A Comparative Analysis of Car Insurance Risk Modelling: Traditional Statistical Models and Bayesian Approaches

Miguel Moreira Marcelino
Master in Computational Statistics and Data Analysis
Department of Mathematics
Faculty of Sciences of the University of Porto
2024

A Comparative Analysis of Car Insurance Risk Modelling: Traditional Statistical Models and Bayesian Approaches

Miguel Moreira Marcelino

Project Report carried out as part of the Master in Computational Statistics and Data Analysis

Department of Mathematics

2024

Supervisor
Margarida Maria Araújo Brito, Associate Professor, FCUP

Abstract

We develop an effective motor tariff model by evaluating various approaches, including frequentist statistical methods and Bayesian models. In particular, we estimate the pure premium by modeling frequency and severity of claims, through a data set from an actual Spanish insurance company. According to the data structure, we applied Generalized Linear Mixed Models under both approaches.

The relevant models were fitted in R and thereafter assessed using a combination of evaluation metrics. We manage to find adequate models for both frequency and severity, which may then be used to estimate the pure premium.

Keywords: Car insurance, Pricing, Generalized Linear Mixed Models, Bayesian models, risk modeling, predictive analytics.

Contents

Acknowledgments

To my parents and siblings: your support and encouragement have been the cornerstone of my path over the past five years of studying. I owe you an immeasurable debt of gratitude. Your belief in me, your sacrifices, and your boundless love have shaped me into the person I am today.

To my grandma: your generosity and kindness have been fundamental to my educational journey. Your caretaking nature and support was one of the main pillars of my education. You made sure that I always had a stable and nurturing environment to pursue my studies, a home away from home. For that, I am eternally thankful.

To my friends: my companions on this academic journey. From late-night study sessions to fun nights out, each memory we've shared has made my journey all the more enjoyable and meaningful.

To my advisor: your guidance and support have been invaluable all throughout the creation of this document. Your expertise and encouragement have helped me navigate the challenges and complexities of my academic pursuits, and for that, I am truly grateful.

List of Figures

List of Tables

List of Abbreviations

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

CI Credibility Interval

DIC Deviance Information Criterion

GDP Gross Domestic Product

GLMM Generalized Linear Mixed Models

LRT Likelihood Ratio Test

MAE Mean Absolute Error

MCMC Markov Chain Monte Carlo

ML Maximum Likelihood

OR Odds Ratio

QQ-Plot Quantile-Quantile Plot

REML Restricted Maximum Likelihood

RMSE Root Mean Squared Error

VIF Variance Inflation Factor

ZINB Zero-Inflated Negative Binomial

ZIP Zero-Inflated Poisson

Chapter 1.

Introduction

In the last few decades, the insurance industry has increased in both popularity and in necessity. As people increasingly seek protection against certain events that could bring financial hardship, insurance has become indispensable, even mandatory in some situations. The influence of this industry is now so significant that their insolvency could potentially impact millions of people and even destabilize the economy of the country in which they operate. Therefore, it is in everyone's best interest to ensure that insurance companies manage their assets and liabilities effectively.

To meet the required obligations, an insurance company must establish a **reserve**. Given these substantial costs, it's crucial for insurance companies to ensure they have the necessary funds to constitute the reserve and guarantee their ability to pay for claims according to the insurance contracts. The reserve needs to be constantly assessed by applying proper actuarial methods to assess the adequacy of reserves and ensure compliance with regulatory requirements.

1.1 Gross and net premiums, written premium and Unearned Premium Reserve (UPR)

When a policyholder carries out a contract with an Insurance company, they agree to pay a periodic fee in exchange for partial or total coverage of a risk or a set of risks. This fee, commonly known as the **premium**, protects against potentially costly events that could lead to financial ruin. The premium paid by the customer is referred as the **gross premium**. This amount includes commissions paid to agents who sell the policies, legal expenses associated with settlements,

taxes, and reinsurance costs, where insurance companies transfer some of their risk to other insurers. After deducting these fees, we arrive at the **net premium** which, informally, is the actual value considered in estimating reserves and other financial calculations.

There are three additional concepts of interest when dealing with premium: **written premium**, **earned premium**, and **unearned premium reserve (UPR)**.

- **Written Premium**: This represents the total amount customers agree to pay for insurance policies sold during an accounting period. It reflects the revenue the company expects to earn over the policy term.
- **Earned Premium**: This is the portion of the written premium that corresponds to the coverage provided during the accounting period. Insured policyholders pay premiums in advance, so insurers do not immediately recognize the entire premium as profit. The premium is considered earned only when the insurer has fulfilled its obligation.
- **Unearned Premium Reserve (UPR)**: This reserve accounts for the portion of the premium that has not yet been earned because the coverage period extends into future accounting periods. The UPR represents the liability for future coverage that the insurer still owes to policyholders.

To better understand these three concepts, consider the following example: suppose an insurance company underwrites a one-year policy with a premium of 800€, starting on April 1st. On the day the policy is written, the written premium is 800€. Recall that the earned premium refers to the portion of the premium which has been consumed since the start of the policy up to the present moment. By the end of the year, on December 31st, nine out of twelve months (equivalently three out of four quarters) of the policy's term has elapsed, so three quarters of the premium is earned.

$$\text{Earned Premium} = 800€ \times \frac{9}{12} = 600€$$

The remaining value constitutes the UPR, which is the portion of the premium that has yet to be earned:

$$\text{UPR} = 800€ - 600€ = 200€$$

In this project, we will pay more attention to the net premium.

1.2 Loss ratios

An insurance company is required to keep track of its financial status at all times, which is achieved by analyzing several key indicators. Some of the most critical indicators are the **loss ratios**, given by the following expression:

$$\text{Loss Ratios} = \frac{\text{Claims Paid} + \text{Adjustment expenses}}{\text{Earned Premium}} \tag{1.1}$$

This ratio represents the percentage of earned premiums used to pay for claims and adjustment expenses. Thus, an increase in paid claims/adjustment expenses and/or a decrease in earned premiums results on a higher loss ratio, which could be a signal of financial distress. There are three possible scenarios:

1. $\text{Loss ratio} > 100\%$: The insurance company is paying out more in claims that it earns from premiums, indicating a financial loss;

2. $\text{Loss ratio} = 100\%$: The insurance company uses all collected premiums to pay claims, resulting in neither profit nor loss from its policies;

3. $\text{Loss ratio} < 100\%$: The insurance company has surplus premiums after paying for claims, indicating profitability.

Loss ratios are valuable for evaluating the need to adjust tariffs, comparing the profitability of different lines of business and with other insurance companies.

1.3 Pricing in car insurance

The task of calculating premiums that accurately reflect the risk profiles of policyholders falls under the branch of **Pricing**. This involves gathering relevant data over a period of time and using computational models to predict the frequency and cost of claims for each policyholder. On one hand, premiums need to be sufficient to cover the required claims. However, setting premiums too high might deter potential and existing policyholders to prefer other companies. Therefore, it's imperative that the company finds a balance between affordability and sufficiency. It's not just about the frequency of claims but also their potential severity. The **premium** is the monetary amount which the policyholder is contractually required to pay in order to be covered by the policy. It's also important to note that even if the characteristics of claims remain consistent over several years, premiums will still fluctuate due to inflation, since it impacts the cost of potential claims. By accounting for inflation, insurers ensure that premiums remain adequate to cover future claims.

1.4 Typical approach for modeling risk in car Insurance

The most common practice in the day-to-day life in insurance companies is to evaluate the risk profile of the policyholder using a set of covariates, usually **categorical variables**. These can vary depending on the kind of risks the insurance company is covering. In the case of motor insurance, the typical considered variables refer to both policyholder's characteristics (e.g. age group, time since driver's license, sex, etc.) and the insured vehicle (type of fuel, main area of transit, etc). For every unique combination of classes among the covariates, there is a **tariff cell**. All policies within the same tariff cell output the same premium, adjusted for these different risk profiles.

Due to the nature of the response variables used to determine the frequency and severity of claims, the most common used models are the Generalized Linear Models (GLMs). The statistical validity of the structure of the Insurance Company's dataset used in GLM regression rely on three assumptions (Mosawi, 2017):

1. **Policy independence.** Let n be the number of policies and X_i be the response for policy i. Then $X_1, \cdots, X_n$ are independent.

2. **Time independence.** Let n be the number of disjoint time intervals and X_i be the response for policy i. $X_1, \cdots, X_n$ are independent.

3. **Homogeneity.** Let X_1 and X_2 be the response for two any policies in the same tariff cell with same exposure. Then X_1 and X_2 have the same probability distribution.

However, one could argue that these assumptions are not always verified:

1. A crash between two vehicles insured by the same company violate the policy independence;

2. Policyholders who have had more claims tend to drive more safely, lowering claim frequency and thus violating time independence;

3. Frequency/ severity of claims could vary throughout the year due to seasonal influences, even if we consider the same time of exposure, violating homogeneity.

1.5 Bayesian perspective on Insurance

In Bayesian Statistics we think of *credibility* in place of *probability*, in the sense that we have convictions that a certain event will happen and we update our beliefs with gathered knowledge. Even though it constitutes a very powerful branch, Bayesian Statistics isn't often employed in Insurance. Although it has gained popularity in recent years due to revolutionary computational techniques, its discredit as "another solution for simple problems" contributes to a general overlook of Bayesian methods. There is still a large number of hurdles to overcome, such as its inability to efficiently handle large datasets, like Insurance data, as well as harder implementation and interpretability of results when compared with frequentist statistical models. Therefore, the latter ones are usually preferred over Bayesian statistical models, which are often way heavier theoretically and practically. However, with the advancement of computational power, nowadays we can easily overcome difficulties of many complex problems through simulation, which allows

us to solve them quickly and easily. It also serves as a tool for inference with impressive precision when done right.

1.6 Outline of project

The steps taken in both approaches share the same statistical foundation:

1. Conducting a descriptive and graphical analysis, extracting the main characteristics of the data;
2. Recapping theoretical knowledge about statistical models employed in this project; more specifically, Generalized Linear Mixed Models and Bayesian statistical models;
3. For each model, identifying the underlying parameters that we want to estimate;
4. Achieving the best possible model through careful evaluation of the results;
5. Comparing model performance among all estimated models, in both approaches.

This work is organized in 5 chapters, beginning with a description of the dataset, review of Statistical models, risk modeling and finishing with a conclusion.

Chapter 2.

Dataset structure and preparation for analysis

Ideally, insurance portfolios are structured so as to appropriately reflect the **exposure** of the policy within each year, which is the duration that a policy is active during a time period. However, due to lack of details regarding the dates of occurrence of claims, we couldn't follow the aforementioned structure. Instead, we have chosen to accommodate the dataset's original structure and characteristics by considering random effects, entailing the usage of a different kind of models which will be mentioned in the next chapter.

Moving on to the data itself, the variables we can usually find in a dataset and that are crucial for modeling are of four types:

1. Identification variable, usually the policy number or some other form of identification;
2. Policy data, which encompasses the characteristics of the policy, such as dates of last and next renewals, number of claims, premium, etc.;
3. Policyholder's data, such as their age, sex, antiquity of license, etc.
4. Risk data, which in the case of motor insurance refers to the characteristics of the vehicle, such as type of fuel, year of matriculation, brand, weight, etc.

2.1 General information about our dataset

The data set used in this project was provided by an anonymous Spanish insurance company, which can be accessed here. The data encompasses key activities within the Comprehensive

Motor Insurance segment, a branch of the non-life motor insurance. It spans a period of three years and two months, from the beginning of November 2015 to the end of December 2018, comprising 105.555 observations and 30 variables, each belonging to their respective types mentioned above. The original variables in the datas et are presented in Table 2.1, whose description can also be consulted here. Each row represents a policy transaction (eq. one full year of exposure, since contract renovation until expiring date), while each column corresponds to a specific variable. The main focus of the analysis will be on cars, which represent most of the business in this branch of the insurance company. In this portfolio, the data are **not balanced**, since policyholders do not all have the same number of observations.

To better understand the variables of interest to the task at hand, it is convenient to know their meaning and their type:

- **ID**: Internal identification number assigned to each annual contract formalized by an insured. Each policyholder can have multiple rows in the dataset, representing different annuities of the product.
- **Date_last_renewal**: Date of last contract renewal (DD/MM/YYYY).
- **Date_next_renewal**: Date of the next contract renewal (DD/MM/YYYY).
- **Date_birth**: Date of birth of the insured declared in the policy (DD/MM/YYYY).
- **Date_driving_licence**: Date of issuance of the insured person's driver's license (DD/MM/YYYY).
- **Seniority**: Total number of years that the insured has been associated with the insurance entity, indicating their level of seniority.
- **Premium**: Net premium amount associated with the policy during the current year.
- **Cost_claims_year**: Total cost of claims incurred for the insurance policy during the current year.
- **N_claims_year**: Total number of claims incurred for the insurance policy during the current year.

- **R_Claims_history**: Ratio of the number of claims filed for the specific policy to the total duration (whole years) of the policy in force. It provides an indication of the policy's claims frequency history.
- **Type_risk**: Type of risk associated with the policy. Each value corresponds to a specific risk type: 1 for motorbikes, 2 for vans, 3 for passenger cars and 4 for agricultural vehicles.
- **Area**: Dichotomous variable indicates the area. 0 for rural and 1 for urban (more than 30,000 inhabitants) in terms of traffic conditions.
- **Second_driver**: 1 if there are multiple regular drivers declared, or 0 if only one driver is declared.
- **Value_vehicle**: Market value of the vehicle on 31/12/2019.
- **Type_fuel**: Specific kind of energy source used to power a vehicle. Petrol (P) or Diesel (D).

The remaining variables aren't useful to our analysis and/or were scrapped during the cleaning of the data.

Table 2.1: Classification of Variables by Type

Variable Name	Type of variable			
	Identification	**Date**	**Categorical**	**Quantitative**
ID	X			
Date_start_contract		X		
Date_last_renewal		X		
Date_next_renewal		X		
Date_birth		X		
Date_driving_licence		X		
Date_lapse		X		
Distribution_channel			X	
Max_policies			X	
Max_products			X	
Lapse			X	
Payment			X	
Type_risk			X	
Area			X	
Second_driver			X	
Type_fuel			X	
Seniority				X
Premium				X
Cost_claims_year				X
N_claims_year				X
N_claims_history				X
R_Claims_history				X
Year_matriculation				X
Power				X
Cylinder_capacity				X
Value_vehicle				X
N_doors				X
Length				X
Weight				X
Policies_in_force				X

2.2 Treatment of the data

In its first stage, the raw information of the dataset must be treated so that it fits our needs and objectives. We begin by asserting variable classes, ensuring that each variable belongs to its intended class, such as making sure categorical variables are correctly classified. Dates, which were not useful in their original format, were transformed to extract important information. Specifically, three essential variables were created: `Age`, `Age_since_drivers_license` and `Exposure`. The variable `Age` was calculated by computing the difference between the date of the last renewal and the date of birth, whereas `Age_since_drivers_license` was calculated by computing the difference between the date of the last renewal and the date of issuance of their driver's license. The most crucial variable, `Exposure`, was calculated for each row by computing the duration between the initiation of each contract and the conclusion of the observation period. It was also taken into consideration the fact that the datasets dates of December 31st, 2018, since there were a few policies whose renovation dates were expected in 2019. In these cases, instead of considering the contract renewal date, we considered the date December 31st, 2018.

One other important aspect to address is **inflation**, which is often overlooked and is a factor that can severely impact the incurred costs an Insurance company is expecting. Even though the difference might be of just some euros for a single policyholder, the accumulated amount throughout the whole portfolio could rise to millions of euros, which can be debilitating for the long-term stability of reserves. Thus, adjusting our monetary variables, such as incurred costs and premiums, is crucial, so that they reflect the inflation trend.

Gavin Thompson (2009) explains how we can proceed with the inflation adjustment for non-consecutive years: for years x and y, assuming $x \leq y$, consider the Gross Domestic Product (GDP) Deflator indexes i_x and i_y for those respective years. Then the monetary value from year y adjusted for year x prices is given by

$$\text{Adjusted value} = \frac{i_y}{i_x} Value \tag{2.1}$$

In our case, since the data is as of December 31st 2018, we must calculate the indexes in function of the 2018 index so that we obtain inflation rates as of 2018. Some websites, such as `https://www.dadosmundiais.com`[1], provide the necessary indexes and we can even compute the inflation rate directly in the website.

In our dataset, every policy has an exposure of a whole year after being renewed. Naturally, a portion of the risk lies within the year of renewal and the other portion in the consecutive year. For this reason, costs and premiums were adjusted using a weighted average of their adjusted values as shown in 2.1, where the weights were determined by the proportion of exposure time in each of those years. In other words, if d_i is the number of days covered by the policy in year i,

$$\begin{aligned} \text{Value}_{adj} &= \frac{365 - d_1}{365} \times \frac{i_{2018}}{i_1}\text{Value} + \frac{d_2}{365} \times \frac{i_{2018}}{i_2}\text{Value} \\ &= \text{Value} \times \frac{i_{2018}}{365 \times i_1 i_2} \times [(365 - d_1) \times i_2 + d_2 i_1] \end{aligned} \tag{2.2}$$

2.3 Missing data

It's relatively common to find missing data in Insurance datasets. A lot of the times, the reason for this lies behind the fact that some variables can take on some values which are incompatible with another variable. For example, if a vehicle doesn't have any horsepower, such as bicycles, it doesn't need fuel, thus it can't take on a value for that variable.

The dataset used in this project originally wasn't exempt of missing data. In fact, they're present in two and only two variables, `length` and `Fuel_type`. After analyzing the values of all other variables for the observations with NAs in at least one of those two variables, we came to the conclusion that the NAs in `Fuel_type` are due to vehicles that don't use any type of fuel. Due to the severe lack of representation of these vehicles, motorbikes and agricultural vehicles, it was decided that these vehicle categories were to be removed. In the end, `Type_risk` is left with two types of vehicles: vans and passenger cars.

[1] For Spain, the website is `https://www.dadosmundiais.com/europa/espanha/inflacao.php`

Chapter 3.
Background review on statistical models

3.1 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMMs) are an extension to both usual linear mixed models and generalized linear models. While on one hand the extension of the former allows the response variable to follow non-normal distributions, it also extends the latter in the way that it overcomes the assumption of independence between observations, thereby allowing correlation, a common occurrence in hierarchical and longitudinal studies.

Let $Y_i \in \mathbb{R}^{T_i}$ be the response variable vector of the i-th subject, $X_i \in \mathbb{R}^{T_i \times (p+1)}$ the design matrix of the observed covariates, $\beta \in \mathbb{R}^{p+1}$ the vector of fixed effects, $Z_i \in \mathbb{R}^{T_i \times (q+1)}$ the model matrix of the random effects and $u_i \in \mathbb{R}^{q+1}$ the vector of random effects. Just like linear mixed models, GLMMs are characterized by two components:

1. A random component, allowing the response components $Y_{ij}, j = 1, \cdots, T_i$, **conditioned on** u_i to follow any distribution from the exponential family. The probability density function of this family can be rewritten as

$$f(y_{ij}; \theta, \phi) = \exp\left(\frac{y_{ij}\theta - b(\theta)}{a(\phi)} + c(y_{ij}, \phi)\right), \tag{3.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions, ϕ is the dispersion parameter, and θ is the canonical parameter of the distribution. Therefore, if Y_{ij} is a random variable whose probability density function belongs to the exponential family, then $E(Y_{ij}) = \mu_{ij} = \frac{\mathrm{d}b}{\mathrm{d}\theta}$ and

$\mathrm{Var}(Y_{ij}) = \frac{\mathrm{d}^2 b}{\mathrm{d}\theta^2} a(\phi)$ (Mosawi (2017)). The function $\frac{\mathrm{d}^2 b}{\mathrm{d}\theta^2}$ is called the **variance function**.

The random effects are added to capture the within-subject variability and correlations. They follow a multivariate normal distribution centered at 0 and with variance matrix V. We assume that $Y_{i1}|u_i, \cdots, Y_{iT_i}|u_i$ are pairwise independent and that the random effects are also pair-wise independent. (Goldburd (2020)).

2. A systematic component, where the linear predictor

$$\eta_{ij} = X_{ij}\beta + Z_{ij}u_i \tag{3.2}$$

is linked to the mean μ_{ij} through a monotonic and differentiable function g called the **link function**, that is, $g(\mu_{ij}) = \eta_{ij}$. The purpose of the link function is to allow more flexibility in associating the model predictions with the explanatory variables. In the particular case of insurance, it enables insurers to build **multiplicative models** using the *log* link function, which usually best fits reality. Some advantages to these kinds of models include (Goldburd (2020)):

   - Their simple implementation;
   - They don't allow negative premiums
   - Their intuitive appeal: It doesn't make much sense to say that having a claim should increase the auto premium by $500, regardless of whether the base premium is $1,000 or $10,000. It makes more sense to consider the relative increase, for example, saying that there's been a 10% increase in premium.

Each distribution may have several link functions which can work with it, however for each distribution there is a special link function called **canonical link function**, such that $\eta_{ij} = \theta = g(\mu_{ij})$, where θ is the canonical parameter. These link functions are the most desirable due to their mathematical and computational convenience, however one could argue that, in some situations, other link functions would be contextually more suitable (Faraway (2006)).

Table 3.1 summarizes this information.

Distribution	$E(Y)$	**Canonical Link Function**	$Var(Y)$
Normal	μ	$g(\mu) = \mu = \theta$	σ^2
Binomial	p	$g(p) = \log\left(\frac{p}{1-p}\right)$	$np(1-p)$
Gamma	μ	$g(\mu) = -\frac{1}{\mu}$	$\alpha\mu^2$
Exponential	μ	$g(\mu) = \frac{1}{\mu}$	μ^2
Geometric	p	$g(p) = \log\left(\frac{1-p}{p}\right)$	$\frac{1-p}{p^2}$
Poisson	μ	$g(\mu) = \log(\mu)$	μ
Negative Binomial	μ	$g(\mu) = \log(\mu)$	$\mu + \frac{\mu^2}{\alpha}$
Inverse Gaussian	μ	$g(\mu) = \frac{1}{\mu^2}$	$\frac{\mu^3}{\lambda}$

Table 3.1: Distributions, their mean, canonical link functions, and variance functions. Source: `https://www.researchgate.net/figure/The-canonical-link-function-of-the-exponential-family-distribution_tbl1_351788190`

3.1.1 Estimation of parameters

Goldburd (2020) states that GLMMs are a two-step process: firstly, the fixed effects are estimated; then, the second step estimates the random effects, related to the probability distribution that their coefficients follow.

Let $f_v(u)$ be the density function of random effects and $f_\beta(y|u)$ the response density given the random effects. To simplify notation, let's consider $\theta = (\beta^T, v^T)^T$. Then, the joint density of random effects and response can be written as

$$f_\theta(u, y) = f_\beta(y|u) f_v(u) \tag{3.3}$$

Consider a GLMM as represented in 3.1, with link function $g(\cdot)$. The likelihood function comes as

$$\begin{aligned} L(\theta|y) &\coloneqq \int_{\mathbb{R}^q} f_\beta(y|u) f_v(u) du_1 \cdots du_q \\ &= \int_{\mathbb{R}^q} f_\theta(u, y) du_1 \cdots du_q \end{aligned} \tag{3.4}$$

and the log-likelihood is $\ell(\theta|y) = \log L(\theta|y)$. However, we usually cannot express this integral in a closed form, motivating the use of numerical methods that either determine the maximum likelihood (e.g. Monte Carlo Expectation-Maximization) or approximate the likelihood (e.g Penal-

ized Quasi-Likelihood). Knudson (2016) provides a detailed overview of these methods and a few other. Through different paths, all these methods have the common goal of obtaining the maximum likelihood estimator of the parameters, variance components and random effects.

Estimating both the parameters and the variance components is typically performed using Maximum Likelihood (ML) estimation. However, a challenge arises when estimating the variance components, as the likelihood function is conditional on the estimated $p + 1$ regression coefficients. Since the ML estimator assumes these parameters are known, it does not account for the degrees of freedom lost due to the estimation of these coefficients, leading to biased estimates of the variance components (McCulloch and Casella (1992, 2006)).

To address this issue, Restricted Maximum Likelihood (REML), extended to mixed models by Patterson and Thompson (1974), provides a solution. REML works by eliminating the fixed effects from the likelihood function, leaving it dependent solely on the variance components. This approach involves using part of the data to estimate the fixed effects and the remaining data to estimate the variance components. The residuals obtained after estimating the fixed effects are then used to estimate the variance components. For a clear explanation of this process, we may refer to McCulloch and Casella (1992, 2006). Despite its good properties, there are some situations where REML underperforms and might not be preferred over ML. However, if the sample size is large enough, then either estimator tends to perform well since both are consistent. As this is the case with most Insurance portfolios, we don't have to worry about bias in the estimates.

3.1.2 Goodness-of-fit statistics

In standard linear models, the response variable can be expressed by the following equation:

$$y = \hat{\mu} - (y - \hat{\mu}), \tag{3.5}$$

which represents the observed values as the difference between the fitted values and the residuals (Nelder and Wedderburn (1972)). Figure 3.1 shows the geometric interpretation of these residuals: they correspond to the projection of the observed values on the subspace generated by the predictors, minimizing the distance between the observed values and the fitted values $(X\hat{\beta})$.These residuals are a primary method for evaluating the model fit. However, in GLMMs,

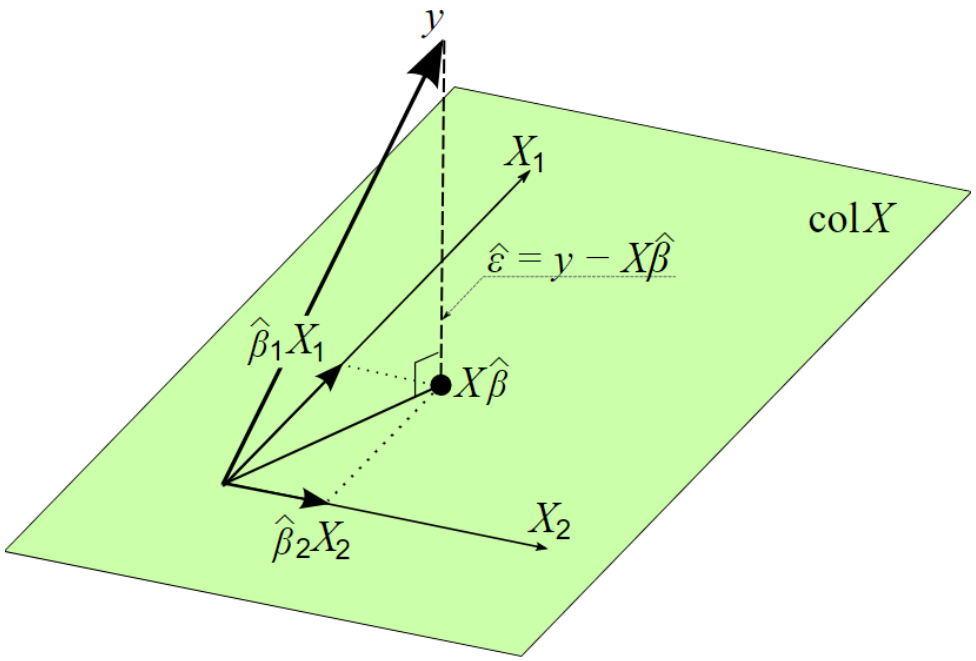


Figure 3.1: Geometric interpretation of the main concepts in regression. Source: `https://commons.wikimedia.org/wiki/File:OLS_geometric_interpretation.svg`

the residuals are given by $r = Y - g^{-1}(X\beta)$, making them more complex to interpret due to the transformation provided by the link function.

Instead of residuals, we use the deviance, a generalization of the sum of squares of residuals. The deviance quantifies how well the model $\mathcal{M}$ fits the data by comparing it with the saturated model $\mathcal{S}$, which is the best possible model and estimates a parameter for each observation, achieving the highest possible likelihood (Mosawi (2017)). The **deviance** is defined as

$$D = -2\phi(\ell_{\mathcal{M}}(\beta) - \ell_{\mathcal{S}}(\beta)), \tag{3.6}$$

where $\ell_{\mathcal{M}}(\beta)$ is the log-likelihood of the fitted model, $\ell_{\mathcal{S}}(\beta)$ is the log-likelihood of the saturated model, and ϕ is a scaling factor, which may or may not be unknown. This statistic asymptotically follows a $\chi^2(n-k)$, where k is the rank of the matrix of predictors X and n is the number of observations.Since $\ell_{\mathcal{M}}(\beta) \leq \ell_{\mathcal{S}}(\beta)$, we have $D \geq 0$. The higher the deviance, the greater the disparity between the model and the data, while a deviance equal to 0 signifies a perfect fit, as is the case with the saturated model.

As stated before, analyzing residuals is a great way of seeing if all assumptions made about

the model are met. However, since we're not limited to just Gaussian errors anymore, it doesn't make sense to use the regular definition of residuals mentioned above at 3.5, since they are too complex to interpret in that form. Therefore, an extension of the definition of residuals is needed, aiming to serve the same purpose as the standard residuals. Nelder and Wedderburn (1972) propose three definitions, however we will focus on two of the simpler and most commonly used ones.

One of the widely used definitions is the **Pearson Residuals**, which are simply the regular residuals but scaled by the estimated variance function, comparable to standardized residuals for linear models. It is defined as

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{3.7}$$

where $V(\mu) = \frac{d^2 b}{d\theta^2}$, as presented in the Introduction. However, the disadvantage of this statistic is that, for non-normal distributions, their distribution tends to be skewed, and they might fail to exhibit the same properties as regular residuals. Due to these limitations, a preferable alternative is the **deviance residuals**. Recalling that the deviance is a generalization of the sum of squares of residuals, each observation contributes with a quantity d_i such that $\sum d_i = D$, where D is the total deviance given by 3.6. Deviance residuals can then be defined as:

$$r_{Di} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{3.8}$$

With this definition, r_D increases with $y_i - \mu_i$ and satisfies $\sum_i r_{D_i}^2 = D$.

The last goodness-of-fit statistics of interest in this project are the **Akaike Information Criterion (AIC)** and the **Bayesian Information Criterion (BIC)**. Introduced by Akaike in 1974, the AIC is derived from the concept of Kullback-Leibler divergence and is formulated as

$$\text{AIC}_M = -2\ell_M + 2p_M,$$

where ℓ_M is the log-likelihood and p_M is the number of parameters. The AIC is widely used for

comparing non-nested models, although it tends to favor more complex models, particularly in small samples. On the other hand, BIC, developed by Schwarz in 1978, is an approximation of the Bayes factor and is expressed as

$$\text{BIC}_M = -2\ell_M + p_M \times \log(n).$$

The BIC introduces a stronger penalty for model complexity, especially as the sample size n increases, making it more conservative than the AIC and more likely to select simpler models (Flossain (2002)).

While both AIC and BIC are useful in model comparison, they have limitations:

1. Their assessment of model quality is relative, rather than absolute;

2. Neither method gives the probability that the model is correct;

3. BIC's sensitivity to sample size can lead to different model selections in different datasets.

Ultimately, the choice between AIC and BIC depends on the specific context, with AIC being more flexible and BIC more strict in model selection for large samples. In general, we tend to use both in model comparison, not exclusively just one or the other. Zuur et al. (2009) advises that an AIC/BIC based on REML is not comparable with an AIC/BIC obtained by ML.

3.1.3 Multicollinearity

When conducting a regression analysis, it is often overlooked that collinearity can occur among predictors, meaning they may exhibit approximate or near-linear dependencies. When this happens, the variance of the estimated coefficients can increase dramatically, reducing their precision and altering the interpretation of the significance of the parameters.

A key method for evaluating multicollinearity is through **Variance Inflation Factors** (VIFs), which assess each predictor's contribution to the multicollinearity problem. In a multiple regression with p predictors, VIFs are derived from the diagonal elements of the inverse of the correlation matrix of the p predictors. The VIF for the ith predictor is given by:

$$\text{VIF}_i = r_{ii} = \frac{1}{1 - R_i^2}, \quad i = 1, \ldots, p$$

where R_i^2 is the multiple correlation coefficient of the regression between X_i and the remaining $p - 1$ predictors.

There is not a universally accepted cutoff for what constitutes a "high" VIF, although some researchers have suggested cutoff values of 5 or 10 (Murray et al. (2012), which seems to be the common rule nowadays.

The presence of correlated **continuous** variables can also be evaluated using corrplots, which calculate pairwise correlations between each of these variables. Dormann et al. (2013) suggest that only when the values of correlation exceed 0.7 does it become problematic.

3.1.4 Relevant tests for the analysis and comparison of models

Finding the optimal GLMM involves finding two optimal components, that of the fixed effects and random effects. Zuur et al.(2009) suggests a protocol for achieving the optimal model, based on a top-down procedure:

1. Start with a "beyond optimal" model, where the fixed component contains all explanatory variables and as many interactions as possible. If this is unfeasible, use a selection of explanatory variables that are likely to have a significant effect on the optimal model;

2. Find the optimal structure of the random component. Two models with nested random structures **must be compared using REML**, either through LRT or AIC/BIC, but always using REML;

3. Find the optimal structure of the fixed component. Two models with nested fixed structures **must be compared using ML**;

4. Present the final model using REML estimation.

After fitting a model, we're interested checking if an explanatory variable has a significant effect on the response, which can be done by carrying out the appropriate hypothesis tests. In this section, two of the most commonly used tests shall be addressed, the **Wald Test** and **Log-Likelihood Ratio Test (LRT)**, which both have their univariate and multivariate variations. However, since the multivariate case is the generalization of the univariate one, only the multivariate version will be discussed.

Let β_j be some parameter associated to a predictor X_j employed in the regressed model, and A a set of nonnegative integers such that $i \in A \implies 0 \leq i \leq p+1$, where p is the number of predictors. The test is formulated as follows:

- $H_0 : \forall j \in A, \beta_j = 0$
- $H_1 : \exists j \in A : \beta_j \neq 0$

In Wald's test, under H_0, the test statistic is

$$W = \hat{\beta}^\top \Sigma^{-1} \hat{\beta} \sim \chi^2(k), \tag{3.9}$$

where $\hat{\beta}$ is the vector of estimated coefficients and Σ is the covariance matrix of the estimates. Since this is a computationally intensive test to perform, usually only the univariate test is performed. In fact, by default, the univariate test for each parameter are present in the summary of the output of the fitted model.

In R, the p-values for the fixed effects observed in the summary output of the fitted model are obtained through Inference, based on Wald statistic. One could also compute confidence intervals for the fixed effects based on Wald test.

Let $B = \{b_1, \cdots, b_m\} \subseteq A$ such that $i \neq j \implies b_i \neq b_j$. The formulation of the test makes us note that it is equivalent to comparing the goodness-of-fit of M_0 and M_1, where M_0 is the model with $\beta_{b_1} = \cdots = \beta_{b_m} = 0$ (i.e, the predictors $X_{b_1}, \cdots, X_{b_m}$ aren't included) and M_1 is the model with $\beta_{b_j} \neq 0,\ j = 1, \cdots, m$. This statement is the foundation of the **LRT**.

To compute the test statistic, we find the log-likelihoods of each model. Let $\ell(M_0)$ be the log-likelihood of the nested model M_0 and $\ell(M_1)$ be the log-likelihood of the complex model M_1. The test statistic is given by:

$$\text{LR}(M_0, M_1) = -2\left[\ell(M_0)/\ell(M_1)\right] \sim \chi^2\left(|B|\right), \tag{3.10}$$

where $|B|$ denotes the cardinal of subset B. In other words, the statistic follows a chi-squared distribution with degrees of freedom equal to the difference in the number of free parameters between M_1 and M_0.

In R, this test can be performed using the *anova()* function, where you input the two models (first M_1 and then M_2) and specify the test type with `test = "LRT"`. Additionally, the function *confint()* can be used to obtain confidence intervals for the fixed effects, utilizing LRT as the default method for these computations.

An important issue with the LRT must be mentioned. Recall the second step of the protocol for finding the optimal model, which involves obtaining the optimal random component — one of the ways of doing it being via LRT. In this situation, Zuur et al. (2009) warns us about **testing on the boundary**. The convergence in this case is to a mixture of χ^2 distributions, making p-values unreliable. To address this, it is suggested to either divide the outputted p-value by 2 or consider a significance level of $\alpha = 0.1$ instead of $\alpha = 0.05$.

3.1.5 Overdispersion in Poisson Regression

For a Poisson distribution, the variance is equal to the mean, which implies a dispersion parameter ϕ of 1, so $Var(Y) = \mu$. A common issue with Poisson regression for count data is **overdispersion**, which occurs when the observed variability exceeds the model's assumptions, meaning the variance is greater than the mean. One must bear in mind that overdispersion can occur in other models, thus when we're referring to overdispersion in a Poisson distribution, it should be explicitly mentioned as "Poisson overdispersion". Underdispersion is theoretically possible, but as it is much less frequent, it won't be addressed in this project.

Hilbe (2014) distinguishes between apparent and real overdispersion. Apparent overdispersion can result from factors such as missing covariates or interactions, outliers, non-linear effects modeled as linear, or an inappropriate choice of link function. Real overdispersion occurs when the true variance of the data exceeds the mean, or due to clustering, correlation among observations, or an excess of zeros, though the latter may not always lead to overdispersion. If the model fails to adjust for overdispersed data, **standard errors become biased** and thus their associated explanatory variables **might falsely appear significant** .

The Poisson model's dispersion statistic, as stated by Hilbe (2014), is given by

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}}{n - (p+1)}, \tag{3.11}$$

which is simply the sum of the squared Pearson residuals divided by the residual degrees of freedom. The numerator of 3.11 can also be seen as Pearson's χ^2 statistic, which follows a $\chi^2(n-(p+1))$. By dividing it by $n-(p+1)$, we adjust for the degrees of freedom, providing an estimate of the dispersion parameter. A value of $\hat{\phi}$ close to 1 suggests no overdispersion, indicating that the variance is roughly equal to the mean. However, if it is notably greater than 1, it signals the presence of overdispersion.

Introducing a dispersion parameter inflates the standard errors of the estimates by $\sqrt{\phi}$ (Gelman and Hill (2007) Thus, if parameters are highly significant, a $\hat{\phi}$ ratio slightly above 1 may not be problematic. Conversely, for parameters near the significance threshold, even a $\hat{\phi}$ of just slightly over 1 might render the parameter non-significant due to the increased standard error. Generally, a $\hat{\phi}$ value greater than 1.5 indicates that corrective action is required. Zuur et al. (2009) advises to use the negative binomial distribution if $\hat{\phi}$ values exceed 15 or 20.

3.1.6 Zero-inflated models

In several situations, we notice a bigger presence of zeros than what we would expect. This is known as **zero inflation**. According to Zuur et al.(2009), we can distinguish between two types

of zeroes: *true zeroes*, which are generated by the count model, and *false zeroes*, which are all others not generated by it and may appear due to several attributable causes. In the case of claim counts, *true zeroes* refer to the actual absence of claims, whereas *false zeroes* are linked to the presence of claims, but that weren't reported by the policyholder to the insurance company. However, neither Poisson nor Negative Binomial models discern between these two kinds of zeros. Assuming all of these are *true zeroes* may lead to two likely consequences: overdispersion and biased estimated parameters and standard errors.

Zero-inflated models are based on mixture models with two components, a probability mass at $y = 0$ and a count distribution $f(y;\theta)$, in order to address both kinds of zeroes. Essentially, this model aims to generate zeros through two processes, a **binomial process** and a **count process**. The binomial process models the probability of obtaining *false zeroes*, whereas the count process generates counts and *true zeroes*.

The probability of obtaining a false zero, π_i, can be modeled through a logistic regression using a set of covariates $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{iq})$. This probability is linked to the predictors through the *logit* function as follows:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_q Z_{iq}, \tag{3.12}$$

where γ_i are the regression coefficients. Therefore, π_i can be expressed as

$$\pi_i = \frac{e^{\gamma_0+\gamma_1 Z_{i1}+\cdots+\gamma_q Z_{iq}}}{1+e^{\gamma_0+\gamma_1 Z_{i1}+\cdots+\gamma_q Z_{iq}}}, \tag{3.13}$$

The most commonly used zero-inflated models are **Zero-Inflated Poisson (ZIP)** and **Zero-Inflated Negative Binomial (ZINB)**, in accordance with the most used claim counts models.

Zero-Inflated Poisson models address the scenario where data exhibit a greater number of zeros than would be expected under a standard Poisson distribution. Mathematically, the probability

mass function of a ZIP model is expressed as (cf. Zuur et al.(2009)):

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i}, & \text{if } y_i = 0, \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, & \text{if } y_i > 0, \end{cases} \tag{3.14}$$

where π_i represents the probability of a *false zero* and μ_i is the mean of the Poisson distribution. The mean and variance of the ZIP model are given by (cf. Zuur et al.(2009)):

$$\mathbb{E}(Y_i) = \mu_i(1-\pi_i), \tag{3.15}$$

$$\text{Var}(Y_i) = \mu_i(1-\pi_i)(1+\pi_i\mu_i). \tag{3.16}$$

If $\pi_i = 0$, the ZIP model is reduced to the standard Poisson model, where the mean and variance are equal. When $\pi_i > 0$, the variance exceeds the mean, indicating a potential overdispersion due to the excess number of zeros.

Zero-Inflated Negative Binomial (ZINB) models extend the ZIP model to accommodate overdispersion. The probability mass function for the Negative Binomial distribution is (cf. Zuur et al.(2009)):

$$f(y_i \mid \mu_i, k) = \frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} \left(\frac{k}{k+\mu_i}\right)^k \left(1 - \frac{k}{k+\mu_i}\right)^{y_i}. \tag{3.17}$$

Therefore, the probability mass function of a ZINB model is:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{k}{k+\mu_i}\right)^k, & \text{if } y_i = 0, \\ (1-\pi_i)\frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)}\left(\frac{k}{k+\mu_i}\right)^k\left(1-\frac{k}{k+\mu_i}\right)^{y_i}, & \text{if } y_i > 0. \end{cases} \tag{3.18}$$

where k is the overdispersion parameter.

The mean and variance of the ZINB model are (cf. Zuur et al.(2009)):

$$\mathbb{E}(Y_i) = \mu_i(1-\pi_i) \tag{3.19}$$

$$\text{Var}(Y_i) = (1-\pi_i)\left(\mu_i + \frac{\mu_i^2}{k}\right) + \mu_i^2(\pi_i^2 + \pi_i) \tag{3.20}$$

When $\pi_i = 0$, the variance corresponds to that of the Negative Binomial distribution. As k increases, the variance converges to μ_i, approximating the Poisson distribution. Thus, smaller values of k indicate greater overdispersion.

3.2 Bayesian Modelling

3.2.1 Motivation

Suppose we are interested in studying a population that follows a distribution depending on an unknown parameter θ (note that θ doesn't necessarily have to be a scalar; it could also be a vector or matrix, for example). Most statisticians would employ frequentist methods, such as confidence intervals or hypothesis tests, to find the most likely values for this parameter using a sample from the population. In contrast, Bayesian statistics treats θ as a random variable rather than a fixed unknown. Consequently, the approaches to this task taken in Bayesian statistics differ significantly from those of frequentist statistics, introducing new methods for inference.

The foundation of Bayesian statistics is Bayes' Theorem, which states that if A and B are events such that $P(B) \neq 0$, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.21}$$

The interpretation of *probability* in itself shapes the interpretation of Bayes' Theorem. In Bayesian statistics, it represents the certainty that some statement/proposition is true. The theorem associates this certainty before and after updating our beliefs, as stated by Brewer.

3.2.2 Main concepts

Before we move on to the modeling, we first review the basic concepts of Bayesian modeling.

A **statistical model** is a collection of distributions indexed by a parameter θ belonging to a parameter space Θ, which defines a probability model as

$$M = \{f(y|\theta), \theta \in \Theta\} \tag{3.22}$$

In Bayesian statistics, the goal is to determine the probability model $f(y \mid \theta^*) \in M$ that gener-

ated the data, where θ^* is the true value of the parameter inherent to the population.

Suppose y is a random sample from the population being studied. The statistical model then becomes a function of both θ and y; in other words, we can view it as a joint probability distribution, denoted by $f(y, \theta)$. Using the properties of densities, we can rewrite it as

$$f(y, \theta) = \pi(\theta) f(y|\theta) \tag{3.23}$$

Since θ is considered to be a random variable which we wish to learn more about, then we can 3.23 on 3.21 to deduce the following:

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} \tag{3.24}$$

From 3.24 we observe that the knowledge we have regarding θ after observing the data is a compromise between the statistical model and the prior information we possess on θ.

To draw conclusions about the population's parameter based on the obtained sample, we start by assigning an initial distribution to θ, known as the **prior distribution**, represented by $\pi(\theta)$. This distribution reflects the prior information we have about θ. If we have little to no prior knowledge about the parameter and its characteristics, we usually choose a distribution with a larger variance so that the possible values for θ don't have a too small probability. Such distributions are called **non-informative distributions**. Conversely, if we have some knowledge about the parameter, we tend to choose a distribution with a smaller variance, which are called **informative distributions**

Having picked a prior distribution, its aggregation with the chosen statistical model originates the **Bayesian model**, which is defined as

$$M = \{f(y|\theta); \theta \sim \pi(\theta)\} \tag{3.25}$$

The Bayesian model is the starting point for the whole process, which is determining the posterior distribution based on the data we have collected.

Once the data is collected, all information about θ is contained in the **likelihood function** $L_y(\theta)$. It is defined as a function of the parameter θ and, in the discrete case, it represents the probability of observing y, assuming that the value of θ is the true one:

$$L_y(\theta) \propto f(y|\theta) \tag{3.26}$$

After observing the data and conditioning θ to it, by applying Bayes' Theorem 3.21 we get the **posterior distribution**

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{f(y)}, \tag{3.27}$$

where $f(y) = \sum_\theta \pi(\theta)f(y|\theta)$ if θ is discrete (resp. $f(y) = \int_\theta \pi(\theta)f(y|\theta)d\theta$ if θ is continuous). This distribution reflects the information about the parameter θ after collecting the data.

Note that $f(y)$ doesn't depend on θ, thus we can write:

$$\pi(\theta)|y) \propto \pi(\theta)f(y|\theta) = \pi(\theta)L_y(\theta) \tag{3.28}$$

One might question why 3.26 has a *proportional* sign and not an *equal* sign. Brewer explains why: the prior times the likelihood cannot be equal to the posterior because it is not normalized, however the shape of the posterior is the same. From 3.28 we deduce that the posterior distribution is nothing more than a compromise between the prior distribution and the likelihood function, in other words, a compromise between the prior knowledge about θ and the evidence.

Using the prior distribution $\pi(\theta)$, we can define the **prior predictive distribution** as

$$f(\tilde{y}) = \int f(\tilde{y}|\theta)\pi(\theta)d\theta \tag{3.29}$$

Analogously, we can define the **posterior predictive distribution** using the posterior distribu-

tion:

$$f(\tilde{y}|y) = \int f(\tilde{y}|\theta)\pi(\theta|y)d\theta \tag{3.30}$$

The posterior distribution reflects everything we know about the behavior of the data. With it, we can make predictions, compute credible intervals, point or interval estimates, hypothesis tests, predictions and more.

3.2.3 Example

Let's use a very common and easy-to-grasp example in statistics: suppose we have a coin whose probability of landing on heads, p, is unknown. Let Y be the number of heads obtained in N coin tosses. Then

$$Y|\, p \sim Bin(N,p) \iff f_{Y|\, p}(k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{3.31}$$

Now, we must pick a prior distribution for p. We mustn't forget that the prior that we pick should have a support over all possible values for the parameter. Since $0 < p < 1$, a good candidate would be $\pi(\alpha,\beta) \sim Beta(\alpha,\beta)$, whose probability density function is $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$. Note that for $\alpha = \beta = 1$, we get $U(0,1)$, which is uninformative. However, from our real-word experience we know that the vast majority of coins are fair. Thus, we pick a $Beta(2,2)$ as our prior. This way, we assert that we expect p to be close to 0.5, but that we're aware that it could also take on any value in the interval $[0,1]$.

Let's say we perform an experiment, in which we toss the coin N times and we observe n_H heads. Recalling 3.28 and noting $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ doesn't depend on p, we have

$$\begin{aligned}\pi(p|y = n_H) &\propto f_{Y|\,p}(n_H) \times \pi(\alpha, \beta) \\ &\propto p^{n_H}(1-p)^{N-n_H} \times p(1-p) \\ &= p^{n_H+1}(1-p)^{N-n_H+1}\end{aligned} \tag{3.32}$$

Note that the expression of the posterior distribution seems quite familiar. That's because it looks exactly like a $Beta(\alpha = n_H + 2, \beta = N - n_H + 2)$ without the normalizing constant $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, which in this case would be $\frac{\Gamma(4)}{\Gamma(2)^2}$. Thus, for this statistical model, the prior distribution belongs to the same family of distributions as the posterior distribution. This means that the $Beta$ distribution is **conjugated** in the Binomial model. These kinds of distributions are often most desirable when choosing a prior due to their stability and efficiency in deriving the posterior distribution.

3.2.4 Markov Chain Monte Carlo (MCMC) and JAGS (Just Another Gibbs Sampler)

Obtaining the posterior distribution only used to be achievable through analytical methods, by solving the integrals mentioned above. However, since θ is often a multidimensional vector, this task is extremely difficult most times, and can even be impossible altogether. Nonetheless, ever since the rise of computers and computational methods, different algorithms have been developed to make this process much more efficient and quick, whilst maintaining an acceptable error rate. Some of the most important algorithms used in a variety of applications are **Markov Chain Monte Carlo (MCMC)**. For a comprehensive overview of these concepts, Martinez (2002) provides an extensive guide to the relevant techniques and methodologies mentioned in this section.

MCMC methods are a set of algorithms used to sample from a probability distribution, based on constructing a Markov chain that has the desired distribution as its stationary distribution. These methods were revolutionary because they allow the simulation of a wide range of distributions, which would otherwise be tedious to obtain analytically.

Within Bayesian Computational Statistics, every MCMC algorithm starts with a given initial value, θ_0, and the subsequent values are iteratively simulated from the previous one. After a certain number of simulations, the chain converges to its stationary value and the simulations

correspond to the posterior distribution. The chain must be:

- Homogeneous: Transition probabilities from one state to another are invariant;
- Irreducible: Every state can be reached from any other state in a finite number of iterations;
- Aperiodic: There are no absorbing states.

The main algorithm of interest to us belongs to the family of **Gibbs Sampling**, which is one of the simpler methods but nonetheless usually very effective. This algorithm is applicable when the joint distribution isn't explicitly known and/or is too hard to sample from, but we know the conditional distribution of each parameter. The fact that the algorithm is applicable does not necessarily mean it is the most appropriate, but in the case of Gibbs Sampling, it is preferred due to its efficiency.

JAGS (Just Another Gibbs Simulator) is a Bayesian software implementable in R. As its name suggests, it relies on Gibbs Sampling to simulate the posterior distribution from a data set and given Bayesian model, which needs to be previously defined. To have access to JAGS we first must download the software and then install it in R as a package. Given a Bayesian model $\{f(\tilde{y}|\theta), \theta \sim \pi(\theta)\}$, where $\theta = (\theta_1, \theta_2, \cdots, \theta_d)$ are the parameters of the model, an initial condition $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_d^{(0)})$ and a sample from the population $y = (y_1, \cdots, y_n)$, the underlying algorithm in JAGS is as follows: for $i = 0, 1, \cdots, n$, where n is the intended number of iterations, generate

1. $\theta_1^{(i+1)}$ from $\pi(\theta_1^{(i)}|\theta_2^{(i)}, \cdots, \theta_d^{(i)})$
2. $\theta_2^{(i+1)}$ from $\pi(\theta_2^{(i)}|\theta_1^{(i+1)}, \cdots, \theta_d^{(i)})$
3. ...
4. $\theta_d^{(i+1)}$ from $\pi(\theta_d^{(i)}|\theta_1^{(i+1)}, \cdots, \theta_{d-1}^{(i+1)})$

After defining the Bayesian model (which already contains the data's distribution and the prior distributions), we just need to provide JAGS with the data y (and other variables $x_1, \cdots, x_n$ should they be needed, for example in regression) to be used in the simulations and the parameters which we aim to simulate their posterior distribution. It must be noted that in JAGS, variance

components are instead written in terms of precision $\tau = 1/\sigma^2$. For example, Normal distributions are described as $N(\mu, \tau)$.

The goal is to correctly implement Markov Chains whose stationary distributions are the respective parameter's posterior distributions. It would be convenient to know at which point we could safely say that the chains converged and all the simulations actually correspond to a sample of the posterior distribution. However, since the generation of the chains depends on an initial condition, in most situations it's usually not possible to know if we started the chain within the admissible region of the posterior distribution. If the chain does converge to the desired distribution, only after a certain number of iterations (sometimes large) does it simulate values that belong in that region. This leads to the MCMC samples in this initial period not being randomly drawn from the posterior distribution. To correct this, we often simulate at least two chains (with different initial conditions) and then "burn" the initial iterations of each chain, known as the **burn-in period**. The number of iterations to discard depends on the moment when the chains converge, which can be analyzed through a **traceplot**, which plots the simulated chains together (Figure 3.2). Through it, we can have a visual confirmation the the chains did converge and a rough indicator of at which iteration the chains mixed together.

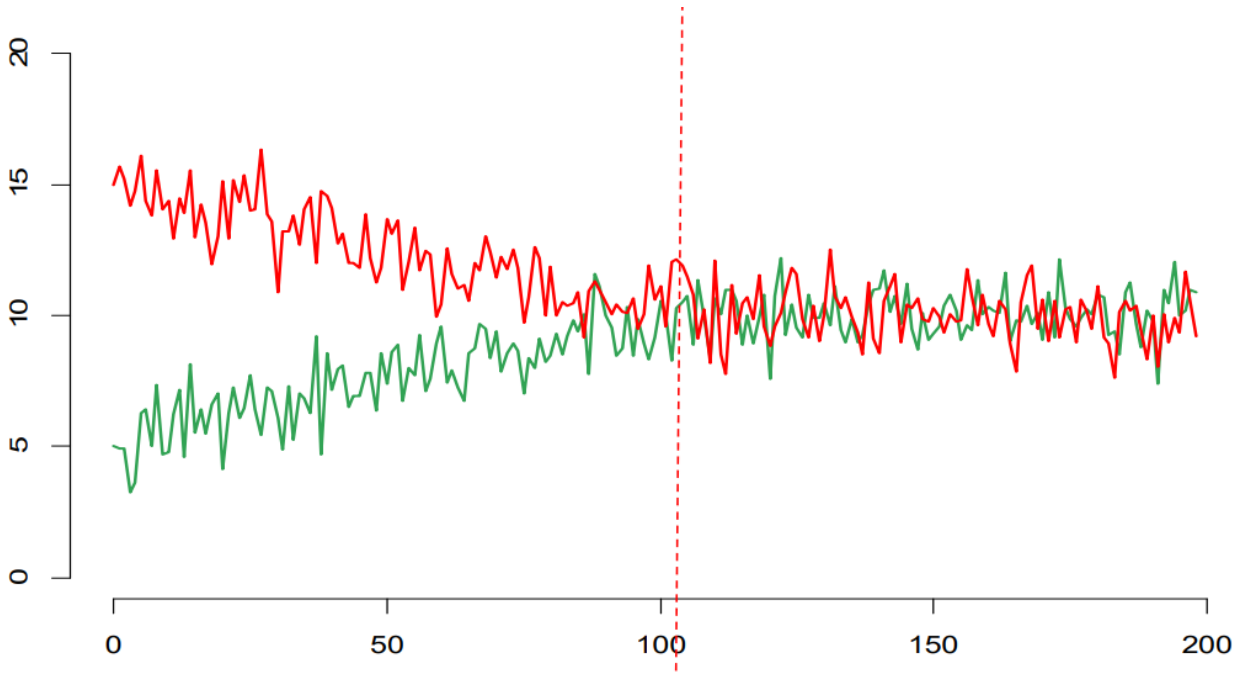


Figure 3.2: Traceplots of two simulated chains converging. Simulations before the red dashed line should be discarded (burned-in). Source: PowerPoint presentation by Xavi Puig Oriol, Departament d'Estadística i I.O., Universitat Politècnica de Catalunya, 2020.

Due to Markov Chains' own nature, the MCMC chains' simulations depend on previous ones to be generated, leading them to be substantially autocorrelated. Therefore, basing them on all simulations of the chains (after the burn-in period) is inadvisable, as we intend the generated chains to represent random draws of posterior distributions. Thus, it is common to perform inference on **thinned** chains, where only a fraction of the simulations will actually be saved, so that they are spaced apart enough to be considered "independent". In practice, the lag between two iterations needed for them to be considered essentially independent can be estimated through the analysis of an ACF, for example.

Despite these two issues, in general we can overcome them by carrying out a large enough number of simulations, which would provide us more liberty on choosing the burn-in period and the thinning. Despite this, if the Bayesian model is too complex and/or the dataset to be employed is very large, we need to balance the overall number of simulations and the burn-in period and the thinning.

Chapter 4.

Risk modeling

As mentioned in the first chapter, the goal behind Pricing is to determine a fair value for the premium that correctly reflects the insured's risk profile. For an insurance company, the pure premium is simply **the product of the frequency and severity of the claims**. However, these two components must be modeled separately, because whereas frequency data is related to count data, severity refers to the costs of claims, which is a continuous variable.

Due to the structure of the data, each policyholder has at least one observation associated with them. The data is hierarchical, meaning that if we fixate on one specific policyholder, we have at least one observation associated to them, introducing correlation between observations. Because independence of observations is an assumption of GLMs, we cannot apply them to the data, as we must tackle this correlation as well. Therefore, a random effect must be introduced at the policyholder's level to address the correlation between observations of the same individual.

The models in this project were fitted in R using the *glmmTMB()* function from the `stats` package, which is a special kind of GLMM function that provides some more stability, since we can choose from a list of optimizers in case the default one does not work. This function allows for various inputs, but the most relevant for our purposes are the model formula, offsets, the intended distribution family, and the link function. If no specific link function is provided, R defaults to using the canonical link function associated with the chosen distribution family.

4.1 Modeling the frequency

The frequency of claims is defined as the number of claims that occurred during the considered time frame. In other words,

$$\text{Frequency} = \frac{\text{Number of claims}}{\text{Exposure}} \tag{4.1}$$

Recall that exposure represents the duration during which a policy is active or the period during which the insured is protected. In our dataset, since each policy typically has a one-year duration, the exposure for most observations is 1. However, in some cases, adjustments were made to account for the data as of December 31, 2018.

Before proceeding with the models, it is important to remark that not all observations have the same exposure. Modeling claim counts without accounting for exposure would be inappropriate, since it is very different to have, for example, two claims in a one-year span or two claims in the course of ten years. To accurately model the claim frequency, we need to incorporate exposure into our models. Then, using 4.1, we have

$$\log\left(\frac{\text{Number of claims}}{\text{Exposure}}\right) = \eta \iff \log(\text{Number of claims}) = \eta + \log(\text{Exposure}) \tag{4.2}$$

If we take the exponential on both sides, we get

$$\text{Number of claims} = \text{Exposure} \times e^{\eta}.$$

Intuitively, this makes sense. For two individuals with identical characteristics, if one has double the exposure of the other, we would expect the first individual to have double the number of claims compared to the second. In other words, the number of claims should increase proportionally with exposure. By the above equation, we see that exposure should enter the model as a logarithmic variable with a coefficient of 1. This can be achieved in R using the *offset* command, which explicitly includes a variable in the model with a fixed coefficient.

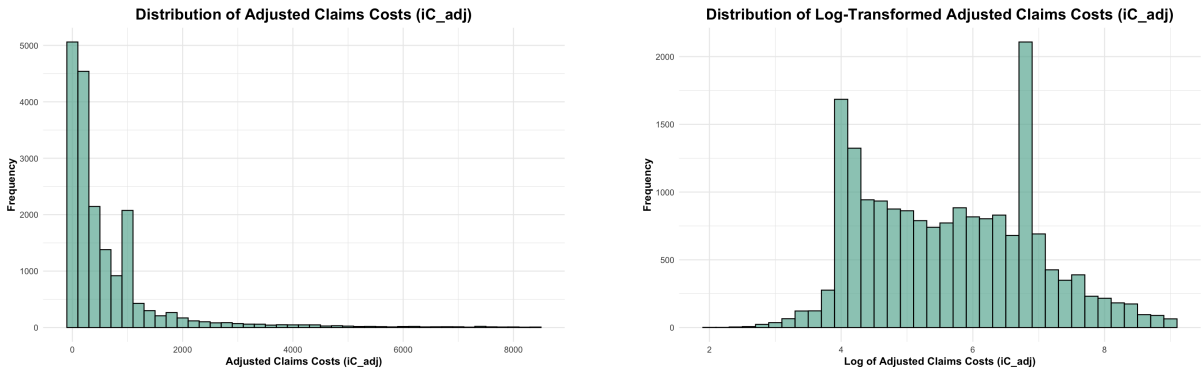
The most common approach to modeling the frequency is using a GLM with a Poisson distribution, usually using the canonical link function $g_c(\mu) = \log(\mu)$

4.2 Modeling the severity

Claim severity is defined as the monetary loss of an insurance claim, seen as an average cost per claim. In other words,

$$\text{Severity} = \frac{\text{Cost of claims}}{\text{Number of claims}} \tag{4.3}$$

The severity of claims is directly related to their cost: higher severity leads to more costly claims and vice versa. Given the substantial variability in costs from policy to policy, insurance companies must accurately model each risk profile to protect their reserves.



Claims Costs Log Claims Costs

Figure 4.1: Distribution of Claim Costs

The distribution of claim costs is typically positively skewed as there are usually more claims with lower severity. This characteristic is on par with what our data show in the plot on the left of figure 4.1. Although several distributions from the exponential family could be appropriate for modeling claim costs, actuaries and many authors often prefer the Gamma distribution. To maintain a multiplicative model structure, the log-link function is commonly used instead of the canonical link function, the inverse function.

4.3 The standard procedure

We begin by applying the traditional methodology commonly used by insurance companies, as detailed in the previous sections. Both models were developed by manually applying a stepwise selection and subsequently removing non-significant variables, whose significance was assessed using the LRT, described in Section 3. The reference categories for the categorical variables were chosen on the basis of their highest frequency:

- **Area**: Rural;
- **Type_risk**: Passenger cars;
- **Second_driver**: No;
- **Type_fuel**: Diesel.

Due to the large size of the dataset, interactions between variables were not considered.

In general, if the predictors present in the model have an estimated positive coefficient, then they are positively associated with the response. Otherwise, if they have an estimated negative coefficient, then those predictors are negatively associated with the response. Thus, a predictor is associated with more frequency/cost of claims if its estimated coefficient in the respective model is positive. Since we are using a $\log$ link, we cannot interpret the coefficients directly, instead we interpret the exponential of the coefficient.

In the frequency model, continuous variables had to be standardized, per the suggestion of the actual GLMM function, whose output would print an error if they were not standardized. Table 4.1 summarizes the results of the frequency model using a Poisson fit with the $\log$ link function, which tells us how each variable influences the frequency of claims. Unfortunately, interpreting standardized continuous variables is not as straightforward as interpreting them normally. For example, the estimated coefficient for `Age` was about -0.05; since $e^{-0.05} \approx 0.95$, then, keeping all other variables constant, for every increase in one standard deviation of `Age`, the frequency of claims decreases about 5%. Regarding categorical variables, take, for example, the effect observed for the dummy "Vans", which is a category within the `Type_risk` variable; since

$e^{0.09567} \approx 1.10$, it means that van drivers have an increased 10% frequency of claims than those who drive passenger cars, keeping every other variable constant.

Table 4.1: Poisson Model Summary w/ $\log$ link function: Estimates, Standard Errors, and p-Values; standardized continuous variables

Variable	Dummy (if applicable)	Estimate	Std. Error	P-Value
Intercept		-2.57041	0.02366	< 2e-16
Age		-0.05087	0.01108	4.38e-06
Age_of_car		-0.13805	0.01155	< 2e-16
Seniority		0.08586	0.01164	1.64e-13
Premium_adj		0.14995	0.00929	< 2e-16
R_Claims_history		1.29471	0.01179	< 2e-16
Type_risk	Vans	0.19839	0.03024	5.36e-11
Area	Urban	-0.0363	0.01671	0.0301
Second_driver	Yes	0.13423	0.03010	8.20e-06
Distribution_channel	Insurance Broker	-0.0458	0.01587	0.0039
σ_u^2		0.1759		
Residual Deviance		83510.6		
AIC		83530.6		
BIC		83621.6		

For severity models, in contrast to frequency models, it was found that standardizing the continuous variables actually resulted in exploding error rate. Table 4.2 summarizes the results of the severity model using a Gamma fit with $\log$ link function, which tells us how each variable influences the cost of claims. For this model, every dummy regarding each categorical variable turned out to be non significant. It is also interesting to note that all common predictors between the two models have the same sign. This means that every variable associated with the higher frequency of claims is also associated with higher costs of claims.

Now that we have an idea of how each variable affects frequency and severity, we can finally estimate the pure premium. It is based on the **relativities** of each variable for both models, which are obtained by applying the inverse of the link function to the coefficients for each model (tables 4.1 and 4.2). If the identity link function is used in any model, the relativities for that model correspond to the coefficients.

Table 4.2: Gamma GLMM with $\log$ link function: Estimates, Standard Errors, and p-Values; standardized continuous variables

Variable	Estimate	Std. Error	P-Value
Intercept	5.583	0.04813	< 2e-16
Age_of_car	-0.02459	0.00222	< 2e-16
Seniority	0.00493	0.00223	0.0275
R_Claims_history	0.2601	0.01312	< 2e-16
Premium_adj	0.00129	0.000075	< 2e-16
σ_u^2	0.693		
Residual Deviance	187377.8		
AIC	187391.8		
BIC	187444.1		

The **tariff relativities** for each variable/category, which are the ones used in the pure premium equation, are calculated by multiplying the relativities from each model, since they are both multiplicative. These combined relativities are presented in Table 4.3 and reflect the following combined tariff:

$$\begin{aligned}\text{Pure Premium} = {} & 20.3400 \times (0.9503)^{\text{Age}} \times (0.8506)^{\text{Age\_of\_car}} \\ & \times (1.0949)^{\text{Seniority}} \times (1.1632)^{\text{Premium\_adj}} \times (4.7315)^{\text{R\_Claims\_history}} \\ & \times (1.2194)^{D_{\text{Vans}}} \times (0.9644)^{D_{\text{Urban}}} \\ & \times (1.1436)^{D_{\text{Yes}}} \times (0.9552)^{D_{\text{Insurance Broker}}}\end{aligned} \tag{4.4}$$

where D_i are dummy variables indicating the presence of each category (1 if the category is present, 0 otherwise).

4.4 Addressing overdispersion and zero inflation

As mentioned in the previous section, overdispersion is one of the main issues to combat when modeling claims counts. In R, it is very easy to check for overdispersion (or underdispersion) in GLMMs, simply by implementing the function *check_overdispersion* from the *performance* pack-

age, which is based on the code from Gelmann and Hill (2007) to test if there is overdispersion. The output contains an estimate of the overdispersion parameter, the Pearson statistic value χ^2 and the test value p. If we are able to reject the null hypothesis (which states that there is overdispersion in the model), then we can assume that the model is not overdispersed. The test applied to the Poisson frequency model renders the following output:

```
# Overdispersion test


       dispersion ratio =     0.358
  Pearson's Chi-Squared = 23883.103
                p-value =         1


No overdispersion detected.
```

For a level α of 0.05, we cannot reject the null hypothesis that there is no overdispersion, suggesting that there is no reason not to consider the Poisson model.

Another important issue we need to address is that of zero-inflation. As stated before, both Poisson and Negative Binomial models generate some zeros, though in this case we are easily able to detect that there is an overwhelming presence of them. Even though we concluded that there is no overdispersion in the Poisson model, neither it nor the Negative Binomial model can address the excess of zeroes in their natural form. Fortunately, in R we have handy the same function used for those models, *glmmTMB()*. However, for Zero-Inflated models, we need to specify two formulas, `formula` and `ziformula`. The former represents the regular count process (Poisson or Negative Binomial) and the latter represents the Binomial process. Thus, if one so chooses, they may employ different predictors for each process. In our case, the manual stepwise approach was also taken with the Binomial process, similarly to the count process and all the models fitted before.

Table 4.3: Tariff Relativities using Poisson model for frequency and Gamma model for severity

Variable	Category	Frequency Relativity	Severity Relativity	Tariff Relativity
Intercept		-2.57041	5.583	20.3400
Age		0.9503	1	0.9503
Age_of_car		0.8712	0.9757	0.8506
Seniority		1.0896	1.0049	1.0949
Premium_adj		1.1617	1.0013	1.1632
R_Claims_history		3.6485	1.2969	4.7315
Type_risk	Passenger Cars	1	1	1
	Vans	1.2194	1	1.2194
Area	Rural	1	1	1
	Urban	0.9644	1	0.9644
Second_driver	No	1	1	1
	Yes	1.1436	1	1.1436
Distribution_channel	Agent	1	1	1
	Insurance Broker	0.9552	1	0.9552

4.5 Comparing fitted frequency models

Apart from the Poisson model, three other models were fitted: Negative Binomial, ZIP and ZINB. Even though the overdispersion test did not allow us to conclude that there is overdispersion in the Poisson model, it is still interesting to compare its performance against the Negative Binomial model. Not only that, but the excess of zeroes motivated the fitting of ZIP and ZINB models.

All models were fitted using the training set, comprising 66640 observations, and tested using the test set, comprising 28561 observations, according to the partition 70% | 30% of the data set. In either partition, the vast majority of observations are associated with an absence of claims (approximately 77% in both sets). The results are displayed in Table 4.4.

Table 4.4: Frequency Model Comparison: Misclassifications, AIC, BIC, Deviance

Model	Misclassifications	AIC	BIC	Deviance
Frequency - Poisson	2302	83530.6	83621	83510.6
Frequency - Negative Binomial	2304	82308.5	82408.7	82286.5
Frequency - ZIP	1576	72180.1	72307.6	72152.1
Frequency - ZINB	1355	71708.1	71844.7	71678.1

Since the response variable is discrete, misclassifications were computed by predicting the response variable and checking if the predicted result differed, at most, of one claim from the true value. In other words, we consider a prediction to be a misclassification if $|\hat{y} - y| > 1$. By analyzing all the metrics in the table above, we can safely say that the ZINB model had the best overall performance. Between the Poisson and Negative Binomial, misclassifications were almost identical in proportion, but the goodness of fit metric suggest that the Negative Binomial model has a slightly better fit than Poisson's.

Despite being the best model in this situation, it's likely that actuaries wouldn't pick it since it loses the property of being multiplicative, due to being a mixture of models. This means that pure premium can't be represented as a function of tariff relativities. Although it can still be calculated, it is not as easy as when we're considering a purely multiplicative model, such as the Poisson or Negative Binomial with $\log$ link.

The estimated coefficients for the ZINB model can be consulted in table 4.5. Interpreting the coefficients in the count process is the same as the Poisson model. However, the coefficients of the binomial process aren't interpreted in the same way.

For a binary predictor X_1 (coded as 0 or 1), the odds ratio (OR) of observing a zero count when X_1 changes from 0 to 1 is given by:

$$\text{OR}(X_1 = 0, X_1 = 1) = \frac{\pi(X_1 = 1)/(1 - \pi(X_1 = 1))}{\pi(X_1 = 0)/(1 - \pi(X_1 = 0))} = e^{\beta_1} \tag{4.5}$$

Thus, the exponential of the coefficient e^{β_1} represents the odds ratio of transitioning from the reference class (e.g., no zero count) to the class where a zero count is more likely, while keeping

all other variables constant.

For a continuous predictor X_1, if X_1 increases by 1 unit, the odds ratio of observing a zero count is given by:

$$\text{OR}(\text{increase in } X_1 \text{ by } 1) = e^{\beta_1} \tag{4.6}$$

In this case, e^{β_1} represents the odds ratio of observing a zero count for a one-unit increase in the continuous predictor X_1, while keeping all other variables constant.

Table 4.5: ZINB Model Summary: Count Process, Inflation Process, and Additional Information

Variable	Dummy (if applicable)	Estimate	Std. Error	P-Value
	Count Process			
Intercept		-2.57041	0.02366	< 2e-16
Age		-0.05087	0.01108	4.38e-06
Age_of_car		-0.13805	0.01155	< 2e-16
Seniority		0.08586	0.01164	1.64e-13
Premium_adj		0.14995	0.00929	< 2e-16
R_Claims_history		1.29471	0.01179	< 2e-16
Type_risk	Vans	0.19839	0.03024	5.36e-11
Second_driver	Yes	0.13423	0.03010	8.20e-06
Type_fuel	Diesel	-0.07661	0.02973	0.0101
	Inflation Process			
Intercept		-0.8502	0.0657	< 2e-16
Age		0.0458	0.0153	0.0020
Age_of_car		0.0289	0.0158	0.0610
Seniority		-0.0324	0.0161	0.0390
Premium_adj		-0.0485	0.0137	0.0006
R_Claims_history		-0.2314	0.0162	< 2e-16
Type_risk	Vans	-0.0543	0.0286	0.0550
Second_driver	Yes	-0.0942	0.0291	0.0012
Type_fuel	Diesel	0.0624	0.0312	0.0430
σ_u^2		0.1858		
Residual Deviance		71678.1		
AIC		71708.1		
BIC		71844.7		

4.6 Fitting Severity Models

We have seen that the distribution of claim counts has a high number of zeros, which implies that for those rows, the associated costs are 0. Thus, for modeling the severity, the data set had to be empty of every row where there was a null cost. Most commonly, insurance companies use the Gamma model with a $\log$ link, although other models can also be employed, for example Lognormal with $\log$ link. In addition to the Gamma GLM with $\log$ link, the Lognormal LMM model was also used.

As shown in Figure 4.1, claim costs are positively skewed, with a long tail to the right. The presence of **large claims** are mostly responsible for these long tails and can make the model estimates not as reliable. To avoid affecting the fit of the model with large infrequent values, Ohlsson and Johansson (2010) suggest truncating the distribution, skipping high claim costs over a certain value C and, when adequate, treating very high losses separately. However, there is no official consensus regarding this value: while on one hand we want it to be as large as possible to cover as many claims as we can, it needs to be small enough to avoid getting unreliable estimations. Certain quantiles of the variable were analyzed to understand how costs are comprised into each of them. They are as follows:

Quantile levels (q)	$F^{-1}(q)$
0.90	96.95
0.925	294.38
0.95	882
0.975	2906.33
0.99	8340.01

Table 4.6: Quantiles and Corresponding Values for Cost of Claims

Table 4.6 illustrates how most claims are associated with relatively low costs. More specifically, 99% of claims have a cost equal to or less than 8340, as also corroborated by the quantile plot in Figure 4.2. Therefore, we define the threshold C as the 99th quantile.

As stated in Section 4.3, most actuaries opt for the Gamma distribution. However, there are many other distributions that are also valid to model the severity, such as Weibull, Pareto and Lognormal, to name a few. In this case, due to the data's characteristics and the limitations that

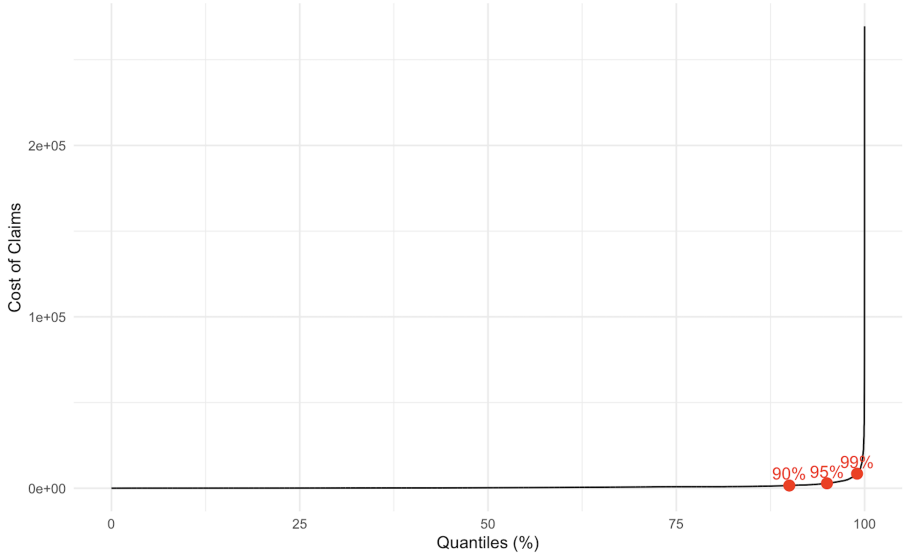
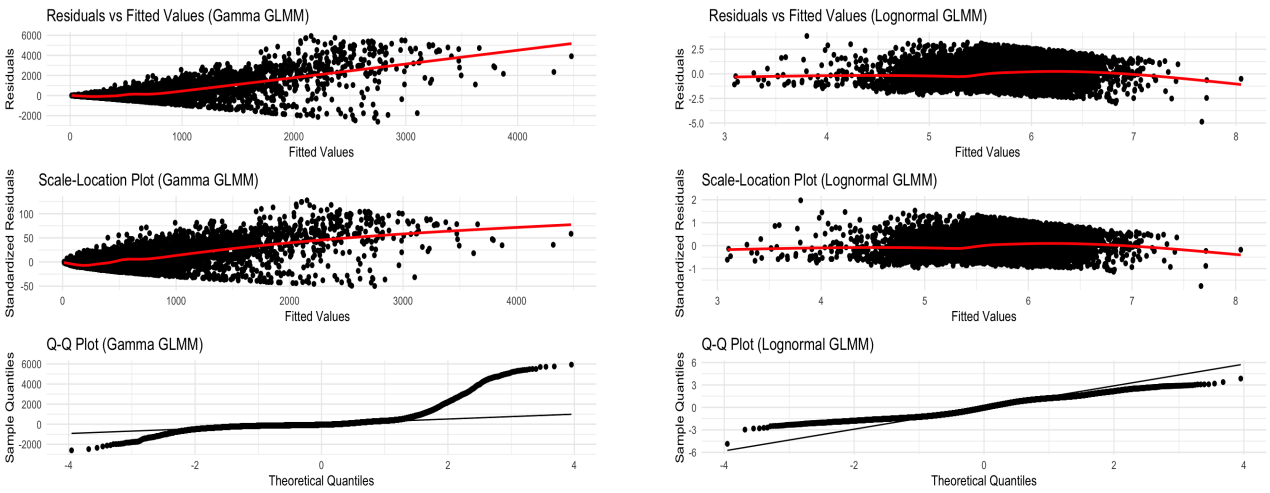


Figure 4.2: Quantile-plot of Cost of Claims

came with them, only the Lognormal distribution was considered.

In Section 4.3 the Gamma model was already fitted, whose results can be consulted in table 4.2. If a variable is assumed to follow a Lognormal distribution, then the logarithm of that variable follows a Normal distribution, which redirects us to an LMM instead. Thus, the variable $\log(Y)$ was fitted with an LMM, whose results can be checked in table 4.7.

After fitting the models, it is imperative to assess their goodness of fit by examining diagnostic plots (figure 4.3), which give us an indication of the model's performance. The residuals in the Gamma model exhibit heteroskedasticity due to their funnel shape, whereas the Lognormal model does not reveal any defined patterns. The QQ-Plots shows that the fit isn't the best in the Lognormal model, however, it is much worse in the Gamma model.



Diagnostic plots - Gamma model | Diagnostic plots - Lognormal model

Figure 4.3: Diagnostic plots for fitted Severity models

Table 4.7: Lognormal GLMM Summary: Estimates, Standard Errors, and p-Values

Variable	Estimate	Std. Error	P-Value
Intercept	5.221	0.04189	< 2e-16
Age_of_car	-0.02423	0.002035	< 2e-16
R_Claims_history	0.1829	0.01165	< 2e-16
Premium_adj	0.001090	0.00007015	< 2e-16
σ_u^2	0.2361		
σ^2	1.41		
Residual Deviance	40646.2		
AIC	40660.2		
BIC	40712.4		

4.7 Comparing fitted severity models

Analogously to the frequency models, all models were fitted using the training set, which comprises 12898 observations, and tested using the test set, comprising 5529 observations, according to the partition 70% | 30% of the data set. By analyzing all the metrics in Table 4.8 and the diagnostic plots in Figure 4.3, we verify that the AIC and BIC of the Lognormal model are much lower than those of the Gamma model. In addition to that, every metric in the Lognormal model is also notably lower, despite having a slightly higher RMSE than the Gamma model. In summary, it is clear that the Lognormal model is the best choice for these data.

Table 4.8: Comparison of Severity Models AIC, BIC, MAE, RMSE, and Median Absolute Error

Model	AIC	BIC	Deviance	MAE	RMSE	Median Absolute Error
Gamma GLMM	187391.8	187444.1	187377.8	294.88	849.93	262.22
Lognormal GLMM	40660.2	40712.4	40646.2	150.85	866.27	5.23

4.8 The Bayesian Approach

Section 3.2 covered the basic concepts of the Bayesian modeling, mentioning the definition of the Bayesian model by specifying the probability model we intend to use as well as the priors for all the parameters. Even though we are dealing with two different approaches, we can still take some notes from the frequentist GLMMs, in the sense that the same steps taken before were also taken here, such as standardization of continuous variables for the frequency models. Due to computational limitations, we were only able to test Poisson and Negative Binomial models for the frequency and Gamma and Lognormal for the severity. The reason for not including Zero-Inflated frequency models lies in the fact that in these cases we are fitting two different processes, each with their own parameters, which paired with the large dataset becomes computationally exhausting. For severity models, the reasons are the same as described in Section 4.6.

Naturally, the first choice we come across is which priors to use for the model's parameters. As stated in Section 3.2, when we do not have much information, if any, the chosen priors must be non-informative and should cover **all possible values**. As was the case, for every frequency and severity model, the priors were chosen as follows:

1. Let $\{\beta_i : i \in \{0, \cdots, p+1\}\}$ be the model parameters. Then,

$$\beta_i \sim N(0, 10^2) \tag{4.7}$$

   This prior's support is $\mathbb{R}$ and attributes a larger probability to the values closest to 0, which is what intend in regression: we don't want them to be too large in absolute value, otherwise relativities can start to lose sense;

2. Let σ_u^2 be the variance of the intercept's random effect and σ^2 the variance of errors in the Lognormal model. We define the priors of σ_u and σ as a truncated Normal distribution, in which its support becomes $\mathbb{R}_0^+$ and still assigns more probability to values close to 0. Then,

$$\sigma_u, \sigma \sim N(0, 10^2)T(0, \infty) \tag{4.8}$$

3. Let θ be the Negative Binomial model's overdispersion parameter. Then

$$\theta \sim Gamma(1, 1) \tag{4.9}$$

   Since $\theta > 0$, this prior's support, $\mathbb{R}^+$, matches this restriction.

4. Let α be the shape parameter for the Gamma model (α is also used in the rate parameter as $rate = \alpha/\mu$, where μ represents the model's mean). Then,

$$\alpha \sim Gamma(1, 1) \tag{4.10}$$

   Since $\alpha > 0$, this prior's support, $\mathbb{R}^+$, also matches this restriction.

Almost all the remarks in the introduction of Section 4.3 regarding the protocol followed to achieve the best possible are valid here as well. However, the significance level of the predictors was assessed differently: for each parameter, a 95% credibility interval (CI) was built based on the quantiles 0.025 and 0.975 of their posterior distribution; if 0 belonged to the CI, the predictor was deemed not significant and was removed from the model.

The output is given in a form of table, for each parameter containing its expected value, standard deviation and quantiles 2.5%, 25%, 50%, 75% and 97.5%. There are some concepts that appear to be foreign but are indeed analogous to frequentist GLMM:

- **Deviance Information Criterion (DIC)**

  DIC balances model fit and complexity, providing a criterion for model comparison, similar to frequentist GLM(M)'s criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

- **Effective Number of Parameters (p_D)**

  p_D represents an estimate of the effective number of parameters in the model, including both fixed and random effects. It is a complexity indicator and helps in model comparison and selection. It is defined as

$$pD = \frac{\text{Var}(\text{Deviance})}{2}$$

- **Deviance**

  Deviance measures the fit of the model. Just like in frequentist GLM(M)s, it is used to compare nested models and assess goodness of fit. It is defined as

$$\text{Deviance} = -2 \times \log(\ell(f(y|\theta))$$

- $\hat{R}$ **(Rhat statistic - Potential Scale Reduction Factor)**

  The Rhat statistic is used to assess the convergence of Markov chain Monte Carlo (MCMC) simulations. Gelman and Rubin (1992) define it as

$$\hat{R} = \sqrt{\frac{\widehat{V}}{W}} \tag{4.11}$$

  where W is the within-chain variance and $\widehat{V}$ is the estimate of the target distribution's variance based on both within-chain and between-chain variances. Values of $\hat{R}$ close to 1 suggest convergence, while values significantly greater than 1 indicate lack of convergence. The threshold for $\hat{R}$ has continued to evolve over time, now fixed at 1.01 as stated by Vehtari (2021). Originally, Gelman and Rubin (1992) proposed a threshold of 1.1, which is less conservative. This change in the threshold marks a clear progression towards more strict

criteria for diagnosing convergence. For this project, the latter value shall be considered as a reference.

- **n.eff (Effective Sample Size)**: The effective sample size reflects the number of independent samples from the posterior distribution.

Comparing models can be achieved by comparing goodness-of-fit statistics present in the output of the model, namely DIC, pD and Deviance. This is similar to the comparison of AIC/BIC/Deviance in frequentist models. In this sense, the best frequency model turned out to be the Negative Binomial model, since the Poisson model had very large values for Deviance, pD and DIC, as shown in Table 4.9.

Effective Number of Parameters (p_D)	4.345321e+13
Deviance Information Criterion (DIC)	4.345322e+13
Mean of Deviance	36478273.437

Table 4.9: Goodness of fit statistics for Poisson JAGS model

The results obtained from JAGS for the chosen frequency model are summarized in table 4.10. One of the first things to analyze in the model's diagnostics is the $\hat{R}$ values, which we have set a threshold of 1.1, as suggested by Gelman and Rubin (1992). All values fall below this threshold, pointing out that all chains have converged. We can also state that Deviance, pD and DIC are much lower compared to those of the Poisson model.

Table 4.10: Frequency JAGS - Negative Binomial

Parameter	mu.vect	sd.vect	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
`Intercept (b0)`	-2.635	0.041	-2.718	-2.662	-2.634	-2.607	-2.557	1.017	99
`Ageofcar (b2)`	-0.054	0.011	-0.076	-0.061	-0.054	-0.046	-0.031	1.003	670
`Valuevehicle (b3)`	-0.076	0.011	-0.097	-0.084	-0.076	-0.069	-0.054	1.001	3000
`Seniority (b4)`	0.121	0.011	0.100	0.114	0.121	0.129	0.142	1.001	3000
`Premiumadj (b5)`	0.335	0.011	0.315	0.328	0.335	0.342	0.357	1.002	1100
`RClaimshistory (b6)`	1.546	0.015	1.517	1.536	1.546	1.556	1.576	1.001	3000
`Typerisk (b7)`	0.155	0.030	0.096	0.136	0.155	0.176	0.214	1.014	120
`Typefuel (b10)`	-0.100	0.024	-0.147	-0.116	-0.100	-0.084	-0.052	1.001	1800
σ_u	0.558	0.021	0.513	0.545	0.561	0.573	0.595	1.032	470
θ	0.830	0.025	0.782	0.813	0.829	0.846	0.879	1.007	1400
Deviance	81581.729	242.832	81129.709	81416.957	81568.130	81740.303	82083.928	1.024	610

Effective Number of Parameters (p_D)	29444.4
Deviance Information Criterion (DIC)	111026.1

The behavior of each simulated chain can be visualized and analyzed using a trace plot (recall Section 3.2). Ideally, the simulated chains should appear mixed and constant with little volatility. For this model, Figure 4.4 confirms that all chains regarding the model's parameters indeed have mixed, are kept relatively constant, and their volatility is reduced. The overdispersion parameter θ exhibits the same behavior, although the mixing is slightly worse. For σ_u, the chains appear quite volatile, which looks like it might have impacted the mixing of the chains. These observations are also supported by the fact that its $\hat{R}$ value is the highest among all parameters. Despite this, since $\hat{R} < 1.1$, there is no reason to worry about them not converging.

The posterior distribution of each parameter in figure 4.5 shows us how the distribution of each of the parameters changed after observing the data. Starting from non-informative priors, we arrive at posteriors that provide good results. We also notice that all parameters except σ_u and θ started with a symmetric prior, and the respective posteriors obtained are also almost perfectly symmetrical. The other two parameters did not start with a symmetric prior, and whereas the shape of σ_u's posterior isn't quite symmetric, that of θ is.

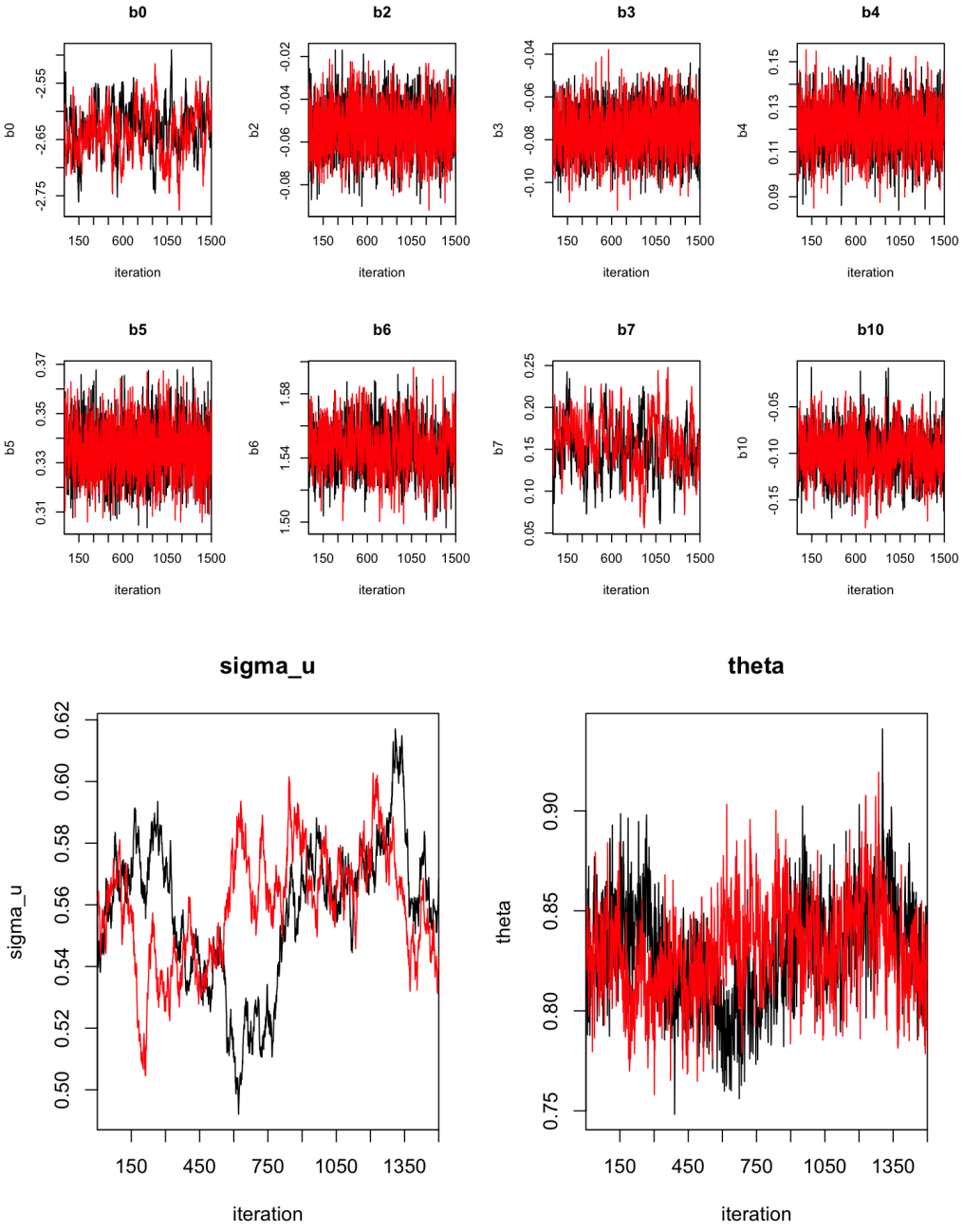


Figure 4.4: Traceplots of simulated chains for each significant coefficient in Negative Binomial model, the random effect's variance and overdispersion parameter

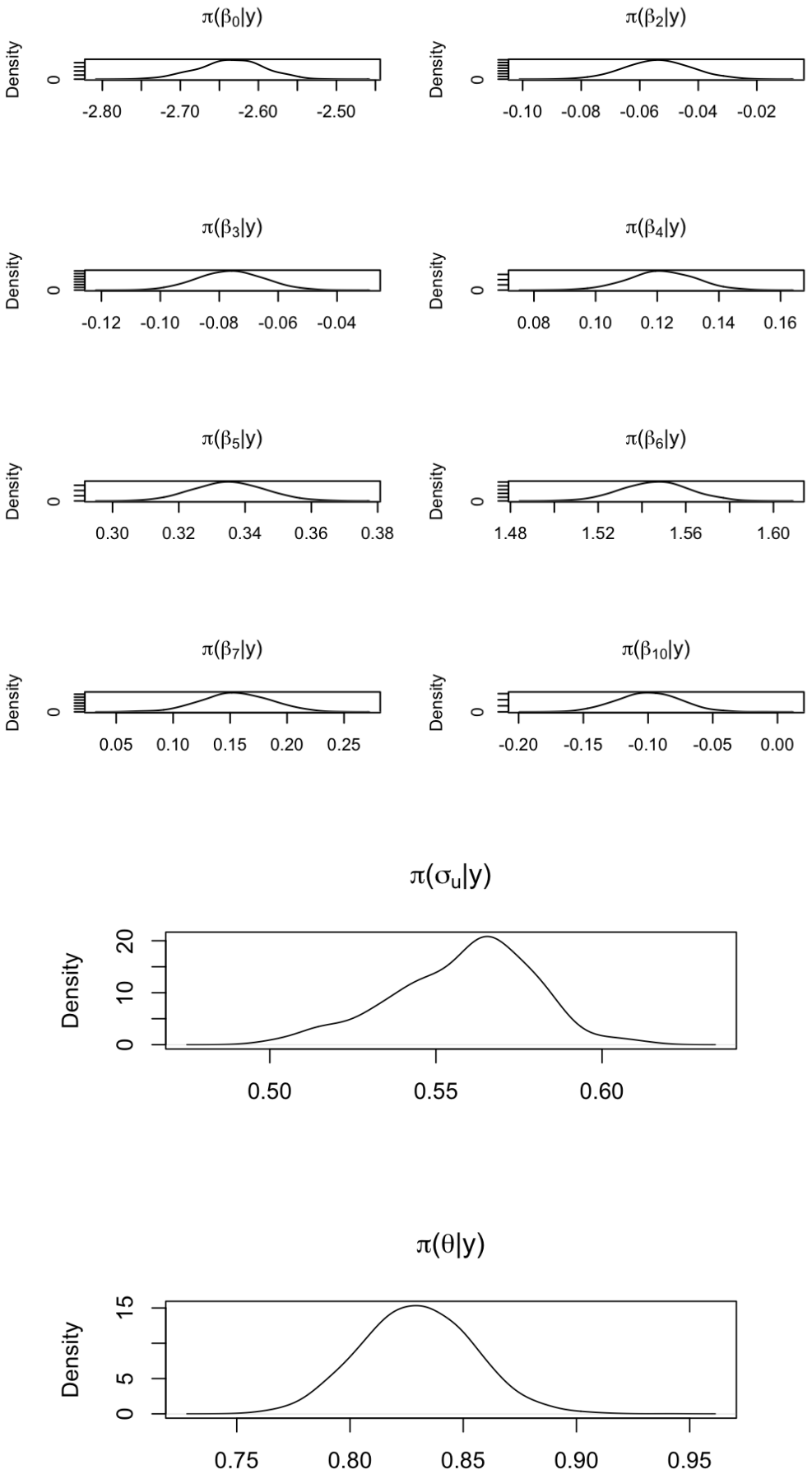


Figure 4.5: Posterior distributions of each parameter - Negative Binomial

For severity models a similar approach was taken in the sense that relevant priors were kept and new parameters were specified accordingly in the Bayesian models. As in frequentist severity models, continuous variables were not standardized.

The best severity model turned out to be the Lognormal model, as the Gamma model had larger values for the mean of Deviance, pD, and DIC, as shown in table 4.11:

Table 4.11: Goodness of fit statistics for Gamma JAGS model

Effective Number of Parameters (p_D)	32128.2
Deviance Information Criterion (DIC)	212531.3
Mean of Deviance	180403.102

The Lognormal model's results can be viewed in table 4.12. Once again, we must first analyze the values of $\hat{R}$, remembering that a threshold of 1.1 is being considered. All values fall below this threshold except for σ, σ_u, and deviance. However, the traceplot of σ_u presents a very similar behavior to that of Figure 4.4 and the traceplot of σ appears to be a bit more volatile than θ of the Negative Binomial model. We can also state that Deviance, pD, and DIC are much lower than those in the Gamma model.

Table 4.12: Severity JAGS - Lognormal

Parameter	**mu.vect**	**sd.vect**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**	**Rhat**	**n.eff**
`Intercept (b0)`	5.130	0.050	5.031	5.097	5.131	5.164	5.226	1.001	5000
`Age of car (b2)`	-0.022	0.002	-0.026	-0.024	-0.022	-0.021	-0.018	1.001	5000
`Value vehicle (b3)`	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1
`Seniority (b4)`	0.004	0.002	0.001	0.003	0.004	0.006	0.008	1.001	5000
`Premium adj (b5)`	0.002	0.000	0.002	0.002	0.002	0.002	0.002	1.001	5000
`R Claims history (b6)`	0.162	0.012	0.139	0.155	0.162	0.170	0.185	1.001	3800
σ	1.176	0.011	1.151	1.169	1.177	1.184	1.196	1.137	16
σ_u	0.242	0.042	0.173	0.208	0.240	0.268	0.338	1.274	9
Deviance	40777.999	198.738	40278.332	40666.115	40810.011	40928.038	41065.309	1.262	10

Effective Number of Parameters (p_D)	17614.1
Deviance Information Criterion (DIC)	58392.1

Similarly to the Negative Binomial model, the posterior distribution of each parameter in Figure 4.6 shows us how the distribution of each of the parameters changed after observing the data. On one hand, the model's coefficients share a very similar shape as that of the Negative Binomial

model. On the other hand, for the two remaining parameters, we see that the posterior distribution of σ_u is non-symmetric because it presents a curvy behavior, whereas the posterior distribution of σ has a bit of a tail to the left, but otherwise seems to be approximately symmetric.

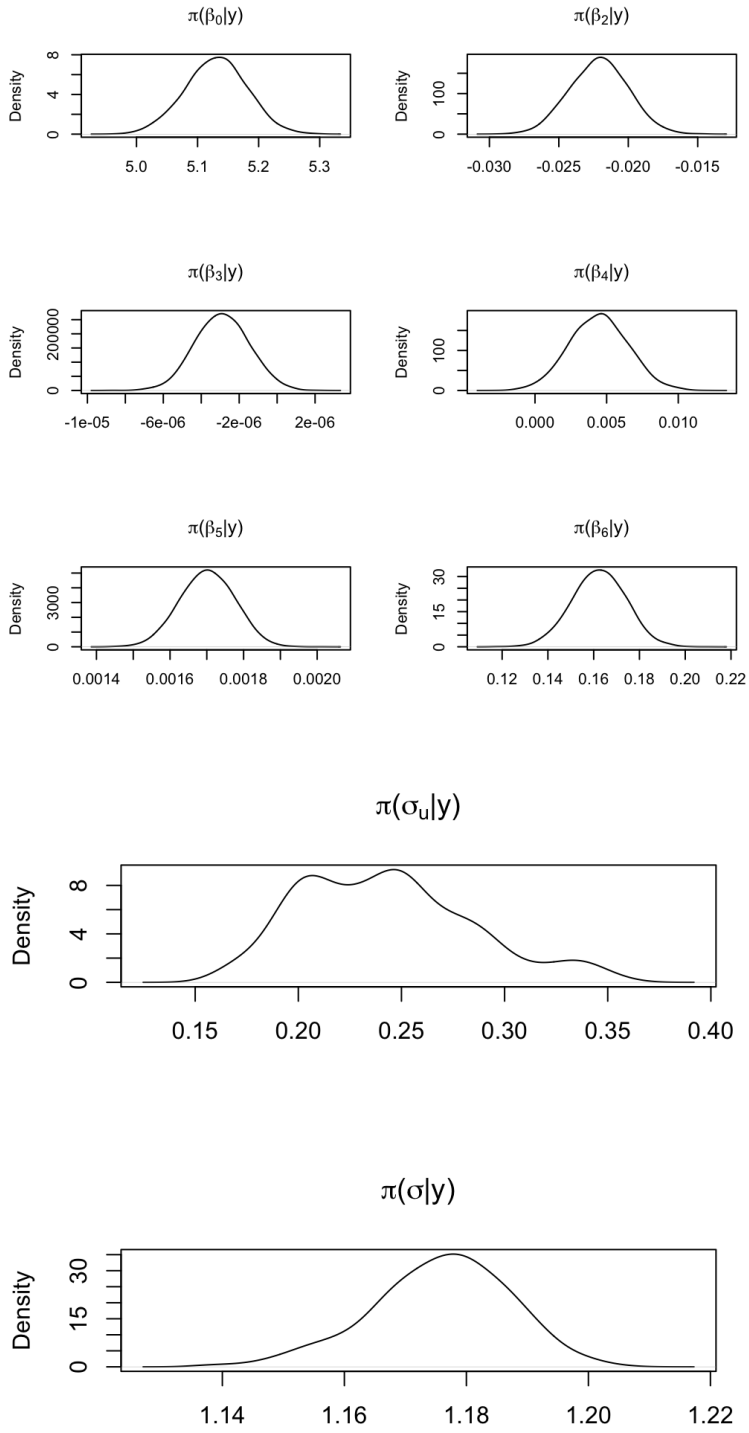


Figure 4.6: Posterior distributions of each parameter - Lognormal

4.9 Predicting claim counts and costs with Bayesian models

The usual method of predicting the response variable for a new set of data with frequentist GLMM is done by substituting the predictors with the data values, after the model has been fit. However, in Bayesian Statistics, the response instead follows a distribution $f(y|\theta)$, where $\theta = (\theta_1, \cdots, \theta_d)$ is a vector containing all the model's parameters. To make predictions using a set of covariates $X = X_1, \cdots, X_p$, we can follow these steps: for $i = 1, \cdots, s$,

1. Obtain the posterior distributions of $\theta_1, \cdots, \theta_d$;
2. Obtain $\theta_1^{(i)}, \cdots, \theta_d^{(i)}$ by sampling from posterior distributions of $\theta_1, \cdots, \theta_d$;
3. Generate $y_{\text{pred}}^{(i)} \sim f(y \mid X, \theta_1^{(i)}, \cdots, \theta_d^{(i)})$.

The posterior predictive distribution is made up of $y_{\text{pred}}^{(1)}, \cdots, y_{\text{pred}}^{(s)}$ and reflects the response behavior of a subject with these characteristics. The number of simulations s is subjective, but it should be large enough to ensure that it mimics the posterior predictive distribution as best as possible. Let us exemplify this by taking the claim cost of a random policyholder. Following the steps above for $s = 1000$ simulations, we obtain the posterior predictive distribution for this person (Figure 4.7).

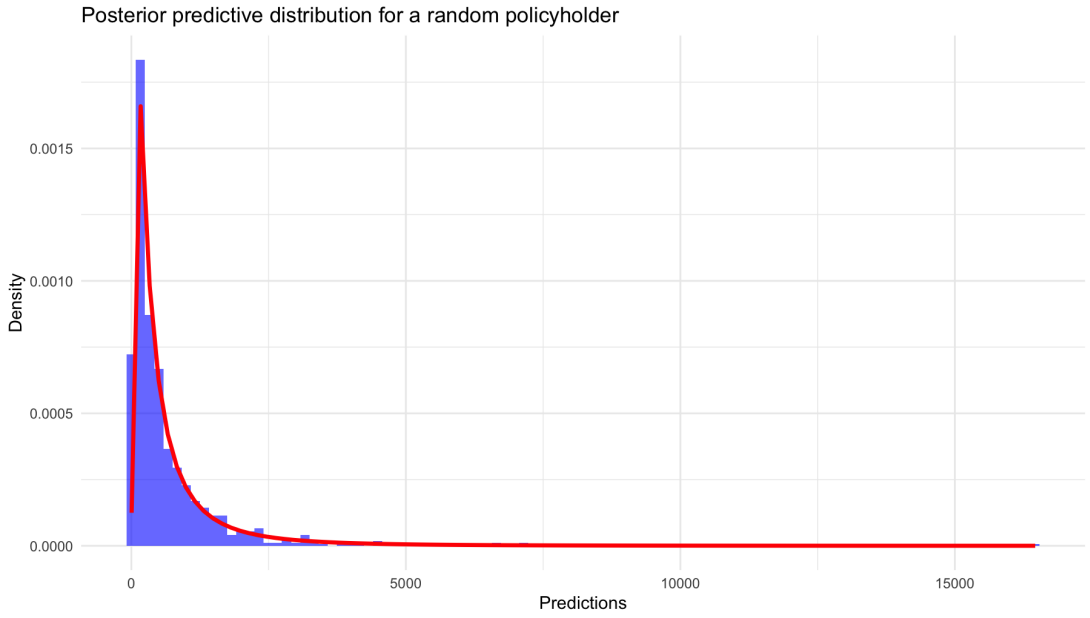


Figure 4.7: Posterior Predictive Distribution of Costs for a random policyholder, fitted with the Lognormal distribution

4.10 Comparison of results - Frequentist vs. Bayesian

This last section aims to compare the results from the best-performing models in both frequentist and Bayesian approaches. Given that goodness-of-fit statistics from these approaches are not directly comparable, we instead focus on alternative performance metrics, which were already computed for frequentist models. In Bayesian models, these metrics were approximated by taking the mean of the posterior predictive distribution for each subject. However, due to the complexity of the Bayesian frequency model, we were unable to compute misclassifications. Consequently, our comparison centers on the best severity model, which was the Lognormal model for both approaches.

Table 4.13: Comparison of Severity Models: Frequentist vs. Bayesian Metrics

Model	Deviance	MAE	RMSE	Median Absolute Error
Frequentist Lognormal GLMM	40646.20	150.85	866.27	5.23
Bayesian Lognormal GLMM	40778.99	528.84	832.48	392.90

Table 4.13 allows us to compare the metrics for each approach. Deviance is included here since the concept is similar in both approaches, though it's worth noting they're not exactly the same since the models were not fitted in the same way. By analyzing the table we note that deviance, MAE and Median Absolute Error are quite lower in the frequentist model when compared to those of the Bayesian model. On the other hand, RMSE is a bit larger in the frequentist model. Therefore, these values suggest that the frequentist model might be a better option to model severity. Despite this, we must bear in mind that the choice of priors greatly affects the fit of the Bayesian model, especially if they are more informative. Thus, if one were to choose different priors for every parameter, they might get a very different fit than what was obtained with the current choice of priors.

Chapter 5.

Conclusion and future work

Frequency and severity were modeled in different stages. For claim frequency, the ZINB model provided the best fit, as it was able to capture the excess zeros in the data more effectively than both the Poisson and Negative Binomial models.

For claim severity, the Lognormal model outperformed the Gamma model. The Lognormal GLMM's flexibility in modeling positive skewed data allowed it to better fit the observed data.

These findings highlight the importance of selecting the appropriate models for both claim frequency and severity. By choosing the right models, insurers can assess risk more effectively and develop more informed pricing strategies. Although the Bayesian approach is computationally demanding, it offers satisfactory parameter estimates and a better understanding of the uncertainties inherent to the models.

Achieving the best possible model is not an easy task, especially since many of the available distributions might not be suitable for application, either due to their characteristics or the data itself, as observed in this project. In Bayesian models, this challenge becomes even greater due to the lack of direct implementation and the need to manually adapt the models to fit the data.

Although Bayesian and frequentist methods are commonly applied separately, there remains an important question: should frequentist models be considered independently from Bayesian models? We argue that they should not. Each approach provides valuable insights that can complement each other. This holds true not only for insurance modeling, but for many other fields in science and humanities as well.

In conclusion, the results of this work provide a clear demonstration of the advantages and trade-offs of using traditional and Bayesian models for car insurance risk modeling. The ZINB model proved to be the most effective for frequency, and the Lognormal GLMM for severity. The

Bayesian approach, though time-consuming, offers a deeper understanding of parameter behavior and uncertainty.

There are other algorithms based on Bayesian statistics that may be more efficient. One such example is Stan, a library for Bayesian inference based on the No-U-Turn Sampler (a variant of Hamiltonian Monte Carlo), which is both faster and more user-friendly than traditional MCMC methods. For future work, it is interesting to consider this algorithm as an alternative to JAGS. We believe that with the ongoing advancement in algorithm efficiency, it is only a matter of time until Bayesian statistics becomes as efficient as frequentist models for large-scale applications such as Insurance Pricing. With further advances in computational methods, it will undoubtedly become a more accessible and efficient tool in actuarial science and beyond.

References

[1] Al-Mosawi, M. (2017). *An Extension of Generalized Linear Models for Dependent Frequency and Severity*, Stockholms Universitet.

[2] Brewer, B. J. (n.d.). *Introduction to Bayesian Statistics*. Retrieved from `http://creativecommons.org/licenses/by-sa/3.0/deed.en_GB`. This work is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License.

[3] Dormann, C. F., J. Elith, S. Bacher, et al. 2013. Collinearity: a review of met hods to deal with it and a simulation study evaluating their performance. *Ecography 36*: pp. 27–46.

[4] Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC. ISBN 1-58488-424-X (Print Edition), ISBN 978-1-58488-424-8 (Print Edition).

[5] Flossain, M. Z. (1995-1998). *AIC and BIC - The two competitive information criteria for model selection in economics and statistics. The Journal of Engineering Review, Part II: Statistical Science*, **XIX-XXII**, 133-140. Engineering University. (Printed in 2002).

[6] Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. ISBN 978-0-521-86706-1 (hardback), ISBN 978-0-521-68689-1 (paperback).

[7] Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, vol. 7, no. 4, pp. 457-511.

[8] Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2020). *Generalized Linear Models for Insurance Rating* (2nd ed.). Casualty Actuarial Society. ISBN 978-1-7333294-3-9 (print edition), ISBN 978-1-7333294-4-6 (electronic edition).

[9] Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press. ISBN 978-1-107-02833-3.

[10] Knudson, C. (2016). *Monte Carlo Likelihood Approximation for Generalized Linear Mixed Models*. PhD dissertation, University of Minnesota. Advisers: C. Geyer and G. Jones.

[11] Lledó, J., & Pavía, J. M. (2023). *Dataset of an Actual Motor Vehicle Insurance Portfolio*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Available at: `https://doi.org/10.3886/E193182V1`.

[12] Martinez, W. L., & Martinez, A. R. (2002). *Computational Statistics: Handbook with MATLAB*. Chapman & Hall/CRC, Boca Raton, FL.

[13] Murray, L., Nguyen, H., Lee, Y.-F., Remmenga, M. D., & Smith, D. W. (2012). *Variance Inflation Factors in Regression Models with Dummy Variables*. Conference on Applied Statistics in Agriculture.

[14] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), pp. 370-384. Royal Statistical Society.

[15] Ohlsson, E., & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, Heidelberg, Dordrecht, London, New York. ISBN 978-3-642-10790-0.

[16] Patterson, H. D., & Thompson, R. (1974). *Maximum likelihood estimation of components of variance*. In *Proceedings of the 8th International Biometric Conference*. Washington DC: Biometric Society.

[17] Searle, S. R., Casella, G., & McCulloch, C. E. (1992, 2006). *Variance Components*. John Wiley & Sons.

[18] Thompson, G. (2009). *Statistical Literacy Guide: How to Adjust for Inflation*. Last updated: February 2009. Available at: `https://researchbriefings.files.parliament.uk/documents/SN04962/SN04962.pdf`

[19] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-

Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, vol. 16, no. 2, pp. 667–718.

[20] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science+Business Media, LLC.

[21] R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: `https://www.R-project.org/`.