Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Ole-Magnus Høiback

Estimation in Credibility Models Using Bayesian Hierarchical Models

Master's thesis in Applied Physics and Mathematics
Supervisor: Håkon Tjelmeland
Co-supervisor: Anne Randi Syversveen
June 2022

NTNU
Norwegian University of Science and Technology

Ole-Magnus Høiback

Estimation in Credibility Models Using Bayesian Hierarchical Models

Master's thesis in Applied Physics and Mathematics
Supervisor: Håkon Tjelmeland
Co-supervisor: Anne Randi Syversveen
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

Credibility theory lays the foundations and provides the original model of how the insurance industry predicts expected loss and determines price levels (Bühlmann & Gisler, 2005). In the context of modern statistics, credibility theory is used to estimate random effects in a generalized linear mixed model. This thesis proposes two alternative models, serving as extensions to the original model. The proposed models extend the usual framework of a generalized linear mixed model, giving Bayesian hierarchical models. The first proposed model extends the distributional assumptions on the random effects by assuming that they are normally distributed. This extends the original model, which is restricted to assumptions only about the first and second moment of the random effect. The second proposed model keeps the distributional assumptions made in the previous model and extends this by adding assumptions on the fixed effects of the general linear mixed model. As with the random effects, we assume that the fixed effects are normally distributed. The last model also assumes a prior distribution on the variance parameters of the fixed and random effects.

Credibility theory is insufficient to estimate the parameters of the proposed models. For the first model, we implement a variation of the Monte Carlo expectation-maximization algorithm, as proposed in Levine and Casella (2001). For the last model, we employ Markov Chain Monte Carlo. Ranking the results from the three models, i.e. the original model and the two proposed extensions, we find that a Bayesian hierarchical model with prior distributions on both the fixed and random effects arguably outperforms the original model. Given the complexity of the Markov Chain Monte Carlo algorithm, we get a computer-intensive method, and compared to the original model and its estimation method, convergence is slow. Finally, we discuss alternatives for estimation and implementation that can yield a method outperforming the original model.

Sammendrag

Kredibilitetsteori gir både fundamentet og den originale modellen for hvordan forsikringsbransjen predikerer forventet tap og prisnivåer (Bühlmann & Gisler, 2005). I moderne statistikk ser vi at kredibilitetsteori gir en metode for estimering av tilfeldige effekter i generaliserte miksede lineære modeller. I denne oppgaven foreslår vi to alternative modeller som begge utvider den originale modellen. De foreslåtte modellene utvider også strukturen til en generalisert mikset lineær modell og betraktes som Bayesianske hierarkiske modeller. Den første modellen vi foreslår antar en normalfordeling på de tilfeldige effektene. Dette går lenger enn den originale modellen som kun gjør antakelser på det første og andre momentet til de tilfeldige effektene. Den andre foreslåtte modellen tar med seg antakelsene til den første modellen og bygger igjen videre på denne ved å anta at de faste effektene også har en sannsynlighetsfordeling. I likhet med de tilfeldige effektene, antar vi at de faste effektene er normalfordelte. I tillegg antar også denne modellen en apriorifordeling på variansen til de tilfeldige og faste effektene.

Kredibilitetsteori alene er ikke nok til å estimere parametere i de foreslåtte modellene. For den første foreslåtte modellen implementerer en variant av Monte Carlo expectation-maximization-algoritmen som er beskrevet i Levine and Casella (2001). For den andre modellen foreslås en implementering av Markov Chain Monte Carlo-algoritmen. Ved rangering av resultater fra den originale modellen, den første og den andre foreslåtte modellen, viser vi at den andre modellen kan slå den originale. Med andre ord, vi ser at en Bayesiansk hierarkisk modell med fordelinger på både de faste og tilfeldige effektene kan gjøre det bedre enn den originale modellen. Markov Chain Monte Carlo-algoritmen gir en beregningstung metode, og sammenlignet med den originale modellen, og dens metode, opplever vi treg konvergens. Avslutningsvis ser vi på alternativer for estimering og implementering som kan gi en metode som slår den originale modellen på resultater og samtidig tar hensyn til beregningstid.

Preface

This master's thesis has been written during the tenth and final semester of a master's degree in industrial mathematics. The thesis counts for 30 SP and is formally the result of the work done in the course TMA4900 - Industrial Mathematics. For the student, the overall goal of the master's thesis is demonstration of in-depth knowledge and understanding of a specific topic and independent production of scientific work. Along with this thesis, the job done over the semester was presented before advisors and master's students in industrial mathematics.

I want to thank my advisor Håkon Tjelmeland for valuable feedback and fruitful discussions throughout the year. I would also like to thank If Insurance and Anne Randi Syversveen for their interest in the project and for making data available to me. As this thesis also concludes my time as a student, I would like to thank my family and friends for their support and inspiration throughout the five years. Finally, I would like to thank my girlfriend Ada for bringing joy, and love to the life of a sometimes stressed and afraid master's student.

Ole-Magnus Høiback,
Trondheim, June 2022

Contents

Chapter 1

Introduction

Determining risk is of high importance when selling insurance. Specifically, one wants to predict what a customer will claim in the future accurately. Is it likely that customer A will make more claims than customer B? If yes, why is that the case? Statistical models allow us to model how often a customer will make a claim and how extensive this claim will be. Some statistical models even will enable us to interpret how they predict different outcomes for different customers.

For many years, generalized linear models and credibility theory have been a cornerstone of insurance pricing. Generalized linear models provide a broad class of statistical models with nice mathematical and statistical properties. On the other hand, credibility theory has grown out of the insurance industry itself and has provided ways of incorporating both individual and collective risk in an insurance portfolio. Using credibility theory, statisticians have distinguished and given value to risks with different exposure or *credibility*.

It is primarily Hans Bühlmann who has got the credit for developing credibility theory in a mathematical setting (Hickman & Heacox, 1999). In Bühlmann and Gisler (2005), the rationale for credibility theory in insurance is thoroughly introduced, along with proposals for real-life applications. At the heart of it is the belief that every risk in the collective has an associated risk profile, where the risk profiles are independent and identically distributed random variables. A typical real-life example could be automobile insurance, where the brand of the car represents the risk profile. So *Saab*, *Volvo*, *Peugot* etc. have different risks associated with them, but the risks are assumed to have been drawn independently from the same probability distribution.

Furthermore, we assume that the number of claims a customer will make is a random variable. The same goes for the size of the claim. Returning to our example, the size of the claim is a random variable, drawn from a distribution which depends on attributes about the driver and the car, including the brand of the car. This constitutes a hierarchical model. The main goal is estimation of the size of

the claim, but getting there we need an estimation of the risk profile as well.

One way of doing this is through the use of empirical Bayes, which is the main method employed by Bühlmann and Gisler (2005). This is not a truly Bayesian analysis, but uses the Bayesian formulation of the problem. Instead of assigning priors and computing posteriors, we estimate the mean of the priors with the available data. Another way of solving the problem is through a pure Bayesian model, or Bayesian hierarchical model (Klugman, 1987). This will be the focus of this thesis. From a pure Bayesian formulation of the model, we estimate parameters using various techniques and compare our results to the traditional empirical Bayes that is so popular today.

Where a pure Bayesian model results in a posterior distribution of one or several parameters, empirical Bayes provides only point estimates of these. Hence, we expect the former to tell us more and provide better estimates. For this reason, we consider it to be an attractive alternative to empirical Bayes. There is no novelty using purely Bayesian models for this purpose, but it becomes increasingly more available as computers grow more powerful. Coupled with specific distributions, we use efficient algorithms to estimate our parameters.

Chapter 2

Data analysis

In this chapter, we explore our data set and look at the covariates we will be using when building our model. Initially, we will spend some time introducing relevant notation and insurance-specific formulations. Where the introduction in Chapter 1 should provide motivation and interest for the problem, an introduction to standard terms and formulations specific to the insurance industry will provide broader insight and understanding of our problem.

2.1 A brief introduction to insurance

An insurance company usually sells a wide range of *insurance policies*. An *insurance policy* is a legal contract between the insurance company and *the insured*, e.g. a person or another company. Talking about an *individual risk*, we refer to the risk tied to one individual policy, i.e. the expected cost that one policy will have for the company. Suppose the insured experiences damage and the policy covers the specific damage. In that case, they can make a *claim* to the insurance company, that is, a request for compensation for their losses. Depending on the fulfillment of requirements stated in the policy, the insured can get their claim validated, which results in a payment to cover their losses. The size of the claim the insured will get is referred to as *claim severity*.

When we talk about the *duration* of a policy, we mean the number of days the policy has been active divided by 365. Therefore, duration is a number between 0 and 1, giving a "weight" to the policy. Making multiple claims on the same policy is also possible, giving us the *number of claims*. We can find the *claim frequency* by dividing the number of claims by the duration. This means that a policy that has been in place between the customer and company for several years, will have one observation per year in effect.

The insured will have to pay for the insurance, a payment that is referred to as the *premium*. The premium is supposed to cover claims made by the insured and other insured people or companies. It is also meant to cover administrative fees

and more. However, we are interested in the *pure premium*, which is the payment necessary to break even with the insurance company's loss on all its claims. Looking at all the insured customers within a group, we can calculate the *average pure premium*, i.e. the average cost of risk for the insurance company. It is the determination of this figure which is of highest statistical interest within insurance. How much loss will we have to cover in the future?

For an individual policy, the pure premium is simply the total claim severity the customer has had. It can also be seen as the number of claims multiplied by the average claim severity. Typically, however, we will model the pure premium as

$$\text{Pure premium} = \text{Claim frequency} \times \text{Average claim severity}.$$

The collection of policies on the same product, e.g. car insurance or life insurance, will be referred to as a *portfolio*. It compromises many different customers, but they are all common in what they insure, e.g. a car or a life.

In the next section, we will look at the specific data set available to us. We can already say something about the notation that we will be using. We aim to model a *response variable*, the variable we aim to predict and understand. This could, for example, be the claim frequency or average claim severity. Next, we have *explanatory variables*, that is, variables used to explain the response variables. The terms should be familiar for all readers who have seen regression models at some point. We use Y to denote the response variable and X to denote the explanatory variables in the form of a design matrix. We also use y_{ijt} and x_i when referring to unique data points. We will return to the notation in Chapter 3.

2.2 The data set

In the following chapters, we go on and build a statistical model. The model's parameters are fitted using real data provided by If. The data consists of over 700000 observations, where each observation is connected to one individual insurance policy. The data set contains insurances on valuables, i.e. usually expensive assets such as the Norwegian national costume "bunad", art, etc. It includes what type of valuable an observation is related to and information about the owner, such as their age and area of residence.

The data was collected between 2012 and 2020, as shown in Figure 2.1. The x-axis represents the year of insurance, while the left side y-axis shows the total duration in each year. The right side y-axis measures the average pure premium in NOK. By request from If, all of the average pure premiums plotted in this chapter have been multiplied with a common arbitrary constant to mask the true value. There has been a downward trend in the total duration, i.e. the number of observations per year, as can be seen by the vertical bars. The scatter plot in the same

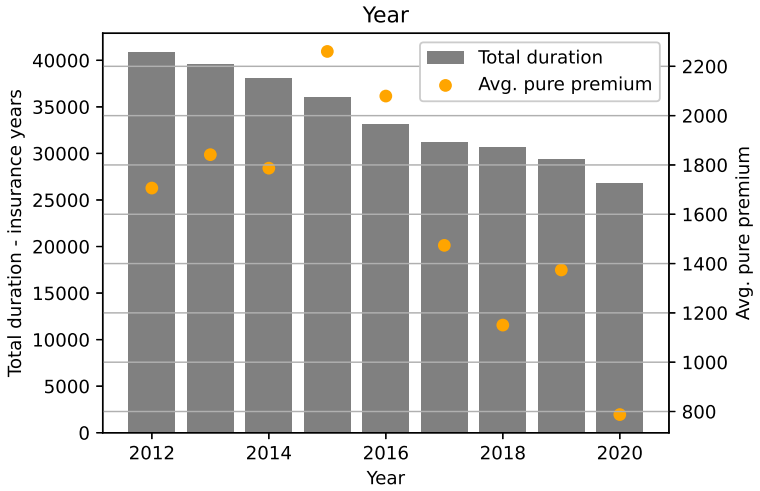


Figure 2.1: The total duration and average pure premium of policies from 2012 to 2020. The x-axis gives the year of insurance. The vertical bars indicate the total duration, measured by the left side y-axis. The scatter plot indicates the average pure premium in NOK, measured by the right side y-axis.

figure tells us how the average pure premium has varied, where the pure premium is the average cost per policy year per insurance. There is no clear trend, but it is reasonable to assume that the outlier that 2020 represents is connected to the COVID-19 pandemic.

2.3 Variable selection

Variable selection, often referred to as feature selection, is essential for statistical analysis. Identifying and keeping the variables best suited for analysis is an art of its own, and many methods and algorithms are available. In this thesis, however, we are interested in how the parameters are estimated, more than how the choice of parameters fits the underlying patterns in the data. Should we conclude that our method is competitive with the existing models, it seems likely that an even better result can be reached by focusing more on variable selection.

We identify five key attributes from the original data that we assume to be important for making predictions about future claims on the object. We want to know what type of *object* the valuable is, i.e. art, hearing-aid, wedding rings, etc. We also want to know the *sum* they have insured for, i.e. the potential amount the customer may claim. We also want to know the *age* of the customer, along with the *region* of Norway they live in. Lastly, the customers has an associated *rating* which, when negative tells us that the customer reports less than expected, and when positive the customer make more claims than expected.

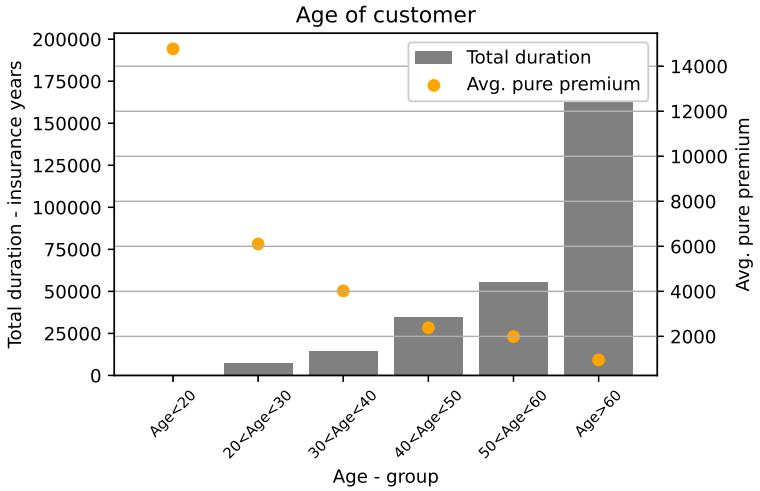


Figure 2.2: The total duration and average pure premium distributed across six age groups. The x-axis gives the groups of age. The vertical bars show that we have an increasing amount of observations in the highest age groups.

Figure 2.2 shows how we have grouped the ages in different intervals. What is also clear is that most of the insurance belongs to older people. The number of things one wants to insure is arguably a number that grows throughout life, and therefore the positive trend in duration for each increasing interval seems reasonable. It should also be noted that there is room for many more people in the interval containing those of age 60 or above. Just as interesting is that the pure premium decreases for each increasing interval of age. From this plot only, it appears that older customers have a lower cost than the younger ones. It should also be pointed out that the exposure, i.e. the total duration, on the youngest intervals is low, making possible outliers very significant.

The following figure, Figure 2.3, shows how we have distributed the observations according to the insured sum. The group containing observations with a *low* level has an insured sum up to 1000 NOK. Subsequently, *medium-low* is up to 2000 NOK, *medium-high* is up to 4000 NOK, and *high* is anything above 4000 NOK. It is no apparent trend between the groups.

Figure 2.4 tells us how customers are distributed, according to the previous experience If has with the customer. Most customers find themselves at level 0 or close to it, while some have largely subceeded or exceeded their expected number of claims. From a rating of -2 and upwards, there is a clear positive trend in the pure premium, which resonates with the natural conclusion that customers making more claims than expected should also be costlier. The pure premium for ratings between -6 and -3 is more confusing. Given the low exposure, it could be

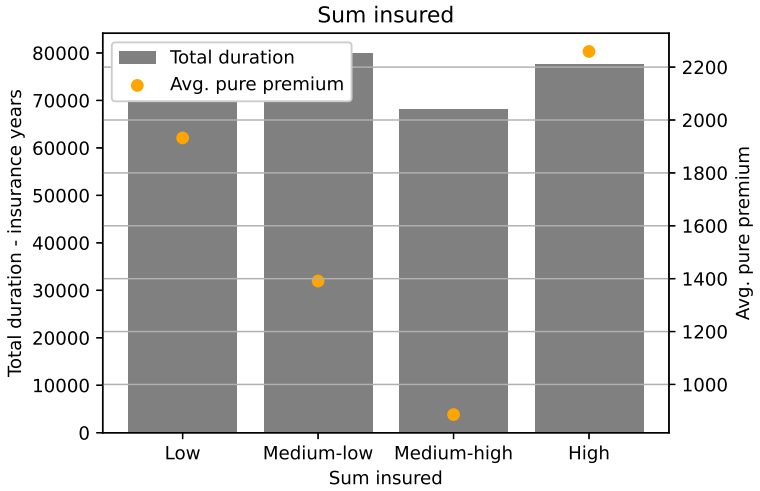


Figure 2.3: The total duration and average pure premium distributed among four levels of the insured sum. The x-axis gives the sum insured by customers. The groups have been divided into "Low" (sums up to 1000 NOK), "Medium-low" (sums from 1000 to 2000 NOK), "Medium-high" (sums from 2000 to 4000 NOK), and "High" (sums over 4000 NOK).

argued that the pure premium does not have to be representative of the expected pure premium. However, it is peculiar that all four levels are elevated.

The region of customers is represented in Figure 2.5. The smallest exposure can be found in *Nordland* (NORDL) and *Troms/Finnmark* (TR/FI), which corresponds well with how the population is distributed in Norway. The highest pure premium can be found in *Oslo* and the lowest in *Østlandet* (OST). Finally, Figure 2.6 shows various exposure and pure premium among 15 different types of object.

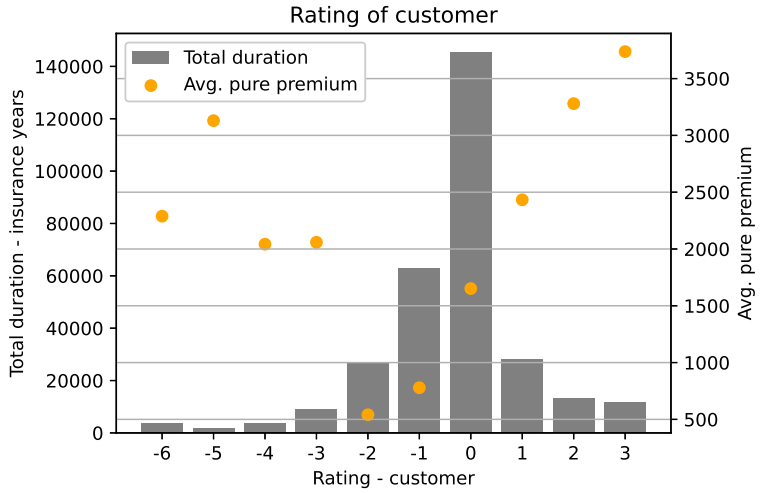


Figure 2.4: The total duration and average pure premium varies over the customer rating. The rating of customers is given by the x-axis. A low number indicate less reported claims than expected and vice versa.

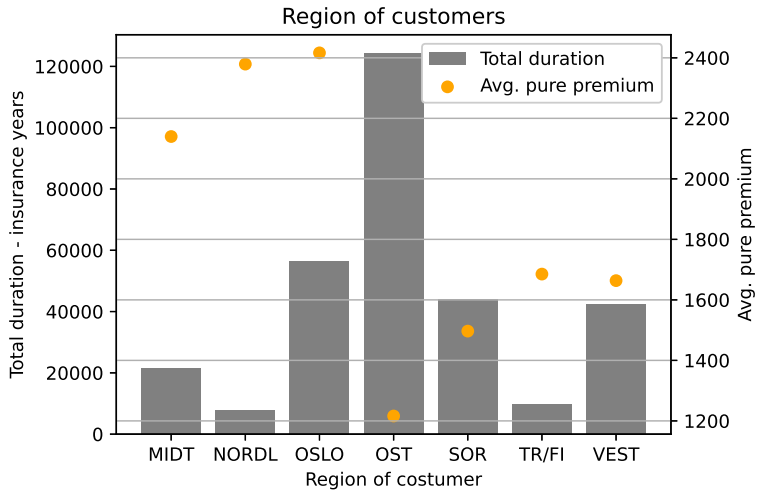


Figure 2.5: The total duration and average pure premium over different regions. The region of the customer is given by the x-axis. Norway has been divided into seven familiar regions.

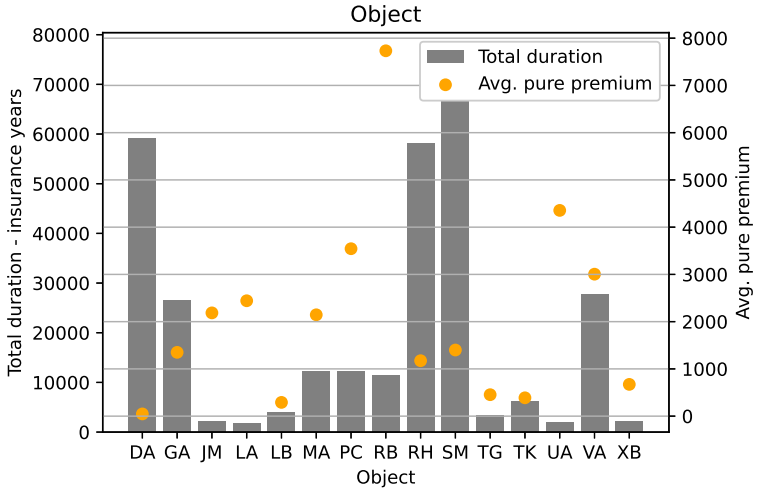


Figure 2.6: The total duration and average pure premium for the type of object. On the x-axis, we have the objects which have been anonymized. To better understand what they represent, one could imagine that DA is art, GA is kitchen equipment, etc.

Chapter 3

Bayesian inference

3.1 Introduction to Bayesian inference

Bayesian inference allows us to associate probability distributions with parameters of a statistical model. It is probably best understood compared to the frequentist conception of statistical inference. It is two schools with one single difference. Bayesian inference lets us include some hypothesis or prior knowledge in a model parameter, while frequentist inference lets observations do all of the job estimating a parameter. Imagine that we want to do inference on a Poisson distribution with rate parameter λ. From a frequentist point of view, we can estimate the value of λ by our observations, e.g. by maximum likelihood estimation (MLE). However, in the Bayesian world, we would assign a *prior* distribution to λ, e.g. a Gamma distribution with so-called *hyperparameters* α and β. Using Bayes' theorem, we can find the *posterior* distribution of λ, incorporating both prior and new knowledge (Givens & Hoeting, 2012).

Theorem 3.1.1 (Bayes' theorem)**.** *Assuming that X and Y are continuously distributed random variables, let $f(x)$ and $f(y)$ denote their marginal distribution and $f(x|y)$ and $f(y|x)$ their conditional distributions. Then we have*

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}.$$

Proof. The conditional probability $f(x|y)$ is defined by

$$f(x|y) = \frac{f(x,y)}{f(y)},$$

where $f(x,y)$ is the joint distribution of X and Y. From this, it is straightforward to show that

$$\begin{aligned} f(x|y) &= \frac{f(x,y)}{f(y)} \\ &= \frac{f(y|x)f(x)}{f(y)}. \end{aligned}$$

Figure 3.1: DAG of our hierarchical model. The hyperparameters α and β determine the distribution of λ, and λ is again a parameter in the distribution of y.

□

Returning to our example of the Poisson distribution, we can assume that $f(y|\lambda)$ is Poisson distributed with rate parameter λ, and that $f(\lambda)$ is Gamma distributed with parameters α and β. A directed acyclic graph (DAG) of the hierarchical model can be seen in Figure 3.1. Generally, a DAG of a hierarchical model provides insight into its structure and possible distributions that can be formulated from it. In literature, it is usual to refer to y as the observed variable, λ as the latent or hidden variable, and α and β as hyperparameters. Usually, we only include the random elements of a model in the DAG, meaning that hyperparameters such as α and β are left out. Applying Bayes' theorem in this case yields

$$f(\lambda|y) = \frac{f(y|\lambda)f(\lambda)}{f(y)},$$

where $f(\lambda|y)$ is called the posterior distribution, $f(\lambda)$ is the prior distribution, and $f(y|\lambda)$ is the likelihood. In the denominator, we find $f(y)$, a function not having λ as a variable. It is thus considered to be a simple constant when evaluating $f(\lambda|y)$. Letting $c = 1/f(y)$, we can reformulate to

$$f(\lambda|y) = cf(y|\lambda)f(\lambda).$$

As c is only a constant, it does not include any information about the shape of the distribution and is thus often ignored. However, it is a crucial component of a probability distribution, as it ensures that the density function integrates to 1. For this reason, it is called the *normalizing constant*. If we drop the normalizing constant, we do not retain the equality between the two distributions but get distributions proportional to one another. We indicate proportionality by the following notation

$$f(\lambda|y) \propto f(y|\lambda)f(\lambda).$$

Knowing $f(\lambda|y)$, it is possible to do inference on λ. We can find quantities of interest, such as expected value, variance, and credibility intervals. Next, we will discuss Bayesian hierarchical models, a natural continuation of Bayesian inference.

3.2 Bayesian hierarchical models

The previous chapter looked at how we could assume a distribution on a parameter λ of Poisson distribution. We assigned hyperparameters α and β to the

Gamma distribution we assumed that λ took. How would it have been had we assigned distributions to α and β as well? This kind of model specification gives rise to Bayesian hierarchical models (Casella & Berger, 2002). Building from the previous example, we can introduce a more complex statistical model with several levels and distributions. As the model will be used in a regression setting later, we will also introduce some necessary notation.

We will refer to the response variable as Y_{ijt}. The index i refers to a certain combination of covariate values, e.g. when $i = 3$ we have a person under 20 living in Oslo, and $i = 4$ could be a person between 20 and 30 living in Oslo. The index j refers to a cluster modeled as a random effect, e.g. $j = 2$ could be a hearing aid, and $j = 3$ could be art. Finally, t refers to the unique observation number for every observation that shares i and j, e.g. we could have several observations having $i = 2$ and $j = 3$. In such a case, we use t to distinguish between them. We could further illustrate this point with the example variables Y_{111} and Y_{112}, i.e. they share the same fixed and random effect values, but Y_{111} is the first observation within this group and Y_{112} the second. As they share i and j, they also share all information we use to model. Hence, it is natural that we predict the same mean for them. Therefore, Y_{111} and Y_{112} will have the same mean μ_{11}. Note also that $q = (q_1, q_2, ..., q_J)$ and y is the vector of observations $y = (y_{111}, y_{112}, ...y_{NJn_{ij}})$. This notation will be used throughout the thesis. We now look at an example of a model where this kind of notation is useful. The model is a generalized linear mixed model, which is something we will look closer at in Chapter 4.

Example 3.2.1. *Let Y_{ijt} be a random variable, which conditional on λ_{ij} is Poisson distributed with rate parameter λ_{ij}. Furthermore, let $\ln(\lambda_{ij}) = x_i^T\beta + q_j$. Next, we let q_j be normally distributed, with zero mean and variance τ^2. Next, τ^2 is a random variable following an inverse gamma prior distribution with shape α and scale ι. We can formalize it in the following form*

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Poisson(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= x_i^T\beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I), \\ \tau^2 &\sim IG(\alpha, \iota), \end{aligned}$$

and a DAG of the hierarchical model can be seen in Figure 3.2. Notice how the q_j's all depend on the τ^2, the cornerstone of credibility theory. The above is a complex model, and an example of a generalized linear mixed model, which introduced and discussed in the next chapter. It illustrates nicely how we can model with several layers and distributions. Using Bayes' theorem, we can also find distributions proportional to $f(q|y, \tau^2)$ and $f(\tau^2|y, q)$ and do inference on them.

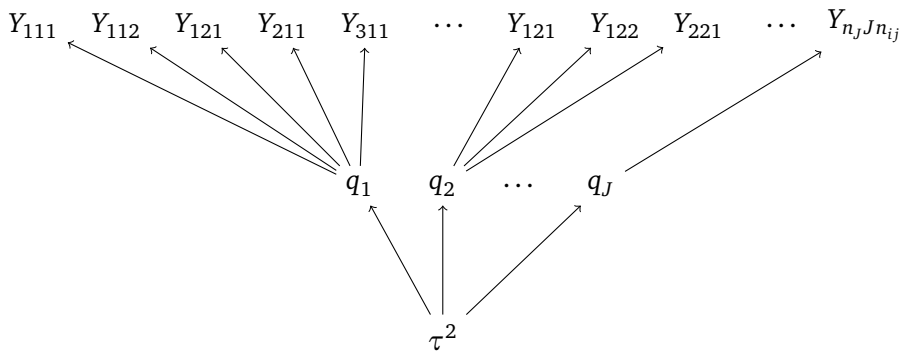


Figure 3.2: DAG of our hierarchical model. The hyperparameters α and ι determine the distribution of τ^2, which in turn is a variance parameter in the distribution of q_j. The distribution of q_j is then used in λ_{ij} which is the parameter working on y_{ijt}.

3.3 Conjugate prior

Given a class F of pdfs or pmfs $f(x,\theta)$, a prior p in the class Π is a conjugate prior for F if the posterior π is in Π for every possible $f \in F$ and $p \in \Pi$ (Casella & Berger, 2002). Knowledge about conjugate priors is advantageous because of their computational convenience.

Example 3.3.1. *This example shows that the inverse gamma distribution is a conjugate prior for the normal distribution. Assume that we have $X_1, X_2, ..., X_n$ independent and identically distributed with distribution $\mathcal{N}(\mu, \sigma^2)$. Furthermore, we assume that μ is known and σ^2 is unknown. We model σ^2 with an inverse gamma distribution with shape α and scale β. We now try to find an expression that is proportional to the posterior distribution π of σ^2. We have*

$$
\begin{aligned}
\pi(\sigma^2|x) &\propto f(x|\mu,\sigma^2)p(\sigma^2|\alpha,\beta) \\
&\propto \left[\prod_{i=1}^{n} f(x_i|\mu,\sigma^2)\right] p(\sigma^2|\alpha,\beta) \\
&\propto \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{2-(\alpha+1)} \exp\left(\frac{-\beta}{\sigma^2}\right) \\
&\propto \left[(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}\right)\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{2-(\alpha+1)} \exp\left(\frac{-\beta}{\sigma^2}\right) \\
&\propto \left(\sigma^2\right)^{-\left(\alpha+1+\frac{n}{2}\right)} \exp\left(-\frac{1}{\sigma^2}\left(\beta + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}\right)\right) \\
&\propto IG\left(\alpha+\frac{n}{2}, \beta + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}\right).
\end{aligned}
$$

We have not made any special assumptions on our parameters, so any inverse gamma distributed prior will result in a posterior inverse gamma when the likelihood is normal with unknown variance.

Chapter 4

Generalized linear mixed models

Generalized linear mixed models (GLMM) are a class of distributions that is easy to work with, both from a statistical and mathematical point of view. They are also attractive because of the wide range of distributions that can be formulated as a GLMM, such as the normal, Poisson, and exponential distributions (J. A. Nelder & Wedderburn, 1972). Every GLMM consists of two main pieces, a probability distribution with a specific form and a so-called *link function*. In the next section, we focus on the former of the two before looking closer at the link function.

4.1 Exponential dispersion model

A requirement for any GLMM is a probability distribution that can be written in the form of an exponential dispersion model. The exact required form of the distribution varies in the literature, and it is not uncommon to require it to be an exponential family. Even if the names differ, the forms are often equivalent when applied to various statistical problems. We will restrict ourselves to one form only, namely the exponential dispersion model.

Definition 4.1.1. *Let Y_{ijt} be a random variable. It belongs to the exponential dispersion models (EDM) if the density of its distribution can be written as*

$$f(y_{ijt}|\theta_{ij}) = \exp\left(\frac{y_{ijt}\theta_{ij} - b(\theta_{ij})}{\phi} w_{ijt} + c(y_{ijt}, \phi, w_{ijt})\right),$$

where θ_{ij} is the natural (or canonical) parameter, $b(\theta_{ij})$ is such that $f(y_{ijt}|\theta_{ij})$ is normalized and the first and second derivatives, $b'(\theta_{ij})$ and $b''(\theta_{ij})$, exists, $c(y_{ijt}, \phi, w_{ijt})$ is a function not having the natural parameter as a variable, ϕ is a dispersion parameter and w_{ijt} a known weight.

In the following, we illustrate how the Poisson and Gamma distributions can be rewritten into the form of an EDM by identifying θ_{ij}, $b(\theta_{ij})$, and $c(y_{ijt}, \phi, w_{ijt})$, thus proving that they belong to the class of EDMs.

Example 4.1.1. *Starting from the most common form of the Poisson distribution, we show how it can be rewritten to an EDM.*

If $Y_{ijt} \sim Poisson(\lambda_{ij})$*, then*

$$\begin{aligned} f_{Y_{ijt}}(y_{ijt}|\lambda_{ij}) &= \frac{\lambda_{ij}^{y_{ijt}}}{e^{-\lambda_{ij}} y_{ijt}!} \\ &= \exp\left(y_{ijt}\ln\lambda_{ij} - \lambda_{ij} - \ln y_{ijt}!\right) \\ &= \exp\left(y_{ijt}\theta_{ij} - b(\theta_{ij}) + c(y_{ijt},\phi,w_{ijt})\right), \end{aligned}$$

where

$$\begin{aligned} \phi &= w_{ijt} = 1, \\ \theta_{ij} &= \ln(\lambda_{ij}), \\ b(\theta_{ij}) &= \exp(\theta_{ij}), \\ c(y_{ijt},\phi,w_{ijt}) &= \ln(1/y_{ijt}!). \end{aligned}$$

Example 4.1.2. *Similarly, we can show how the most common form of Gamma distribution can be rewritten to an EDM.*

If $Y_{ijt} \sim Gamma(\alpha_{ij},\beta_{ij})$*, then*

$$f_{Y_{ijt}}(y_{ijt}|\alpha_{ij},\beta_{ij}) = \frac{1}{\Gamma(\alpha_{ij})}\beta_{ij}^{\alpha_{ij}} y_{ijt}^{\alpha_{ij}-1} e^{-\beta_{ij}y_{ijt}}.$$

We can reparametrize by introducing $\delta_{ij} = \alpha_{ij}$ *and* $\mu_{ij} = \frac{\alpha_{ij}}{\beta_{ij}}$*. We now have*

$$\begin{aligned} f_{Y_{ijt}}(y_{ijt}|\mu_{ij},\delta_{ij}) &= \left(\frac{\delta_{ij}}{\mu_{ij}}\right)^{\delta_{ij}} \frac{1}{\Gamma(\delta_{ij})} {y_{ijt}}^{\delta_{ij}-1} \exp\left(\frac{-\delta_{ij}y_{ijt}}{\mu_{ij}}\right) \\ &= \exp\left(\theta_{ij}y_{ijt} - b(\theta_{ij}) + c(y_{ijt},\phi,w_{ijt})\right), \end{aligned}$$

where

$$\begin{aligned} \phi &= w_{ijt} = 1, \\ \theta_{ij} &= -1/\mu_{ij}, \\ b(\theta_{ij}) &= \log(\mu_{ij}) = -\log(-\theta_{ij}) \\ c(y_{ijt},\phi,w_{ijt}) &= \delta\log(\delta_{ij}) + (\delta_{ij}-1)\log(y_{ijt}) - log(\Gamma(\delta_{ijt})). \end{aligned}$$

A nice property of the EDMs is the simplicity in which we can compute the mean and variance.

Theorem 4.1.1. *If* Y_{ijt} *belongs to an EDM, then*

$$\begin{aligned} E[Y_{ijt}] &= b'(\theta_{ij}), \\ Var[Y_{ijt}] &= \frac{\phi}{w_{ijt}} b''(\theta_{ij}). \end{aligned}$$

Proof. We begin by simplifying the expression for the moment generating function of Y.

$$\begin{aligned}
&M_{Y_{ijt}}(s)\\
=&\mathrm{E}[\exp(sY_{ijt})]\\
=&\int_{y_{ijt}} \exp\left(\frac{y_{ijt}\theta_{ij}-b(\theta_{ij})}{\phi}w_{ijt}+sy_{ijt}+c(y_{ijt},\phi,w_{ijt})\right)dy_{ijt} \qquad (4.1)\\
=&\exp\left(\frac{w_{ijt}}{\phi}b(\theta^*_{ij})-\frac{w_{ijt}}{\phi}b(\theta_{ij})\right)\int_{y_{ijt}} v(\theta^*_{ij},y_{ijt},\phi,w_{ijt})dy_{ijt}, \qquad (4.2)
\end{aligned}$$

where

$$\begin{aligned}
v(\theta^*_{ij},y_{ijt},\phi,w_{ijt}) &= \exp\left(\frac{y_{ijt}(\theta^*_{ij})-b(\theta^*_{ij})}{\phi}w_{ijt}+c(y_{ijt},\phi,w_{ijt})\right),\\
\theta^*_{ij} &= \theta_{ij}+(\phi/w_{ijt})s
\end{aligned}$$

and Equation 4.2 is found by adding and subtracting $b(\theta^*_{ij})$ in the exponential in Equation 4.1. Within the integrand in Equation 4.2, we have the density of another EDM, implying that it integrates to 1.

So we have

$$M_{Y_{ijt}}(s)=\exp\left(\frac{w_{ijt}}{\phi}\left(b(\theta^*_{ijt})-b(\theta_{ijt})\right)\right).$$

Recall that $M'_{Y_{ijt}}(0)=\mathrm{E}[Y_{ijt}]$ and that $M''_{Y_{ijt}}(0)=\mathrm{E}[Y^2_{ijt}]$. Let us now calculate the first and second derivative of $M_{Y_{ijt}}(s)$ and set them equal to zero in order to find expressions for the mean and variance.

$$\begin{aligned}
M'_{Y_{ijt}}(s) &= \frac{w_{ijt}}{\phi}b'(\theta^*_{ij})\frac{\phi}{w_{ijt}}\exp\left(\frac{w_{ijt}}{\phi}\left(b(\theta^*_{ij})-b(\theta_{ij})\right)\right),\\
M''_{Yijt}(s) &= \left(\frac{w_{ijt}}{\phi}b''(\theta^*_{ij})\left(\frac{\phi}{w_{ijt}}\right)^2+\left(\frac{w_{ijt}}{\phi}\right)^2 b(\theta^*_{ij})^2\right)\exp\left(\frac{w}{\phi}\left(b(\theta^*_{ij})-b(\theta_{ij})\right)\right)
\end{aligned}$$

Using that $\theta^*_{ij}=\theta_{ij}$ when $s=0$, we find the expectation and variance,

$$\begin{aligned}
\mathrm{E}[Y_{ijt}] &= M'_{Y_{ijt}}(0)=b'(\theta)\\
\mathrm{E}[Y^2_{ijt}] &= M''_{Y_{ijt}}(0)=\frac{\phi}{w_{ijt}}b''(\theta_{ij})+(b'(\theta_{ij}))^2\\
\mathrm{Var}[Y_{ij}] &= \mathrm{E}[Y^2_{ij}]-\mathrm{E}[Y_{ij}]^2=\frac{\phi}{w_{ijt}}b''(\theta_{ij}).
\end{aligned}$$

□

Having formulated an EDM and looked at the important property on its expectation and variance, we go on to the next important component of a GLMM.

4.2 Link function

When coupled with an EDM, the link function is everything we need to have a fully specified GLMM. The link function is the component that shows the great flexibility and thus attractiveness of using GLMMs.

Definition 4.2.1. *The link function is a function g, which links the linear predictor η_{ij} with the mean $E[Y_{ijt}] = \mu_{ij}$ of the distribution. More specifically, the link function is so that*

$$g(\mu_{ij}) = \eta_{ij},$$

where $\eta_{ij} = x_i^T \beta + q_j$. In literature, we are often referenced to a response function h, rather than the link function g. The response function is simply the inverse of the link function.

Having defined the link function, we can investigate its role in the previous Examples 4.1.1 and 4.1.2.

Example 4.2.1. *In Example 4.1.1, we showed that $b(\theta_{ij}) = \exp(\theta_{ij})$, and we know that $\theta_{ij} = \ln(\lambda_{ij})$. From the property of the mean of an EDM, we also know that $b'(\theta_{ij}) = \mu_{ij} = h(\eta_{ij})$. We have $b'(\theta_{ij}) = \frac{d}{d\theta_{ij}} \exp(\theta_{ij}) = \exp(\theta_{ij}) = \lambda_{ij}$, which is what we would expect knowing the Poisson distribution.*

Choosing that the linear predictor $\eta_{ij} = g(\mu_{ij}) = \theta_{ij}$, we obtain the so-called canonical, or natural, link function. In this case, this would indicate that $\mu_{ij} = b'(\theta_{ij}) = \exp(\theta_{ij}) = \exp(\eta_{ij})$. This yields a so-called multiplicative model, as changes to the linear predictor has a multiplicative effect on the mean.

Example 4.2.2. *In Example 4.1.2, we showed that $b(\theta_{ij}) = \ln(-\theta_{ij})$, and we know that $\theta_{ij} = -1/\mu_{ij}$. From the property of the mean of an EDM, we also know that $b'(\theta_{ij}) = \mu_{ij}$. We have $b'(\theta_{ij}) = \frac{d}{d\theta_{ij}} - \ln(-\theta_{ij}) = -1/\theta_{ij} = \mu_{ij}$, which is what we would expect knowing the Gamma distribution and our chosen reparametrization.*

Choosing that the linear predictor $\eta_{ij} = g(\mu_{ij}) = \theta_{ij}$, we obtain what the canonical link function. In this case, this would indicate that $\mu_{ij} = -1/\eta_{ij}$. Contrary to Example 4.2.1, this link function is not as easy or intuitive to work with. Notice for example that we will require that $\eta_{ij} < 0$, because of the relation $b(\theta_{ij}) = \ln(-\theta_{ij})$. This will require the use of constrained optimization when we want to estimate parameters. A possibility is to drop the natural link function and instead formulate a relation that yields a multiplicative model. Wanting a multiplicative model, we need $\mu_{ij} = \exp(\eta_{ij})$. We also know that $\mu_{ij} = -1/\theta_{ij}$. Hence, a reparametrization of θ_{ij} is sufficient. By letting $\theta_{ij} = -1/\exp(\eta_{ij})$, it is straightforward to show that we will obtain a multiplicative model.

Chapter 5

Methods of parameter estimation and simulation

The vast amount of data available requires computation that is efficient and ideally adapted to the given data and our assumptions. Computational techniques learned in elementary statistical courses are no longer feasible when going from ten data points to several hundred thousand. This chapter will look at algorithms and techniques that are better suited for big data sets and complex statistical models.

5.1 Monte Carlo method

The Monte Carlo method is a class of methods to estimate values numerically with the help of random sampling. It was originally proposed by Nicholas Metropolis and Stanislaw Ulam in 1949. At the time, they both worked in Los Alamos, and the method was developed and used as a tool in computations in mathematical physics (Metropolis & Ulam, 1949). The use and effectiveness of the method have probably far exceeded the expectations of its creators, and it was elected as one of the top 10 algorithms of the 20th century within computing in science and engineering (Dongarra & Sullivan, 2000).

Tossing a coin several times to approximate the probability of getting heads is an example of the Monte Carlo method. The Monte Carlo integration is of particular interest to us, which uses random samples to estimate some mean μ of a probability density function $f(x)$. Let us assume that $\mu = E[g(X)]$, where X follows a probability density function $f(x)$. By the definition of expectation, we have

$$\mu = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

It is not unusual that this integral is difficult or even impossible to compute analytically. It can however be simple to draw random samples $x_1, ...x_n$ from $f(x)$.

Monte Carlo integration is a technique that exploits the possibility of sampling from a distribution when the analytical solution to the integral is intractable. A Monte Carlo estimate $\hat{\mu}$ of μ is then

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} g(x_i).$$

It is straightforward to show that the estimate is unbiased. Furthermore, it is possible to show that $\hat{\mu}$ converges to μ by using the strong law of large numbers (Givens & Hoeting, 2012).

5.2 Rejection sampling

Rejection sampling (RS) is a Monte Carlo method used to generate samples from a distribution $f(x)$ which is difficult to sample directly from. It does so with help from a proposal distribution $g(x)$ which we can sample from. We do not impose any restrictions on $g(x)$ other than that we need to know some constant A such that $Ag(x) \geq f(x), \forall x$ (Gamerman & Lopes, 2006).

The algorithm starts by drawing a proposal x from $g(x)$ and then a sample u from the uniform distribution on $[0, 1]$. If $u \leq f(x)/Ag(x)$ we accept x as a sample from $f(x)$. A pseudocode is provided in Algorithm 1.

Algorithm 1 Rejection sampling

1: Draw $x \sim g(x)$
2: Draw $u \sim U(0, 1)$
3: **if** $u \leq \frac{f(x)}{Ag(x)}$ **then**
4: Accept x as sample from $f(x)$
5: **else**
6: Reject x as sample from $f(x)$
7: **end if**

As is evident from Algorithm 1, we are not guaranteed a new sample from $f(x)$ at every iteration, and one has to rerun the algorithm until one reaches the desired sample size. The algorithm's effectiveness depends on the similarity between $f(x)$ and $g(x)$. The probability of acceptance is given by $f(x)/Ag(x)$, and if $f(x)$ tends to be much smaller than $Ag(x)$ it will spend more time to generate a new sample.

Using this algorithm, we ensure that each sample x is a sample from $f(x)$. We can show this using Bayes' theorem. We let $h(x, u)$ denote the joint density between x and u. We want to find an expression for the conditional distribution $h(x|u \leq f(x)/Ag(x))$, that is, the distribution of an x that is accepted and thus considered

to have been drawn from $f(x)$. We have

$$\begin{aligned} h\left(x|u \leq \frac{f(x)}{Ag(x)}\right) &= \frac{h\left(u \leq \frac{f(x)}{Ag(x)}|x\right) g(x)}{\int h\left(u \leq \frac{f(x)}{Ag(x)}|x\right) g(x)dx} \\ &= \frac{\frac{f(x)}{Ag(x)} g(x)}{\int \frac{f(x)}{Ag(x)} g(x)dx} \\ &= \frac{\frac{1}{A} f(x)}{\int \frac{1}{A} f(x)dx} \\ &= \frac{f(x)}{\int f(x)dx}, \end{aligned}$$

which demonstrates that each accepted x is a sample from the normalized distribution of $f(x)$. For a thorough introduction to RS, see Devroye (1986). Next, we discuss a variation of the algorithm.

Adaptive rejection sampling (ARS) was introduced in Gilks and Wild (1992), and it represents a variation of the RS algorithm. The ARS comes closer to having similar $g(x)$ and $f(x)$ by continually changing the proposal distribution $g(x)$ to approach that of $f(x)$. It requires $f(x)$ to be log-concave and uses this property to approach $f(x)$ from above. It readjusts and learns from every iteration, and thus it will outperform the classical RS as the number of samples grows.

5.3 Markov chain Monte Carlo

We employ Monte Carlo integration in cases where the integral is intractable, but sampling is easy. Markov chain Monte Carlo (MCMC) is another class of methods dealing with the scenario where sampling from $f(x)$ is difficult or impossible. However, it is assumed that we can evaluate $f(x)$, and MCMC methods aim to draw from a distribution that approximates that of $f(x)$, potentially up to a normalizing constant. It does that by the use of Markov chains. Next, we restrict our definition of a Markov chain to a discrete state space, but it is just as easy to define it within a continuous setting.

Given a sequence of random variables $X_1, X_2, ..$, the sequence is referred to as a Markov chain if and only if the probability of event number $k+1$ only depends on the outcome of event number k. That is

$$P(X_{k+1} \in A|X_k = x, X_{k-1} \in A_{n-1}, ..., X_1 \in A_1) = P(X_{k+1} \in A|X_k = x),$$

where $A_0, ..., A_{k-1}, A \subset S$ and $x \in S$ with S being the state space S (Gamerman & Lopes, 2006). Of biggest interest is the limiting properties of the Markov chain, i.e. how does the chain behave when $k \to \infty$?

A Markov chain has a unique limiting distribution $\pi(x)$ if the chain can be shown to be *irreducible*, *aperiodic* and *positive recurrent*. We will not detail these terms, but we can remind ourselves about their meaning and implications on a chain. An irreducible Markov chain is a chain where all states can be reached from all states, i.e. there is a positive probability for reaching state j from state i in a finite number of steps, for all $i, j \in S$. An aperiodic Markov chain is a chain that can return to any given state i in a multiple of d steps, where d is equal to one. A positive recurrent Markov chain means that starting in any state i, the expected mean time to return is finite. A more rigorous and detailed introduction to Markov chains can be found in Ross (2014).

An essential part of any MCMC method is constructing a chain that eventually reaches $\pi(x) = f(x)$. The state space S is already specified by the model and our assumptions. The key is how we transition from one state i to another state j. This is where most MCMC methods branch out, and in the following, we will look closer at two methods.

5.3.1 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a classical MCMC algorithm. After Metropolis' work in 1949 and further improvements presented in Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953), a statistics professor named Hastings provided a generalized method in Hastings (1970). Its goal is to sample from a target distribution $f(x)$ which is otherwise difficult to sample directly from.

At $t = 0$ the algorithms starts at the starting point $X^{(0)} = x^{(0)}$. The starting point is drawn from a proposal distribution g, and it is required that $f(x^{(0)}) > 0$, where $f(x)$ is our target distribution. Given $X^{(t)} = x^{(t)}$, the next step for sampling $X^{(t+1)}$ is given in Algorithm 2.

Algorithm 2 Metropolis-Hastings algorithm

1: Sample a proposal value y from $g(x^{(t)}, y)$.
2: Compute acceptance probability $\alpha(x^{(t)}, y) = \min\left\{1, \frac{f(y)g(y,x^{(t)})}{f(x^{(t)})g(x^{(t)},y)}\right\}$
3: Draw u from Uniform(0,1).
4: **if** $u \leq \alpha(x^{(t)}, y)$ **then**
5: $\quad x^{(t+1)} \leftarrow y$
6: **else**
7: $\quad x^{(t+1)} \leftarrow x^{(t)}$
8: **end if**

How the chain moves from one state to another is given by the *transition kernel* $p(x, y)$, where x is our current state and y is the proposed state. The transition

kernel consists of the product of two elements, namely the *proposal density* $g(x, y)$ and the *acceptance probability* $\alpha(x, y)$. We have

$$p(x, y) = g(x, y)\alpha(x, y), x \neq y.$$

The probability of staying in the same state is somewhat more complicated, but as we shall see, it is not necessary to spend time evaluating it. The proposal density is simply a density *proposing* the next step, i.e. a density we can sample from. The acceptance probability is defined as

$$\alpha(x, y) = \min\left\{1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right\}.$$

The transition kernel is at the heart of every MCMC algorithm and is what makes the algorithms unique. Despite their differences, they share the fact that the transition kernel has to be constructed in a way that ensures convergence to the desired distribution. The Metropolis-Hastings algorithm rests on the so-called *detailed balance equation*, i.e. it requires the chain to fulfill the condition

$$f(x)p(x, y) = f(y)p(y, x).$$

The condition results in a *time-reversible* Markov chain, that is, a chain that is the same going forward and backward. This requirement is sufficient but not necessary to ensure a chain that converges to $f(x)$. This can be demonstrated by using some basic properties of statistics. We have

$$\begin{aligned}\sum_x f(x)p(x, y) &= \sum_x f(y)p(y, x), \forall x, y \\ &= f(y)\sum_x p(y, x) \\ &= f(y),\end{aligned}$$

which gives that $f(x)$ is the stationary distribution of a general Markov chain built using the detailed balance equation.

Furthermore, we want to show that the specific transition kernel used in this case satisfies the detailed balance equation. We then have

$$\begin{aligned}f(x)p(x, y) &= f(x)g(x, y)\alpha(x, y), \forall x \neq y \\ &= f(x)g(x, y)\min\left\{1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right\} \\ &= \min\{f(x)g(x, y), f(y)g(y, x)\} \\ &= f(y)g(y, x)\min\left\{\frac{f(x)g(x, y)}{f(y)g(y, x)}, 1\right\} \\ &= f(y)p(y, x),\end{aligned}$$

where all we do is some manipulation of the minimum function which holds because the probabilities are defined and strictly positive in each iteration. For cases where $x = y$, it is trivially true that the detailed balance equation holds. Thus, $f(x)$ is the chain's stationary distribution, which is what we wanted to show.

As is clear to the reader, we have not specified any restrictions on the distributions, and it is possible to employ distributions with properties that make the algorithm even simpler. An example of a chain that makes the Metropolis-Hastings algorithm simpler is the random walk Metropolis. It imposes a symmetric proposal density around the current state. A consequence of this is that the acceptance probability α is simplified to $\alpha = \min\left\{1, \frac{f(y)}{f(x)}\right\}$ (Givens & Hoeting, 2012).

5.3.2 Gibbs sampling

The Gibbs sampler is yet another MCMC algorithm, named after an American physicist by the two brothers Stuart and Ronald Geman. The algorithm was initially developed for image restoration and is yet another example of a popular statistical method with its roots in another science (Geman & Geman, 1984). As with the Metropolis-Hastings algorithm, there is no lack of flexibility and variations with this algorithm. However, we can identify and highlight a core, unifying every possible *Gibbs sampler*, where a Gibbs sampler is a nickname for the chain that employs the Gibbs sampling algorithm.

The Gibbs sampler is characterized by how it reduces the sampling of high-dimensional blocks to lower-dimensional blocks. It can thus be employed to obtain samples from high-dimensional joint densities (Gelfand, 2000). This makes it an attractive option to the Metropolis-Hastings algorithm, popularly used on lower-dimensional and often only univariate densities. As mentioned, there are many ways to mould the Gibbs sampler, some of which we will look at later. However, for the time being, we will restrict our attention to the most basic Gibbs sampler.

Imagine that we have a joint target distribution $f(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, ..., x_n)$. Direct sampling from the joint and marginal distributions is assumed to be difficult or impossible. The Gibbs sampler proposes to build a Markov chain which samples from conditional distributions. From Bayes' theorem, we have

$$f(x_i|\boldsymbol{x}_{-i}) \propto f(\boldsymbol{x}) \forall i,$$

where $\boldsymbol{x}_{-i}$ is the vector $(x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_{n-1}, x_n)$. Sampling from the univariate conditional distributions is often available in closed form or with simpler algorithms. A very general pseudocode of how the Gibbs sampler works can be seen in Algorithm 3.

It should be clear that this chain is Markov, as each step only uses the last step

Algorithm 3 Gibbs sampler

1: At $t = 0$ we initialize $\boldsymbol{x}^{(0)}$.
2: **while** Not converged **do**
3: Sample from $f(x_1^t | x_2^{t-1}, x_3^{t-1}, ..., x_n^{t-1})$.
4: Sample from $f(x_2^t | x_1^t, x_3^{t-1}, ..., x_n^{t-1})$.
5: $\vdots$
6: Sample from $f(x_n^t | x_1^t, x_2^t, ..., x_{n-1}^t)$.
7: Increment t.
8: **end while**

when drawing new values. Proving that the chain has $f(\boldsymbol{x})$ as its limiting distribution is a bigger task, and the reader is therefore referred to Geman and Geman (1984) for a proof.

Gibbs sampling is often used as a part of an MCMC algorithm, where draws from certain variables can be made as a Gibbs step. This will always be accepted, as the full conditional distribution of the variable is known. This is an implementation that will be used in our method as well. There is a sense of duality here, as we could also call it a Gibbs sampler, where the full conditionals difficult to sample from can be done with the help of an MCMC step. What is important to note is that we use combinations of Gibbs sampling and MCMC.

5.4 Expectation-maximization algorithm

The expectation-maximization (EM) algorithm is an algorithm used when we want to estimate maximum likelihood estimates (MLE) from incomplete data. In our context, missing data refers to some unobserved latent variable. The algorithm first appeared in Dempster, Laird and Rubin (1977), which gives a thorough and sufficient introduction to the algorithm.

Before we look at how the algorithm operates, we introduce a suitable context for the algorithm to work in. Assume that we have a random variable Y which we observe, and an unobserved latent variable Z. The complete data X is the collection of Y and Z, i.e. $X = (Y, Z)$. Seeing that this is an algorithm aiming to estimate parameters, it is natural to have some parameter ψ as well. How these quantities are related will vary. An example of how a model can be structured will follow after a brief demonstration of the motivation of the algorithm and its construction.

Typically, the observed data density $f(y|\psi)$ is what we would use to formulate a likelihood and estimate the MLE $\widehat{\psi}$ of ψ. In some cases, the likelihoods can be challenging to work with, and in such a case, the complete data X and the latent variable Z might ease computation. Let $f(x|\psi) = f(y, z|\psi)$ denote the complete

data density and $f(z|x,\psi)$ the latent variable density.

The EM algorithm works by maximizing the log-likelihood of $f(x|\psi)$ rather than $f(y|\psi)$. Let $Q(\psi|\psi^{(t)})$ denote the expectation of the log-likelihood of the complete data, that is

$$\begin{aligned} Q(\psi|\psi^{(t)}) &= \mathrm{E}\left\{\log L(\psi|x)|y,\psi^{(t)}\right\} \\ &= \mathrm{E}\left\{\log f(x|\psi)|y,\psi^{(t)}\right\}. \end{aligned}$$

It is the structure of the model, with an unobserved z, that gives us the unusual expression of the expectation of a log-likelihood function. Having formulated the above, we can further develop the expression

$$Q(\psi|\psi^{(t)}) = \int \log f(x|\psi) f(z|y,\psi^{(t)}) dz. \tag{5.1}$$

Computation of $Q(\psi|\psi^{(t)})$ results in a function of ψ, which is subsequently maximized to obtain $\psi^{(t+1)}$. Thus, the EM algorithm consists of two distinct and characteristic steps, namely the expectation step of computing $Q(\psi|\psi^{(t)})$ and the maximization step of computing the maximum of $Q(\psi|\psi^{(t)})$. This can be seen in Algorithm 4.

Algorithm 4 EM algorithm

1: At $t = 0$ we initialize $\psi^{(0)}$.
2: **while** Not converged **do**
3: Compute $Q(\psi|\psi^{(t)})$.
4: Maximize $Q(\psi|\psi^{(t)})$ with respect to ψ.
5: $\psi^{(t+1)}) \leftarrow \psi$.
6: Increment t.
7: **end while**

Example 5.4.1. *To properly understand the usefulness of the algorithm, we will return to the example of the hierarchical model of a Poisson distributed random variable Y with rate parameter λ, where λ is Gamma distributed with parameters α and β. In this case, we have $\psi = (\alpha, \beta)$, and we aim to estimate the MLE $\widehat{\psi}$ of ψ. Direct computation through the function $L(\psi|y)$ is not easy, but we can exploit the structure of our model with the EM algorithm.*

The complete data is $x = (y, \lambda)$ and the expected log-likelihood $Q(\psi|\psi^{(t)})$ can be formulated as

$$\begin{aligned} Q(\psi|\psi^{(t)}) &= \int [\log f(x|\psi)] f(\lambda|y,\psi^{(t)}) d\lambda \\ &= \int [\log f(y,\lambda|\psi)] f(\lambda|y,\psi^{(t)}) d\lambda. \end{aligned}$$

The term $\log f(y, \lambda|\psi)$ *can be computed with Bayes' theorem to find a proportional density. The same goes for* $f(\lambda|y, \psi^{(t)})$*. Having done that, we can compute* $Q(\psi|\psi^{(t)})$ *up to some constant and maximize this. It is clear that the missing constant should not change the result in the expectation step.*

The example illustrates how the latent variable λ is used to ease computation. We did not, however, look into how we would integrate the integrand. Even though it would be straightforward to find a proportional expression, it is not evident that we would manage to solve the integral analytically. In the next section, we will look at a possible remedy to this problem.

5.4.1 Monte Carlo expectation-maximization algorithm

A variation of the EM algorithm, named Monte Carlo expectation-maximization (MCEM) algorithm, was developed by Wei and Tanner (1990). It was motivated by the difficulty of analytically evaluating the integral to compute the expected log-likelihood $Q(\psi|\psi^{(t)})$. They proposed that, given $\psi^{(t)}$, one could generate samples from the distribution $f(z|y, \psi^{(t)})$ from Equation 5.1. If the generation of samples is feasible, it is then possible to utilize the Monte Carlo integration to approximate the expectation of the log-likelihood. By generating samples $(z_1, z_2, ..., z_n)$ we have

$$\begin{aligned} Q(\psi|\psi^{(t)}) &= \int \log f(x|\psi) f(z|y, \psi^{(t)}) dz \\ &= \int \log f(y, z|\psi) f(z|y, \psi^{(t)}) dz \\ &\approx \frac{1}{m} \sum_{v=1}^{m} \log f(y, z_v|\psi) \\ &:= Q_m(\psi|\psi^{(t)}). \end{aligned}$$

Knowing how to compute the log-likelihood, we can compute an approximation $Q_m(\psi|\psi^{(t)})$ of $Q(\psi|\psi^{(t)})$ without integration.

An implementation of the algorithm has been proposed by Levine and Casella (2001). This implementation will be the basis of our method, which we return to in the following chapter.

Chapter 6

Credibility theory

In this chapter, we look at credibility theory and its methods of estimating parameters in a GLMM. This is the current way of estimation and the method we aim to build from and challenge. Imagine that we have an insurance portfolio. It contains several individual risks, and our goal is to predict the pure premium y_{ijt}. How can we do this? As a statistician, it is not difficult to imagine suitable models. However, this problem was initially actuarial rather than statistical in its nature. The actuarial solution was and still is credibility estimates. The notion of credibility estimates has existed for over 100 years (Norberg, 2004).

Credibility estimates, and their theory, rest on the assumption that the individual risks within the same portfolio are *similar*. In other words, we want to use information from the other risks in predicting the new value for the individual risk i. A possible prediction of the pure premium can be

$$\widehat{y}_{ijt} = z_{ij.}\bar{y}_{ij.} + (1 - z_{ij.})\mu,$$

where $\widehat{y}_{ijt}$ is the predicted pure premium, $\bar{y}_{ij.}$ is the observed mean of the risks in group $ij.$, $z_{ij.}$ is a credibility weight on $[0, 1]$, and μ is the overall pure premium mean of the portfolio. If $z_{ij.} = 1$, we predict that the next observation will be in line with what we have seen on this group of risks previously. If $z_{ij.}$ is lower, we put more credibility on the portfolio and less trust in previous observations of the risk in the same group. This can be useful if we have made few or no observations on the group $ij.$ before.

The above equation is an early form of credibility estimates. Credibility theory, originally developed by actuaries, has been formalized and further developed to an extent where it is used as a tool in more complex and modern statistical applications. A typical practice today is to use credibility estimates to estimate random effects parameters in generalized linear mixed models (J. Nelder & Verrall, 1997). In this thesis, we approach the problem by investigating how alternative ways of estimating both the fixed and random effects will compare with the more traditional credibility weight used in insurance today. To properly understand how

estimates are computed in the credibility setting, we include a pseudocode and a short discussion of the backfitting algorithm, which employs credibility theory when fitting a GLMM.

6.1 Backfitting algorithm

When employing the backfitting algorithm, we are in the context of estimating parameters in a GLMM. We consider the response variable Y_{ijt}, the fixed effects β, and random effects q. In our setting, β will refer to parameter values of the categorical variables given by age, region, customer rating, and insured sum. The random effect q will refer to the type of valuable the observation is, i.e. is it art, hearing-aid etc.

Estimation of the fixed effects is done using standard GLM fitting techniques. Then we estimate τ^2 and σ^2, which are the assumed variance of q and the within-group variance, respectively. The weight w_{ijt} is given for every observation. Grouping observations over i and t, we denote the total duration of level j by $w_{.j.}$.

The following equations is used estimating the random effects at each level j,

$$\widehat{q}_j = \tilde{z}_j \overline{\tilde{Y}}_{.j.} + (1 - \tilde{z}_j)\mu_0, \tag{6.1}$$

$$\tilde{z}_j = \frac{\tilde{w}_{.j.}}{\tilde{w}_{.j.} + \frac{\sigma^2}{\tau^2}}, \tag{6.2}$$

$$\overline{\tilde{Y}}_{.j.} = \frac{\Sigma_{i.t} \tilde{w}_{ijt} \tilde{Y}_{ijt}}{\Sigma_{i.t} \tilde{w}_{ijt}}. \tag{6.3}$$

Equation 6.1 estimates the random effect of q_j, where μ_0 is some assumed mean of q. Note that μ_0 is used to incorporate possible prior belief about the value of q_j. Equation 6.2 and 6.3 compute the credibility weight $\tilde{z}_j$ and the observed mean $\overline{\tilde{Y}}_{.j.}$ for level j. To give some interpretation to this, we see from Equation 6.1, that the estimated random effect will be a combination of the mean observed in the group $\overline{\tilde{Y}}_{.j.}$, and our assumed mean μ_0. We remark that $\overline{\tilde{Y}}_{.j.}$ is not a mean of the predicted response variable but the random effect only, hence the tilde used in the notation. It is not necessary to fully understand the computation but rather to observe how the algorithm operates, as seen in Algorithm 5.

For a further introduction to model estimation in insurance, see Ohlsson and Johansson (2010). For an introduction to modern Credibility theory, see Bühlmann and Gisler (2005). In the next chapter, we introduce alternatives to the model specification and model estimation.

Algorithm 5 Backfitting algorithm

Initialize $q \leftarrow \mu_0$
while Not converged **do**
    Fit $\widehat{\beta}$ by estimating a GLM with $\ln(q)$ as offset
    Compute $\widehat{\tau}^2$ and $\widehat{\sigma}^2$
    Compute $\widehat{q}$ as in Equation 6.1
end while
return $\widehat{Y} = \exp(X\widehat{\beta} + \widehat{q})$

Chapter 7

Lorenz curves and Gini coefficient

The Lorenz curve was introduced in Lorenz (1905) and provided an intuitive measure and visualization of the distribution of wealth. The original curve was given in a two-dimensional plot. The x-axis represented an ordered proportion of a population from poorest to richest, and the y-axis was the ordered cumulative wealth. An example of such a plot is given in Figure 7.1. If the wealth is evenly distributed over the population, we would expect 10% of the wealth to belong to 10% of the population, 20% to 20%, etc. This would yield the line of equality, i.e. the diagonal line in the figure. The Lorenz curve depicted in this figure does not indicate an evenly distributed wealth but rather an imbalance. For example, we can look at around the 50% mark along the x-axis. At this point, we have the poorest half of the population. Looking at where the Lorenz curve at 50% intersects on the y-axis, it is clear that they have approximately 25% of the cumulative wealth. This procedure of visualizing balance between two distributions has been widely used in economics and is also employed as a tool for rating insurance models.

Formalizing the idea of a Lorenz curve, we imagine a member i of a population with individuals $i = 1, ..., k$. Member i has an associated income X_i, which is assumed to be a random variable with cumulative distribution function $F(x)$. For a given value of x, the resulting value of $F(x)$ tells us what proportion of the population has an income less or equal to x. In line with the idea presented in Gastwirth (1971), we define the inverse $F^{-1}(t)$ of $F(x)$ as

$$F^{-1}(t) = \inf_x \{x : F(x) \geq t\}.$$

For a chosen proportion t of the population, this definition will return the smallest income x that satisfies $F(x) \geq t$. The definition might seem excessive for a continuous cumulative distribution function, but it allows for a similar interpretation for a discrete cumulative distribution function.

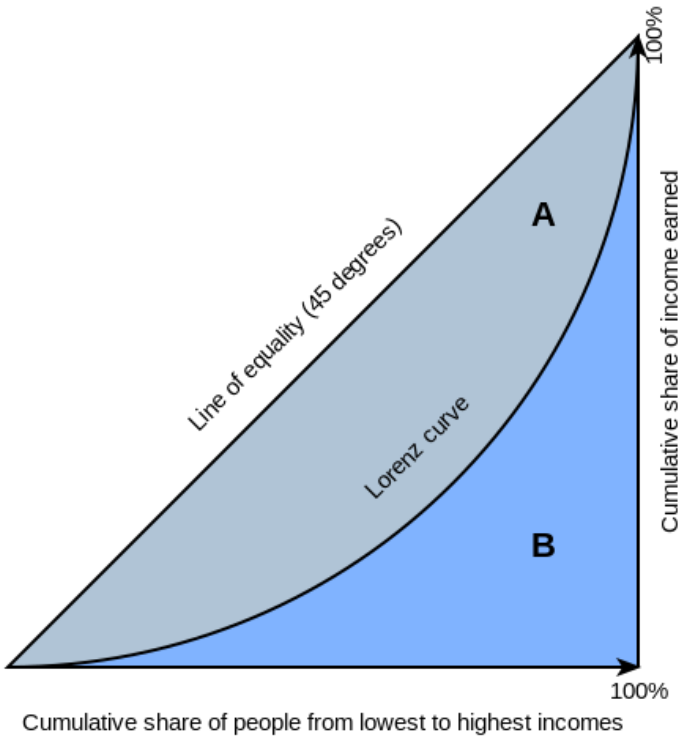


Figure 7.1: An example of the Lorenz curve. On the x-axis we have the sorted proportional share of the population, and on the y-axis we have the cumulative share of wealth.

Next, the Lorenz curve for a random variable X with inverse cumulative distribution function $F^{-1}(t)$ and mean μ is defined as

$$L(p) = \frac{1}{\mu} \int_0^p F^{-1}(t) dt,$$

where p being a proportion, we have $p \in [0, 1]$. It can also be shown that $L(0) = 0$ and $L(1) = 1$ by integration by parts (Frontczak, Jaeger & Schumacher, 2017). In reality, the construction of the Lorenz curve is empirical and based on observed data. We now construct an example with our insurance portfolio. Let P_i denote the paid pure premium, i.e. the cost for the customer, of observation i. We can sort the observations from smallest to largest premium and produce a plot of the corresponding Lorenz curve. The empirical distribution function is

$$\widehat{F}_P(x) = \frac{\sum_{i=1}^n P_i I(P_i \leq x)}{\sum_{i=1}^n P_i}.$$

Figure 7.2 shows the computed Lorenz curve. Looking at the x-axis and the around the fraction equaling 0.75, we see that the Lorenz curve crosses the y-axis as 0.2. This signifies that 75% of the population pays only 20% of the total portfolio, signifying that some customers pay a lot for their insurance. The dotted line represents the line of equality, where the curve would be situated if there was an equality between the fraction of the population and premium.

From the original Lorenz curve, which compares one distribution to the population, we now look at a Lorenz curve that compares two distributions over the same population. This thesis will plot the cumulative pure premium with the cumulative claim amount, i.e. cumulative loss. Given loss y_i for observation i, a prediction for the pure premium P_i, each graph is generated from the empirical cumulative distribution functions of the premium $\widehat{F}_P(x)$ and of the and loss $\widehat{F}_L(x)$, as given by

$$\widehat{F}_P(x) = \frac{\sum_i^k P_i I(P_i \leq x)}{\sum_i^k P_i}$$
$$\widehat{F}_L(x) = \frac{\sum_i^k y_i I(P_i \leq x)}{\sum_i^k y_i}.$$

The Lorenz curve is the graph resulting from $(\widehat{F}_P(x), \widehat{F}_L(x))$.

A standard measure often used with a Lorenz curve is the Gini coefficient. The coefficient is a metric that summarizes the Lorenz curve in a single number (Dorfman, 1979). It is related to the area between the line of equality and the Lorenz curve. Let G denote the Gini coefficient, A denotes the area in grey, and B denotes the area in blue in Figure 7.1. The Gini coefficient G is a measure of A's area in the

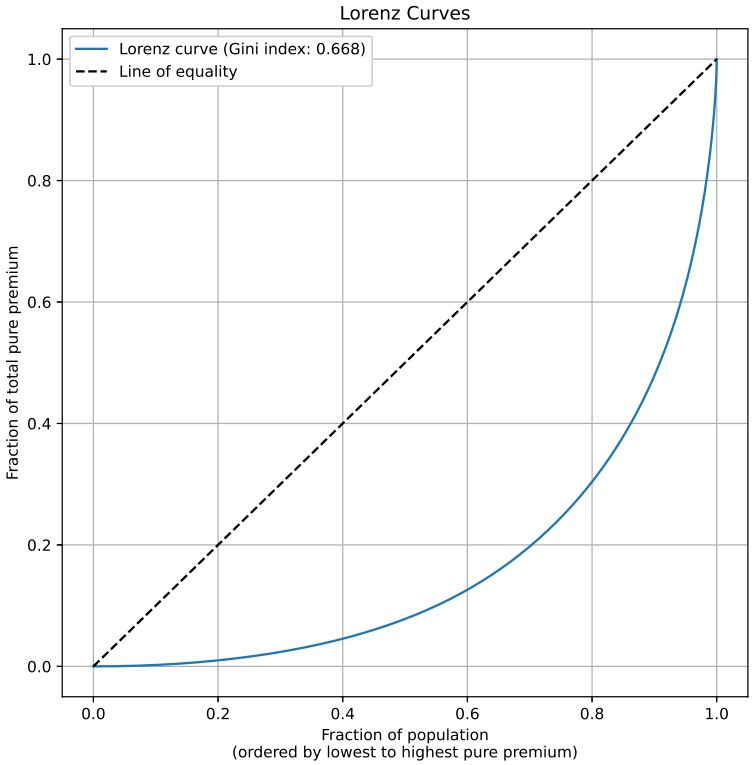


Figure 7.2: An example of the Lorenz curve for the pure premium in our insurance portfolio. On the x-axis we have an ordered proportion of the population, and on the y-axis we have the cumulative pure premium.

total area under the line of equality. Assuming that the x-axis and y-axis are of length 1, the total area given by $A+B$ is 0.5 and the Gini coefficient is given by

$$G = \frac{A}{A+B} = 2A = 1-2B.$$

A positive Gini coefficient indicates that the Lorenz curve is mainly below the line of equality. Conversely, we get a negative number when the curve is mainly above the line of equality. A larger Gini coefficient, both positively and negatively, implies a larger imbalance of the distributions. In the case of insurance, we aim for a high positive Gini coefficient, as this indicates an imbalance where the pure premium is larger than the loss, i.e. a profitable situation for the insurance company. The extreme case is when the Lorenz curve follows the x-axis and y-axis, giving $A = 0.5$ and $G = 1$. This represents a case where the insurance company has no loss but generates pure premiums.

Chapter 8

Model extensions

The overall goal is to predict the pure premium of an insured object, and in this chapter, we aim to model this with hierarchical models. As previously mentioned, the pure premium can be expressed as the product of the claim frequency and the average claim severity. In the following, we present the models of claim frequency and claim severity as they are modeled today, using credibility theory. These will be our baseline models. We also present two possible model extensions to this and later methods to estimate the parameters of these models.

8.1 Models for claim frequency

The claim frequency is normally modeled with the help of a Poisson process (Ohlsson & Johansson, 2010). Let Y_{ijt} denote the response variable, in this case, the claim frequency. The associated weight w_{ijt} is the duration of the policy, and X_{ijt} is the number of claims on the policy. We have $Y_{ijt} = X_{ijt}/w_{ijt}$. We assume that X_{ijt} is Poisson distributed with mean λ_{ij} when the duration $w_{ijt} = 1$. Writing the distribution of X_{ijt} in the form of an EDM gives us

$$f(x_{ijt}|\lambda_{ij}) = \exp(-w_{ijt}\lambda_{ij})\frac{(w_{ijt}\lambda_{ij})^{x_{ijt}}}{x_{ijt}!}, \text{ for } x_{ijt} = 0, 1, 2, ...$$

Knowing the relationship between the claim frequency Y_{ijt} and the claim amount X_{ijt} we can express the distribution of the claim frequency

$$\begin{aligned} f(y_{ijt}|\lambda_{ij}) &= P(X_{ijt} = w_{ijt}y_{ijt}) \\ &= \exp(-w_{ijt}\lambda_{ij})\frac{(w_{ijt}\lambda_{ij})^{w_{ijt}y_{ijt}}}{w_{ijt}y_{ijt}!} \\ &= \exp(w_{ijt}(y_{ijt}\theta_{ij} - \exp(\theta_{ij})) + c(y_{ijt}, w_{ijt})), \end{aligned}$$

where $\theta_{ij} = \ln(\lambda_{ij})$, $b(\theta_{ij}) = \exp(\theta_{ij})$, $\phi = 1$ and $c(y_{ijt}, w_{ijt}) = w_{ijt}y_{ijt}\ln(_{ijt}) - \ln((w_{ijt}y_{ijt}!))$. On this form, it is clear that we can build a GLMM around this distribution. The only thing missing is a link function relating the mean λ_{ij} and

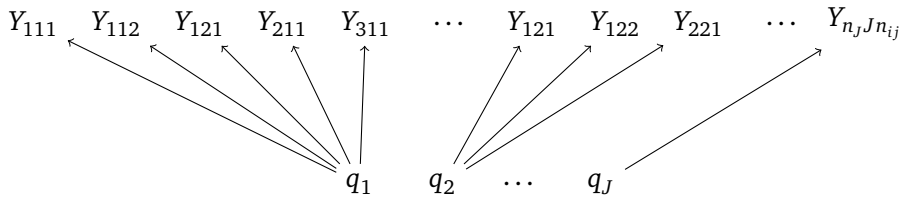


Figure 8.1: DAG of the proposed hierarchical model. The distribution of q is used in λ which is the parameter working on Y.

the linear predictor η_{ij}. The canonical link function can be identified by relating the canonical form $\eta_{ij} = g(\lambda_{ij}) = \theta_{ij}$ to the above equations. The canonical link function is the exponential function, i.e. $\exp(\theta_{ij}) = \lambda_{ij}$. Choosing this link function, we get a multiplicative model, which is a model where changes in the linear predictor will have a multiplicative effect on the mean. This is desirable as it is easy to interpret the effect of a parameter on the prediction, and we, therefore, use this going forward.

So, we now have a Poisson distributed response variable Y_{ijt}, and a link between the rate parameter λ_{ij} and the linear predictor η_{ij}, where the linear predictor is a linear combination of fixed and random effects. At this point, the model can be summarized as

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Poisson(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= x_i^T \beta + q_j. \end{aligned}$$

Our baseline model, which is in effect today, assumes that the random effects are identically and independently distributed with mean $\mathrm{E}[q_j] = 0$ and variance $\mathrm{Var}[q_j] = \tau^2$. However, a specific distribution is not assumed. We propose two extensions to this model. The first model is different only in the assumptions on the random effect. We assume $q_j \sim \mathcal{N}(0, \tau^2)$, instead of assumptions only about its moments. A DAG of the model can be seen in Figure 8.1, and the model can be summarized as

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Poisson(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= x_i^T \beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I). \end{aligned}$$

Such an assumption will allow for Bayesian inference on the posterior distribution of q_j. The reason for choosing the normal distribution is not that obvious, and there are reasonable arguments for choosing both Gamma and Beta distributions to model the random effect (Bühlmann & Gisler, 2005). We have however restricted ourselves the normal distribution because of its nice mathematical properties.

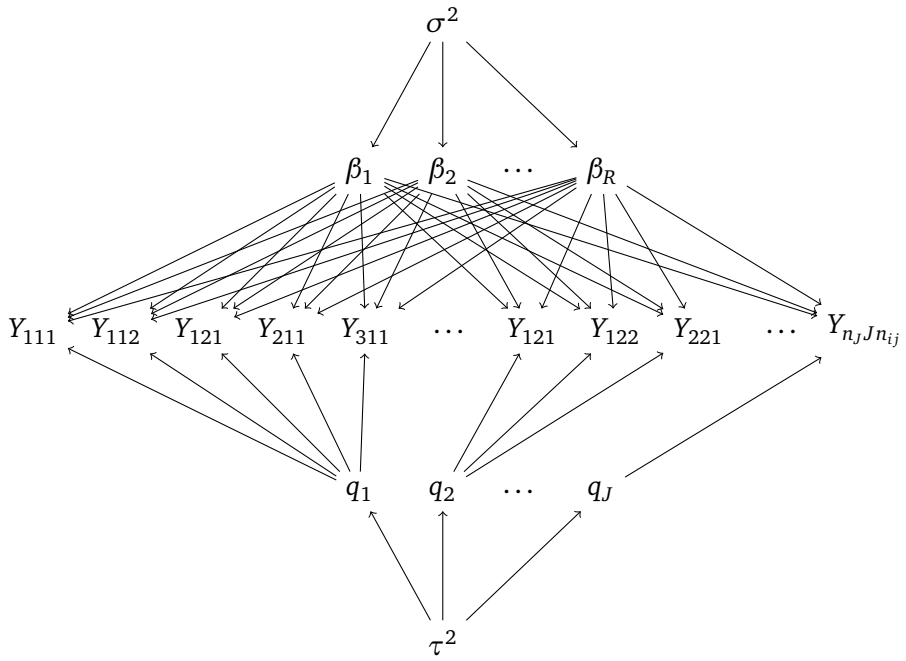


Figure 8.2: DAG of the proposed hierarchical model. The variance parameters σ^2 and τ^2 determine the distribution of β and q, respectively. These are in turn used to determine the distribution of Y_{ijt}.

The second model extends the previous where we also include distributional assumptions on the fixed effects and variance parameters. A DAG of this model can be seen in Figure 8.2, and the model can be summarized as

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Poisson(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= x_i^T \beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I), \\ \beta &\sim \mathcal{N}(0, \sigma^2 I), \\ \tau^2 &\sim IG(\alpha_\tau, \iota_\tau), \\ \sigma^2 &\sim IG(\alpha_\sigma, \iota_\sigma). \end{aligned}$$

As with the previous model, the choice of distributions is based chiefly on what is normal to find in the literature, as hierarchical models specified for insurance are not that common (Gamerman & Lopes, 2006). We have little prior information about what the variance parameters might look like, and we, therefore, want an objective prior distribution for these. The inverse gamma distribution is suitable as it is relatively objective when the shape and scale parameters are small. The inverse gamma distribution is also a conjugate prior for the normal distribution, which is helpful for the calculations to come.

8.2 Models for claim severity

Claim severity is often modeled with the help of a Gamma distribution (Ohlsson & Johansson, 2010). The distribution has support on $(0, \infty)$, and therefore we use only observations with previous claims when modeling. Let X_{ijt} denote the total cost of an observation with duration w_{ijt}. We then have the claim severity $Y_{ijt} = X_{ijt}/w_{ijt}$. In Example 4.1.2, we showed that the Gamma distribution can be written into the form of an EDM with the correct parametrization. Example 4.2.2 showed that the relation $\theta_{ij} = -1/\exp(\eta_{ij})$ is needed to obtain a multiplicative model.

The baseline model for claim severity is similar to the one seen for claim frequency. Once again, we assume the random effects q_j are identically and independently distributed with mean $\mathrm{E}[q_j] = 0$ and variance $\mathrm{Var}[q_j] = \tau^2$. The model can be summarized in the following form

$$\begin{aligned} f(Y_{ijt}|\mu_{ij}, \delta_{ij}) &\sim Gamma(\mu_{ij}, \delta_{ij}), \\ \ln(\mu_{ij}) &= x_i^T \beta + q_j, \end{aligned}$$

where δ_{ij} is assumed to be fixed and common for every combination of i and j. Our proposed model extension is identical to the one made for the claim frequency. The simpler model has the form

$$\begin{aligned} f(Y_{ijt}|\mu_{ij}, \delta_{ij}) &\sim Gamma(\mu_{ij}, \delta_{ij}), \\ \ln(\mu_{ij}) &= x_i^T \beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I). \end{aligned}$$

and the more complex model can be summarized as

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Gamma(\mu_{ij}, \delta_{ij}), \\ \ln(\mu_{ij}) &= x_i^T \beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I), \\ \beta &\sim \mathcal{N}(0, \sigma^2 I), \\ \tau^2 &\sim IG(\alpha_\tau, \iota_\tau), \\ \sigma^2 &\sim IG(\alpha_\sigma, \iota_\sigma). \end{aligned}$$

8.3 Parameter estimation

Having proposed three different models for our problem, we now identify suitable ways of estimating parameters in each model. A reminder of where we want to end up might be of aid at this stage. Remember that we want a prediction for the pure premium. The pure premium can be expressed as the product of claim frequency and average claim severity, i.e.

$$\text{Pure premium} = \text{Claim frequency} \times \text{Claim severity},$$

for a given observation. The models above are all multiplicative, meaning that changes in the linear predictor will have a multiplicative effect on the mean. Let $\widehat{Y}_{ijt}$ denote the prediction of the pure premium Y_{ijt}, following from the estimation of parameters in our models. Furthermore, let $\widehat{Y}^f_{ijt}$ and $\widehat{Y}^s_{ijt}$ denote the predicted claim frequency and severity, respectively. We choose to model the prediction of the pure premium as the prediction of the claim frequency multiplied by the prediction of the claim severity. More formally, we write it as

$$\begin{aligned}\widehat{Y}_{ijt} &= \widehat{Y}^f_{ijt} \times \widehat{Y}^s_{ijt} \\ &= \widehat{e^{\eta^f_{ij}}} \times \widehat{e^{\eta^s_{ij}}},\end{aligned}$$

where η^f_{ij} and η^s_{ij} are the linear predictor of the frequency and severity models, respectively. Writing out the linear predictor, we have

$$\widehat{Y}_{ijt} = \widehat{e^{\beta^f_0+\beta^f_1 I+\ldots+\beta^f_R I+q^f_j}} \times \widehat{e^{\beta^s_0+\beta^s_1 I+\ldots+\beta^s_R I+q^s_j}}.$$

In order to produce predictions of frequency and claim, we need estimates of β and q in both cases. Exactly how this is resolved will differ from model to model, so we return to this in the next chapters.

8.3.1 Claim frequency

We have presented three models for the claim frequency. The baseline model which is in use today, the simple model which adds distributional assumptions to the random effects of the baseline model, and a more complex model with additional distributional assumptions on fixed effects and variance parameters.

Baseline model and the backfitting algorithm

Estimation of parameters in the baseline model is done with the backfitting algorithm. We estimate the fixed effects with usual GLM fitting techniques, which is usually some variation of the least squares method. The random effects are estimated with the so-called credibility weights. This was introduced in Section 6.1.

The simple model and the MCEM algorithm

The next model added assumptions on the distribution of the random effect q. The model was stated as follows

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Poisson(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= x_i^T\beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I). \end{aligned}$$

For this case, we propose an implementation of the MCEM algorithm, as described in 5.4.1. We identify the vector of parameters ψ we aim to estimate. In our given context, $\psi = (\beta^T, \tau^2)$. Before the first iteration, we initialize the vector $\psi^{(0)} = (0^T, 0.1)$.

We need to sample from $f(q|y, \psi^{(t)})$, as the samples will be used to generate Monte Carlo estimates of the expected log-likelihood, which was defined as

$$Q_m(\psi|\psi^{(t)}) = \frac{1}{m}\sum_{v=1}^{m} \log f(y, q^v|\psi),$$

where q^v is sample number v of the vector q and m is the number of samples of q. Levine and Casella (2001) propose using importance sampling to reduce the computational expense. This will relieve them from sampling at every iteration of the algorithm, but instead, compute importance weights on an initial sample. The idea is proposed in a general setting, but we propose a possibly better way with our given model specification. It can be shown that $f(q|y, \psi^{(t)})$ can be written as

$$f(q|y, \psi^{(t)}) \propto f(q_1|y, \psi^{(t)}) f(q_2|y, \psi^{(t)}) ... f(q_J|y, \psi^{(t)}).$$

So q_r and q_l are conditionally independent given y and $\psi^{(t)}$ for $r \neq l$. We can show this by first finding an expression for $f(q|y, \psi^{(t)})$. Letting $h(q|\psi^{(t)})$ denote the assumed joint normal distribution of q, we have

$$\begin{aligned} & f(q|y, \psi^{(t)}) \\ \propto & f(y|q, \psi^{(t)}) h(q|\psi^{(t)}) \\ = & \prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} f(y_{ijt}|q_j, \psi^{(t)}) \prod_{j=1}^{J} h(q_j|\psi^{(t)}) \\ = & \prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} f(y_{ijt}|q_j, \psi^{(t)}) \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\tau^{(t)^2}}} \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \\ = & \left(\frac{1}{\sqrt{2\pi\tau^{(t)^2}}}\right)^J \prod_{j=1}^{J} \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} f(y_{ijt}|q_j, \psi^{(t)}). \end{aligned}$$

Having an expression for the conditional distribution of q, we can show that we can rewrite it on the form $f(q|y,\psi^{(t)}) = \rho f(q_1|y,\psi^{(t)})f(q_2|y,\psi^{(t)})...f(q_J|y,\psi^{(t)})$,

$$\begin{aligned}
&f(q|y,\psi^{(t)}) \\
&\propto \left(\frac{1}{\sqrt{2\pi\tau^{(t)2}}}\right)^J \prod_{j=1}^{J} \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} f(y_{ijt}|q_j,\psi^{(t)}) \\
&= \rho \prod_{j=1}^{J}\left[\exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} f(y_{ijt}|q_j,\psi^{(t)})\right] \\
&= \rho \prod_{j=1}^{J} f(q_j|y,\psi^{(t)}),
\end{aligned}$$

where $\rho = \left(1/\sqrt{2\pi\tau^{(t)2}}\right)^J$. For every j in $1,...,J$ we can sample q_j individually, by using

$$\begin{aligned}
&f(q_j|y,\psi^{(t)}) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} f(y_{ijt}|q_j,\psi^{(t)}) \\
&= \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} \exp\left(w_{ijt}(y_{ijt}\theta_{ij}^{(t)} - \exp(\theta_{ij}^{(t)})) + c(y_{ijt},w_{ijt})\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} \exp\left(w_{ijt}(y_{ijt}\theta_{ij}^{(t)} - \exp(\theta_{ij}^{(t)}))\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} \exp\left(w_{ijt}(y_{ijt}(x_i\beta^{(t)} + q_j) - \exp(x_i\beta^{(t)} + q_j))\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} \exp\left(w_{ijt}(y_{ijt}q_j - \exp(x_i\beta^{(t)} + q_j))\right).
\end{aligned}$$

There is a possibility to sample q by individually sampling $q_1,...,q_J$. These are univariate distributions, and MCMC might not offer the best solution. Adaptive rejection sampling is a possibility, but it requires $f(q_j|y,\psi^{(t)})$ to be log-concave for every j. We can find an expression of $\log(f(q_j|y,\psi^{(t)}))$ and then show that this is concave, which it is if and only if the second derivative is non-positive. We have

$$\begin{aligned}
&\log(f(q_j|y,\psi^{(t)})) \\
&= \log\left[\rho' \exp\left(-\frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2\right) \prod_{i=1}^{N}\prod_{t=1}^{n_{ij}} \exp\left(w_{ijt}(y_{ijt}q_j - \exp(x_i\beta^{(t)} + q_j))\right)\right] \\
&= \log(\rho') - \frac{1}{2}\left(\frac{q_j}{\tau^{(t)}}\right)^2 + \sum_{i=1}^{N}\sum_{t=1}^{n_{ij}} \left(w_{ijt}(y_{ijt}q_j - \exp(x_i\beta^{(t)} + q_j))\right),
\end{aligned}$$

where ρ' is some normalizing constant. Furthermore, we find

$$\frac{d}{dq_j}\log(f(q_j|y,\psi^{(t)})) = -\frac{q_j}{\tau^{(t)^2}} + \sum_{i=1}^{N}\sum_{t=1}^{n_{ij}} w_{ijt}(y_{ijt} - \exp(x_i\beta^{(t)} + q_j))$$
$$\Rightarrow \frac{d^2}{dq_j^2}\log(f(q_j|y,\psi^{(t)})) = -\frac{1}{\tau^{(t)^2}} - \sum_{i=1}^{N}\sum_{t=1}^{n_{ij}} w_{ijt}\exp(x_i\beta^{(t)} + q_j).$$

Both $\tau^{(t)^2}$ and w_{ijt} are positive quantities because of assumptions made about the initial distribution. Hence, $\frac{d^2}{dq_j^2}\log(f(q_j|y,\psi^{(t)}))$ is non-positive. Adaptive rejection sampling is, therefore, an available option. A possible algorithm is proposed in Gilks and Wild (1992). Having generated samples from $f(q|y,\psi^{(t)})$ at every iteration, we can compute the approximation to the expected log-likelihood $Q(\psi|\psi^{(t)})$.

The function $Q(\psi|\psi^{(r)})$ is not possible to compute analytically, and we proposed the function $Q_m(\psi|\psi^{(r)})$ as an approximation to this. We will have to find a workable expression of the function

$$Q_m(\psi|\psi^{(t)}) = \frac{1}{m}\sum_{v=1}^{m}\ln f(y,q_v|\psi).$$

We have

$$\begin{aligned}
&Q_m(\psi|\psi^{(t)})\\
=&\frac{1}{m}\sum_{v=1}^{m}\ln f(y,q_v|\psi)\\
=&\frac{1}{m}\sum_{v=1}^{m}\left[\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}}\left(a(\theta_{ij}^{(v)},y_{ijt},w_{ijt})\right) - \sum_{j=1}^{J}\ln\sqrt{2\pi\tau^2} - \frac{1}{2}\left(\frac{q_j^{(v)}}{\tau}\right)^2\right],
\end{aligned}$$

where $a(\theta_{ij}^{(v)},y_{ijt},w_{ijt}) = w_{ijt}(y_{ijt}(\theta_{ij}^{(v)}) - \exp(\theta_{ij}^{(v)})) + c(y_{ijt},w_{ijt})$. This is the expression we will use in the implementation, to maximize $Q_m(\psi|\psi^{(t)})$. As we aim maximize with respect to β and τ^2, we can split the above function into to separate optimization problems. We will optimize

$$Q_m^1(\psi|\psi^{(t)}) = \frac{1}{m}\sum_{v=1}^{m}\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}}\left(a(\theta_{ij}^{(v)},y_{ijt},w_{ijt})\right)$$

and

$$Q_m^2(\psi|\psi^{(t)}) = -\frac{1}{m}\sum_{v=1}^{m}\sum_{j=1}^{J}\ln\sqrt{2\pi\tau^2} - \frac{1}{2}\left(\frac{q_j^{(v)}}{\tau}\right)^2,$$

where $Q_m(\psi|\psi^{(t)}) = Q_m^1(\psi|\psi^{(t)}) + Q_m^2(\psi|\psi^{(t)})$. $Q_m^1(\psi|\psi^{(r)})$ is a function not having τ^2 as a variable. In fact, maximization of this function is equivalent to finding the MLE of the regression parameters for a generalized linear model with an offset (Chen, Zhang & Davidian, 2002). Maximizing $Q_m^2(\psi|\psi^{(t)})$ is reduced to maximizing with respect to only τ^2, reducing it to a univariate optimization problem.

A pseudocode, summarizing the mentioned steps, can be seen in Algorithm 6. The number of samples u generated at every iteration affects both the accuracy of the approximation and the computational time. Levine and Casella (2001) propose several approaches, but having $u = 10$ at the first iterations and increasing it as we get closer to convergence is recommended. They also discuss ways to estimate and reduce the unavoidable Monte Carlo error. This is pertinent, but we have looked away from this and instead kept a high u to minimize the damage done by this error.

Having run the algorithm, we are left with an approximated MLE $\widehat{\psi}$ and a sample of size u for each q_j. Seeing that the generations of q_j come from the conditional distribution $f(q_j|y, \psi^{(t)})$, we recognize that the draws are dependent on what we have observed and thus in line with what credibility theory aims for. An estimator of the random effect can be the mean of the samples from the last iteration of the algorithm.

Algorithm 6 MCEM algorithm for claim frequency

1: At $t = 0$ we initialize $\psi^{(0)}$.
2: **while** Not converged **do**
3: Sample $(q_j^{(1)}, q_j^{(2)}, ..., q_j^{(u)}) \sim f(q_j|y, \psi^{(t)})$ for $j = 1, ..., J$ using adaptive rejection sampling.
4: Maximize $Q_m^1(\psi|\psi^{(t)})$ as a usual GLM with offset.
5: Maxmimize $Q_m^2(\psi|\psi^{(t)})$ with some optimization algorithm for univariate functions.
6: $\psi^{(t+1)}) \leftarrow \psi$.
7: Increment t.
8: **end while**

The complex model and the MCMC algorithm

The complex model extends the previously defined model with the addition of a normally distributed β, and we can write the model in the following way

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Poisson(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= x_i^T\beta + q_j, \\ q &\sim \mathcal{N}(0, \tau^2 I), \\ \beta &\sim \mathcal{N}(0, \sigma^2 I), \\ \tau^2 &\sim IG(\alpha_\tau, \iota_\tau), \\ \sigma^2 &\sim IG(\alpha_\sigma, \iota_\sigma). \end{aligned}$$

Using MCMC and Gibbs sampling, as introduced in Section 5.3, we can sample from the full conditionals of q, β, τ^2 and σ^2. At the convergence of the chain, we have samples from the joint distribution of our variables and can use this to do inference. The mean will be especially useful when predicting new values for a given response variable with known covariates.

There are four distributions we aim to sample from with this model, namely the four full conditionals of q, β, τ^2 and σ^2. We go through each of them and look at how we can sample from them, before we summarize it all in a pseudocode that demonstrates how the algorithm has been implemented.

Letting $\theta = (\alpha_\tau, \iota_\tau, \alpha_\sigma, \iota_\sigma)$, a distribution proportional to the joint distribution $f(y, \beta, q, \sigma^2, \tau^2|\theta)$ can be found by Bayes' theorem. We have

$$f(y, \beta, q, \sigma^2, \tau^2|\theta) \propto f(y|\beta, q) f(\beta|\sigma^2) f(q|\tau^2) f(\sigma^2|\theta) f(\tau^2|\theta).$$

Using the fact that the full conditional distributions are proportional to the joint distribution, we can express distributions proportional to the full conditionals as well. For $f(\beta|y, q, \sigma^2, \tau^2, \theta)$ we have

$$f(\beta|y, q, \sigma^2, \tau^2, \theta) \propto f(y|\beta, q) f(\beta|\sigma^2).$$

Similarly, for the full conditional of q, we have

$$f(q|y, \beta, \sigma^2, \tau^2, \theta) \propto f(y|\beta, q) f(q|\tau^2).$$

Reusing the methods from Section 8.3.1, it is straightforward to show that both distributions are log-concave. Adaptive rejection sampling is thus a viable option for sampling from the two distributions.

Left are sampling from the full conditional distributions of σ^2 and τ^2. Here, we can exploit the concept of conjugate prior in Bayesian statistics, as introduced in Section 3.3. When we use the inverse gamma distribution as a prior distribution,

and the likelihood function is a normal distribution, the posterior will also be inverse gamma. We showed this in Example 3.3.1. Therefore, the posterior distribution of both σ^2 and τ^2 are inverse gamma.

A pseudocode, summarizing how we run this MCMC method is given in Algorithm 7. Exiting the algorithm, we are left with samples from a distribution proportional

Algorithm 7 MCMC algorithm for claim frequency

1: At $t = 0$ we initialize values for $\beta^{(0)}, q^{(0)}, \sigma^{2^{(0)}}, \tau^{2^{(0)}}, \theta$.
2: **while** Not converged **do**
3: Sample $q_j^{(t)} \sim f(q_j|y, \beta^{(t-1)}, \sigma^{2^{(t-1)}}, \tau^{2^{(t-1)}}, \theta)$ for $j = 1, ..., J$ using adaptive rejection sampling.
4: Sample $\beta_r^{(t)} \sim f(\beta_r|y, q^{(t)}, \sigma^{2^{(t-1)}}, \tau^{2^{(t-1)}}, \theta)$ for $r = 1, ..., R$ using adaptive rejection sampling.
5: Sample $\sigma^2 \sim IG(\alpha_\sigma + \frac{R}{2}, \iota_\sigma + \sum_r^R \frac{\beta_r^{(t)^2}}{2})$
6: Sample $\tau^2 \sim IG(\alpha_\tau + \frac{J}{2}, \iota_\tau + \sum_j^J \frac{q_j^{(t)^2}}{2})$
7: Increment t.
8: **end while**

to the limiting distribution of our chain. In this case that is the joint distribution $f(y, \beta, q, \sigma^2, \tau^2|\theta)$.

8.3.2 Claim severity

Estimation of parameters in the case of claim severity is not much different from the case of claim frequency. The difference between the models is the assumed distribution of the response variable, which now is Gamma rather than Poisson. As we shall see, this does have some implications on the implementation, as the densities that were log-concave in the previous section no longer possess this property.

Baseline model and the backfitting algorithm

The backfitting algorithm is reused for the baseline model. The only difference from Section 8.3.1 is that the response is modeled as a Gamma distribution rather than a Poisson distribution.

The simple model and the MCEM algorithm

The simple model was summarized as

$$\begin{aligned} f(Y_{ijt}|\mu_{ij},\delta_{ij}) &\sim Gamma(\mu_{ij},\delta_{ij}),\\ \ln(\mu_{ij}) &= x_i^T\beta + q_j,\\ q &\sim \mathcal{N}(0,\tau^2 I). \end{aligned}$$

We will still rely on the MCEM algorithm, but we are forced to make some adaptions. We do not describe every step as extensively as with the claim frequency in Section 8.3.1, but we can highlight some of the main differences.

The first thing that can be shown is that the q_j's are still independent. Their distribution is, however, not log-concave, and adaptive rejection sampling is therefore not an option. In this case, we have resorted to the random walk Metropolis chain, as introduced in 5.3.1. Remember that we only model claim severity with observations having made some claims already due to the support of the Gamma distribution. Hence, the data set is considerably smaller than in the case of modeling claim frequency, and the computational burden is not that large. Aside from that, the maximization step can be done in the same way by optimizing to functions Q_m^1 and Q_m^2.

The complex model and the MCMC algorithm

The more complex model was given as

$$\begin{aligned} f(Y_{ijt}|\lambda_{ij}) &\sim Gamma(\mu_{ij},\delta_{ij}),\\ \ln(\mu_{ij}) &= x_i^T\beta + q_j,\\ q &\sim \mathcal{N}(0,\tau^2 I),\\ \beta &\sim \mathcal{N}(0,\sigma^2 I),\\ \tau^2 &\sim IG(\alpha_\tau,\iota_\tau),\\ \sigma^2 &\sim IG(\alpha_\sigma,\iota_\sigma). \end{aligned}$$

As with the claim severity in the case of the simple model, we cannot show that the full conditional distributions of β and q are log-concave. We will therefore resort to random walk Metropolis when sampling from these distributions. As the conditional distribution of Y_{ijt} has no impact on the posterior distribution of the variance parameters σ^2 and τ^2, we can still use a Gibbs step to generate samples from their distributions. A pseudocode is provided in Algorithm 8.

8.4 Implementation

The baseline model of the claim frequency will be used together with the baseline model of the claim severity. Similarly, the simple models will be used together,

Algorithm 8 MCMC algorithm for claim severity

1: At $t = 0$ we initialize values for $\beta^{(0)}, q^{(0)}, \sigma^{2(0)}, \tau^{2(0)}, \theta$.
2: **while** Not converged **do**
3: Sample $q_j^{(t)} \sim f(q_j|y, \beta^{(t-1)}, \sigma^{2(t-1)}, \tau^{2(t-1)}, \theta)$ for $j = 1, ..., J$ using random walk Metropolis.
4: Sample $\beta_r^{(t)} \sim f(\beta_r|y, q^{(t)}, \sigma^{2(t-1)}, \tau^{2(t-1)}, \theta)$ for $r = 1, ..., R$ using random walk Metropolis.
5: Sample $\sigma^2 \sim IG(\alpha_\sigma + \frac{R}{2}, \iota_\sigma + \sum_r^R \frac{\beta_r^{(t)^2}}{2})$
6: Sample $\tau^2 \sim IG(\alpha_\tau + \frac{J}{2}, \iota_\tau + \sum_j^J \frac{q_j^{(t)^2}}{2})$
7: Increment t.
8: **end while**

and the complex models with each other. To ease understanding in the coming chapters, we will refer to the models by the name of their methods of estimation, i.e. the baseline model will be called the credibility model, the simple model is the MCEM model, and the complex model will be called the MCMC model. This should also highlight the most significant differences between the models and ways of estimation.

Section 8.3 discussed how we would use our methods to estimate parameters and predictions of frequency and severity. In the case of the credibility model, the algorithm leaves us with estimates of all parameters, e.g. we get $\widehat{\beta}_0, ..., \widehat{\beta}_R, \widehat{q}_1, ..., \widehat{q}_J$ when modeling. This can be directly plugged in and used to predict.

For the MCEM model, we are left with estimates for the fixed effects, i.e. $\widehat{\beta}_0, ..., \widehat{\beta}_R$, and samples from the posterior distribution of $q_1, ..., q_J$. We have resolved this by using the mean of these distributions to estimate $\widehat{q}_1, ..., \widehat{q}_J$.

Finally, we have the MCMC model, which produces samples for both the fixed and random effects. Given k samples, a way to predict the claim frequency or severity is given by

$$\begin{aligned}\widehat{Y}_{ijt} &= \widehat{e^{\eta_{ij}}} \\ &= \frac{1}{k}\sum_{a=1}^{k} e^{\eta_{ij}^{(k)}} \\ &= \frac{1}{k}\sum_{a=1}^{k} e^{\beta_0^{(k)} + \beta_1^{(k)} I + ... + \beta_R^{(k)} I + q_j^{(k)}}.\end{aligned}$$

A disadvantage with this approach is the computational time needed when k is large. An alternative approach is to compute means from every distribution and

use these directly. That is

$$\begin{aligned}\widehat{\widehat{Y}}_{ijt} &= \widehat{\widehat{e^{\eta_{ij}}}} \\ &= e^{\widehat{\beta}_0+\widehat{\beta}_1 I+\ldots+\widehat{\beta}_R I+\widehat{q}_j}.\end{aligned}$$

The former is ideal, but the latter is faster. Testing the two estimators on 100000 observations, we have not found a big difference in their mean squared and absolute error. More testing could have been carried out, but it has provided some reassurance about the usefulness of $\widehat{\widehat{Y}}_{ijt}$. In the following chapter, we use $\widehat{\widehat{Y}}_{ijt}$ for the predictions related to the MCMC model.

Chapter 9

Results and findings

We have looked at the methods necessary to predict claim frequency and claim severity with the MCEM and MCMC models. All methods have been written and implemented on real data in Python, yielding predictions that we will investigate in this chapter. The original data set has been divided into a training and a test set, where the parameters were estimated with the training set, and our results come from predictions made on the test set.

Initially, we look into the correctness of our algorithms with the help of simulations. Next, we dissect our results by methods frequently used in the insurance industry. By doing this, we gain insight into both similarities and differences between the methods that can be thought to be significant in practice. Finally, we look into some of the assumptions made in our methods and test these.

9.1 Simulation study

In a simulation study, we aim to test the algorithm's ability to estimate true parameter values. In reality, it can be hard to distinguish flaws in our assumptions and our method of estimating parameters. By doing a simulation study, we consider our assumptions to be true, meaning that a bad outcome must be linked to problems with the algorithm and its implementation.

For the MCEM model, we chose somewhat arbitrary values of β and the variance parameter τ^2, while q is chosen to reflect the variance τ^2. For the MCMC model, both β and q have assumed distributions, and we, therefore, chose values for σ^2 and τ^2 and generate independent samples from the normal distribution for β and q. That is, we draw $\beta \sim \mathcal{N}(0, \sigma^2)$ and $q \sim \mathcal{N}(0, \tau^2)$, for chosen values of σ^2 and τ^2. Modeling the claim frequency, we draw Y^f_{ijt} from a Poisson distribution with mean $\lambda_{ij} = \exp(x_i^T \beta + q_j)$, where x_i is the corresponding covariate values of the fixed effects of observation Y_{ijt}. Similarly, we draw Y^s_{ijt} from a Gamma distribution with mean $\mu_{ij} = \exp(x_i^T \beta + q_j)$ for the claim severity. Having done that,

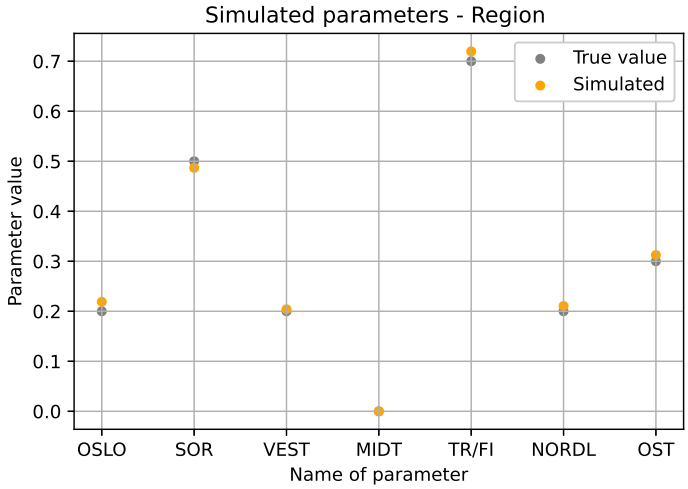


Figure 9.1: We see how the simulated region parameters match the true ones in the MCEM model for claim frequency.

we can multiply the two quantities Y_{ijt}^f and Y_{ijt}^s to obtain the pure premium Y_{ijt} of every observation.

Running the algorithms, we are left with estimated parameter values that can be compared to the true values set initially. In Figure 9.1 we see how the values estimated for the MCEM model match the true values. An equivalent figure for the MCMC model can be seen in Figure 9.2. A major difference is, of course, that the β's estimated for the MCEM model are an estimation of actual parameters. In contrast, for the MCMC model, the estimated parameters are the mean of the posterior distribution of β. The estimated parameter values are used for the same purpose, but their origin is fundamentally different.

Figure 9.3 shows the estimated random effect parameter values in the MCEM model. The simulated values come very close to the true values, but what is especially interesting is the consistent tendency to overestimate the true value slightly. This has been recurring in the simulation study, where simulated values are close to, but not perfectly in line with the true values. We believe that it is connected to the choice of base level and that a different base level would give a slightly different perception of the simulated values. Looking at the plots, it is clear that the methods do well in finding the true value, even though they sometimes miss by a small constant.

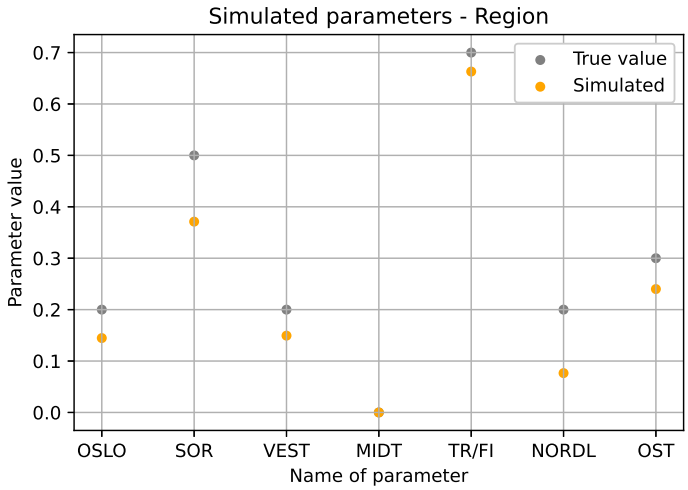


Figure 9.2: As with the previous figure, we see how the simulated region parameters match the true ones in the MCMC model for claim frequency.

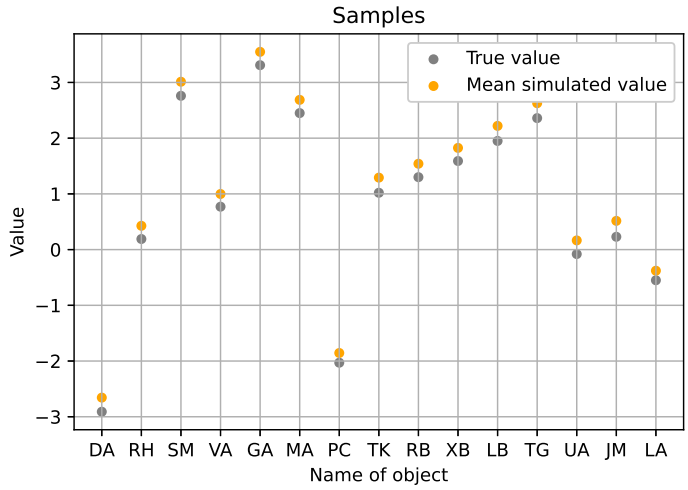


Figure 9.3: We see how the mean of the posterior distribution of the samples match the true parameter values in the MCEM model. The x-axis gives the group, and the y-axis indicate the parameter value.

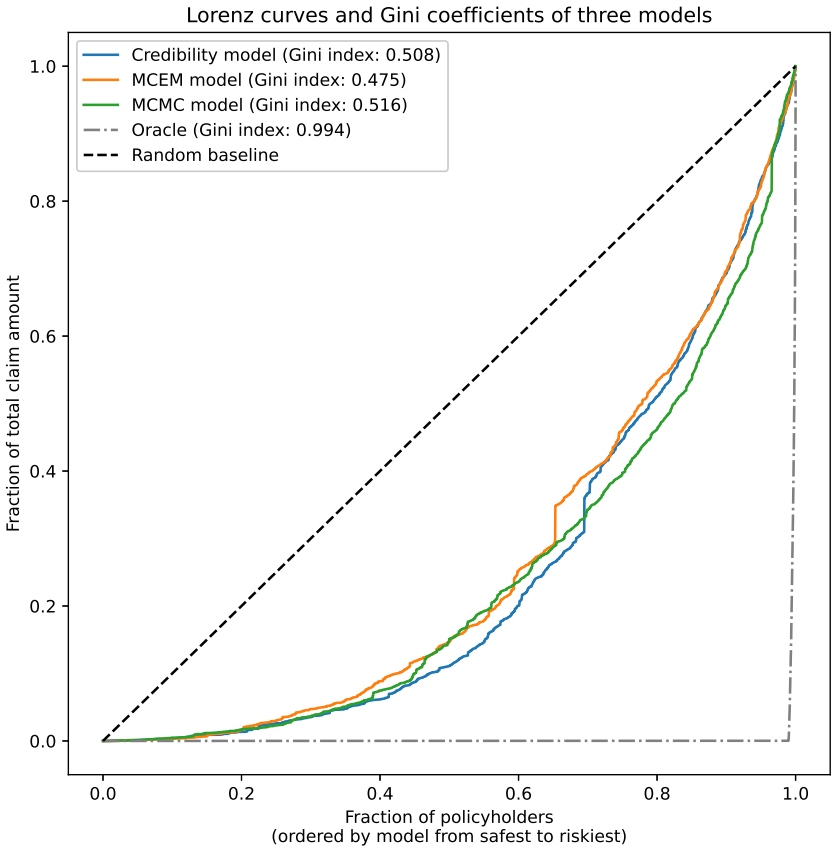


Figure 9.4: The resulting three Lorenz curves from the credibility, MCEM and MCMC models. The Gini coefficient of each curve is given in the legend.

9.2 Lorenz curve and Gini coefficient

This section uses the Lorenz curve and Gini coefficient to distinguish the models' ability to rank the risks. Figure 9.4 plots the classical Lorenz curves coming from the credibility, MCEM, and MCMC models. They are very similar, which is what we would expect, considering the almost identical assumptions that have been made in every model. Looking at the curves directly, it can be challenging to distinguish and point out a "best" curve, i.e. a curve creating the biggest area between itself and the diagonal. However, the Gini coefficients are of help, and using them, the MCMC method gives the best result, followed by the credibility model. Despite their similarities, we can, from this test, conclude that the MCMC model is just as good or even better at ranking the risk than the traditional credibility model. Next, we look at how the actual predictions match the true outcome and use this

to assess the three models further.

9.3 Mean error and risk ratio

In this section we investigate how the predictions of test data compares to the true values. A common metric in this case is the mean squared error (MSE) and mean absolute error (MAE), which allow us to rank the three models by a single number. A single number allow us to quickly distinguish the models, but it does not reveal the nuisances between three similar models. For this very reason, we will also use *risk ratio*, which will be defined later on, to investigate how the three models compare when looking at each parameter of the model.

The mean squared error is given by

$$MSE = \frac{1}{I \times J \times n_{ij}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} (y_{ijt} - \widehat{y}_{ijt})^2,$$

where y_{ijt} is the observed pure premium and $\widehat{y}_{ijt}$ our prediction of the pure premium. Computing this metric for each of our three models, we get an MSE equal to $1.14 * 10^7$ for the credibility model. The MCEM model yields an MSE of $1.17 * 10^7$ and the MCMC model an MSE of $1.13 * 10^7$. Comparing the three models, it is easy to conclude that the MCMC model outperforms the other two. The difference is however marginal and does not provide a conclusive answer in any way. The MSE gets very big because we have some very large claims that we in no way are able to predict. Squaring these terms results in a high overall MSE, and it naturally gives a lot of weight to the largest claims. An alternative is the MAE, which is given by

$$MSE = \frac{1}{I \times J \times n_{ij}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{n_{ij}} |y_{ijt} - \widehat{y}_{ijt}|.$$

This metric will not be as influenced by the largest claims, and may give a better understanding of how the models actually differ. The resulting MAE of the credibility model is 424.9, where the MAE of the MCEM and MCMC models are 424.0 and 425.9 respectively. The MCEM model outperforms the other two, but yet again, the difference is too small to give any real feeling of their differences.

The *risk ratio* is a quantity, defined as the ratio between the true pure premium divided by the predicted pure premium. If we predict a pure premium of 200 for costumer A, and the customer has a realized pure premium of 100, the risk ratio is 0.5. Such a situation is desirable for the insurance company, as opposed to a case where the risk ratio is above one and will result in a loss. We now use the risk ratio to measure how well the models model each parameter in the model, i.e. all β's and q's.

In a very simple world, we could use our predictions for the pure premium as the price we set on the insurance. A perfect model would thus give a risk ratio of 1. In the real world however, there are other costs included in the price, other than just the predicted loss. These are often referred to as operational costs. Thus, our aim should not be a risk ratio of 1, but a ratio so low that the total ratio sums to 1 when operational costs are included. The goal of this thesis is not the determination of this level, so to avoid this problem, we use the observed risk ratio on the current model to level out our three new models. That is, the insurance company may have a risk ratio of 0.6 on the existing portfolio. We assume that this is the target ratio for all of our models. If we sum all predicted pure premiums in a model, we get the total predicted pure premium. Summing all observed pure premiums, we get the total cost of the observations. Dividing the total cost of the observations over the total predicted pure premium, we get the risk ratio of that model. We expect this number to differ for all three models, but in reality, we want them to be the same, so that we compare them on the same ground. That is, we do not rank the models on their ability to predict the total pure premium, but rather on their ability to classify the risks. Hence, we multiply all the predicted pure premiums of a model by a constant to ensure a common risk ratio for all models, e.g. a risk ratio equal to 0.6. By doing this, we ensure that all three models predict the same total pure premium. Their only difference is in their distribution of pure premium across the observations.

In the following, we will look at the risk ratio for each parameter of our model, e.g. the risk ratio of the parameter for "20<Age<30". That is we look at all observations where the age is equal to ""20<Age<30". This will give greater insight into how the model distributes the risk over the portfolio. To make the results more intuitive, we divide the risk ratio for a single parameter by the common risk ratio of all models. As an example, the common risk ratio might be 0.6, as mentioned before. Taking the credibility model and its predictions and observations of pure premium in the group of "20<Age<30", we can compute a risk ratio solely for this parameter. We can then divide the risk ratio of "20<Age<30" by the common risk ratio, and produce what we would call the *normalized risk ratio*. If the normalized risk ratio is equal to 1, we have the same risk ratio for the parameter "20<Age<30" as we have for the entire portfolio. This is desirable. To understand why we would want this, we could imagine the opposite case. Summing over the entire predicted pure premiums of a model, we know that the normalized pure premium will be equal to 1. This follows from our multiplication of the predictions by a constant that ensures the total cost divided by the total pure premium to equal the common ratio. Thus, if we have a parameter, e.g. "20<Age<30", with a normalized risk ratio of 0.5, there must necessarily be another parameter which compensates for this deviation. Looking at the plots which follow, it is thus desirable with normalized risk ratios close to 1.

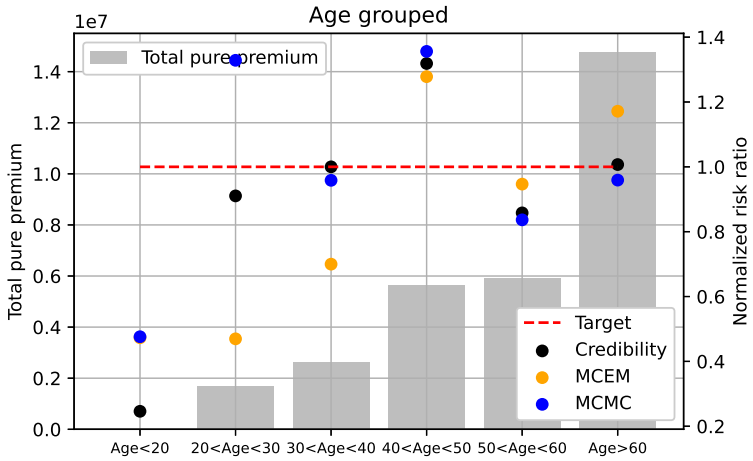


Figure 9.5: The normalized risk ratio of the three models for the age parameters of the model. The red dotted line signifies the desired level equal to 1, and the dots represent each model. The total pure premium is given by the grey bars which correspond to the left side y-axis, while the normalized risk ratios match the right hand y-axis.

Age - normalized risk ratio	Mean absolute error	Mean squared error
Credibility model	0.22	0.17
MCEM	0.31	0.13
MCMC	0.24	0.09

Table 9.1: The MAE and MSE of the normalized risk ratio of the three models for the age parameter.

Figure 9.5 portrays the normalized risk ratio of all age parameters for all three models. There is a lot of information in the plot, but looking at each level for itself, e.g. "20<Age<30" we can determine which model that outperforms the other two. For the level "20<Age<30" it is clear that the credibility model outperforms both the MCEM and MCMC models. The MCEM model scores just under 0.4, signifying that the risk ratio is lower than our desired ratio which again signifies that this parameter is more expensive than it could have been. On the contrary, the MCMC model is too cheap at this level. Looking at the plot, it is easy to determine how the model does for single parameters, but it can be difficult to draw any conclusion for all parameters related to age. As every observation belongs to some age-group, we know that the mean normalized risk ratio will be 1. We can however compute the distance from the dots of each model to the red dotted line. We can use means of both absolute and squared distance to quantify the performance in a plot. Table 9.1 shows the scores related to Figure 9.5. The models are quite similar, but it can be argued that the MCMC model outperforms the two other

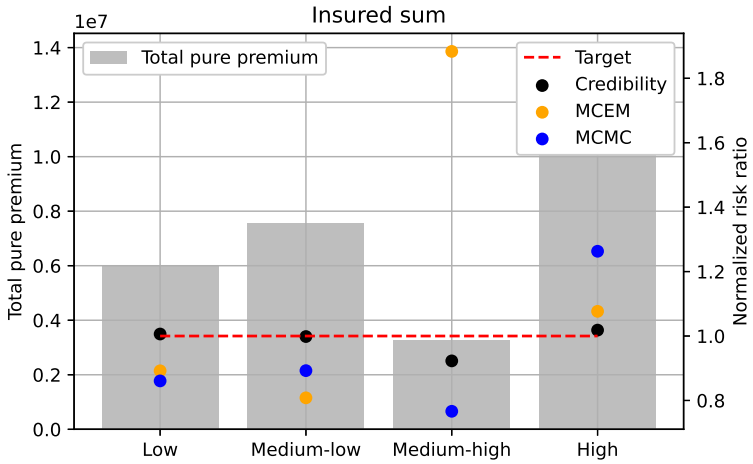


Figure 9.6: The normalized risk ratio of the three models for the insured sum parameters of the model. The red dotted line signifies the desired level equal to 1, and the dots represent each model. The total pure premium is given by the grey bars which correspond to the left side y-axis, while the normalized risk ratios match the right hand y-axis.

Insured sum - normalized risk ratio	Mean absolute error	Mean squared error
Credibility model	0.03	0.01
MCEM	0.31	0.21
MCMC	0.19	0.04

Table 9.2: The MAE and MSE of the normalized risk ratio of the three models for the insured sum parameter.

when considering both the quantities.

A similar plot and table for the insured sum can be found in Figure 9.6 and Table 9.2. Simply looking at Figure 9.6, it is obvious that the credibility model vastly outperforms the other two. This is impression can also be confirmed by looking at Table 9.2. We also note that the MCMC model is closer to the credibility model, and that the MCEM model struggle somewhat.

The parameters for the customer rating can be seen in Figure 9.7. As with the insured sum, it is quite easy to find the best model from the figure alone. The big difference is that it is the MCMC model which gives the best result. Table 9.3 confirms this, and also highlight the bad predictions made by both the credibility model and the MCEM model. It is strange that they struggle so much for the lower rated customers, while the MCMC model manage much better. A possible

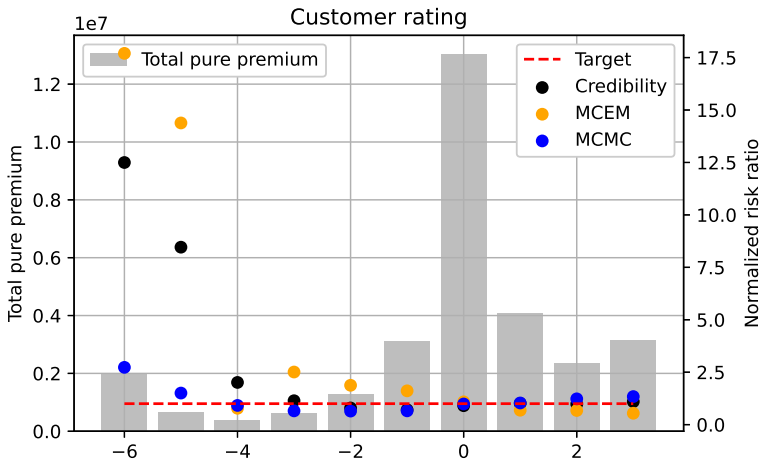


Figure 9.7: The normalized risk ratio of the three models for the rating parameters of the model. The red dotted line signifies the desired level equal to 1, and the dots represent each model. The total pure premium is given by the grey bars which correspond to the left side y-axis, while the normalized risk ratios match the right hand y-axis.

Rating - normalized risk ratio	Mean absolute error	Mean squared error
Credibility model	2.08	18.90
MCEM	3.45	46.19
MCMC	0.40	0.38

Table 9.3: The MAE and MSE of the normalized risk ratio of the three models for the customer rating parameter.

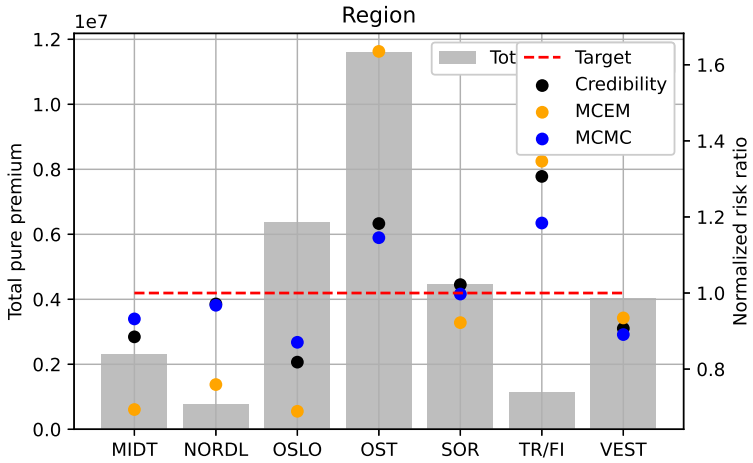


Figure 9.8: The normalized risk ratio of the three models for the region parameters of the model. The red dotted line signifies the desired level equal to 1, and the dots represent each model. The total pure premium is given by the grey bars which correspond to the left side y-axis, while the normalized risk ratios match the right hand y-axis.

Region - normalized risk ratio	Mean absolute error	Mean squared error
Credibility model	0.13	0.03
MCEM	0.28	0.11
MCMC	0.10	0.01

Table 9.4: The MAE and MSE of the normalized risk ratio of the three models for the region parameter.

explanation could be that the credibility model and MCEM model are both estimated with usual GLM fitting techniques, e.g. some least squares procedure, but the MCMC parameter has some more flexibility with it being a mean of some known distribution.

Next in line is the region parameters, represented in Figure 9.8. Reading the plot, it once again becomes clear that the MCEM model struggles to match the credibility and MCMC model. It is difficult to separate the credibility and MCMC model, but looking at Table 9.4 we see that the MCMC just outperform the credibility model.

Finally, the models ability to predict the objects can be seen in Figure 9.9. There are 15 objects, giving 45 dots in the plot. It is therefore difficult to make any conclusion just by looking at the plot. Table 9.5 can be of some aid. The MCEM model has some trouble matching its two competitors. The credibility and MCMC model

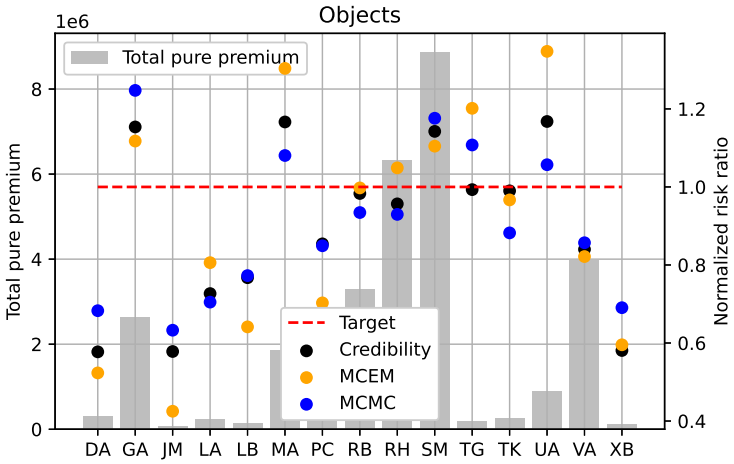


Figure 9.9: The normalized risk ratio of the three models for the object parameters of the model. The red dotted line signifies the desired level equal to 1, and the dots represent each model. The total pure premium is given by the grey bars which correspond to the left side y-axis, while the normalized risk ratios match the right hand y-axis.

Object - normalized risk ratio	Mean absolute error	Mean squared error
Credibility model	0.19	0.05
MCEM	0.24	0.09
MCMC	0.18	0.04

Table 9.5: The MAE and MSE of the normalized risk ratio of the three models for the object parameter.

on the other hand continues to match each other. The MCMC model performs slightly better, but the difference is almost negligible.

9.4 Interpretation of the results

Taking a step back, we now spend some time reflecting on the results from a broader perspective. The previous sections have demonstrated both similarities and differences between the models. The question is now if we can draw some conclusions about the possible superiority of one of the models.

Section 9.2 used Lorenz curves and Gini coefficients to compare the results from the three models. Because of the similarity, reading the plot was not straightforward, but the Gini coefficient revealed that the MCMC model was slightly better than the credibility and MCEM models. Section 9.3 gave us insight into the models' ability to predict using traditional statistical tools, such as the mean absolute and squared error. It became clear that the credibility and MCMC model often has an almost identical performance, while the MCEM model tended to do worse. That is somewhat counter-intuitive, as the MCEM model lies somewhere in between the credibility and MCMC model in its assumptions. Deciding on a "best" model based on the above sections, it seems reasonable to conclude that the MCMC model outperforms the other two. However, the difference between the MCMC and credibility model is so tiny that some reflection about other factors than just their ability to predict is necessary.

The credibility model is popular for several reasons, one of which is its superiority in computational efficiency. Computation of point estimates is faster than the generation of samples from a distribution we cannot sample from directly. In light of this, the usefulness of the MCMC model will depend on the user's interests and necessities. If one is to update the pure premium prediction frequently and thus needs to run the MCMC algorithm often, it can be problematic with how the method is proposed in this thesis. If one wishes to update the predictions less frequently, say once every week or month, the MCMC model should not be overlooked.

If one concludes that the MCMC model is superior to the credibility model, there is also an array of possibilities to improve computational time. Looking away from trivial optimization of the current implementation, there are methods of parameter estimation that are more efficient than the traditional MCMC algorithm. Integrated nested Laplace approximation (INLA) is an attractive option, but it also forces some specific requirements on the structure of the model (Rue, Martino & Chopin, 2009). Without going into too much detail, the requirements should not pose any problem with our current model specification, thus making INLA an available option that is likely to reduce the computational time.

There is also a case to be made about the little exploration that has been made in the MCMC model. We have not explored or truly tested the choice of prior distributions or the value of the hyperparameters. It is therefore not impossible that further improvements can be made with the MCMC model. There is also the possibility of generating more samples and thus reducing the Monte Carlo error, introduced when we compute the mean of the different distributions. All in all, there is a case to be made for the MCMC model. It seems likely that further development of the model and implementation of modern parameter estimation methods can yield better predictions and reduce the computational burden of the Bayesian hierarchical model.

Chapter 10

Closing remarks

Prediction of the pure premium is of high importance to insurance companies. Prediction of a pure premium that is too high will likely lead customers to find insurance with a better price elsewhere. If the prediction is too low, the company risks attracting bad customers and high risk. This thesis has proposed two model extensions to the current model. We have also proposed methods of estimating the two models.

Chapter 6 introduced credibility theory and explained how it had found its place in modern statistics. We argued that it is a way of estimating what are essentially random effects in a GLMM. Building on this perspective, we proposed two model extensions in Chapter 8. Where the credibility theory only made assumptions on the first and second moment of the random effects, we extended the idea by initially assuming the random effects to be normally distributed with zero mean and common variance τ^2. This resulted in the simple model, later referred to as the MCEM model. In the next model, we assumed the fixed effects to be random and come from a normal distribution with zero mean and common variance σ^2. We also added a prior distribution to the variance parameters σ^2 and τ^2. This resulted in the complex model, later referred to as the MCMC model.

Having decided on models and methods of estimation, we tested the models' ability to predict the pure premium on test data. Chapter 9 presented the results and ranked the models by various metrics. The big question is if we have managed to propose a model that beats the current model. The short answer to that question is maybe. Section 9.4 discussed and presented a more nuanced answer to the question. Suppose the sole goal is to make the best possible prediction of the pure premium. In that case, the MCMC model appears to be suitable and competitive with the credibility model employed today. However, in a broader perspective, one might ask if the small gains achieved by using the MCMC model might be absorbed by the potential cost of implementing and maintaining this model. Another remark is that training and test set choice will always influence the parameter estimation, predictions, and interpretation of the results. With the results being so

close, one could wonder how the results would have been on a new data or a different portfolio.

As we have decided to work with extensions of the credibility model, one might argue that other models or statistical methods should be tested on insurance data. The models and methods we have studied are, after all, several decades old. Our counterargument is that the multiplicative GLMMs are explainable and very intuitive, not only for statisticians and analysts but also for their colleagues and customers. For this reason, we recommend insurance companies using the credibility model to look at the possibility of also building and testing Bayesian hierarchical models on different portfolios. It is viable to construct models as we have proposed in this thesis, but if the models are to be used on a bigger scale, we would recommend optimizing the estimation methods.

References

Bühlmann, H. & Gisler, A. (2005). *A Course in Credibility Theory and its Applications*. Springer-Verlag.

Casella, G. & Berger, R. (2002). *Statistical Inference*. Duxbury Pacific Grove, CA.

Chen, J., Zhang, D. & Davidian, M. (2002). A Monte-Carlo EM algorithm for generalized mixed models with flexible random effects distribution. *Biostatistics (Oxford, England)*, *3*, 347-60.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag.

Dongarra, J. & Sullivan, F. (2000). Guest editors introduction to the top 10 algorithms. *Computing in Science & Engineering*, *2*, 22-23.

Dorfman, R. (1979). A formula for the Gini coefficient. *The Review of Economics and Statistics*, *61*, 146–149.

Frontczak, R., Jaeger, M. & Schumacher, B. (2017). From power curves to discriminative power: Measuring model performance of LGD models. *Journal of Mathematical Finance*, *07*, 657–670.

Gamerman, D. & Lopes, H. F. (2006). *Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica*, *39*, 1037–1039.

Gelfand, A. E. (2000, December). Gibbs sampling. *Journal of the American Statistical Association*, *95*, 1300–1304.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*, 721-741.

Gilks, W. R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *41*, 337–348.

Givens, G. H. & Hoeting, J. A. (2012). *Computational Statistics*. Wiley.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Hickman, J. C. & Heacox, L. (1999). Credibility theory. *North American Actuarial Journal*, *3*, 1–8.

Klugman, S. (1987). Credibility for classification ratemaking via the hierarchical normal linear model. *Proceedings of the Casualty Actuarial Society*, *74*, 272-321.

Levine, R. A. & Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, *10*, 422–439.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*, 209–219.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953, June). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087–1092.

Metropolis, N. & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335–341.

Nelder, J. & Verrall, R. (1997). Credibility theory and generalized linear models. *ASTIN Bulletin*, *27*, 71–82.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*, 370–384.

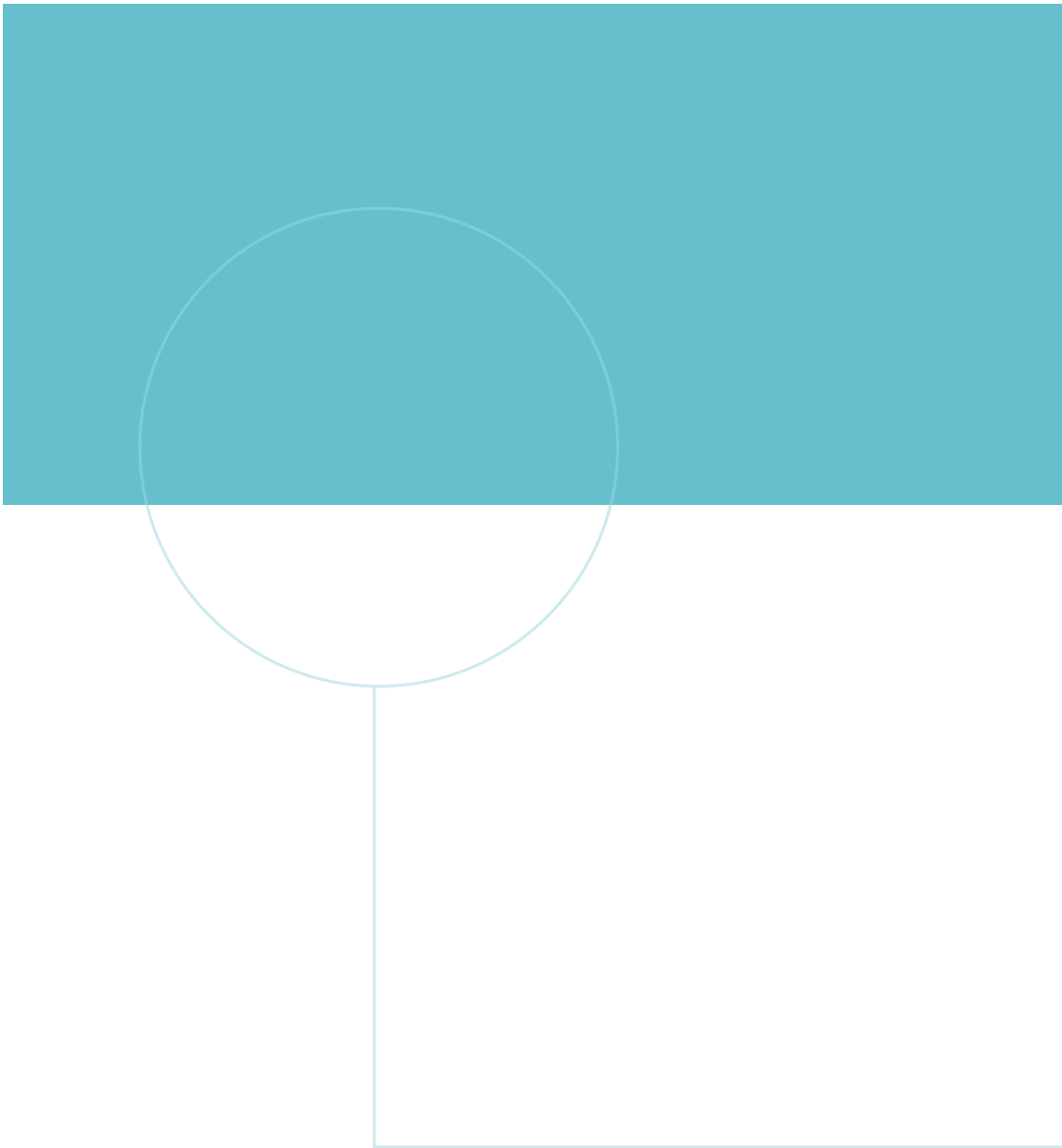
Norberg, R. (2004). Credibility theory.

Ohlsson, E. & Johansson, B. (2010). *Non-life Insurance Pricing with Generalized Linear Models*. Springer Berlin Heidelberg.

Ross, S. (2014). *Introduction to Probability Models*. Academic Press.

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

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*, 699–704.



NTNU

Norwegian University of
Science and Technology