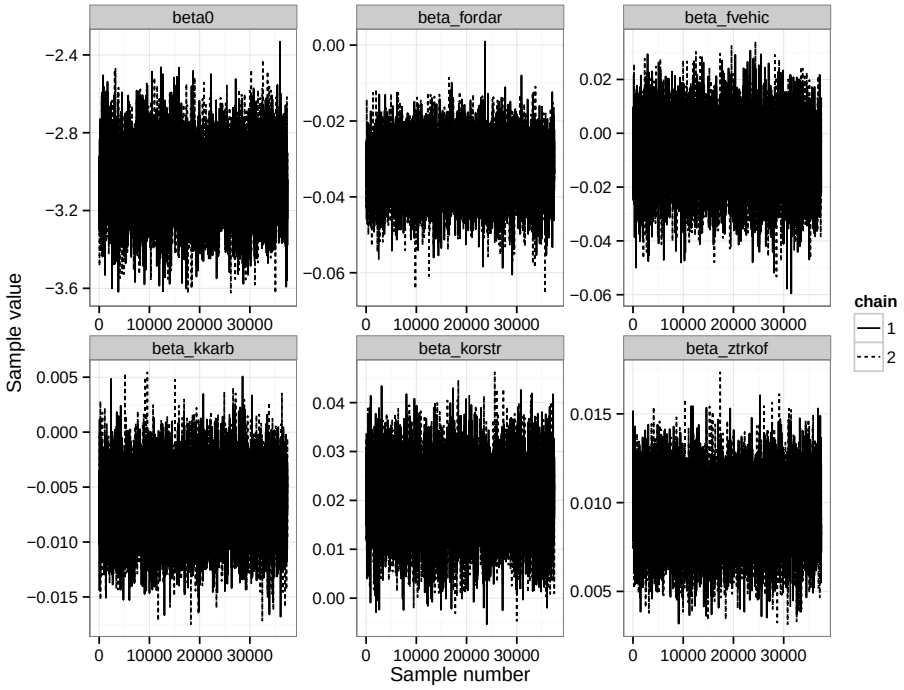

Figure 6.1: Chain sample plots for the GLM model on data subset 1.

Next, we consider the estimated posterior distributions of the parameters $\beta_0, \beta_{\text{fordar}}, \ldots, \beta_{\text{ztrkof}}$. We estimate these posterior distributions by a Gaussian kernel density estimate calculated from the posterior samples obtained from the MCMC simulations. Figure 6.4 shows the distributions of the elements of $\boldsymbol{\beta}$ for the GLM model, with one panel for each combination of β and data subset. We find that the distributions of a parameter, e.g. β_0 are quite similar for each of the six data subsets, especially in terms of variance, but that there is some difference in the subset means. The same thing is seen in the corresponding plot for the GLMM model, Figure 6.5. The estimated posterior distributions of σ_u from the GLMM model are shown in Figure 6.6. It can be seen that these differ substantially between the data subsets. Studying the chain trajectories in Figure 6.3 indicates that convergence has not been fully reached for data subsets 4 and 5, and that mixing is slow overall. However, as mentioned previously, this is not a primary concern, as it does not affect $\boldsymbol{\beta}$ much, and our aim is to predict claim frequencies rather than understand the distribution of the u_i's. As previously mentioned, sampling took approximately 70 hours to complete, making it infeasible to run the models longer to improve convergence.

The fact that the distributions of $\boldsymbol{\beta}$ differ somewhat between the different

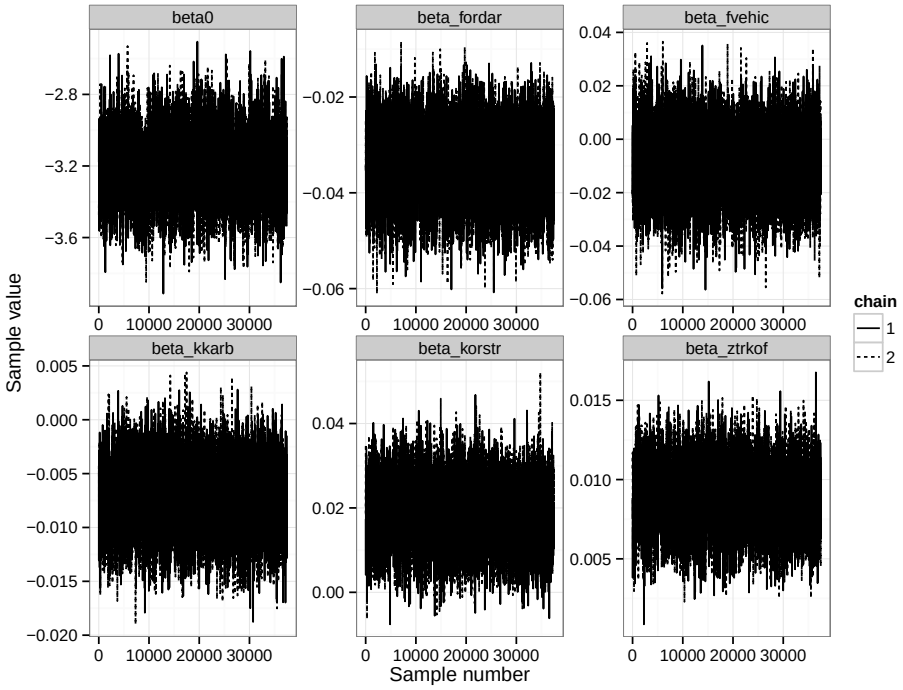

Figure 6.2: Chain sample plots for the GLMM model on data subset 1.

data subsets raises the question of how to obtain a single estimate of $\boldsymbol{\beta}$ from the six that we receive as output from the MCMC simulations. One ad-hoc possibility is to simply pool the samples from all six simulations. The pooled posterior distributions for $\boldsymbol{\beta}$ under the GLM and GLMM models can be seen in Figure 6.7. We note that the resulting posterior distributions are satisfactory in that they are smooth and unimodal, with little skewness. It can also be seen that the posterior distributions for $\beta_{\text{fordar}}, \ldots, \beta_{\text{ztrkof}}$ under the GLM and GLMM models are remarkably similar to each other. For β_0, there is some difference, with the GLMM model's posterior distribution being shifted towards lower values. This can be understood by looking at Figure 6.8, which shows the distribution of the sample means of the random intercepts $u_i, i = 1, \ldots, 12000$. We see that the distribution is very skewed, with a strong peak around zero and slightly below, but with a number of u_i's that are significantly higher than zero. The mean of the means is just below zero, at $-4.5 \cdot 10^{-4}$, while the median is lower, at $-2.8 \cdot 10^{-3}$. Recalling from Table 1.1 that most policies in the data have zero claims while the rest have 1–5 claims during the entire duration of the data, the distribution of means of u_i is not too surprising. The policies with no claims have u_i's close to zero, while those with some claims create the smaller peaks visible in Figure 6.8. Since the most common scenario is for a policy to have zero claims, the β_0

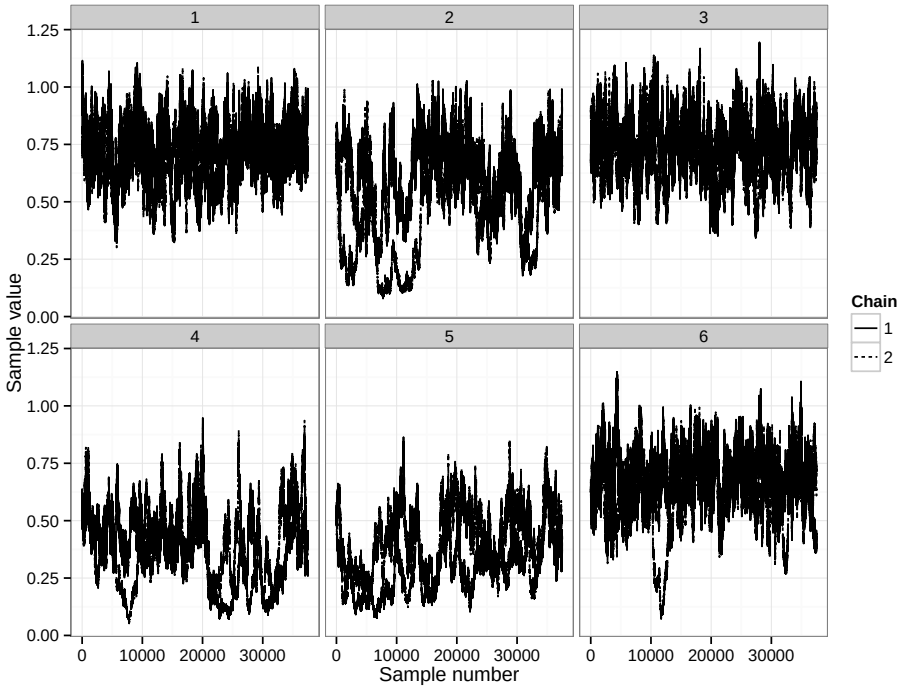

Figure 6.3: Chain sample plots for σ_u in the GLMM model, one panel per data subset.

is weighted toward this case, and so the u_i's have a skew distribution with a heavy right tail. This indicates that the model assumption that $u_i \sim \mathrm{N}(0, \sigma_u^2)$ could be improved upon by letting u_i follow some distribution that more closely resembles Figure 6.8.

6.3 Comparison metrics

After fitting the models and generating the estimated posterior predictive distributions as well as point predictions for all observations in the evaluation data sets, the model comparison methods discussed in Section 2.4, and the tariff comparison methods described in Chapter 5, were applied to the results, in order to determine if the GLMM model is an improvement over the reference GLM model. The Lorenz curves, Gini scores, and quotient test rely on point predictions of the claim frequencies for the observations in the evaluation data. As point prediction for an observation, we used the mean of the samples drawn from the posterior predictive distribution for this observation.

As mentioned in the end of Section 6.2, we do not have a single posterior distribution for the parameters $\boldsymbol{\beta}$, but actually one per data subset, six in

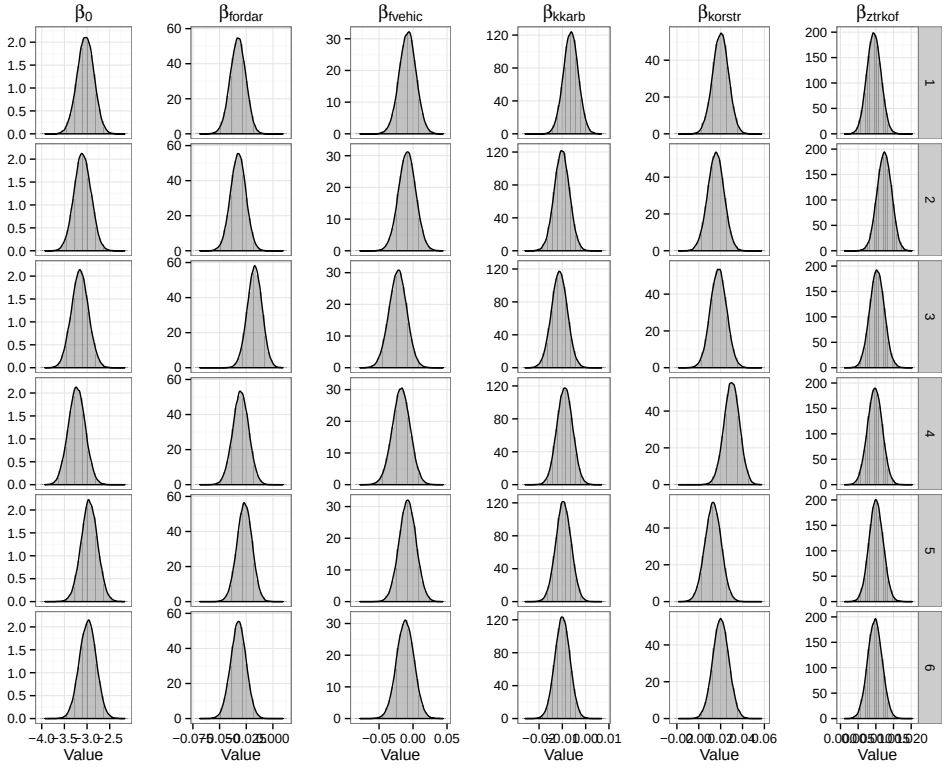

Figure 6.4: Plots of the estimated posterior distribution of $\boldsymbol{\beta}$ for the GLM model, one row per data subset. The plots show histograms with an overlaid Gaussian kernel density estimate.

total. For the purposes of comparing the GLM and GLMM models, however, we are interested mainly in the predictive distributions of future claims for the policies. For this, we do not need to combine the different $\boldsymbol{\beta}$ vectors, as the GLM and GLMM model for each data subset delivers predictions for the policies included in that subset. In particular, we obtain estimates for the posterior predictive distributions of the evaluation data subset directly from the MCMC simulations of the models for that subset. Since the posterior predictive samples of each observation in the evaluation data is generated by the models for that specific data subset only, we can compare the GLM and GLMM models over the entire dataset by simply pooling the predictions for each subset into a total evaluation dataset.

6.3.1 Prediction

Our model comparison metrics are based on predictions of the claim counts for the observations in the evaluation data subsets. Some of the metrics need point predictions, while other use full posterior predictive distributions.

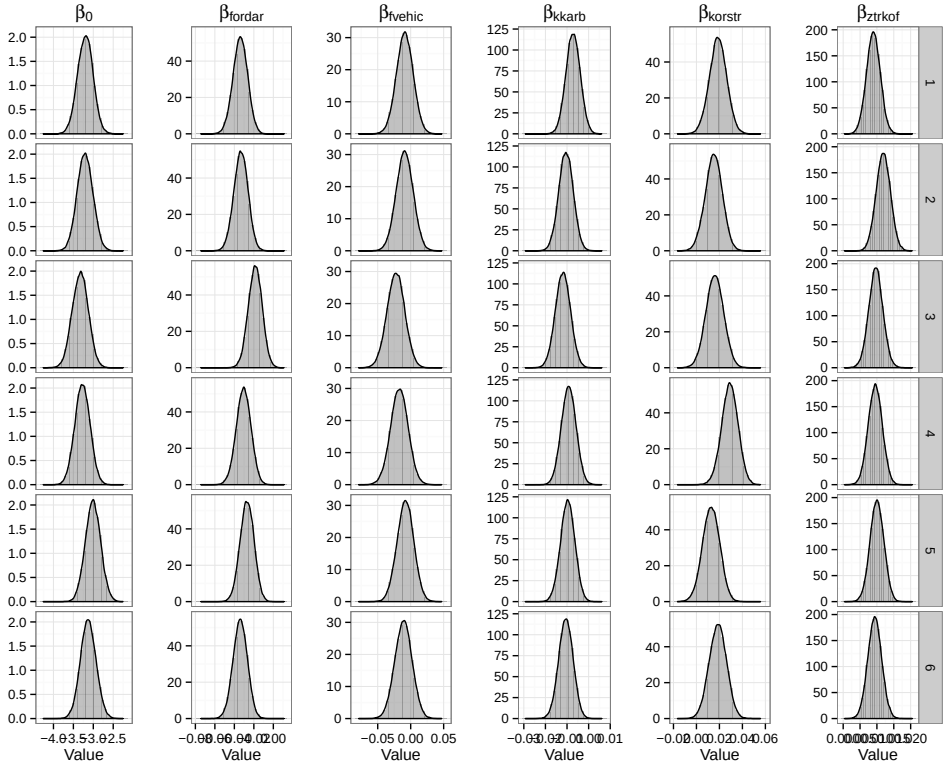

Figure 6.5: Plots of the estimated posterior distribution of $\boldsymbol{\beta}$ for the GLMM model, one row per data subset. The plots show histograms with an overlaid Gaussian kernel density estimate.

The output from the MCMC simulations, however, are not the posterior predictive distributions themselves, but rather samples drawn from them. We will now briefly describe how to go from this output to the point and distribution predictions we need for our model comparisons.

Let $l = 1, \ldots, L$ index the observations in one of the evaluation data subsets, and let $q = 1, \ldots, Q$ index the iterations of the simulation. We denote by $z^\star_{l,q}$ the sampled claim count of observation l for iteration q, i.e. the sample realisation of Z_l. Note that since the evaluation data subsets are out-of-sample data, the actual observed value z_l has not been used in the estimation of the model. Our point prediction for the value of Z_l is then given by the mean of $z^\star_{l,q}$:

$$\widehat{Z}_l = \frac{1}{Q} \sum_{q=1}^{Q} z^\star_{l,q} = \frac{1}{75000} \sum_{q=1}^{75000} z^\star_{l,q}. \tag{6.5}$$

An estimate of the full posterior predictive distribution of Z_l is given by the probability mass function of the samples, i.e. an estimator $\widehat{p}_{l,m}$ of the

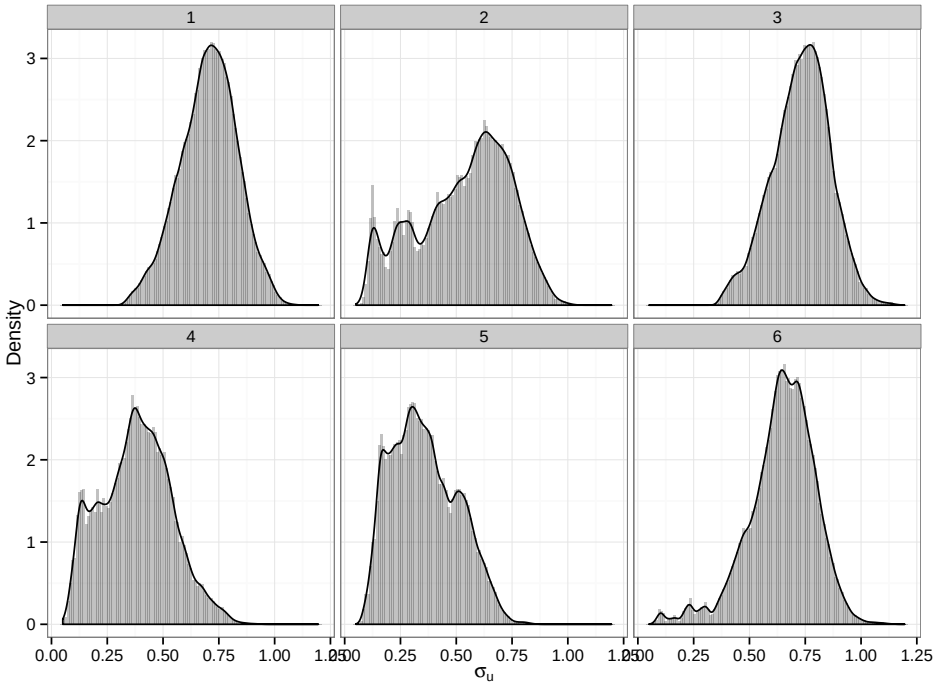

Figure 6.6: Plots of the estimated posterior distribution of σ_u for the GLMM model, one panel per data subset. The plots show histograms with an overlaid Gaussian kernel density estimate.

probability $\mathbb{P}[Z_l = m]$ for some $m = 0, 1, 2, \ldots$ is given by the proportion of the samples $z^\star_{l,q}$ that are equal to m:

$$\widehat{p}_{l,m} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{1}_{\{z^\star_{l,q} = m\}}, \quad m = 0, 1, 2, \ldots \tag{6.6}$$

It can be noted that the claim counts are technically defined on the support $\mathbb{N} = 0, 1, 2, \ldots$, but the reality in our case is that the sampled $z^\star_{l,q}$ are always in $\{0, 1, 2, 3, 4, 5\}$. This is due to an observation period being a maximum of one risk year long, and the claims being rare events, cf. Section 1.2. Therefore, the estimated posterior predictive distribution of observation Z_l can be represented by a vector of probabilities $(\widehat{p}_{l,0}, \widehat{p}_{l,1}, \ldots, \widehat{p}_{l,5})$. As an example, consider the GLMM estimated posterior predictive distribution of the first observation in evaluation data subset 1, seen in Table 6.2. The point prediction for this observation is 0.0396, while the observed value was 0.

Due to low numbers of claims in our data set, and thus also in the samples, we may encounter a situation where for some observation l with the outcome $z_l = m$, there were no samples with $z^\star_l = m$, which leads to the estimator $\widehat{p}_{l,m} = 0$. In other words, the probability distribution estimated

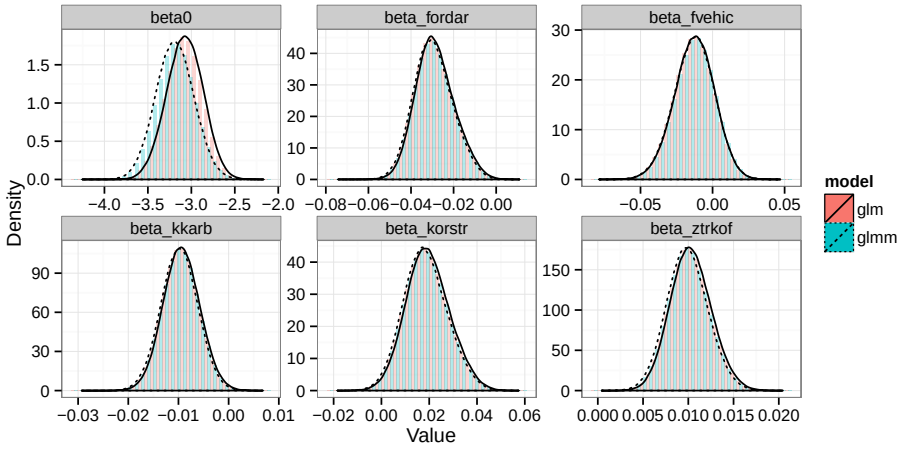

Figure 6.7: Plots of Gaussian kernel density estimates of the posterior distribution of $\boldsymbol{\beta}$ for the GLM and GLMM models, with samples from all six data subsets merged.

Claims	0	1	2	3	4	5
Samples	72 105	2 823	66	6	0	0
Probability	0.96140	0.03764	0.00088	0.00008	0.00000	0.00000

Table 6.2: Estimated posterior predictive distribution from GLMM model for observation 1 in evaluation data subset 1, shown for illustrative purposes.

claims that the outcome observed is impossible. In such a case, the likelihood of the data becomes 0 in the calculations of our partial Bayes factor, leading to likelihood ratios that are either zero or undefined. To avoid this situation, we will in such cases replace $\widehat{p}_{l,m} = 0$ with $\widehat{p}_{l,m} = 1/Q$, i.e. we will act as if there was a single sample where $z_l^\star = m$.

6.3.2 Partial Bayes factor and deviance information criterion

Section 2.4 describes some metrics by which Bayesian models may be compared. Calculation of the Bayes factor itself is difficult based on the MCMC simulations we use to estimate our models, but the partial Bayes factor, defined in (2.13), is easy to calculate. Letting $\mathbf{z}_1$ denote the training data set, and $\mathbf{z}_2$ the evaluation data set for a data group, the partial Bayes factor of the GLMM model as alternative to the GLM model is given by

$$\mathrm{BF}(\mathbf{z}_2|\mathbf{z}_1) = \frac{p(\mathbf{z}_2|\mathbf{z}_1, \mathrm{M}_{\mathrm{GLMM}})}{p(\mathbf{z}_2|\mathbf{z}_1, \mathrm{M}_{\mathrm{GLM}})}. \tag{6.7}$$

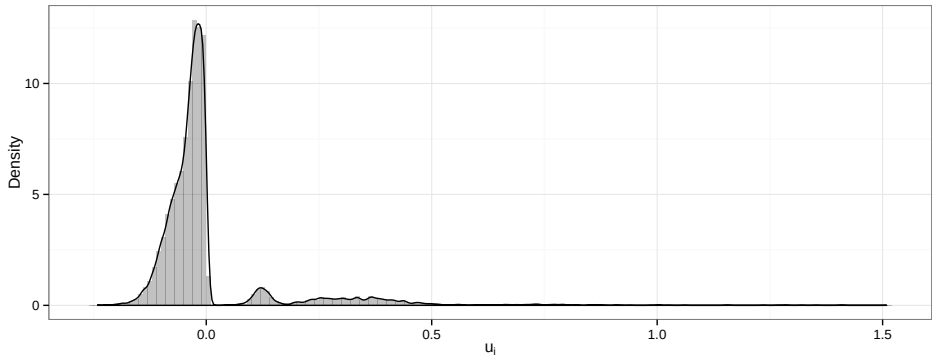

Figure 6.8: Histogram of the distribution of sample means of the random intercepts $u_i, i = 1, \ldots, 12000$. A Gaussian kernel density estimate is overlaid. The mean of the means is just below zero, at $-4.5 \cdot 10^{-4}$, while the median is lower, at $-2.8 \cdot 10^{-3}$.

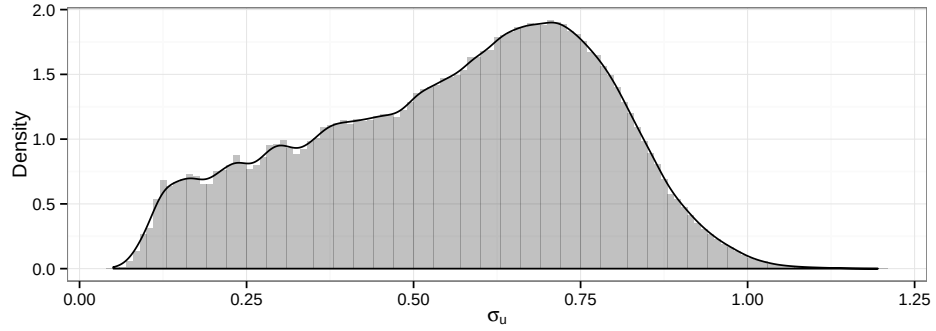

Figure 6.9: Histogram of the posterior distribution of σ_u, with samples from all six data subsets merged. A Gaussian kernel density estimate is overlaid.

This is simple to calculate, as the posterior likelihood $p(\mathbf{z}_2|\mathbf{z}_1, \mathrm{M}_{\mathrm{model}})$ is obtained from the posterior predictive distributions of the evaluation data, and the observed values of $\mathbf{z}_2$. Using the notation of Section 6.3.1, we can index the observations of our evaluation data $z_l, l = 1, 2, \ldots, L$. For each observation, we have an estimated posterior predictive distribution in the form of a vector of probabilities $(\widehat{p}_{l,0}, \widehat{p}_{l,1}, \ldots, \widehat{p}_{l,5})$, as described in Section 6.3.1. Given that the observed value of Z_l is z_l, the likelihood of this is then $\widehat{p}_{l,z_l}$. Since the observations $z_l, l = 1, \ldots, L$ are independent given the model, the estimated posterior likelihood of the evaluation data $\mathbf{z}_2$ conditional on the training data $\mathbf{z}_1$ is

$$\widehat{p}(\mathbf{z}_2|\mathbf{z}_1) = \prod_{l=1}^{L} \widehat{p}_{l,z_l}. \tag{6.8}$$

Calculating this for both the GLMM and GLM models for each data subset, the ratios of the results are then the partial Bayes factors of (6.7). Table 6.3 shows the partial Bayes factors for the six subsets of data. Recalling from Section 2.4 that a BF > 3.2 is considered substantial evidence in favour of an alternative model, we find here that subsets 2–5 display such evidence, while subset 6 falls slightly short of this. Subset 1 has a partial Bayes factor less than 1, indicating that the GLM model is better than the GLMM model there, although not by enough to meet the substantial evidence criterion. Subset 3 has a very high partial Bayes factor. There is no obvious reason for this, although it may be noted that group 3 will stand out in the other metrics as well as one where the GLMM model is particularly good compared to the GLM model.

Data subset	1	2	3	4	5	6
Partial BF	0.61	6.35	6 666.25	4.41	12.34	2.22

Table 6.3: Partial Bayes factors of GLMM model over GLM for each of the six evaluation data subsets.

The deviance information criterion, DIC, is also easily obtained as an output of a MCMC simulation by the method described in Section 2.4. Table 6.4 shows the value of $\text{DIC}_{\text{GLMM}} - \text{DIC}_{\text{GLM}}$ for each of the six data subsets, as well as a total value obtained by the sum of the individual subset differences. Unlike our other comparison metrics, which are all based on out-of-sample predictive performance on the evaluation data subsets, the DIC is a goodness-of-fit measure based on the in-sample training data. The DIC is lower for a better fitting model, so negative differences indicates that the GLMM model has a better fit to the data than the GLM model. Carlin and Louis (2009) suggest that the DIC difference between two models should be at least 3 to 5 to be of interest. Looking at Table 6.4, we find that the GLMM model has a lower DIC for all data subsets, but that subsets 4 and 5 have small enough DIC differences that it is not significant. All other subsets point to the GLMM model outperforming the GLM model, as does the total DIC difference over the entire training data.

Group	Overall	1	2	3	4	5	6
ΔDIC	-62.22	-18.46	-12.56	-15.91	-1.07	-2.44	-11.78

Table 6.4: Deviance information criterion differences between GLMM and GLM models, for the overall training data and each of the six subsets.

We find that both the partial Bayes factors and the deviance information criterion indicate that the GLMM model with random intercepts per policy is an improvement over the GLM model. We will now move on to compare the models using the insurance tariff comparison methods described in

6.3.3 Lorenz curves and Gini scores

As described in Section 5.1, we can use a form of Lorenz curves to estimate to what extent the GLMM model is capturing risk differentiation that the GLM model is missing. Recall that the Lorenz curve is based on ordering all policies by the ratio P_{alt}/P_{ref}, where P_{alt} is the alternative model's predicted claim frequency, i.e. the GLMM model prediction, and P_{ref} the reference model's predicted claim frequency, i.e. the GLM prediction. The Lorenz curve shows the accumulated fraction of claims as a function of the accumulated fraction of reference predicted claims, when the policies are ordered by the ratio. The interpretation of the plot is that the policies are ordered from those the alternative model would lower the prediction of the most, to those that would receive the highest increase. If the alternative model is an improvement on the reference model, then the Lorenz curve should fall below the dashed equality line, as more reference claim predictions than actual claims is accumulated in the beginning of the curve.

Data subset	Overall	1	2	3	4	5	6
Gini score	0.051	0.015	0.006	0.113	0.060	0.082	0.064

Table 6.5: Gini scores for the overall evaluation data, and each of the six subsets.

The Gini score summarises the information in the Lorenz curve by taking two times the area between the equality line and the Lorenz curve, as described in Section 5.1. If the curve is mostly below the line, then the alternative model is an improvement on the reference, and the Gini score will be positive. If instead the curve is mostly above the line, then the reference model is better, and the Gini score will be negative. Figure 6.10 shows the Lorenz diagrams for the six data subsets. For subsets 1 and 2, the curve indicates only a slight improvement compared to the reference model, while subsets 3–6 exhibit more clear evidence of the alternative model being an improvement. The corresponding Gini scores are shown in Table 6.5. They also indicate that there is overall an improvement in the risk identification of the GLMM model over the GLM model in all six subsets of the data, but that it is not large in subsets 1 and 2. A Gini score for the entire evaluation data has also been calculated by simply merging all the predictions, and it too shows an improvement. We may also note that subset 3 again stands out, with a particularly high Gini score.

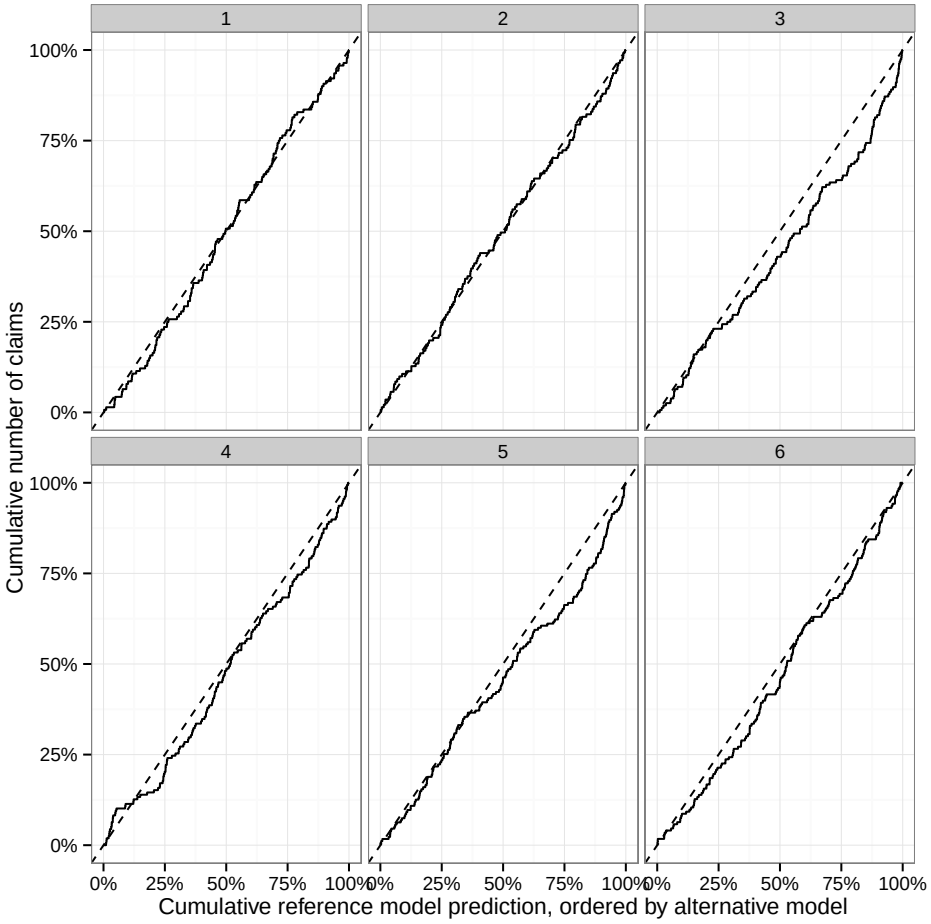

Figure 6.10: Lorenz diagrams of GLMM model vs GLM model, one panel per evaluation data subset.

6.3.4 Quotient test

The quotient test, described in Section 5.2, divides the policies into two groups: those that receive a lower claim frequency prediction under the alternative model, and those that receive a higher prediction. The former group is thus those that the alternative model has identified as being "overpriced" by the reference model, while the latter group is "underpriced". For each group, the ratio of observed claim counts divided by predicted claim counts is calculated for both alternative and reference models. If the alternative model is an improvement over the reference model, then its predictions for each group should be closer to the observed outcome, and the ratios should be closer to 1.

Table 6.6 shows the quotient test performed on each data subset sepa-

rately, as well as on the entire evaluation data together. For each subset, a winning model is designated, being the one with predictions closer to observed outcome. We see that subsets 2–6, the alternative GLMM model has predictions closer to observed outcome than the reference GLM model. In subset 1, it is instead the GLM model that performs better. It can be noted that the alternative model correctly identifies the groups as over- and underpriced, but then sometimes overcompensates in the other direction, underpricing those that were formerly overpriced, and vice versa. We may also note that subset 3 again stands out with the GLMM model being particularly good compared to the GLM.

Data subset	Change in prediction from reference to alternative	Ratio observed/predicted Alternative	Reference	Winner
Overall	decrease	1.01	0.94	alt
	increase	0.97	1.17	alt
1	decrease	1.16	1.04	ref
	increase	0.61	0.84	ref
2	decrease	1.03	0.96	alt
	increase	0.94	1.13	alt
3	decrease	0.95	0.86	alt
	increase	1.12	1.53	alt
4	decrease	1.01	0.97	alt
	increase	0.98	1.06	alt
5	decrease	0.93	0.89	alt
	increase	1.14	1.23	alt
6	decrease	1.02	0.93	alt
	increase	0.96	1.24	alt

Table 6.6: Quotient test values for overall evaluation data and each subset.

6.3.5 Proper scoring rules

The quadratic, logarithmic, ranked probability, and spherical scoring rules, described in Section 5.3, were applied to the data by calculating the score for the GLM and GLMM models' posterior predictive distributions of each observation in an evaluation data subset, and then summing these scores per model. A total score for the entire evaluation data set is obtained by simply summing the scores for the six evaluation data subsets.

Using the notation of Section 6.3.1, we let $l = 1, \ldots, L$ index the observations in the evaluation data set, and let z_l be the observed number of claims for observation l. We denote by $\widehat{p}_{l,m,\mathrm{GLMM}}$ the estimated posterior

predictive probability that $Z_l = m$, as predicted by the GLMM model, and similarly with $\widehat{p}_{l,m,\text{GLM}}$. The estimated posterior predictive distribution of observation l from a given model is then represented by the vector of probabilities $\widehat{\mathbf{p}}_{l,\text{model}} = (\widehat{p}_{l,0,\text{model}}, \ldots, \widehat{p}_{l,5,\text{model}})$. The score for each model and scoring rule on an evaluation data subset is then calculated as

$$S_{\text{model,subset}} = \sum_{l=1}^{L} S(\widehat{\mathbf{p}}_{l,\text{model}}, z_l). \tag{6.9}$$

By then taking $S_{\text{GLMM,subset}} - S_{\text{GLM,subset}}$ for each scoring rule and evaluation data subset, we obtain a difference where a positive value means that the GLMM model received a higher score, and vice versa. We are using definitions of the scores for which a higher score represents a better prediction, so positive differences indicate that the GLMM model outperforms the GLM model. The score differences were calculated for each evaluation data subset separately, and also added together for an overall score. The results are presented in Table 6.7.

The table shows that the scores were unanimous in preference for the GLMM model for subsets 3–5. For subset 1, the logarithmic and quadratic scores indicates that the GLM model performs better, while the quadratic and ranked probability show preference to the GLMM. For subsets 2 and 6, only the logarithmic score prefers the GLMM model. When summed over the entire evaluation data set, the scores all agree that the GLMM model is the better performer. We note that the differences in score between the GLM and GLMM models were largest for every score in subset 3, which also stood out in the partial Bayes factor, Lorenz curves, and Gini coefficients as the subset where the GLMM model was especially good in comparison to the GLM model.

Group	Quadratic	Logarithmic	Ranked probability	Spherical
Overall	0.56	14.96	0.38	0.08
1	0.04	-0.49	0.12	-0.01
2	-0.11	1.85	-0.08	-0.04
3	0.65	8.80	0.32	0.17
4	0.12	1.48	0.06	0.02
5	0.37	2.51	0.23	0.09
6	-0.49	0.80	-0.27	-0.16

Table 6.7: Differences in scores between GLMM and GLM models.

6.3.6 Summary of comparisons

We have applied a variety of comparison metrics to our GLMM and GLM models: partial Bayes factors, DIC, Lorenz curves, Gini scores, quotient test

and proper scoring rules, and conclude that overall, the metrics agree that the GLMM model performs better than the GLM model at predicting the claim frequencies of our TPL motor insurance data.

Chapter 7

Conclusion

The aim of this thesis was to study a certain form of credibility model for non-life insurance tariffs, namely a generalised linear mixed model with a random intercept per policy. The specific setting chosen for evaluating the performance of this model was the prediction of claim frequencies in third party liability motor insurance, where an analysis data set was provided by the Swedish insurance company Trygg Hansa. A secondary aim was to use Bayesian inference, as it allows for use of distributional predictions instead of only point predictions. This allowed the use of proper scoring rules, which are a powerful tool for evaluating and comparing predictive performance.

The purpose of an insurance tariff is to predict as accurately as possible the loss due to claims that an insurance policy will generate. Certainly an understanding of the underlying factors that drive the risk, and the methods by which they do so, is helpful to the actuarial practitioner, but it is not the purpose of a tariff. An example is that vehicles with a higher engine power have increased risk of traffic accidents. This is almost certainly because engine power is correlated with something that does in fact cause accidents—a hypothesis would be that more aggressive drivers are more likely to purchase high-powered vehicles. To understand traffic risk, one would then be better served attempting to measure aggressiveness, as it is clear that engines do not on their own cause accidents, even if they have the horsepower of an entire cavalry regiment. However, to predict insurance claim costs, engine power is a very helpful rating factor, as the data is quantifiable, easily available to use in a tariff, and, most importantly, has predictive power. Thus, the focus of the model evaluation in this thesis is on predictive performance, and not on inference on model parameters themselves.

A standard Poisson generalised linear model was chosen as a reference model, and both this and the generalised linear mixed model were estimated under the Bayesian paradigm, using Markov Chain Monte Carlo methods. The results of this data analysis, described in Chapter 6, show that for the third party liability motor insurance data set from Trygg Hansa, the GLMM

model with a random intercept per policy has better predictive performance than a GLM model with the same fixed effects, but without the policy-level random intercepts. The models were compared using an ensemble of metrics, mostly based on point predictions, but the Bayesian estimation of the models also allowed the use of proper scoring rules (Section 5.3), which compare models based on full predictive distributions.

The fact that the GLMM based credibility model outperformed the reference makes it of interest to apply it in a full scale context. Unfortunately, implementation of such models in practice may prove difficult for several reasons. We found that the JAGS (Plummer, 2003) software used for MCMC inference was unable to deal with the entire analysis data set in one go, forcing us to split the data into six subsets and estimate models for each subset. Since the analysis data set is of very modest size (44 186 risk years) compared to those common at insurance companies, this means that alternative means of estimation need to be considered in order to use such models in practice. An attempt was made to estimate the models using the more computationally efficient INLA method described in Section 2.2, but due to time constraints and lack of documentation in the software, the attempt was not further pursued. The promise of INLA to speed up estimation is very attractive, and should be investigated further. It must be noted that these are difficulties of implementation, and are not inherent to the model itself.

The advantage of the GLMM model is that it improves the predicted claim frequency for a given policy when more data is available for that specific policy. This means that such a tariff model is useful when insurance policies are renewed, as the premium can then be determined for each policy based on the most up-to-date information available for that policy, ensuring that each policy receives a fair price.

The Trygg Hansa TPL insurance data set used contains data over a time period of 4.5 years, and due to the low frequency of TPL claims, most policies encounter few claims during this period. Despite this, the GLMM model is able to use the policy-level random intercepts to improve the predictive distributions. Further studies might consider applying the GLMM model in settings with higher claim frequencies, such as motor hull insurance, or attempt to span over longer time frames. One way to accomplish the latter is to include a random intercept on the customer level rather than on the policy level. A given motor insurance policy might not stay in force for many years, but it may well be that the company has insured the same customer multiple times previously, giving a longer time frame to estimate the customer's risk profile. Bayesian hierarchical models can be used in such a setting. Another interesting topic for future study is to consider ways to create customer risk estimates by bringing together information from multiple insurance branches, e.g. using both motor and home insurance information to determine the risk of a motor insurance policy.

Summing up our results, we have found that based on an ensemble of

metrics, a Poisson generalised linear mixed model with policy-level random intercepts improves claim frequency prediction in third party liability motor insurance, compared to a reference Poisson generalised linear model. This predictive improvement was demonstrated on a representative actual data set obtained from Trygg Hansa. The GLMM model studied in this thesis is able to perform in the claims-sparse setting of private TPL insurance, thus offering a promising way of using credibility models in private lines motor insurance.

Appendix A

JAGS source code

The JAGS code used for estimation of the GLM and GLMM models is shown here. The variable `antsk` is the number of claims for an observation, while the random intercepts, called u_i in the thesis, are referred to in the code as `u`. Note that the indexing is slightly different; in the JAGS code we use `k` to index observations in the training data set, and `l` to index observations in the evaluation data set. `i` indexes the policies. In the thesis in general, we use i for policies, and j for repeated observations of a given policy and index a given observation using the combination of i and j.

Apart from the JAGS code, approximately 1 500 lines of supporting R code were used to perform data manipulation, run the JAGS simulations, and generate figures for the report.

Listing A.1: The JAGS code for the GLM model.

```
# GLM model
model
{
  for (k in 1:n_obs) {
    mu[k] <- riskar[k]*exp(beta0 + beta_fordar*fordar[k]
       + beta_fvehic*fvehic[k] + beta_kkarb*kkarb[k] +
       beta_korstr*korstr[k] + beta_ztrkof*ztrkof[k])
    antsk[k] ~ dpois(mu[k])
  }

  # specify distribution for fixed effects
  beta0       ~ dnorm(-2.978, 1/(0.327^2))
  beta_fordar ~ dnorm(-0.032, 1/(0.011^2))
  beta_fvehic ~ dnorm(-0.011, 1/(0.019^2))
  beta_kkarb  ~ dnorm(-0.010, 1/(0.005^2))
  beta_korstr ~ dnorm( 0.020, 1/(0.011^2))
  beta_ztrkof ~ dnorm( 0.010, 1/(0.003^2))

  # predict test data
  for (l in 1:n_test_obs) {
```

```
    mu_t[l] <- riskar_t[l]*exp(beta0 +
       beta_fordar*fordar_t[l] +
       beta_fvehic*fvehic_t[l] + beta_kkarb*kkarb_t[l]
       + beta_korstr*korstr_t[l] +
       beta_ztrkof*ztrkof_t[l])
    antsk_t[l] ~ dpois(mu_t[l])
  }
}
```

Listing A.2: The JAGS code for the GLMM model.

```
# GLMM model
model
{
  for (k in 1:n_obs) {
    mu[k] <- riskar[k]*exp(beta0 + beta_fordar*fordar[k]
       + beta_fvehic*fvehic[k] + beta_kkarb*kkarb[k] +
       beta_korstr*korstr[k] + beta_ztrkof*ztrkof[k] +
       u[kundnr_mapped[k]])
    antsk[k] ~ dpois(mu[k])
  }

  # specify distribution for fixed effects
  beta0       ~ dnorm(-2.978, 1/(0.327^2))
  beta_fordar ~ dnorm(-0.032, 1/(0.011^2))
  beta_fvehic ~ dnorm(-0.011, 1/(0.019^2))
  beta_kkarb  ~ dnorm(-0.010, 1/(0.005^2))
  beta_korstr ~ dnorm( 0.020, 1/(0.011^2))
  beta_ztrkof ~ dnorm( 0.010, 1/(0.003^2))

  # specify distribution for random effects
  for (i in 1:n_policies) {
    u[i] ~ dnorm(0, taub)
  }
  taub ~ dgamma(0.01, 0.01)

  # predict test data
  for (l in 1:n_test_obs) {
    mu_t[l] <- riskar_t[l]*exp(beta0 +
       beta_fordar*fordar_t[l] +
       beta_fvehic*fvehic_t[l] + beta_kkarb*kkarb_t[l]
       + beta_korstr*korstr_t[l] +
       beta_ztrkof*ztrkof_t[l] + u[kundnr_mapped_t[l]])
    antsk_t[l] ~ dpois(mu_t[l])
  }
}
```

Bibliography

K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40:58–76, 2007.

H. Bühlmann and A. Gisler. *A course in credibility theory and its applications*. Springer, 2005.

H. Bühlmann and E. Straub. Glaubwürdichkeit für Schadensätze. *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker*, 1970.

M. J. Brockman and T. S. Wright. Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries*, 119, 1992.

B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall, 3rd edition, 2009.

C. Czado, T. Gneiting, and L. Held. Predictive model assessment for count data. *Biometrics*, 65:1254–1261, 2009.

M. Denuit, X. Maréchal, S. Pitrebois, and J-F. Walhin. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Wiley, 2007.

E. W. Frees, G. Meyers, and A. D. Cummings. Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association*, 106 (495), September 2011.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of The American Statistical Association*, 102(477), March 2007.

B. Jørgensen. *The theory of dispersion models*. Chapman & Hall, 1st edition, 1997.

B. Jørgensen and M. C. P. De Souza. Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1:69–93, 1994.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

M. Lavine and M. J. Schervish. Bayes factors: what they are and what they are not. *The American Statistician*, 53(2):119–122, 1999.

C. E. McCulloch and S. R. Searle. *Generalized, linear and mixed models*. Wiley, 2001.

K. P. Murphy, M. J. Brockman, and P. K. W. Lee. Using generalized linear models to build dynamic pricing systems for personal lines insurance. In *Casualty Actuarial Society Winter 2000 Forum*, 2000.

E. Ohlsson and B. Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, 2010.

M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, version 3.4.0, 2003.

M. Plummer. *rjags: Bayesian graphical models using MCMC*, 2014. URL `http://CRAN.R-project.org/package=rjags`. R package version 3-13.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

RStudio Team. *RStudio: Integrated Development Environment for R, version 0.98.1091*. RStudio, Inc., Boston, MA, 2012. URL `http://www.rstudio.com/`.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, series B*, 71(2):319–392, 2009.

D. J. Spiegelhalter, N. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, series B*, 64(4):583–639, 2002.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.

H. Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011. URL `http://www.jstatsoft.org/v40/i01/`.