UiO : Department of Mathematics
University of Oslo

Modelling Car Insurance Data with Individual Effects

Hanne Tresselt

Master's Thesis, Spring 2020

This master's thesis is submitted under the master's programme *Stochastic Modelling, Statistics and Risk Analysis*, with programme option *Statistics*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 30 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

We show how both Poisson regression and recurrent events models can be used to model the number of claims to expect on a car insurance policy. We also show that the same is true when these models are extended to include a random effect/frailty. We then look at the effect of different assumptions made regarding the distribution of this random effect/frailty, through simulated data sets, one where we do not know the true distribution and several where we controlled the distribution and variance of the random effect/frailty. The results showed that the choice of frailty did seem to have an impact on the estimation of expected number of claims. They also indicated that the choice of distribution to use for the frailty was more important for data with a higher degree of heterogeneity than for data with a lower degree of heterogeneity.

Acknowledgements

First and foremost, a massive thanks goes out to my supervisor, Prof. Ørnulf Borgan, for keeping me on track, and always having great suggestions for improvements. This thesis would never have been what it is today without his guidance and supervision.

I also have to thank Julie Heggelund and Cecilie Haugstvedt for proofreading and making the probability of still finding mistakes a little smaller.

A special thank you also goes to Christian-Magnus Mohn for putting up with me over the last few months, even when I mostly had my head glued to the computer screen and was only physically present. Finishing this master's degree would not have been possible without your support.

And last, but not least, a special thanks to SARS-CoV-2 for making sure all distractions were kept to a minimum for the past three months. Also, thanks to everyone in study hall B802 and everyone in Realistforeningen for making Blindern a great place to be, in the days when campus was still open.

Hanne Tresselt
Oslo, June 2020

Contents

List of Figures

List of Tables

1

Introduction

One of the main contributing factors when setting a premium to be paid by a holder of a car insurance policy is the number of claims the policyholder is expected to make. This number of claims is most often modelled using a Poisson regression model, with covariates describing the policyholder and the car (de Jong and Heller 2008, ch. 6). An alternative approach is to use a recurrent events model with piecewise constant baseline intensity, and the covariates as proportional effects (Aalen et al. 2008, ch. 5).

In their standard form these models assume independence between policyholders and between different claims from the same policyholder. This implies that two policyholders with the same covariates, e.g., the ages of the policyholder and of the car, will have the same number of expected claims. Capturing all the information about the policyholders and their cars, that may have an effect on their risk of reporting a claim, is not practically feasible. In effect two policyholders with the same covariates may still have different risks.

To account for this heterogeneity the Poisson regression can be expanded to a generalized linear mixed model (GLMM) by adding a random effect to the model (Agresti 2015, ch. 9). Similarly, the recurrent events model accounts for this by the use of a proportional frailty model (Aalen et al. 2008, ch. 6).

The subject of this thesis is to study the two different modelling approaches. The main focus will be the effects of assumptions made regarding the distribution of the random effect/frailty, and how they affect the estimation of the number of claims to expect from a policyholder.

This is done by first presenting the two approaches of modelling in the homogeneous case, i.e., without any random effect/frailty. We then present the approaches including the random effect/frailty, i.e., the effect of heterogeneity, and different distributions to use for these. The same example data set is used throughout this part. In the end we study our own simulated data.

1.1 Outline

The rest of the text is organized as follows:

Chapter 2 introduces the Poisson regression model and the recurrent events model, and shows that they have proportional likelihoods. An example on vehicle insurance claims, to illustrate the results, is also introduced here.

Chapter 3 introduces the addition of a random effect/frailty to the models, an unobserved random variable for each policyholder, i.e., where the number of claims for a policyholder are no longer independent between time periods. This chapter covers the case where this random effect/frailty has a lognormal distribution. How to make the data compatible with a clustered survival setting is also covered.

Chapter 4 looks at two more distributions, gamma and inverse Gaussian, to use on the frailty(/random effect). It also covers the standardization of the lognormally distributed frailty(/random effect), making the results for this distribution comparable with the other two. The model in focus is now only the recurrent events model, and the distribution of the accompanying frailty.

Chapter 5 covers simulation of data and their use in exploring the importance of the choice of distribution for the frailty(/random effect) when fitting a model. The main focus is to see how this choice of distribution affects the expected number of claims.

Chapter 6 contains some concluding remarks and a discussion of the results.

Appendices A and B contain some additional results, and computer code used in the simulations.

1.2 Main R packages used

The fitting of generalized linear models was done using the packages `glmmML` (Broström 2019) and `lme4` (Bates et al. 2015), while the recurrent events models were fitted using the `parfm` package (Munda et al. 2017).

All tables that were exported from R to LaTeX were exported using the package `xtable` (Dahl et al. 2019). All plots were made using `ggplot2` (Wickham 2016), with the colour palette from `viridis` (Garnier 2018). `patchwork` (Pedersen 2019) was used for collecting plots in a grid. A lot of the transformation of data was done using functions from the `dplyr` package (Wickham et al. 2020).

2

The homogeneous models

In this chapter we will look at two different approaches to modelling counts, for instance the number of claims on an insurance policy. In this setting we will also assume independence between policyholders and between claims from the same policyholder, i.e., for two policyholders with the same given covariates, e.g., age and type of vehicle, the expected number of claims will be the same.

We will introduce the Poisson regression model and the recurrent events model, and show that they have proportional likelihoods such that methods for both generalized linear models and survival analysis can be used for modelling claim counts. We will illustrate the results with an example on vehicle insurance claims.

2.1 Poisson regression

We have a portfolio with n insurance policies and look at the number of claims each one of them reports during each of K years. We assume that policyholder i, where $i = 1, \dots, n$, is insured for a part E_{ik} of the k-th year for $k = 1, \dots, K$. E.g., if the policyholder is insured for six months, then $E_{ik} = 0.5$.

Then letting Y_{ik} be the observed number of claims for insurance policy i in year k, $Y_{ik} \sim \text{Pois}(\mu_{ik})$ is a reasonable assumption. Thus

$$\mu_{ik} = \text{E}(Y_{ik})$$

and

$$f(y_{ik}) = P(Y_{ik} = y_{ik}) = \frac{\mu_{ik}^{y_{ik}}}{y_{ik}!} e^{-\mu_{ik}}.$$

Now μ_{ik} will depend on E_{ik} and covariates that describe the policyholder and the car. We introduce the covariates $x_{ik1}, x_{ik2}, \dots, x_{ikp}$ with coefficients β_j, $j = 1, \dots, p$. If we include an intercept β_0, we may on vector form write $\mathbf{X}_{ik} = (1, x_{ik1}, x_{ik2}, \dots, x_{ikp})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$, so $\boldsymbol{\beta}^T \mathbf{X}_{ik} = \sum_{j=0}^{p} \beta_j x_{ikj}$.

We have that

$$\mu_{ik} = E_{ik} \mu(\mathbf{X}_{ik})$$

where $\mu(\mathbf{X}_{ik})$ is the expected number of claims in one year for a policyholder with covariates $\mathbf{X}_{ik}$. We now assume that

$$\mu(\mathbf{X}_{ik}) = e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} = \exp\left\{\sum_{j=0}^{p} \beta_j x_{ikj}\right\}.$$

This then gives

$$\mu_{ik} = E_{ki}\mu(\mathbf{X}_{ik}) = \exp\left\{\sum_{j=0}^{p} \beta_j x_{ijk} + \log E_{ik}\right\}. \tag{2.1}$$

In this section, we will assume that the number of claims Y_{ik} for policyholder i for the years $k = 1, \ldots, K$ are independent. We then get the following likelihood

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{k=1}^{K} \frac{\mu_{ik}^{y_{ik}}}{y_{ik}!} e^{-\mu_{ik}} = \prod_{i=1}^{n}\prod_{k=1}^{K} \frac{[E_{ik}\mu(\mathbf{X}_{ik})]^{y_{ik}}}{y_{ik}!} e^{-E_{ik}\mu(\mathbf{X}_{ik})},$$

which is proportional to

$$\prod_{i=1}^{n}\prod_{k=1}^{K} \mu(\mathbf{X}_{ik})^{y_{ik}} e^{-E_{ik}\mu(\mathbf{X}_{ik})}. \tag{2.2}$$

From the general theory on generalized linear models (GLM) (see, e.g., Agresti (2015)) we can now see that the model satisfies the three components of a GLM; the Poisson distribution belongs to the exponential family, we have a linear predictor η_{ik}, and a link function $g(\mu_{ik})$. So when, as in this case, using a logarithmic link function we get that the mean and the linear predictor are linked by

$$\eta_{ik} = \log(\mu_{ik}) = \boldsymbol{\beta}^T \mathbf{X}_{ik} + \log E_{ik} = \sum_{j=0}^{p} \beta_j x_{ikj} + \log E_{ik}.$$

The term $\log E_{ik}$ in (2.1) is often called an offset, and is, in this case, used to correct for differing time periods of observation (de Jong and Heller 2008, p. 67).

If we then look at a covariate, x_{ik1}, an increase of one unit will have a multiplicative effect of e^{β_1} on the mean,

$$\mu_{ik} = e^{\log E_{ik} + \beta_0 + \beta_1(x_{ik1}+1)} = e^{\log E_{ik} + \beta_0 + \beta_1 x_{ik1}} e^{\beta_1}.$$

So, when the covariate is at base level the expected number of claims will be $E_{ik}e^{\beta_0}$. When comparing level j with the base level we also get the multiplicative effect of e^{β_j}, giving an expectation of $E_{ik}e^{\beta_0}e^{\beta_j}$. (de Jong and Heller 2008, ch. 6)

2.1.1 Example: Vehicle insurance claims

Since actual insurance data are not readily available on the individual level, we will use a data set simulated by de Jong and Heller (2008) to illustrate the methods. The data set contains data on $n = 40\ 000$ different policies for $K = 3$ periods, each period being one year (i.e., $E_{ik} = 1$). The driver's age and the value of the vehicle are categorical covariates, each divided into six categories from youngest/cheapest to oldest/most expensive. The number of claims is listed for each policy for each of the three different periods, giving a data set of $3 \times 40\ 000 = 120\ 000$ observations.

Table 2.1: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-1.5900	0.0154	0.2039	0
Age category				
Group 1	0.2636	0.0214	1.3017	1.006e-34
Group 2	0.0842	0.0177	1.0879	1.996e-06
Group 3	0.0341	0.0172	1.0347	0.04741
Group 4	0.0000		1.0000	
Group 5	-0.1682	0.0203	0.8452	1.199e-16
Group 6	-0.0883	0.0234	0.9155	0.0001639
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1856	0.0148	1.2040	6.33e-36
$50 - 75$	0.1547	0.0417	1.1673	0.0002067
$75 - 100$	-0.6741	0.2183	0.5096	0.002018
$100 - 125$	-0.1749	0.2674	0.8396	0.5131
> 125	-1.4381	0.5001	0.2374	0.004029
Period				
1	0.0000		1.0000	
2	0.1062	0.0149	1.1121	8.51e-13
3	0.2344	0.0144	1.2641	2.239e-59

The model was fitted assuming independence both between policyholders, and between each time period for policyholder i. This was done using the `glm` function in R (R Core Team 2019), with a Poisson family and a log-link. The groups with the highest number of policies in them were used as a baseline, i.e., a policyholder in age category 4 with a vehicle worth less than 25 000$ in time period 1.

It can be seen from Table 2.1 that the baseline claim rate is $e^{\beta_0} = e^{-1.5900} = 0.2039$. It can also be seen that a younger person will have a higher claim rate than an older person. A person in age group 1 will for instance have $\frac{1.3017-0.8452}{0.8452} = 54\%$ higher claim rate than someone in age group 5. Similarly, a policy holder with a vehicle valued between 25 - 50 000$ will have a 20.4% higher claim rate than the ones with a vehicle with a value below 25 000$. The ones with higher value vehicles on the other hand will have 49.04%, 16.04% and 76.26% lower claim rates for vehicles valued between 75 - 100 000$, 100 - 125 000$ and above 125 000$ respectively, so the claim rate is not strictly decreasing with increasing value of the vehicle. There also seem to be a slight increase in the claim rates for each time period.

2.2 The recurrent events model

We now consider counting processes $N_i(t)$, $i = 1, \dots, n$, where $N_i(t)$ counts the number of claims for policyholder i in a time period $[0, t]$. The counting process $N_i(t)$ has a corresponding intensity process $\lambda_i(t)$, i.e., the probability of a claim occurring in a small time frame, $[t, t + dt]$, given all known information about the past until this time frame equals $\lambda_i(t)dt$. Under the assumption that the time frame is small enough such that it contains at most one claim, the number of claims occurring in $[t, t + dt]$, $dN_i(t)$, will be either zero or one, and we have $\lambda_i(t)dt = P(dN_i(t) = 1|past)$. (Aalen et al. 2008, p. 26)

A general likelihood valid for counting processes is given by

$$L = \left[\prod_{i=1}^{n} \prod_{0<t\leq\tau} \lambda_i(t)^{\Delta N_i(t)} \right] \exp\left\{ - \int\limits_0^{\tau} \lambda_{\bullet}(t)dt \right\} \tag{2.3}$$

where $\lambda_{\bullet}(t) = \sum_{i=1}^{n} \lambda_i(t)$ is the intensity process of the aggregated counting process $N_{\bullet}(t) = \sum_{i=1}^{n} N_i(t)$, and τ is the maximum observation time (Aalen et al. 2008, p. 212). Also $\Delta N_i(t)$ is the jump of $N_i(t)$ at time t, i.e., $\Delta N_i(t) = 1$ if policyholder i has a claim at time t, otherwise $\Delta N_i(t) = 0$.

We now assume that the counting process has the intensity process

$$\lambda_i(t) = Y_i(t)\alpha_i(t)$$

where $Y_i(t)$ is an at-risk indicator, i.e., $Y_i(t) = 1$ if policyholder i is under observation just before time t, and $Y_i(t) = 0$ otherwise. Note that if $Y_i(t) = 1$ for all t, then $N_i(t)$ will be a Poisson process with intensity $\alpha_i(t)$ (Aalen et al. 2008, p. 33). We will assume that the intensity, $\alpha_i(t)$, is given by

$$\alpha_i(t) = \alpha_0(t)e^{\boldsymbol{\beta}^T \mathbf{X}_i}$$

where the baseline intensity, $\alpha_0(t)$, is piecewise constant as a function of time.

We will now look at an interval of K years, and assume that the baseline intensity is constant for each year, i.e.,

$$\alpha_0(t; \boldsymbol{\theta}) = \sum_{i=1}^{K} \theta_k I_k(t)$$

where $I_k(t)$ is an indicator for subinterval k, i.e., $k - 1 < t \leq k$.

We now have the intervals $(k - 1, k]$, $k = 1, \dots, K$, and look at

$$O_{ik} = \int\limits_0^K I_k(t)dN_i(t) = N_i(k) - N_i(k - 1),$$

i.e., the number of claims for policyholder i in year k. We also define

$$E_{ik} = \int\limits_0^K I_k(t)Y_i(t)dt = \int\limits_{k-1}^{k} Y_i(t)dt,$$

which is the time under observation for policyholder i in year k. Then $\alpha_i(t)$ is equal to $\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}}$ for $k-1 < t \leq k$ and

$$\begin{aligned}\int_0^K \lambda_{\cdot}(t)dt &= \int_0^K \sum_{i=1}^n Y_i(t)\alpha_i(t)dt \\ &= \sum_{i=1}^n \sum_{k=1}^K \int_{k-1}^k Y_i(t)\alpha_i(t)dt = \sum_{i=1}^n \sum_{k=1}^K \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik}\end{aligned}$$

Hence we can write (2.3) as

$$\begin{aligned}L(\boldsymbol{\beta}) &= \left[\prod_{i=1}^n \prod_{k=1}^K \left(\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}}\right)^{O_{ik}}\right] \exp\left\{-\sum_{i=1}^n \sum_{k=1}^K \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik}\right\} \\ &= \prod_{i=1}^n \prod_{k=1}^K \left[\left(\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}}\right)^{O_{ik}} \exp\left\{-\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik}\right\}\right] \qquad (2.4)\end{aligned}$$

which we see is of the same form as the likelihood given in (2.2) for the Poisson regression. We can therefore look at recurrent events models using tools for Poisson regression, and vice versa.

2.3 Summary of chapter

In this chapter we have presented the two modelling approaches, Poisson regression and the recurrent events model. We have also seen how they have proportional likelihoods, and hence can both be used for modelling the number of claims on an insurance policy. We have also introduced a set of data that is used for examples also in chapters 3 and 4. In this first example we have looked at the results of fitting a model for this data using a generalized linear model with a Poisson distribution.

3

The heterogeneous models

When we consider each insurance policy as a cluster, the claims within this cluster will not be independent. In this chapter we will look at two ways of accounting for these correlations. While the previous chapter only accounted for the effects between the clusters, we will in this chapter also look at the effects within each cluster.

3.1 Generalized linear mixed model

The generalized linear mixed model (GLMM) is an extension of the generalized linear model (GLM) including an unobserved random variable for each cluster in the linear predictor. A cluster will in our situation be a single policyholder i, followed over several years, $k = 1, \ldots, K$. We will restrict ourselves to looking at models with a random intercept, i.e., models where we have a random $Z_i = e^{U_i}$. Given $Z_i = z_i$ (or $U_i = u_i$) we have that, the number of claims $Y_{i1}, \ldots, Y_{iK}$ for policyholder i, are independent and Poisson distributed with

$$\begin{aligned} \mathrm{E}(Y_{ik}|z_i) &= z_i \mu_{ik} = z_i E_{ik} \mu(\mathbf{X}_{ik}) \\ &= \exp(\log z_i + \boldsymbol{\beta}^T \mathbf{X}_{ik} + \log E_{ik}) \\ &= \exp(u_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ikj} + \log E_{ik}), \end{aligned} \tag{3.1}$$

with $\boldsymbol{\beta}$ being the fixed effects of the explanatory variables and u_i are the random intercepts with a particular probability distribution, often assumed to be independent observations from a $N(0, \sigma_u^2)$-distribution. Equivalently we may assume that the random effects Z_i are lognormally distributed. (Agresti 2015, ch. 9)

As given in Günther et al. (2014) the likelihood is now given by

$$L = \prod_{i=1}^{n} \left[\int_0^\infty \prod_{k=1}^{K} \left\{ \frac{(z_i E_{ik} \mu(\mathbf{X}_{ik}))^{y_{ik}}}{y_{ik}!} \right\} \exp\{-z_i E_{ik} \mu(\mathbf{X}_{ik})\} g(z_i) dz_i \right],$$

where $g(z_i)$ is the distribution of the random effects Z_i. We now introduce the Laplace transform,

$$\mathcal{L}(c) = \int_0^\infty e^{-cz_i} g(z_i) dz_i,$$

of the random effects. The q-th derivative of the Laplace transform is given by

$$\mathcal{L}^{(q)}(c) = (-1)^q \int\limits_0^\infty z_i^q e^{-cz_i} g(z_i) dz_i. \tag{3.2}$$

We can then write the likelihood as

$$L = \prod_{i=1}^{n} \left[\left(\prod_{k=1}^{K} \left\{ \frac{(E_{ik}\mu(\mathbf{X}_{ik}))^{y_{ik}}}{y_{ik}!} \right\} \right) (-1)^{y_{i\cdot}} \mathcal{L}^{(y_{i\cdot})}(\Lambda_i) \right]. \tag{3.3}$$

where $\Lambda_i = \sum_{k=1}^K E_{ik}\mu(\mathbf{X}_{ik})$ is the sum over all years of the fixed part of the mean (3.1) for policyholder i and $y_{i\cdot} = \sum_{k=1}^K y_{ik}$ is the total number of claims during K insured years for policyholder i.

Focusing, for now, on the case where the random intercept has a normal distribution, the integral in the likelihood does not have a closed form and we have to use numerical approximation methods to find the maximum likelihood estimates. Two such methods are:

Laplace approximation is a method that uses a second-order Taylor series expansion of the exponent of a function. Now consider a one dimensional integral

$$I_n = \int\limits_{-\infty}^{\infty} e^{-nh(u)} du$$

where $h(u)$ is a smooth convex function with minimum at $u = \tilde{u}$, i.e., the point where the first derivative of $h(\tilde{u})$ is zero. A second-order Taylor series expansion of $h(u)$ around $\tilde{u}$ is then

$$h(u) \approx h(\tilde{u}) + \frac{1}{2}\frac{d^2 h(\tilde{u})}{du^2}(u - \tilde{u})^2,$$

which leads to

$$I_n \approx e^{-nh(\tilde{u})} \int\limits_{-\infty}^{\infty} e^{-nh_2(u-\tilde{u})^2/2} du = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})}$$

where h_2 is the second derivative of $h(u)$ evaluated at the minimum $\tilde{u}$, and the equality obtained by using that the normal density has unit integral and the substitution $v = (nh_2)^{1/2}(u - \tilde{u})$. (Davison 2003, ch. 11.3.1)

In the case where Z_i has a lognormal$(0, \sigma_u^2)$ distribution, (3.2) can be written as

$$\mathcal{L}^{(q)}(c) = (-1)^q \frac{1}{\sqrt{2\pi\sigma_u^2}} \int_0^\infty z_i^q \exp(-z_i c) \frac{1}{z_i} \exp\left(-\frac{(\log(z_i))^2}{2\sigma_u^2}\right) dz_i$$

or, by a change of variable, $u_i = \log(z_i)$, as

$$\mathcal{L}^{(q)}(c) = (-1)^q \frac{1}{\sqrt{2\pi\sigma_u^2}} \int_{-\infty}^{\infty} \exp\left\{ qu_i - \exp(u_i)c - \frac{u_i^2}{2\sigma_u^2} \right\} du_i, \tag{3.4}$$

which can be approximated using a Laplace approximation with

$$h(u_i) = -qu_i + \exp(u_i)c + \frac{u_i^2}{2\sigma_u^2},$$

giving

$$\mathcal{L}^{(q)}(c) = (-1)^q \frac{1}{\sqrt{\sigma_u^2 h_2}} e^{-h(\tilde{u})}.$$

A more detailed description of this calculation can be found in Munda et al. (2017).

Gauss–Hermite quadrature is a method that approximates the integral of a function multiplied by a scaled normal density function. The quadratures are defined by integrals of the form

$$\int\limits_{-\infty}^{\infty} h(u)e^{-u^2} du,$$

which is then approximated by

$$\sum_{r=1}^{m} w_r h(u_r),$$

were the quadrature points, u_r are the roots of the m-th order Hermite polynomials and w_r are corresponding weights (Agresti 2015, ch. 9; Liu and Pierce 1994). Because of the "curse of dimensionality" getting good approximations using this method becomes more difficult as the dimension of u_r increases. In our case, it is one-dimensional. The adaptive version of Liu and Pierce (1994) is more efficient and greatly reduces the number of necessary quadrature points needed to approximate the integral. In this version the order one Gauss–Hermite quadrature also becomes the Laplace approximation.

3.1.1 Example: Vehicle insurance claims continued

The model was now fitted assuming independence only between the policyholders, and assuming correlation between time periods for policyholder i. As for the GLM case a Poisson family with a log-link was used, but this time the model was fitted using the `glmmML` package in R (Broström 2019) as well as the `glmer` function in the `lme4` package (Bates et al. 2015), which is the package most commonly used for mixed models. Both Laplace approximation and Gauss–Hermite quadrature methods were tried. In both of these packages the random intercept, U_i has a $N(0, \sigma_u)$-distribution.

Table 3.1 on the next page shows the results from a run with `glmmML` and Laplace approximation. The estimates of the fixed effects are fairly similar to

Table 3.1: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Laplace approximation with `glmmML` on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-3.1032	0.0353	0.0449	0
Age category				
Group 1	0.2644	0.0506	1.3027	1.715e-07
Group 2	0.0419	0.0400	1.0427	0.295
Group 3	0.0457	0.0383	1.0467	0.2328
Group 4	0.0000		1.0000	
Group 5	-0.1830	0.0434	0.8327	2.493e-05
Group 6	-0.1369	0.0508	0.8721	0.007037
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1977	0.0345	1.2186	9.521e-09
$50 - 75$	0.0749	0.0981	1.0778	0.4452
$75 - 100$	-0.5947	0.3815	0.5517	0.119
$100 - 125$	-0.4509	0.5871	0.6371	0.4425
> 125	-1.1965	0.7013	0.3022	0.088
Period				
1	0.0000		1.0000	
2	0.1062	0.0149	1.1121	8.512e-13
3	0.2344	0.0144	1.2641	0
Standard error of random intercept				
$\hat{\sigma}_u$	1.7982	0.0206		0

the results from the fixed effects model in Section 2.1.1 except for the intercept which is lower. This lower intercept can be explained by the fact that the expected value of the lognormal random effect, $Z_i = e^{U_i}$, is not equal to one but $\exp\left\{\frac{\sigma_u^2}{2}\right\}$, which is estimated to be $\exp\left\{\frac{1.7982^2}{2}\right\} = 5.04$. The standard errors on the other hand are now higher, the p-values are also mostly higher. A fairly high $\sigma_u (= 1.7982)$ indicates there is a significant correlation between the time periods for policyholder i. Using Gauss–Hermite with 8, 15 and 20 quadrature points was also tested (see Appendix A.1). These results were very similar to the ones obtained using Laplace approximation, although using 15 quadrature points was a slight improvement to using 8 points there was no further improvement by adding more quadrature points. As can be seen from Table 3.2 on the facing page using more points also increases the computational time from seconds to several minutes.

Using `glmer` the models took significantly longer to fit (see Table 3.2 on the next page), using the default `Nelder-Mead` optimizer, with computational times from 15 minutes to over half an hour and neither obtaining convergence. We then tried changing the optimizer from the default `Nelder-Mead` to `optimx` with method `nlminb`, and added start values, this significantly reduced the computational times. All further model fits using `glmer` can, unless otherwise stated, be assumed to have been done using this optimizer. This change in optimizer also partly solved the problem with obtaining convergence, at least when using 8 Gauss–Hermite quadrature points. The rest of the methods were

Table 3.2: *Approximate run times(in minutes), for one run of each,`glmmML` and `glmer` (default optimizer and `nlminb`) with four different approximation methods,on the full data set.*

	`glmmML`	`glmer`(default)	`glmer(nlminb)`
Laplace	0.5	14	10
Gauss–Hermite:			
8 quadrature points	1.1	24	13
15 quadrature points	1.8	17	15
20 quadrature points	2.2	39	18

also now very close to being within the tolerance levels of achieving convergence. Adding more than 8 quadrature points did not improve the model fit, and all the results produced were very close to each other as well as to the result obtained using `glmmML` with Laplace approximation (see Appendix A.2). All these computations were done on a MacBook pro mid 2014 with operating system 10.14.6, a 2.6 GHz Intel Core i5 processor and 8 GB 1600 MHz DDR3 memory.

Because of these long run times we decided to make a reduced version of the data set, containing $n = 5000$ policies and 15 000 observations, for easier computation. This was done by drawing 3000 and 1171 random policies from the two lowest value categories respectively, while all the observations from the other value categories were kept.

The computational times were still a bit long for quick testing (especially when using the `parfm` command in section 3.2.2), so we made a further reduced version of the data set by drawing 64 random policies from each of the three lowest value categories and keeping the rest (from the already reduced data set), i.e., now containing $n = 300$ policies and 900 observations. This smallest data set will in the following be referred to as the reduced data set.

The result of fitting the model, on the reduced data set, using `glmmML` is shown in Table 3.3 on the following page. Very similar results were also obtained when fitting using `glmer`. Comparing these results with the ones obtained from using the full data set the standard errors are now much higher for all variables, except for the highest value categories which are only slightly higher. Since all the data from these categories were kept in the reduced data set that is to be expected. The estimate of the intercept is fairly similar to the estimate on the full data set. The estimates for the three highest value categories are only a bit (≈ 0.1) lower for the reduced data than for the full data set. The two remaining value categories also have a bit lower estimates than for the full data set, but only slightly for value category 2 (values between 25 and 50 000$). The three lowest age categories all have estimates that are a bit (≈ 0.3) higher than for the full data set while category 5 is about the same amount lower and category 6 a bit higher. The effect, on the expected number of claims, of having the insurance policy for three years is now a bit higher, and having it for two years is a bit lower than for the full data set. The standard error of the random intercept, σ_u, is also a little bit higher than for the model on the full data set. Although, since we are mainly interested in the estimated random effects, $\hat{z}_i$, the large difference in estimates for the reduced data set and the full data set is not of great importance.

Table 3.3: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Laplace approximation with* `glmmML` *on the reduced data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-3.0849	0.5111	0.0457	1.583e-09
Age category				
Group 1	0.6367	0.5865	1.8902	0.2776
Group 2	0.3621	0.5256	1.4364	0.4908
Group 3	0.3336	0.4592	1.3960	0.4675
Group 4	0.0000		1.0000	
Group 5	-0.5225	0.5636	0.5930	0.3539
Group 6	0.0033	0.7570	1.0033	0.9965
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1278	0.4619	1.1364	0.782
$50 - 75$	-0.6791	0.4944	0.5071	0.1696
$75 - 100$	-0.7486	0.5134	0.4730	0.1448
$100 - 125$	-0.6215	0.6946	0.5371	0.3709
> 125	-1.3484	0.7772	0.2597	0.08274
Period				
1	0.0000		1.0000	
2	0.0377	0.1943	1.0384	0.8461
3	0.3393	0.1815	1.4040	0.0615
Standard error of random intercept				
$\hat{\sigma}_u$	1.8194	0.2711		7.513e-50

3.1.2 The random effect

As given in Günther et al.(2014) the conditional distribution of the random effect Z_i given the number of claims for policyholder i, is

$$\begin{aligned}
&f(z_i|Y_{ik}=y_{ik};k=1,\dots,K)\\
&\quad=\frac{\prod_{k=1}^{K}\{(z_iE_{ik}\mu(\mathbf{X}_{ik}))^{y_{ik}}/y_{ik}!)\}\exp\{-z_iE_{ik}\mu(\mathbf{X}_{ik})\}g(z_i)}{\int_0^\infty\prod_{k=1}^{K}\{(z_iE_{ik}\mu(\mathbf{X}_{ik}))^{y_{ik}}/y_{ik}!)\}\exp\{-z_iE_{ik}\mu(\mathbf{X}_{ik})\}g(z_i)dz_i}\\
&\quad=\frac{\prod_{k=1}^{K}\{(z_iE_{ik}\mu(\mathbf{X}_{ik}))^{y_{ik}}/y_{ik}!)\}\exp\{-z_iE_{ik}\mu(\mathbf{X}_{ik})\}g(z_i)}{\prod_{k=1}^{K}\{(E_{ik}\mu(\mathbf{X}_{ik}))^{y_{ik}}/y_{ik}!)\}\,(-1)^{y_{i\cdot}}\mathcal{L}^{(y_{i\cdot})}(\Lambda_i)}\\
&\quad=\frac{z_i^{y_{i\cdot}}\exp\{-z_i\Lambda_i\}g(z_i)}{(-1)^{y_{i\cdot}}\mathcal{L}^{(y_{i\cdot})}(\Lambda_i)}.
\end{aligned}$$

By a similar argument we find that the conditional distribution of U_i, given the number of claims for policyholder i, is given by

$$g(u_i|Y_{ik}=y_{ik};k=1,\dots,K)$$

$$= C_0 \prod_{k=1}^{K} \left[\frac{(e^{u_i} E_{ik} \mu(\mathbf{X}_{ik}))^{y_{ik}}}{y_{ik}!} \exp\left\{-e^{u_i} E_{ik} \mu(\mathbf{X}_{ik})\right\} \right] \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left\{ -\frac{u_i^2}{2\sigma_u^2} \right\},$$

where C_0 is a constant. Many methods, like `glmmML` and `glmer`, for fitting the models return the posterior mode, $\tilde{u}_i$, of u_i. This is the value of u_i where $g(u_i|Y_{ik} = y_{ik}; k = 1, \ldots, K)$ is largest. We now have that

$$\log g(u_i|Y_{ik} = y_{ik}; k = 1, \ldots, K) = C_1 + y_{i\bullet} u_i - e^{u_i}\Lambda_i - \frac{u_i^2}{2\sigma_u^2},$$

where C_1 is another constant. The posterior mode is then found at the point where the derivative of this is equal to zero, i.e., by solving the equation

$$y_{i\bullet} - e^{u_i}\Lambda_i - \frac{u_i}{\sigma_u^2} = 0,$$

which can be solved numerically.

Further we find that the conditional mean of the random effect, Z_i, given the number of claims for policyholder i, is

$$\begin{aligned} \hat{z}_i &= \mathrm{E}(Z_i|Y_{ik} = y_{ik}; k = 1, \ldots, K) \\ &= \int_0^\infty z_i f(z_i|Y_{ik} = y_{ik}; k = 1, \ldots, K) dz_i \\ &= \int_0^\infty z_i \frac{z_i^{y_{i\bullet}} \exp\{-z_i\Lambda_i\} g(z_i)}{(-1)^{y_{i\bullet}} \mathcal{L}^{(y_{i\bullet})}(\Lambda_i)} dz_i \\ &= -\frac{\mathcal{L}^{(y_{i\bullet}+1)}(\Lambda_i)}{\mathcal{L}^{(y_{i\bullet})}(\Lambda_i)}, \end{aligned} \tag{3.5}$$

where $\mathcal{L}^{(q)}(\Lambda_i)$ is given by (3.4) and can be calculated by numerical integration. We can then find the expected number of claims for policyholder i in a year k^* later than K by combining this with (3.1)

$$\begin{aligned} \mathrm{E}(Y_{ik^*}|Y_{ik} = y_{ik}; k = 1, \ldots, K) &= \hat{z}_i E_{ik^*} \mu(\mathbf{X}_{ik^*}) \\ &= -\frac{\mathcal{L}^{(y_{i\bullet}+1)}(\Lambda_i)}{\mathcal{L}^{(y_{i\bullet})}(\Lambda_i)} \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ik^*j} + \log E_{ik^*}). \end{aligned}$$

3.1.3 Example: Vehicle insurance claims continued

We now look further at the model for the reduced data set. Figure 3.1 on the next page shows the estimated random effects, $\hat{z}_i$, for all $n = 300$ policyholders from the reduced data set. It also includes the sum of the fixed part of the mean, $\hat{\Lambda}_i$, with the average number of claims over the $K = 3$ years, for each policyholder, i.

Note that for two policyholders with the same fixed effects, i.e., are in the same age group and have a vehicle in the same value category, have the same random effect if they also have the same number of claims in the period. If, on the other hand, they have different number of claims, the policyholder with the higher amount of claims will have a higher random effect, $\hat{z}_i$. We also see

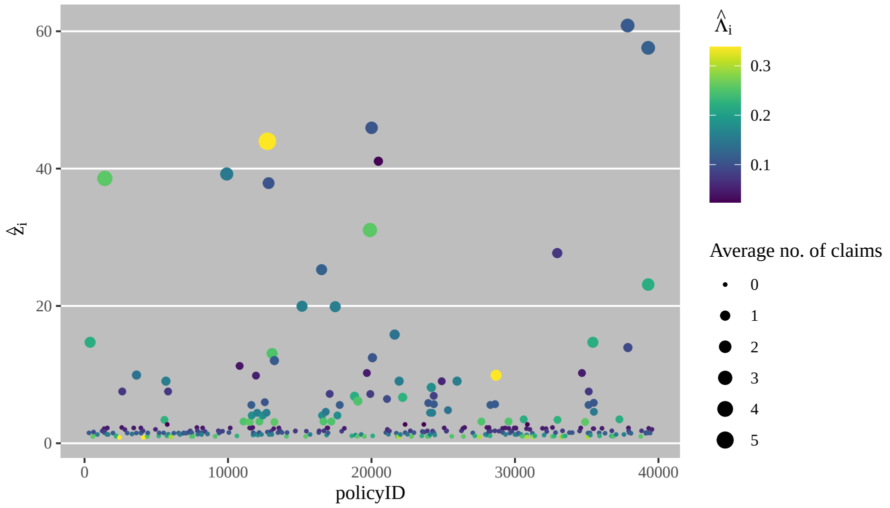

Figure 3.1: *Random effects, $\hat{z}_i$ with lognormal distribution, for the reduced data set. The colour scale indicates the sum over all the years of the fixed part of the mean, $\hat{\Lambda}_i$, and the size scale indicates the average number of claims reported for policyholder i.*

that in general a higher number of claims will result in a larger random effect. Although the variation in the random effects is quite substantial most of them are small, with a few very large ones.

Grouping the random effects, $\hat{z}_i$, by the sum of the fixed part of the mean, $\hat{\Lambda}_i$, and the average number of claims we see these patterns more clearly (figure 3.2). We also see that if a policyholder is in a group with lower probability of having a claim, i.e., low $\hat{\Lambda}_i$, just having a claim will have a large impact on the random effect. A policyholder with a high probability of a claim, on the other hand, will need more claims to have the same impact on the random effect.

The influence of the random effect is illustrated in figure 3.3 by plotting the ratio of the total expected number of claims from the model with random effects and the model without random effects versus the average number of claims observed. Including a random effect in the model, i.e., including the policyholders claims history, will reduce the expectation of claims for those with average number of claims close to zero and increase it for the policyholders with increasing number of average claims.

When comparing the estimated random effects, $\hat{z}_i$, to taking the exponential of the posterior modes, $e^{\hat{u}_i}$, we see from figure 3.4 that the random effects are slightly higher than what is obtained by just taking the exponential of the posterior modes. The difference between them is more noticeable for the policyholders with fewer reported claims on average, as illustrated more clearly by figure 3.5.

3.2 The frailty model for recurrent events

A common term for the unobserved heterogeneity in survival analysis is frailty, i.e., a term to describe the variation between individuals that is not described by the covariates. When we now look at the situation with recurrent events

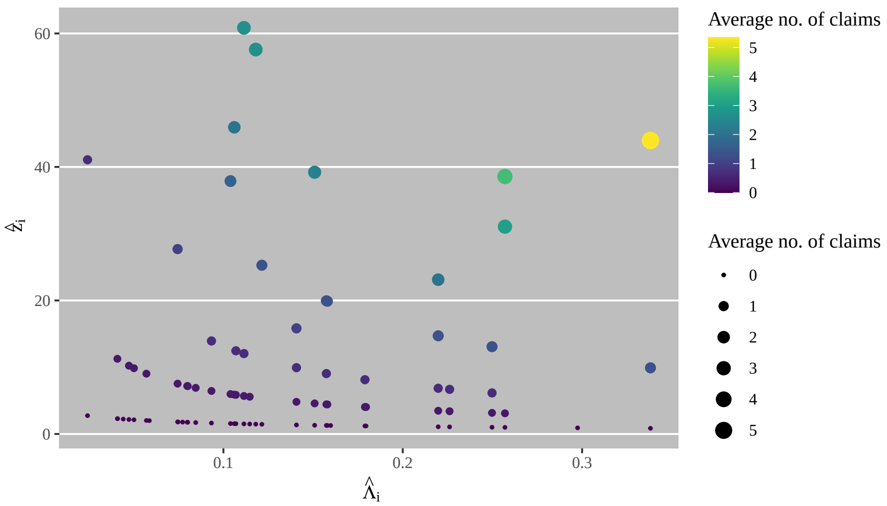

Figure 3.2: *Plot of the random effects, $\hat{z}_i$, of the K years against the sum of the fixed part of the mean, $\hat{\Lambda}_i$. With size and colour scales as indications of the average number of claims reported per year.*

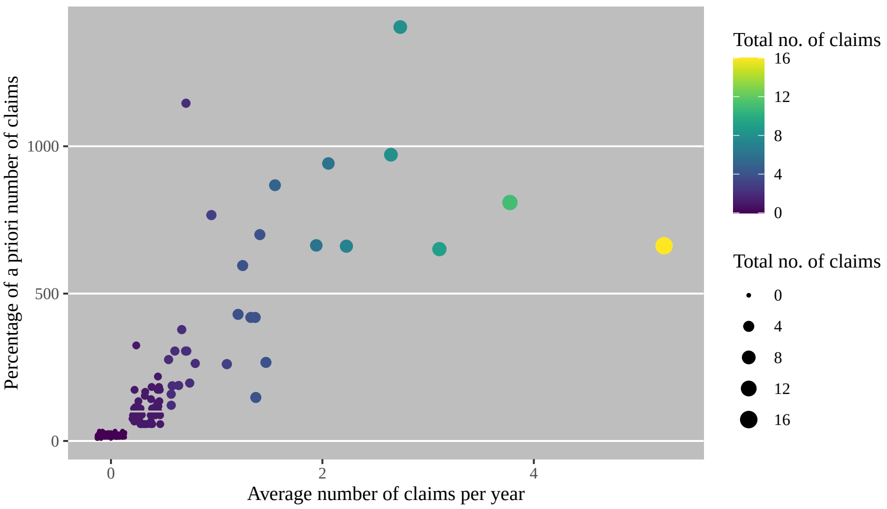

Figure 3.3: *Ratio in percentage of estimated expected number of claims from the model with random effects divided by the estimated expected number of claims from the model without random effects versus the mean number of observed claims.*

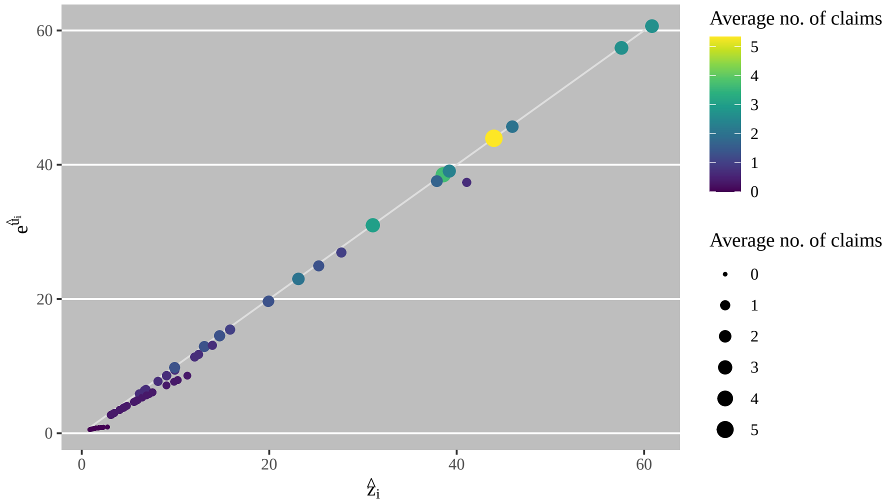

Figure 3.4: *Random effects, $\hat{z}_i$, versus exponential of the posterior modes, $\exp\{\hat{u}_i\}$, for all the policyholders, i, with the colour scale indicating the average number of claims observed.*

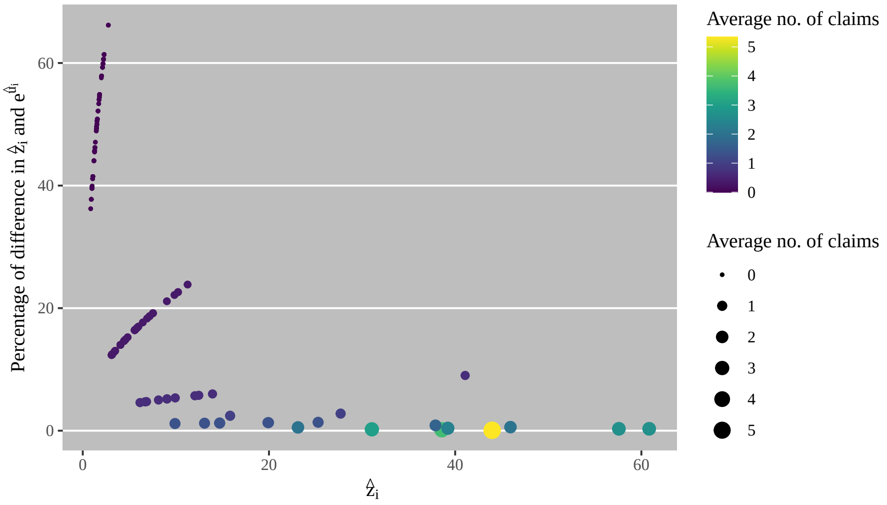

Figure 3.5: *Percentage of difference between $\hat{z}_i$ and $\exp\{\hat{u}_i\}$, with the colour scale indicating the average number of claims observed.*

from section 2.2, we will assume that for each $i = 1, \dots, n$ we have a frailty Z_i, and that given $Z_i = z_i$, we have a counting process $N_i(t)$ with intensity process

$$\lambda_i(t|z_i) = z_i Y_i(t) \alpha_i(t)$$

Here, as in section 2.2, we have that

$$\alpha_i(t) = \alpha_0(t) e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}}$$

where $\alpha_0(t) = \alpha_0(t; \boldsymbol{\theta})$ is piecewise constant, i.e.,

$$\alpha_0(t; \boldsymbol{\theta}) = \sum_{i=1}^{K} \theta_k I_k(t).$$

We then set H_i to be the information we have about the process $N_i(t)$ for the interval $[0, K]$. This could then be information, e.g., about when a policyholder i reports a claim, and how many they have reported during the time of observation, i.e., the period they are insured for. Similarly to (2.3) we then have that

$$P(H_i | Z_i = z_i) = \prod_{0<t\leq K} \lambda_i(t|z_i)^{\Delta N_i(t)} \exp\left\{ - \int_0^K \lambda_i(t|z_i) dt \right\}.$$

By rewriting this in a similar fashion to what was done in section 2.2 we now have that

$$\begin{aligned} P(H_i | Z_i = z_i) &= \prod_{k=1}^{K} \left[\left(z_i \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \right)^{O_{ik}} \right] \exp\left\{ - \sum_{k=1}^{K} z_i \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik} \right\} \\ &= \prod_{k=1}^{K} \left[\left(\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \right)^{O_{ik}} \right] z_i^{O_{i\bullet}} \exp\left\{ -z_i \sum_{k=1}^{K} \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik} \right\}, \end{aligned}$$

where O_{ik} is the number of claims in period k for policyholder i, as before, and $O_{i\bullet} = \sum_{k=1}^{K} O_{ik}$. We now let $g(z_i)$ be the density of the frailty Z_i, assuming the Z_is are independent and identically distributed. Then, by taking the expectation with respect to Z_i, we get that

$$\begin{aligned} P(H_i) &= \int_0^\infty P(H_i | Z_i = z_i) g(z_i) dz_i \\ &= \prod_{k=1}^{K} \left[\left(\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \right)^{O_{ik}} \right] \int_0^\infty z_i^{O_{i\bullet}} \exp\left\{ -z_i \sum_{k=1}^{K} \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik} \right\} g(z_i) dz_i \\ &= \prod_{k=1}^{K} \left[\left(\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \right)^{O_{ik}} \right] (-1)^{O_{i\bullet}} \mathcal{L}^{(O_{i\bullet})}(\Lambda_i), \end{aligned}$$

where $\Lambda_i = \sum_{k=1}^{K} \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik}$. The likelihood then becomes

$$L = \prod_{i=1}^{n} P(H_i) = \prod_{i=1}^{n} \left\{ \prod_{k=1}^{K} \left[\left(\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \right)^{O_{ik}} \right] (-1)^{O_{i\bullet}} \mathcal{L}^{(O_{i\bullet})}(\Lambda_i) \right\}, \tag{3.6}$$

which is corresponding to the likelihood for the generalized linear mixed model (3.3).

The estimation of Z_i is also similar to what was done for the mixed model. From Aalen et al. (2008, p. 278) we have that the conditional density of Z_i given H_i becomes

$$
\begin{aligned}
f(z_i|H_i) &= \frac{P(H_i|Z_i = z_i)g(z_i)}{P(H_i)} \\
&= \frac{\prod\limits_{k=1}^{K}\left[\left(\theta_k e^{\boldsymbol{\beta}^T\mathbf{X}_{ik}}\right)^{O_{ik}}\right] z_i^{O_{i\cdot}} \exp\left\{-z_i \sum\limits_{k=1}^{K} \theta_k e^{\boldsymbol{\beta}^T\mathbf{X}_{ik}} E_{ik}\right\}}{\prod\limits_{k=1}^{K}\left[\left(\theta_k e^{\boldsymbol{\beta}^T\mathbf{X}_{ik}}\right)^{O_{ik}}\right](-1)^{O_{i\cdot}}\mathcal{L}^{(O_{i\cdot})}(\Lambda_i)} g(z_i) \\
&= \frac{z_i^{O_{i\cdot}} \exp\{-z_i\Lambda_i\}}{(-1)^{O_{i\cdot}}\mathcal{L}^{(O_{i\cdot})}(\Lambda_i)} g(z_i)
\end{aligned}
$$

From this we find that

$$
\begin{aligned}
\hat{z}_i &= \mathrm{E}[Z_i|H_i] \\
&= \int\limits_0^\infty z_i f(z_i|H_i)dz_i \\
&= \frac{\int\limits_0^\infty z_i^{O_{i\cdot}+1} \exp\{-z_i\Lambda_i\}\, g(z_i)dz_i}{(-1)^{O_{i\cdot}}\mathcal{L}^{(O_{i\cdot})}(\Lambda_i)} \\
&= -\frac{\mathcal{L}^{(O_{i\cdot}+1)}(\Lambda_i)}{\mathcal{L}^{(O_{i\cdot})}(\Lambda_i)},
\end{aligned}
\tag{3.7}
$$

which corresponds to (3.5), for the mixed model.

3.2.1 Clustered survival data

The command `parfm` in the R package with the same name (Munda et al. 2012), fits frailty models for clustered survival data, and it allows for a number of distributions of the frailty. It is possible to use `parfm` to fit models for recurrent events.

In order to do so, we have to adapt the data for the recurrent events to a clustered survival data setting. To this end we do the following. For each policyholder i and period k where there are two or more claims (i.e., $O_{ik} \geq 2$), we replace the original observation with $n_{ik} = O_{ik}$ artificial "observations" where $D_{ikl} = 1$ and $\tilde{T}_{ikl} = E_{ik}/O_{ik}$ for $l = 1, \ldots, n_{ik}$. For the policyholders i and periods k where $O_{ik} \leq 1$ we just set $n_{ik} = 1$ and let $D_{ikl} = O_{ik}$ and $\tilde{T}_{ikl} = E_{ik}$. The code to achieve this may be found in listing B.1 in appendix B.

We may then consider the observations $(\tilde{T}_{ikl}, D_{ikl}; l = 1, \ldots, n_{ik}, k = 1, \ldots, K)$ as observations from cluster i, $(i = 1, \ldots, n)$. Further we assume that given frailty $Z_i = z_i$, the intensity for the uncensored survival times, T_{ikl}, corresponding to the $\tilde{T}_{ikl}$'s are given as

$$
\alpha_{ik}(t) = \theta_k e^{\boldsymbol{\beta}^T\mathbf{X}_{ik}} \tag{3.8}
$$

with cumulative intensity

$$A_{ik}(t) = \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} t. \tag{3.9}$$

The likelihood for such artificial clustered survival data now becomes (Aalen et al. 2008, section 7.2.2)

$$L = \prod_{i=1}^{n} \left\{ \prod_{k=1}^{K} \prod_{l=1}^{n_{ik}} \left[\alpha_{ik}(\tilde{T}_{ikl}) \right]^{D_{ikl}} (-1)^{D_{i\bullet\bullet}} \mathcal{L}^{(D_{i\bullet\bullet})} \left(\sum_{k=1}^{K} \sum_{l=1}^{n_{ik}} A_{ik}(\tilde{T}_{ikl}) \right) \right\},$$

where $D_{i\bullet\bullet} = \sum_{k=1}^{K} \sum_{l=1}^{n_{ik}} D_{ikl}$. If we insert (3.8) and (3.9) into this expression, and note that $\sum_{l=1}^{n_{ik}} D_{ikl} = O_{ik}$ and $D_{i\bullet\bullet} = \sum_{k=1}^{K} O_{ik} = O_{i\bullet}$, the likelihood may be written

$$L = \prod_{i=1}^{n} \left\{ \prod_{k=1}^{K} \left[\theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \right]^{O_{ik}} (-1)^{O_{i\bullet}} \mathcal{L}^{(O_{i\bullet})} \left(\sum_{k=1}^{K} \sum_{l=1}^{n_{ik}} \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \tilde{T}_{ikl} \right) \right\}.$$

Now we have $\tilde{T}_{ikl} = E_{ik}/n_{ik}$, so that

$$\sum_{k=1}^{K} \sum_{l=1}^{n_{ik}} \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} \tilde{T}_{ikl} = \sum_{k=1}^{K} \theta_k e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}} E_{ik}.$$

Thus we see that the likelihood for the (artificial) clustered survival data is the same as (3.6). This shows that we may use `parfm` for clustered survival data to fit frailty models for recurrent event data.

3.2.2 Example: Vehicle insurance claims continued

The model was now fitted using the `parfm` package (Munda et al. 2017) in R, with an exponential distribution on the baseline intensity and lognormal distribution on the frailty. The default optimizer `nlminb` was also used. As can be seen from table 3.4 the results are very similar to those for the mixed model (table 3.3). The resulting frailties were also very close to the random effects of the mixed model.

This model does not have an intercept like the mixed model, but it does have the baseline intensity parameter,θ_k which in this case corresponds to e^{β_0}, i.e., the intercept would be $\beta_0 = \log(\theta_k) = \log(0.0457) = -3.08497$, which is very close to the intercept of the mixed model.

Running this with the full data set with $n = 40\,000$ took around 2.5 hours to run, while using the larger data set, with $n = 5000$ policies, took a little under 17 minutes to run. When reducing the data set to $n = 300$ policies the execution time was reduced to a little over 40 seconds. Comparing the run time for the full data set with those of table 3.2 we see that `parfm` takes significantly longer to run. Seeing as the clustered survival model includes the extra level $l = 1, \ldots, n_{ik}$, it is not unreasonable that the calculations take a bit longer.

Table 3.4: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the survival model with* `parfm` *on the reduced data set. With lognormally disributed frailties*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Age category				
Group 1	0.6365	0.5862	1.8898	0.2776
Group 2	0.3621	0.5258	1.4364	0.491
Group 3	0.3336	0.4591	1.3960	0.4674
Group 4	0.0000		1.0000	
Group 5	-0.5226	0.5642	0.5930	0.3543
Group 6	0.0034	0.7571	1.0035	0.9964
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1280	0.4619	1.1365	0.7817
$50 - 75$	-0.6789	0.4943	0.5072	0.1695
$75 - 100$	-0.7485	0.5140	0.4731	0.1453
$100 - 125$	-0.6216	0.6933	0.5371	0.3699
> 125	-1.3481	0.7768	0.2597	0.08265
Period				
1	0.0000		1.0000	
2	0.0377	0.1944	1.0385	0.8461
3	0.3392	0.1816	1.4038	0.06175
Variance of frailty				
$\hat{\sigma}_u^2$	3.3097	0.8070		
Baseline intensity parameter				
$\hat{\theta}$	0.0457	0.0226		

3.3 Summary of chapter

In this chapter we have seen that the generalized mixed model and the frailty model for recurrent events have corresponding likelihoods, and so can both be used for modelling the number of claims to expect on an insurance policy. In this chapter we have focused on random effects/frailties with a lognormal distribution. In the coming chapters we will take a closer look at other possible distributions. We have also seen that fitting a frailty model to a clustered survival data set takes a lot longer than fitting a mixed model, likely due to the extra layer to do computations over.

4

Distributions of the frailty

In the previous chapter we looked at both the random effect and frailty, with a lognormal distribution. We will now look at a few more distributions, but now only focus on the recurrent events model and the distribution of the frailty.

4.1 Standardizing the lognormal frailty

In order to make comparisons more fair we need to standardize the lognormally distributed frailties to have expectation equal to one, like the distributions we will compare it with in this chapter. With $U \sim N(0, \sigma_u^2)$, the expectation of the frailty, Z_i, is

$$\mathrm{E}[Z_i] = \exp\left\{\frac{\sigma_u^2}{2}\right\}. \tag{4.1}$$

To standardize the lognormal frailties to have expectation equal to one we therefore have to divide them by (4.1), and also consequently we need to multiply the sum of the fixed part of the expected number of claims, $\hat{\Lambda}$, with (4.1).

The variance of the frailty is then

$$\mathrm{Var}(Z_i) = \sigma_z^2 = e^{\sigma^2{}_u} - 1,$$

so the variance of the frailty for the clustered survival model, fitted with lognormal frailty, is then $\sigma_z^2 = e^{3.3097} - 1 = 26.38$.

The parameters of the lognormal distribution is the mean, μ_u, and standard deviation, σ_u, of U_i, and not the mean, μ_z, and standard deviation, σ_z, of Z_i. It is well known that for the lognormal distribution we have $E(Z_i) = \exp\{\mu_u + \sigma_u^2/2\}$ and $\mathrm{Var}(Z_i) = \exp\{2\mu_u + \sigma_u^2\}[\exp\{\sigma_u^2\} - 1]$. So in order to produce a lognormally distributed frailty Z_i, with mean 1 and variance σ_z^2 we need to use the following parameters

$$\mu_u = -\frac{\log(\sigma_z^2 + 1)}{2}$$

and

$$\sigma_u^2 = \log(\sigma_z^2 + 1).$$

Figure 4.1 shows the probability densities for the lognormal distribution with mean 1 and different variances.

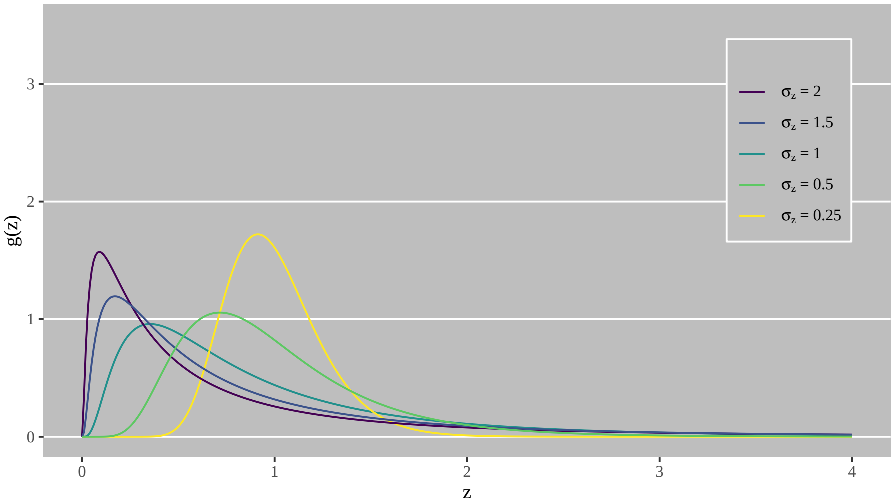

Figure 4.1: *Densities, $g(z)$, with mean $= 1$ and five different $\sigma_z s$, for the lognormal distribution.*

4.2 Gamma distributed frailty

With $Z_i \sim \text{Gamma}(\alpha, \beta)$, the probability density function is given by

$$g(z_i) = \frac{\beta^\alpha}{\Gamma(\alpha)} z_i^{\alpha-1} e^{-\beta z_i}.$$

It is common to fix the mean equal to 1 giving $\alpha = \beta$, and variance, $\sigma_z^2 = 1/\beta$. This variance is used as a measure of the degree of heterogeneity (Aalen et al. 2008, ch. 6.2.2). This gives the Laplace transformation

$$\mathcal{L}(c) = (1 + \sigma_z^2 c)^{-\frac{1}{\sigma_z^2}},$$

and the q-th derivative is then

$$\begin{aligned}\mathcal{L}^{(q)}(c) &= (\sigma_z^2)^q (-1)^q (1 + \sigma_z^2 c)^{-\frac{1}{\sigma_z^2} - q} \prod_{s=1}^{q} \left(\frac{1}{\sigma_z^2} + s - 1 \right) \\ &= (-1)^q (1 + \sigma_z^2 c)^{-q} \left[\prod_{s=0}^{q-1} 1 + s\sigma_z^2 \right] \mathcal{L}(c).\end{aligned}$$

Using this with (3.7) we get the estimated frailty. (Aalen et al. 2008; Munda et al. 2012)). Figure 4.2 shows the probability densities for the gamma distribution with mean 1 and different variances.

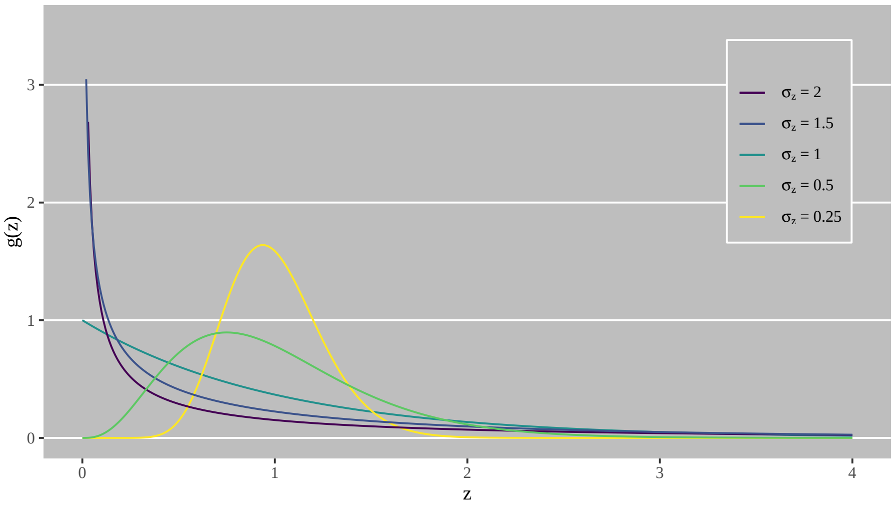

Figure 4.2: *Densities, $g(z)$, with mean $= 1$ and five different σ_z's, for the gamma distribution.*

4.2.1 Example: Vehicle insurance claims continued

Fitting the model with a gamma distributed frailty using `parfm` gives the results presented in table 4.1. From this we see that the estimates of the coefficients are mostly higher than with a lognormal distribution on the frailty (see table 3.4), except for the estimates for the periods, which are the same. The standard errors and p-values on the other hand are mostly lower or the same. The variance of the frailty is much lower indicating a lower degree of heterogeneity.

From figure 4.3 we see that the frailties, $\hat{z}_{i,gamma}$, follow a similar pattern to the frailties with a lognormal distribution (see figure 3.2).

We see from figure 4.4 that the sum of the fixed part of the expected number of claims when the frailty follows a gamma distribution, $\hat{\Lambda}_{i,gamma}$, are mostly a bit lower than the ones when the frailty follows the lognormal distribution, $\hat{\Lambda}_{i,lognorm}$. From figure 4.5 there does not seem to be any clear pattern to the differences.

We also see from figure 4.6 that the frailties with gamma distribution, $\hat{z}_{i,gamma}$, are mostly slightly higher than the ones with lognormal distribution, $\hat{z}_{i,lognorm}$. The difference between them seems to be larger for higher number of average claims. Although from figure 4.7 we see that, in percentage of difference from $\hat{z}_{i,lognorm}$, the frailties for policyholders with few average claims is also quite large.

Then, looking at the expected number of claims in total, i.e., $\hat{z}_i\hat{\Lambda}_i$, for the models with gamma and lognormaly distributed frailties, we see from figure 4.8 that they both seem to give fairly similar results. Taking a closer look at the differences, in figure 4.9, we see that there is quite a large variation in differences in the estimated expected total number of claims where the number of observed claims are small, while it seems to decrease a bit with increasing number of claims.

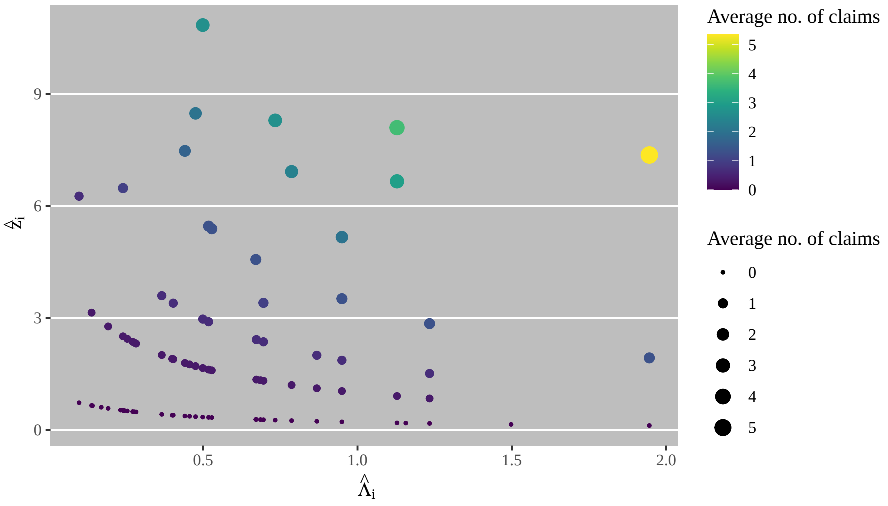

Figure 4.3: *Frailties, $\hat{z}_{i,gamma}$, against the sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,gamma}$. With colour and size scales indicating the average number of claims reported per year.*

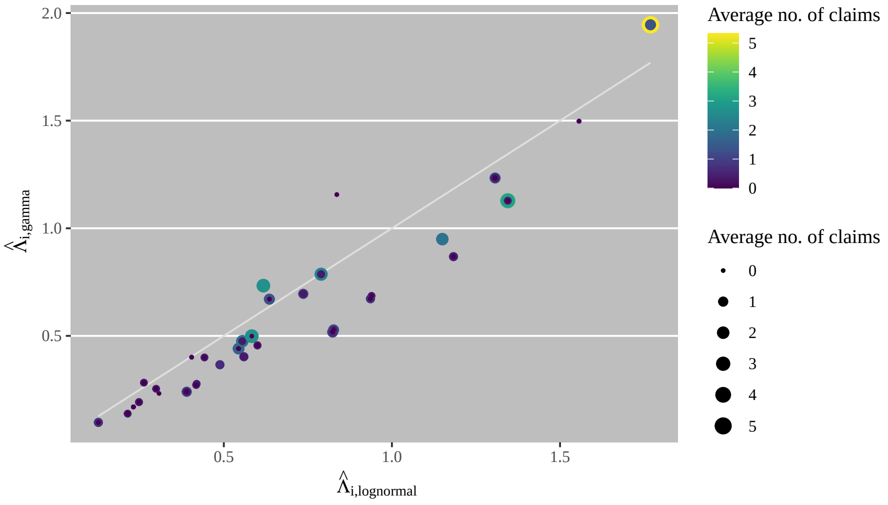

Figure 4.4: *The sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,gamma}$ for gamma distributed frailties, $\hat{z}_{i,gamma}$, versus the sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,lognorm}$ for lognormally distributed frailties, $\hat{z}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

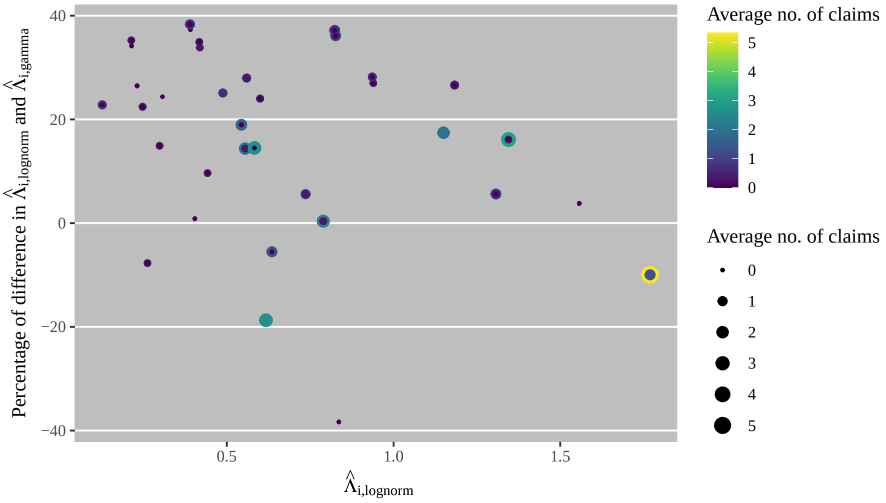

Figure 4.5: *Percentage of difference between $\hat{\Lambda}_{i,lognorm}$ and $\hat{\Lambda}_{i,gamma}$, in proportion to $\hat{\Lambda}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

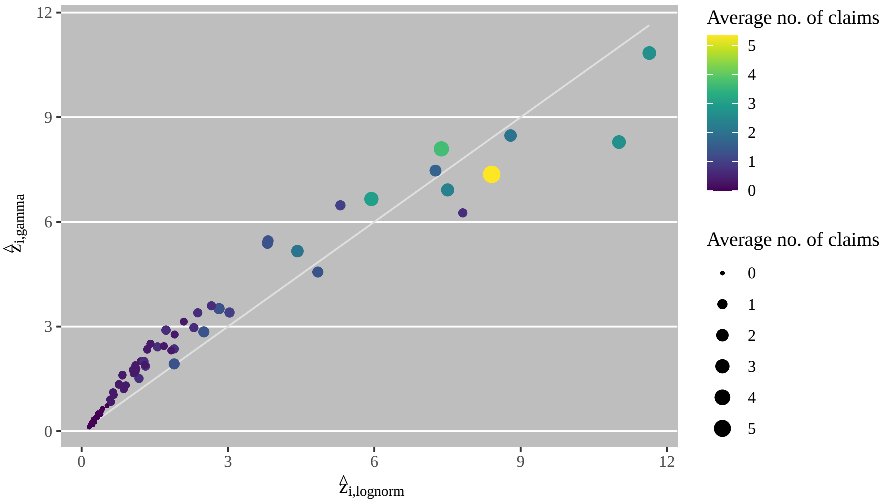

Figure 4.6: *The gamma distributed frailties, $\hat{z}_{i,gamma}$ versus the lognormally distributed frailties, $\hat{z}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

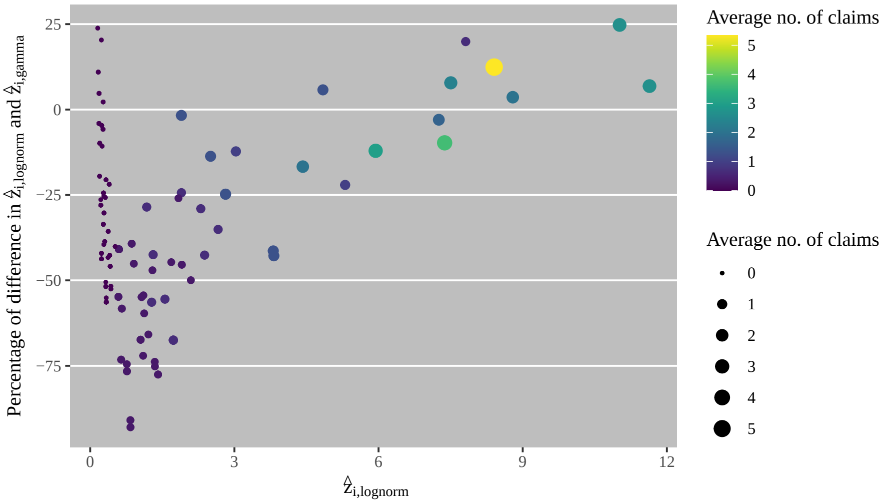

Figure 4.7: *Percentage of difference between $\hat{z}_{i,lognorm}$ and $\hat{z}_{i,gamma}$, in proportion to $\hat{z}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

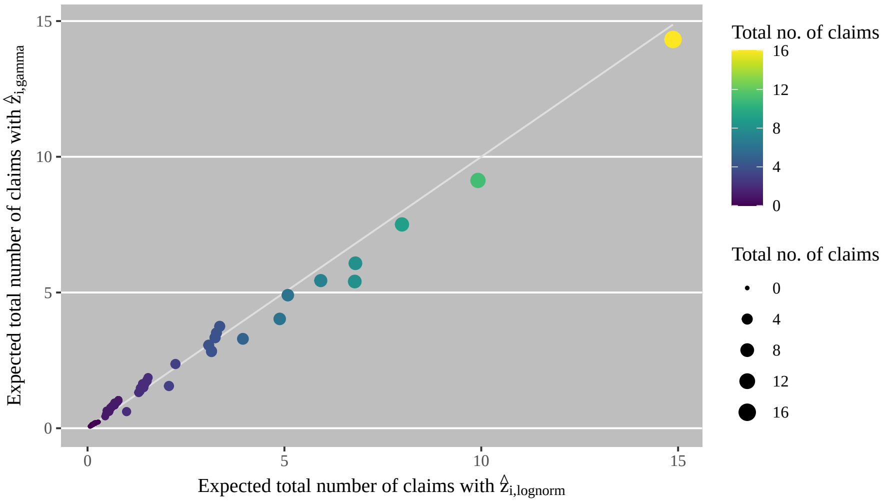

Figure 4.8: *Estimated expected number of claims for the total of the $K = 3$ years with lognormal and gamma distributed frailties. With colour and size scales indicating the total number of claims reported.*

Table 4.1: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the survival model with* `parfm` *on the reduced data set, with gamma distributed frailty*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Age category				
Group 1	1.0625	0.5254	2.8936	0.04315
Group 2	0.5174	0.4784	1.6776	0.2795
Group 3	0.6068	0.4163	1.8346	0.145
Group 4	0.0000		1.0000	
Group 5	-0.3469	0.5045	0.7069	0.4918
Group 6	0.0198	0.6815	1.0200	0.9768
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.2617	0.4305	1.2992	0.5433
$50 - 75$	-0.6440	0.4484	0.5252	0.151
$75 - 100$	-0.7673	0.4619	0.4643	0.09673
$100 - 125$	-0.2583	0.5836	0.7723	0.6581
> 125	-1.3180	0.7251	0.2677	0.0691
Period				
1	0.0000		1.0000	
2	0.0377	0.1943	1.0385	0.846
3	0.3392	0.1815	1.4038	0.06157
Variance of frailty				
$\hat{\sigma}_z^2$	3.7916	0.7363		
Baseline intensity parameter				
$\hat{\theta}$	0.1503	0.0613		

4.3 Inverse Gaussian distributed frailty

The inverse Gaussian distribution is a special case of the generalized inverse Gaussian distribution. With $Z_i \sim GIG(p, a, b)$ the density is

$$g(z_i) = \frac{(\frac{a}{b})^{\frac{p}{2}}}{2K_p(\sqrt{ab})} z^{p-1} e^{-\frac{az+\frac{b}{z}}{2}}, \quad z_i > 0,$$

where $K_p(\omega)$ is the Bessel function (Hougaard 2000, sec. A.4.2)

$$K_p(\omega) = \frac{1}{2} \int_0^\infty t^{p-1} \exp\left\{-\frac{\omega}{2}\left(t + \frac{1}{t}\right)\right\} dt.$$

When $p = -\frac{1}{2}$, $a = \frac{\lambda}{\mu^2}$ and $b = \lambda$ we get the inverse Gaussian distribution. Fixing the mean, μ, to be 1, and setting the variance, $\frac{1}{\lambda}$, to σ_z^2 we get that the density of $Z_i \sim IG(\sigma_z^2)$, by also using the fact that $K_{1/2}(\omega) = K_{-1/2}(\omega) = \sqrt{\frac{\pi}{2\omega}} \exp(-\omega)$, is

$$g(z_i) = \frac{1}{\sqrt{2\pi\sigma_z^2}} z_i^{-\frac{3}{2}} e^{-\frac{(z_i-1)^2}{2z_i\sigma_z^2}}.$$

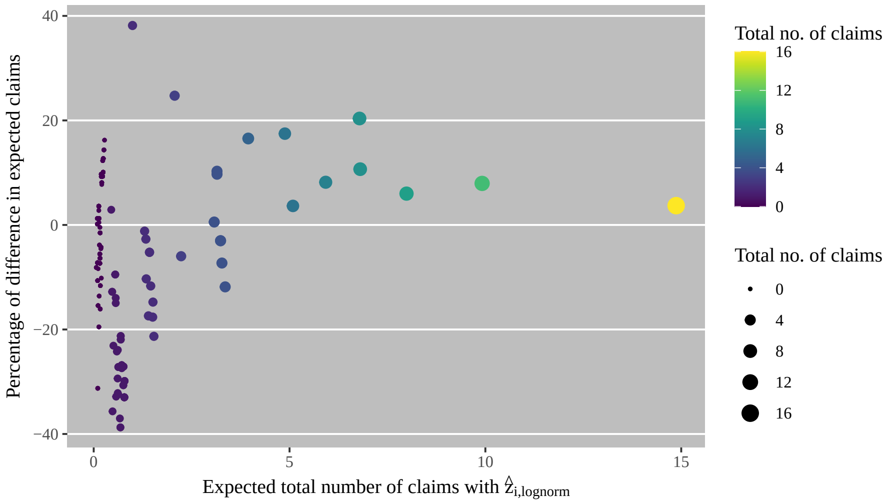

Figure 4.9: *Percentage of difference between expected total number of claims with lognormal and gamma distributed frailties, in proportion to the expected total number of claims with lognormal frailties. With colour and size scales indicating the total number of claims reported.*

The Laplace transform is then (Aalen et al. 2008, sec. 6.2.3; Munda et al. 2012, sec. 2.3)

$$\mathcal{L}(c) = \exp\left\{ \frac{1}{\sigma_z^2} \left(1 - \sqrt{1 + 2\sigma_z^2 c} \right) \right\}, \quad c \geq 0,$$

which has the q-th derivative (Munda et al. 2012, sec. 2.3)

$$\mathcal{L}^{(q)}(c) = (-1)^q (2\sigma_z^2 c + 1)^{-\frac{q}{2}} \frac{K_{q-(1/2)}\left(\sqrt{2(\sigma_z^2)^{-1}(c + \frac{1}{2\sigma_z^2})} \right)}{K_{1/2}\left(\sqrt{2(\sigma_z^2)^{-1}(c + \frac{1}{2\sigma_z^2})} \right)} \mathcal{L}(c).$$

As before, using this with (3.7) will get us the estimated frailties. In figure 4.10 we see plots of the density for different variances, all with the same mean of 1.

4.3.1 Example: Vehicle insurance claims continued

When we fit the model with an inverse Gaussian distributed frailty, $z_{i,ingau}$, using `parfm` we get results presented in table 4.2. We see from this that the effects of the periods do not seem to change with the distribution of the frailty. The estimates of the coefficients are mostly pretty close to those of the lognormally distributed frailties (see table 3.4), although where there are differences the ones where the frailties are inverse Gaussian seem to be a bit higher. The estimated standard errors are placed somewhere in between those from the models with lognormal and gamma distributed frailties. The p-values are close to those of the model with lognormal frailty, while the baseline intensity is closer to that of the model with gamma distributed frailty. Also, the variance of the frailty is slightly higher than for the gamma model, but still much lower than for the lognormal model.

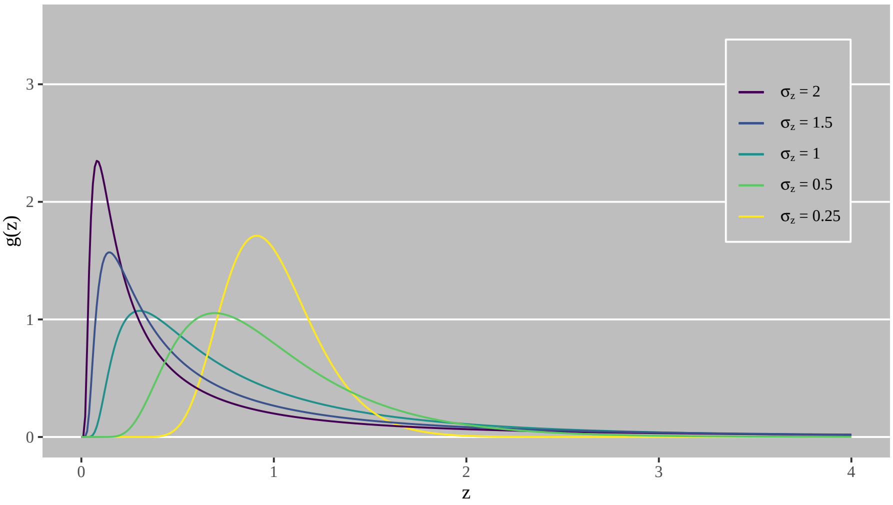

Figure 4.10: *Densities, $g(z)$, with mean $= 1$ and five different σ_zs, for the inverse Gaussian distribution.*

We see from figure 4.11 that frailties follow a similar pattern to those with a gamma and lognormal distribution, i.e., higher impact of having a claim, on the frailty for policyholders with few expected claims based on the fixed effects than for those with more expected claims based on the fixed effects.

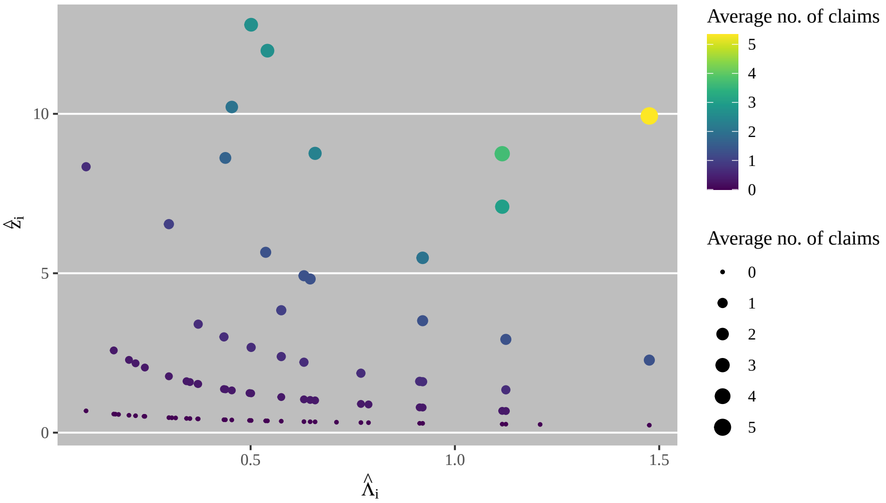

Figure 4.11: *Frailties, $\hat{z}_{i,ingau}$, against the sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,ingau}$. With colour and size scales indicating the average number of claims reported per year.*

From figure 4.12 we see that the sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,ingau}$, follow a relatively linear pattern not too different from the lognormal, $\hat{\Lambda}_{i,lognorm}$, but it is slightly lower. The difference seems to

Table 4.2: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the survival model with `parfm` on the reduced data set, with inverse Gaussian distributed frailty*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Age category				
Group 1	0.6506	0.5418	1.9166	0.2298
Group 2	0.3710	0.5016	1.4492	0.4596
Group 3	0.3789	0.4379	1.4606	0.3869
Group 4	0.0000		1.0000	
Group 5	-0.5283	0.5421	0.5896	0.3298
Group 6	0.0238	0.7374	1.0241	0.9742
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1998	0.4315	1.2211	0.6434
$50 - 75$	-0.6082	0.4659	0.5443	0.1917
$75 - 100$	-0.7427	0.4927	0.4758	0.1317
$100 - 125$	-0.5317	0.6229	0.5876	0.3933
> 125	-1.3398	0.7459	0.2619	0.07244
Period				
1	0.0000		1.0000	
2	0.0377	0.1943	1.0385	0.846
3	0.3392	0.1815	1.4038	0.06157
Variance of frailty				
$\hat{\sigma}_z^2$	5.8505	1.8431		
Baseline intensity parameter				
$\hat{\theta}$	0.1832	0.0835		

increase with increasing $\hat{\Lambda}_i$. Although we see from figure 4.13 that in percent, in proportion to $\hat{\Lambda}_{i,lognorm}$, the difference does not seem to follow any specific pattern and ranges between twelve and twenty four percent.

We see from figure 4.14 that the frailties, $\hat{z}_{i,ingau}$ follow a similar linear pattern to the fixed effects, only these are slightly higher than the lognormally distributed frailties, $\hat{z}_{i,lognorm}$, and the difference seems to increase with increasing number of claims. Although in percent (figure 4.15), in proportion to $\hat{z}_{i,lognorm}$, the difference is largest for claim numbers closer to zero with low frailties and smaller for higher number of claims and larger frailties.

Also, looking at figure 4.16 we see that the expected number of claims, i.e., $\hat{z}_i\hat{\Lambda}_i$, with both inverse Gaussian and lognormally distributed frailties are very close to each other and the observed number of claims. From figure 4.17 we see that the largest differences in expected total number of claims, in proportion to the expected total number of claims with lognormal frailties, occur when the number of claims is close to zero and the difference passes twenty percent. The other expectations have a difference below ten percent.

4.4 Summary of chapter

In this chapter we have first covered the standardization of the lognormal frailty, to make the results comparable to those of the gamma and inverse Gaussian

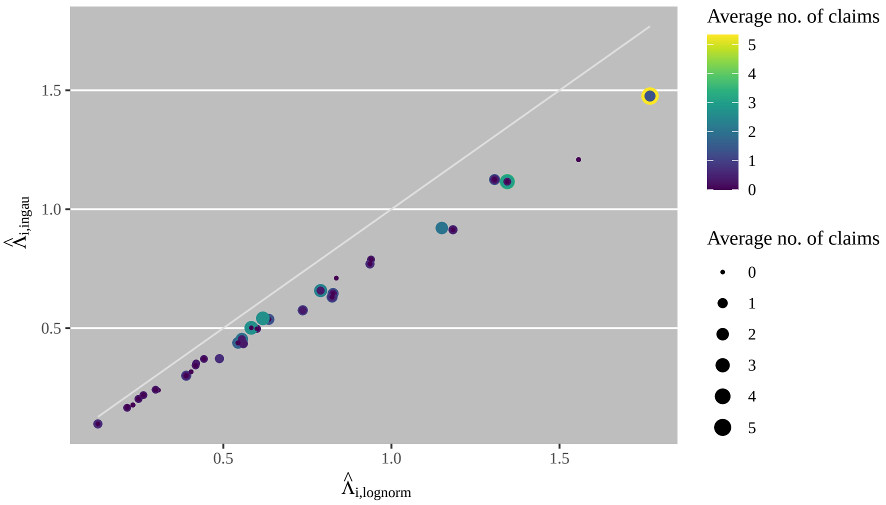

Figure 4.12: *The sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,ingau}$ for inverse Gaussian distributed frailties, $\hat{z}_{i,ingau}$, versus the sum of the fixed part of the expected number of claims, $\hat{\Lambda}_{i,lognorm}$ for lognormally distributed frailties, $\hat{z}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

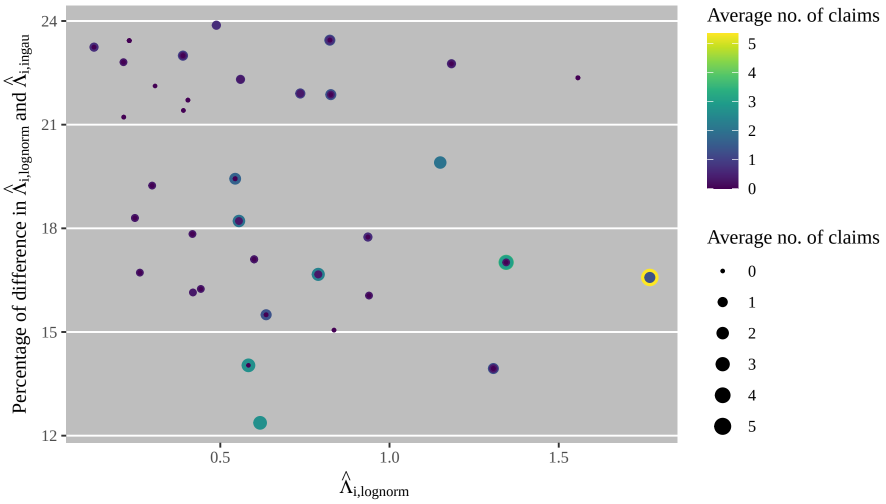

Figure 4.13: *Percentage of difference between $\hat{\Lambda}_{i,lognorm}$ and $\hat{\Lambda}_{i,ingau}$, in proportion to $\hat{\Lambda}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

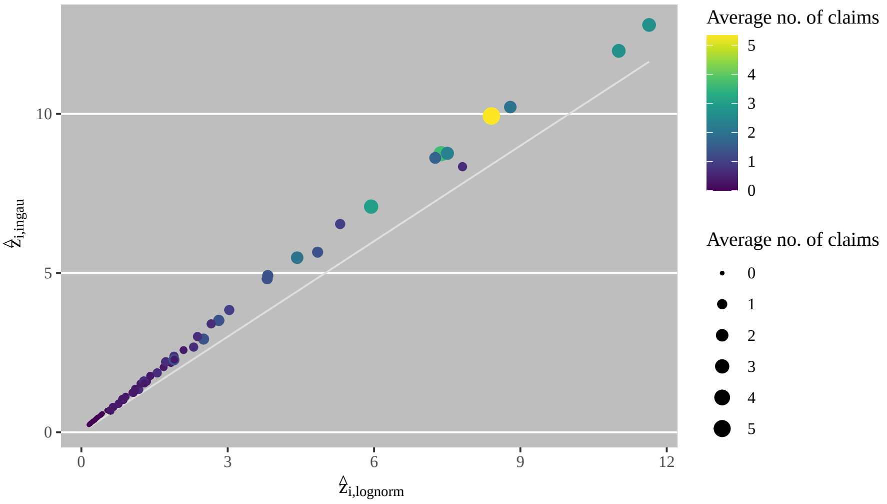

Figure 4.14: *The inverse Gaussian distributed frailties, $\hat{z}_{i,ingau}$ versus the lognormally distributed frailties, $\hat{z}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

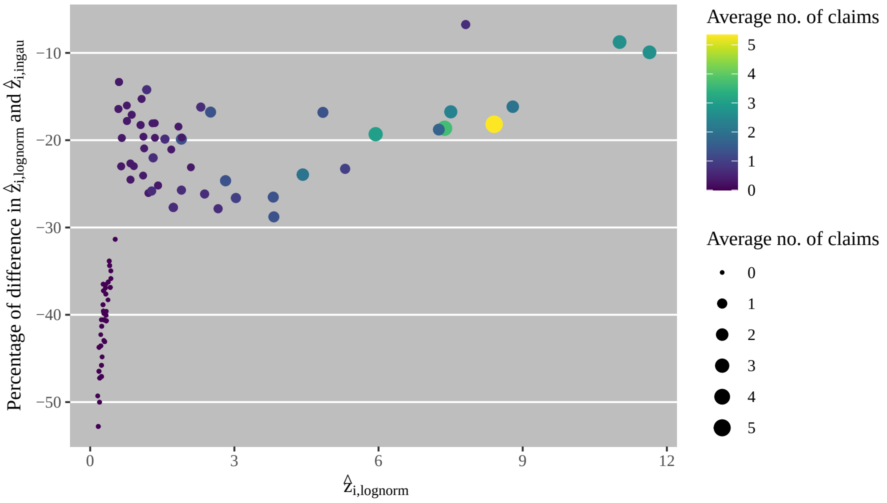

Figure 4.15: *Percentage of difference between $\hat{z}_{i,lognorm}$ and $\hat{z}_{i,ingau}$, in proportion to $\hat{z}_{i,lognorm}$. With colour and size scales indicating the average number of claims reported per year.*

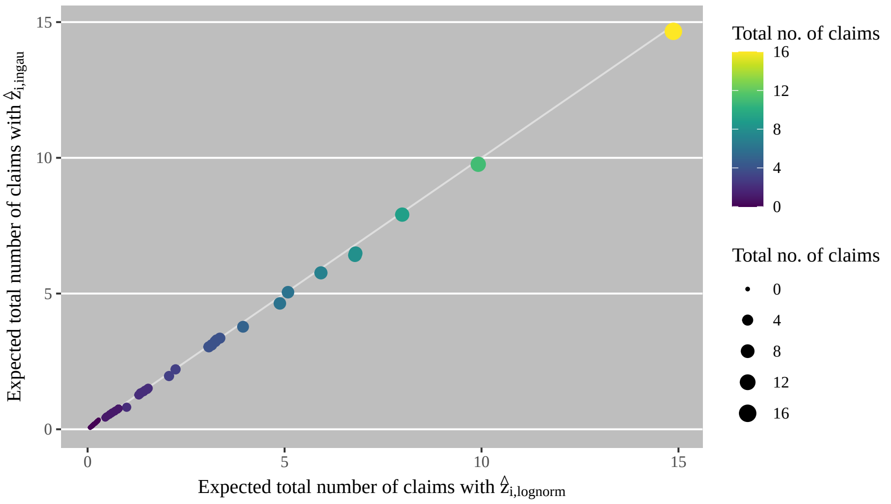

Figure 4.16: *Estimated expected number of claims for the total of the $K = 3$ years with lognormal and inverse Gaussian distributed frailties. With colour and size scales indicating the total number of claims reported.*

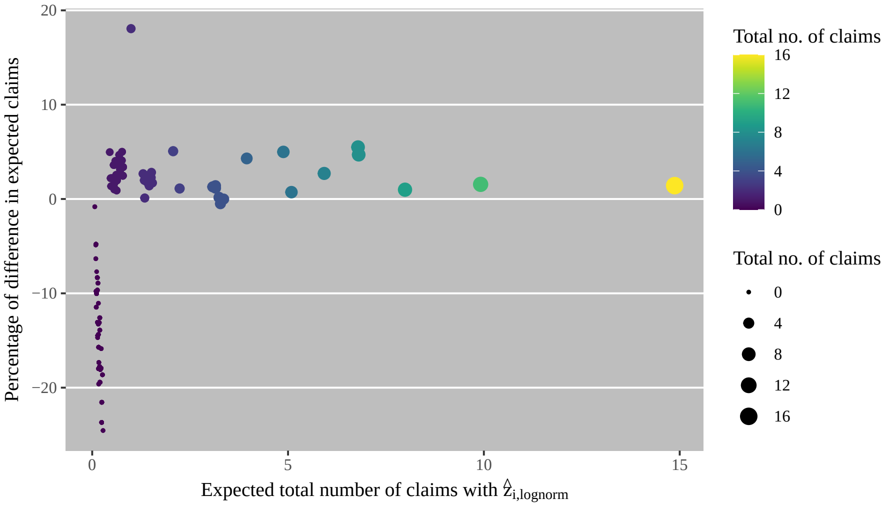

Figure 4.17: *Percentage of difference between expected total number of claims with lognormal and inverse Gaussian distributed frailties, in proportion to the expected total number of claims with lognormal frailties. With colour and size scales indicating the total number of claims reported.*

distributed frailties. We have then presented the gamma and inverse Gaussian distributed frailties. The examples cover comparisons of the results, from fitting the data using these distributions for the frailty, to those (a standardized version) of the fit with lognormal frailty from the results presented in the previous chapter.

From these results we have seen that the expected number of claims are fairly similar for all the distributions of frailty, although the ones predicted by using gamma distribution of frailty are mostly slightly lower than those resulting from fitting with lognormal and inverse Gaussian frailties. The differences between them seem to be increasing with higher numbers of observed claims, although in percentage the differences turn out to be larger for policyholders with fewer reported claims. This is because these numbers are so small that a small change makes a bigger difference here than a slightly larger change does for an already large number of claims. The differences are generally smaller between the results from using a lognormal frailty and an inverse Gaussian frailty, than between using a lognormal and a gamma distributed frailty.

5

Importance of the distribution of the frailty

In the examples used in the previous chapters we have used a data set where we do not know the distribution of the frailty used to generate the number of claims. In order to be able to say something more about the importance of the choice of distribution of the frailty when fitting a model, we will in this chapter simulate our own data where we control the distribution of the frailty.

5.1 Simulation of data

To simulate data for n policyholders for K periods, where each period is one year(i.e., $E_{ik} = 1$), we use as covariates the age of the policyholder, the age of the car and the mileage of the car. The ages of the policyholder, x_1, and the car, x_2, are numeric covariates while the mileage is categorical with 4 levels, giving 3 dummy variables, x_3, x_4 and x_5. The age of the policyholder was sampled from a discrete sequence of ages from 18 to 100. The probability of drawing each age was set based on the proportion of the population for each age in Norway in 2020 (Statistisk sentralbyrå 2020a). Figure 5.1 shows the distribution of age of policyholders.

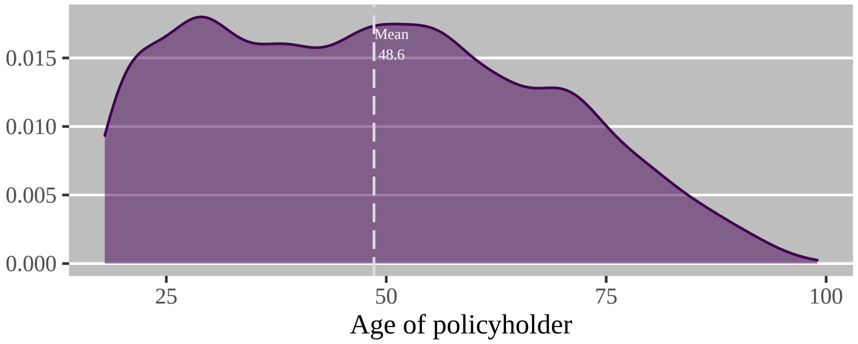

Figure 5.1: *Distribution of age of policyholders.*

The ages of the cars were sampled in a similar manner, from a sequence of ages from 0 (new car) to 21, where 21 includes any car aged >20 years old. The data on ages of cars in Norway (for 2019) (Statistisk sentralbyrå 2020b) was given for groups of ages, so to set the probabilities of drawing a single age, the proportions for each group was divided equally among the ages in the group. Figure 5.2 shows the distribution of age of cars.

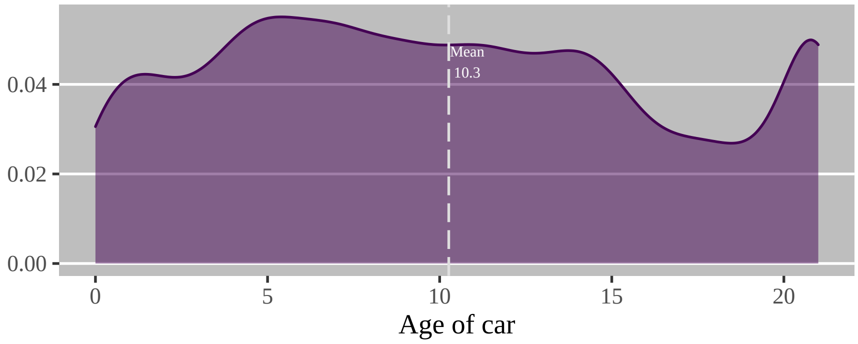

Figure 5.2: *Distribution of age of cars.*

All ages, for both cars and policyholders, were increased by one for each period, from period 2 onwards. The mileage of each car was sampled from four categories, with equal probabilities, where category 1 indicates low mileage and category 4 indicates high mileage. We draw one set of covariates for $n = 250$ policyholders and $K = 4$ periods, which is then used throughout all the simulations. The accompanying code may be found in Appendix B Listing B.2

Some suitable values for the coefficients, β, are also chosen, including a baseline intensity parameter, θ. They were set to $\theta = e^{-2.7} \approx 0.067$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (0.005, 0.009, \log(1.2), \log(1.4), \log(1.6))$. The fixed effects are then $\theta e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}}$, where $\mathbf{X}_{ik} = (x_{ik1}, x_{ik2}, x_{ik3}, x_{ik4}, x_{ik5})^T$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. The first two, x_{ik1} and x_{ik2} increase by one for each period, k, while the rest are kept the same for all periods, i.e., does not change with k.

Then, based on the desired distribution, the frailty,Z, is drawn from either a lognormal distribution, a gamma distribution or an inverse Gaussian distribution, such that they have expectation one and the desired variance. Here we use two different variances on the frailty, one low of $\sigma_z^2 = 0.75^2 = 0.5625$ and one high of $\sigma_z^2 = 2^2 = 4$. The number of claims for each policyholder in each period is then drawn from a Poisson distribution with expectation (and variance) given by the frailty times the fixed effects, $Z_i \theta e^{\boldsymbol{\beta}^T \mathbf{X}_{ik}}$. Together with the id for policyholder, the periods and the covariates, the number of claims make up the data set. The data is then converted to a clustered survival data format to be compatible for use with `parfm` (see Listing B.1 in Appendix B).

The data, with the number of claims generated based on a given distribution of the frailty, is then fitted using `parfm` for each of the three frailties we are looking at. We then aggregate the data by summing up the total number of claims "observed" over all the $K = 4$ periods, for each policyholder, i. To this resulting set of data we also add the predicted frailties, $\hat{z}_i$, for the three model fits, as well as the sums of fixed effects, $\hat{\Lambda}_i = \sum_{k=1}^{K} \hat{\theta} e^{\hat{\boldsymbol{\beta}}^T \mathbf{x}_{ik}}$, and expected number of total claims, $\hat{z}_i \hat{\Lambda}_i$. We also compute deviations of the fitted models from the model fitted using the same distribution of the frailty as was used to generate the data, e.g., $\hat{z}_{i,gamma} - \hat{z}_{i,lognorm}$ when the model was fitted using gamma distributed frailty and the data was generated using a lognormally distributed frailty. These deviations were computed for all three types of resulting data, i.e., frailties, sums of fixed effects and the number of expected claims. The complete code for this function can be found in Appendix B Listing B.3.

For each of the 6 combinations of frailty distribution and frailty variance, we performed 10 replications. We chose to use only 10 replications since one replication took a bit less than a minute to run, and this amounts to almost an hour to run all the $6 \times 10 = 60$ replications. Because of the limited time frame of this thesis, adding many more replications would take too long to run. Also, for our purpose of looking at the effect of assumptions made on the distribution of the frailty, 10 replications are enough. For each of the 10 replications we get estimates of sums of fixed effects, $\hat{\Lambda}_i$, frailty, $\hat{z}_i$, and expected number of claims, $\hat{z}_i \hat{\Lambda}_i$ for 250 policyholders. So in total we get these estimates for 2 500 policyholders.

5.2 Results using low variance on frailty

Before we look at all the 10 replications, we will first take a look at the results from one simulation with low variance ($\sigma_z = 0.75$). Table 5.1 shows how many of the $n = 250$ policyholders reported different numbers of claims, in total over the $K = 4$ periods, for the different frailty distributions used to generate the data. From this we see that at most 5 total claims are reported, but only one or two policyholders reported this many claims. The majority of policyholders reported 0 claims, with quite a few reporting 1 or 2 claims.

Table 5.1: *The number of policyholders, of the $n = 250$, with the number of total claims over the $K = 4$ periods for the different distributions of frailty for one simulation, with low variance.*

No. of claims	0	1	2	3	4	5
Lognormal	158	68	17	6	1	
Gamma	162	62	18	5	1	2
Inverse Gaussian	160	59	23	6	1	1

From figure 5.3 column 1 we see that the sums of fixed effects, $\hat{\Lambda}_i$ are very similar for all the fitted distributions, but covers a slightly larger range when the data is generated with the gamma distributed frailty. For the predicted frailties, $\hat{z}_i$, the values are all similar when the number of claims is low, with mostly increasing deviation with increasing number of claims (figure 5.3 column 2). When the data is generated using a lognormal frailty (plot A.2 in figure 5.3) the predicted frailties when fitting with gamma and inverse Gaussian distributions are fairly similar to each other, and lower than the fit lognormal distribution, and starts deviating from the fit with lognormal distribution with two or more reported claims. From plot B.2 (in figure 5.3) we see that the predicted frailties when fitting the model with lognormal and inverse Gaussian distributions, with gamma distribution used to make the data, are similar to

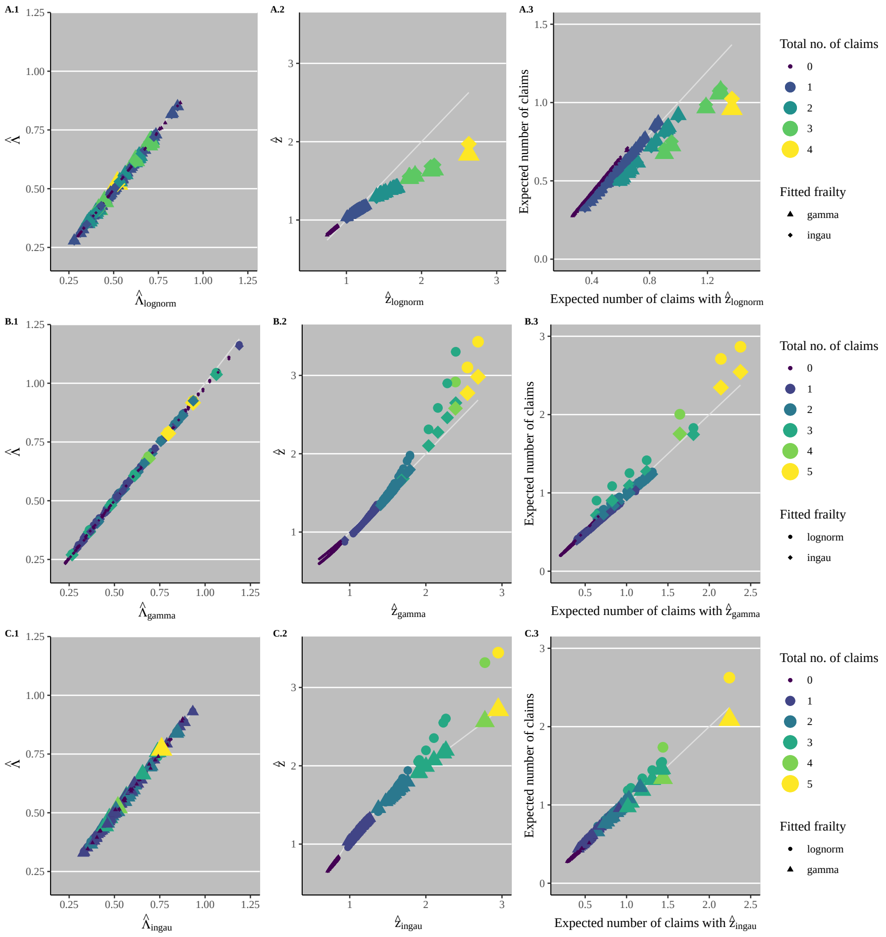

Figure 5.3: *Results of one simulation with low variance. Each row, A, B and C, containing results using lognormal, gamma and inverse Gaussian frailty respectively to generate the data. Columns 1, 2 and 3 contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and size gradients indicate the total number of claims observed, and shape indicating the distribution used for the frailty when fitting the model to the data. The grey line indicate the result when fitting with the same distribution that was used to generate the data.*

the predicted frailties when fitting with gamma distribution when fewer than three claims are reported. With more than three claims they are a bit higher, with lognormal predictions a bit above the inverse Gaussian ones. In the case where the data was generated using an inverse Gaussian distributed frailty (plot C.2 in figure 5.3) the predicted frailties when fitting the model with gamma distributed frailty is very similar to that predicted when fitting with inverse Gaussian distributed frailty. Fitting with lognormally distributed frailty in this case is larger and deviates more from the other two predicted frailties with increasing number of claims.

Column 3 of figure 5.3 contains the expected number of total claims over all $K = 4$ periods. For all cases we see that a low number of observed claims give very similar results of expected number of claims, and deviating more with increasing number of observed claims. Plot A.3 presents the results of fitting the model with the different distributions of frailty to data generated using the lognormal distribution on the frailty. Here both gamma and inverse Gaussian fits expects fewer claims than the lognormal fit when the observed number of claims are higher, and they all expect fewer claims than what is observed. When the data is generated using a gamma distributed frailty (plot B.3) both the lognormal and inverse Gaussian fits expect more claims than the gamma fit for higher numbers of observed claims. Using inverse Gaussian distributed frailty to produce the data set (plot C.3) yields very similar expected number of claims for all three distributions of frailty when fitting the model, although there is some deviation for the highest numbers of observed claims.

Regardless of the distribution used to generate the data we see a similar pattern of expected number of claims, with very similar expectations for all fitted distributions when the observed number of claims is low. When the observed number of claims go higher there is some deviation between the fits, with a general pattern of the gamma and inverse Gaussian being very similar with gamma slightly lower, and the lognormal fit being a bit above the other two. This increased deviation when the observed number of claims increase may be due to the limited amount of data points for these numbers of claims.

We will now take a closer look at the deviations from all the replications. Figure 5.4 shows the deviations of gamma and inverse Gaussian fits from the lognormal fit, when the data is generated using a lognormal distribution on the frailty. It shows that for the sums of fixed effects(column A), $\hat{\Lambda}_i$, the deviations are spread over a bit larger area when fitted with gamma distributed frailty than when fitted with inverse Gaussian frailty, but the tendency for both is that they stay fairly close to zero. For the predicted frailties(column B), $\hat{z}_i$, the tendency is that the deviations descend from zero at a slightly quicker rate for the gamma fit than for the inverse Guassian fit. The deviations of the expected number of claims(column C), $\hat{z}_i\hat{\Lambda}_i$, follow a similar pattern to that of the frailties, but the deviations are lower.

We see from figure 5.5 that the deviations of lognormal and inverse Gaussian fits from the gamma fit, when the data is generated with a gamma distributed frailty, has a similar pattern to that of the lognormal data for the sums of fixed effects(column A), $\hat{\Lambda}_i$, although they are stretched over a slightly wider area. The deviations of the predicted frailties(column B) and the expected number of claims(column C) follow a similar pattern for both fitted distributions of increasing with increasing values (on the x-axis), although with a slightly steeper ascent for the lognormal than for the inverse Gaussian. The deviations of the

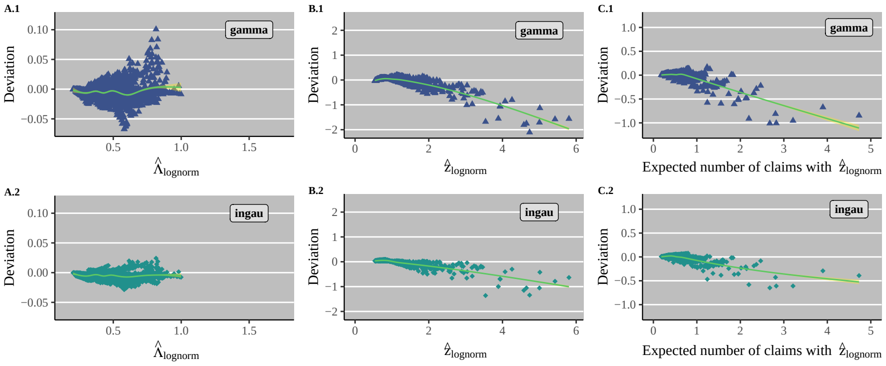

Figure 5.4: *Deviations of ten simulations with low variance, for data generated using lognormally distributed frailty. Rows, 1 and 2, containing results using gamma and inverse Gaussian frailty respectively to fit the data. Columns A, B and C contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and shape indicating the distribution used for the frailty when fitting the model to the data. The line indicate the result of using a generalized additive smoothing function on the deviances.*

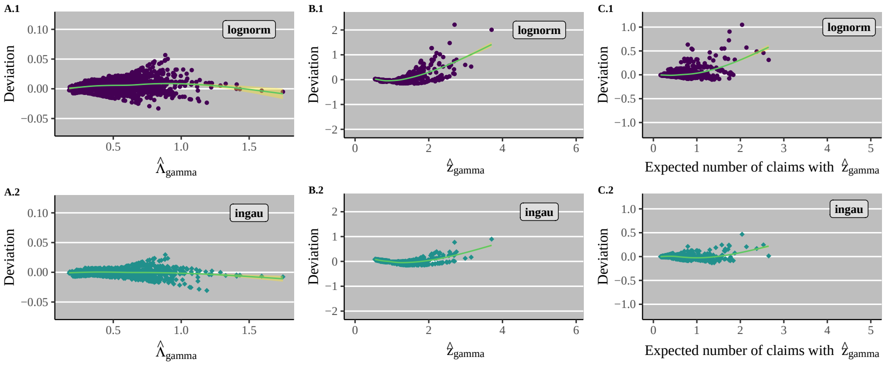

Figure 5.5: *Deviations of ten simulations with low variance, for data generated using gamma distributed frailty. Rows, 1 and 2, containing results using lognormal and inverse Gaussian frailty respectively to fit the data. Columns A, B and C contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and shape indicating the distribution used for the frailty when fitting the model to the data. The line indicate the result of using a generalized additive smoothing function on the deviations.*

frailties are also higher than those of the expected claims.

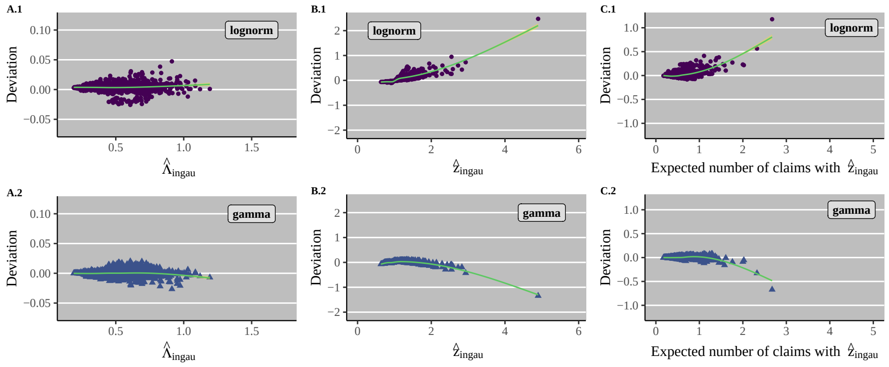

Figure 5.6: *Deviations of ten simulations with low variance, for data generated using inverse Gaussian distributed frailty. Rows, 1 and 2, containing results using lognormal and gamma frailty respectively to fit the data. Columns A, B and C contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and shape indicating the distribution used for the frailty when fitting the model to the data. The line indicate the result of using a generalized additive smoothing function on the deviations.*

We then look at the case where the data was generated using an inverse Gaussian distribution on the frailty (figure 5.6). The deviations of the sums of fixed effects follow similar pattern to the previous two cases, although with a bit less variation and staying within 0.05 of the fit when using inverse Gaussian frailty. The predicted frailties deviate in opposite directions for the two fits, with the one for the lognormal frailty having more of a gradual ascent the gamma one just drops off towards the end. The expected number of claims follow a similar pattern, but with smaller deviations. In general, the confidence bands of the smoothing curves are larger at the end with higher values due to fewer data points.

5.3 Results using high variance on frailty

We will now go back to looking at the results for only one simulation, this time with high variance ($\sigma_z = 2$) on the frailty. Table 5.2 shows how many of the $n = 250$ policyholders reported different numbers of claims in total over the $K = 4$ periods, for the different distributions of frailty used in the simulation of data. From this we see that the largest total number of claims observed now increase to 13, and that we now observe more policyholders with zero claims than we did with the low variance on the frailty. This can be explained by the fact that the maximum of the density of the distributions used for the frailties move to the left of the mean with increasing standard deviations, with longer tails making it possible to draw higher values of frailties (see figures 4.1, 4.2 and 4.10).

From column 1 in figure 5.7 we now see that the sums of fixed effects, $\hat{\Lambda}_i$, are spread out a bit more when the data are generated using gamma and inverse Gaussian distributions than they were with the low frailty. The values are

Table 5.2: *The number of policyholders, of the $n = 250$, with the number of total claims over the $K = 4$ periods for the different distributions of frailty, with high variance.*

No. of claims	0	1	2	3	4	5	6	7	9	10	11	13
Lognormal	182	46	14	1	3	1	1	1		1		
Gamma	189	33	13	5	4	2	1		1		1	1
Inverse Gaussian	177	44	9	8	5	2	2		1	1	1	

in a similar range as before. The predicted frailties (column 2), $\hat{z}_i$, and the expected number of total claims (column 3), $\hat{z}_i\hat{\Lambda}_i$, both show similar patterns to what they did when using a low variance when generating the data. When the observed number of claims are lower the values are similar for all fitted distributions of the frailty. The deviations are increasing with increasing number of observed claims, but the values are now larger. There is also still a trend that the expected number of claims are fewer than the number observed across the board. The pattern of the lognormal fit producing higher values than the inverse Gaussian fit, which in turn is producing higher values than the gamma fit, is also still present.

We will now revert to looking at all the 10 simulations again, starting with the deviations when the data are generated using a lognormally distributed frailty (figure 5.8). The deviations of the sums of fixed effects now show trends of descending from zero for larger values, for both fitted frailties. The deviations for the frailties and expected number of claims mostly have similar trends to the low variance case (figure 5.4), but with larger deviations. The exception being for the predicted frailty when fitting with inverse Gaussian frailty, where the trend is now more of a straight line at zero.

Then when the data is generated using gamma distribution of the frailty we see from figure 5.9 that the trends are similar to those as for the low variance case (figure 5.5), except the deviations are much higher. The same is true when the data are generated using the inverse Gaussian distribution on the frailty (figure 5.10).

5.4 Summary of chapter

In this chapter we have looked at the effects of the choice of distribution on the frailty when fitting a model for the number of claims on car insurance. We did so by simulating a set of covariates, that were kept the same throughout. The number of claims, to combine with the covariates to make up the data set, was generated from a Poisson distribution with $z_i\Lambda_i$ as the expectation (and variance). The only variable changing here is the frailty, z_i, which was drawn from three different distributions, always with expectation 1 and a choice of two different variances. All six combinations of frailty distribution and variance was then used to make data sets, which in turn were fitted using all three distributions of frailty as a modelling assumption.

From this we found that the choice of distribution of the frailty when fitting a model for the number of claims on car insurance does matter. When the variance of the frailty is high the deviations between the choice of frailty to use when fitting the model is larger, deviating by as much as three claims for the

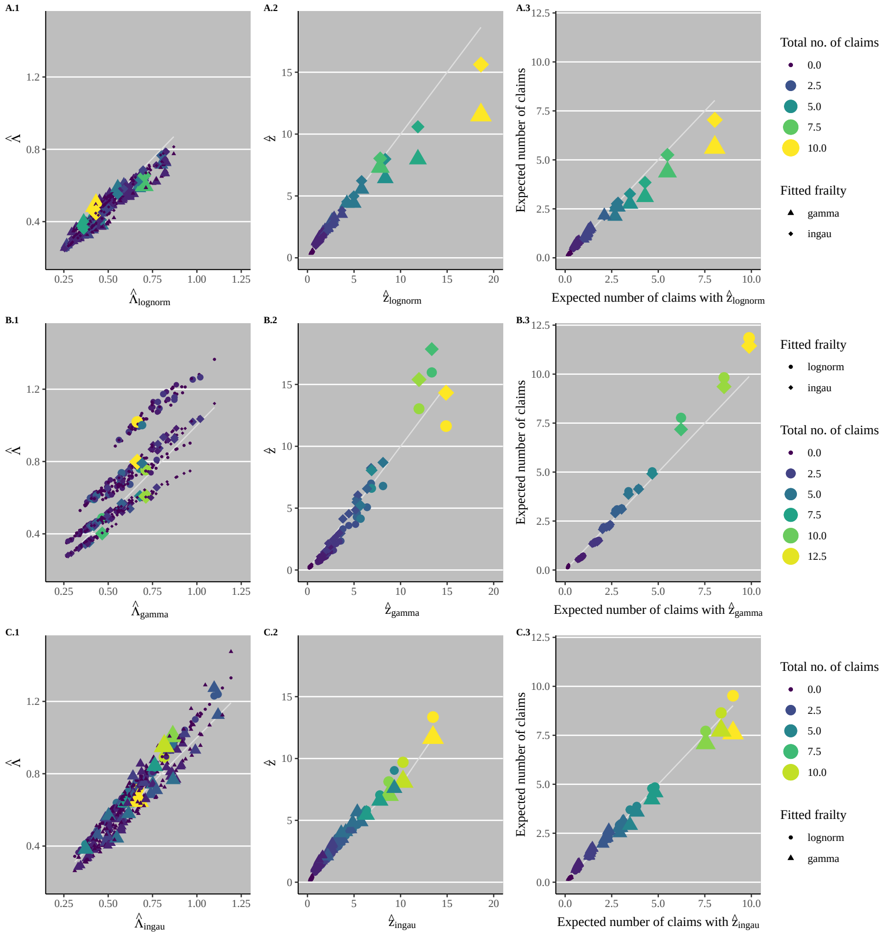

Figure 5.7: *Results of one simulation with high variance. Each row, A, B and C, containing results using lognormal, gamma and inverse Gaussian frailty respectively to generate the data. Columns 1, 2 and 3 contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and size gradients indicate the total number of claims observed, and shape indicating the distribution used for the frailty when fitting the model to the data. The grey line indicate the result when fitting with the same distribution that was used to generate the data.*

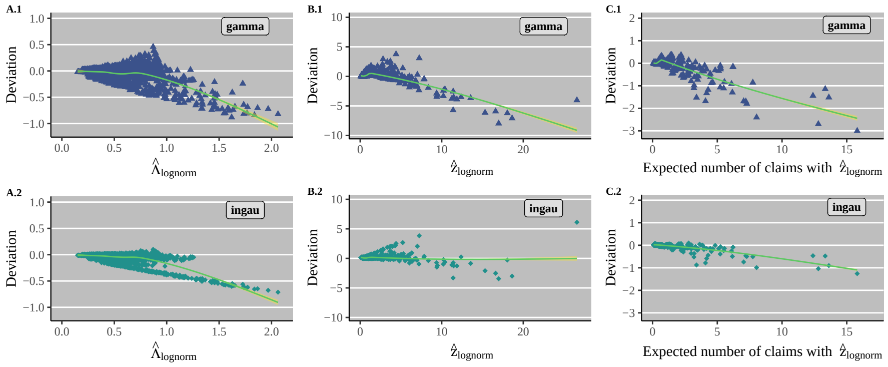

Figure 5.8: *Deviations of ten simulations with high variance, for data generated using lognormally distributed frailty. Rows, 1 and 2, containing results using gamma and inverse Gaussian frailty respectively to fit the data. Columns A, B and C contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims ,$\hat{z}_i\hat{\Lambda}_i$. Colour and shape indicating the distribution used for the frailty when fitting the model to the data. The line indicate the result of using a generalized additive smoothing function on the deviations.*

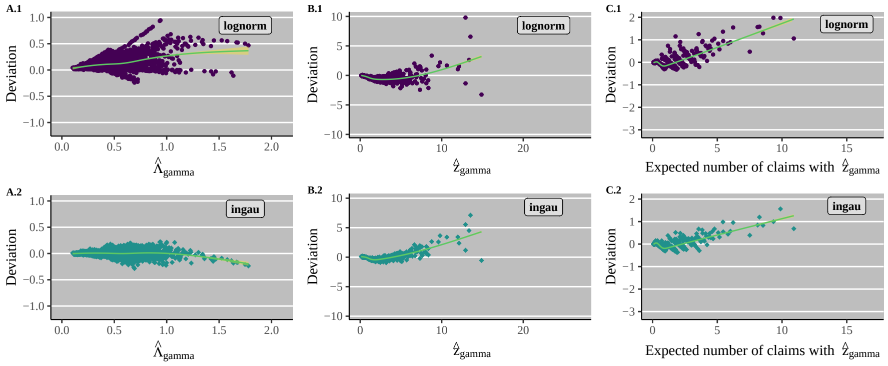

Figure 5.9: *Deviations of ten simulations with high variance, for data generated using gamma distributed frailty. Rows, 1 and 2, containing results using lognormal and inverse Gaussian frailty respectively to fit the data. Columns A, B and C contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and shape indicating the distribution used for the frailty when fitting the model to the data. The line indicate the result of using a generalized additive smoothing function on the deviations.*

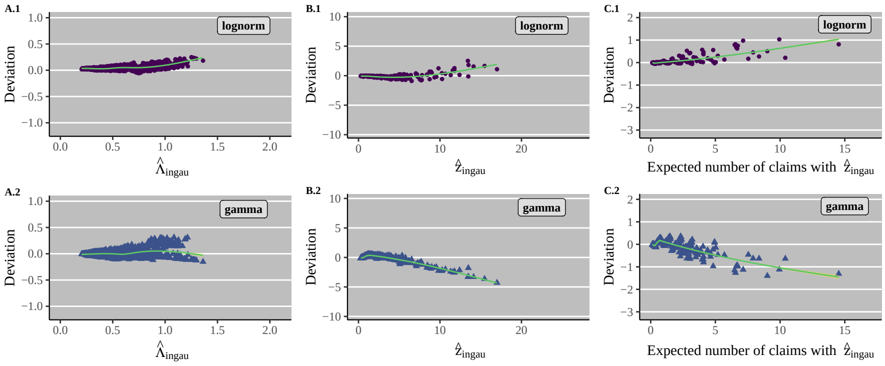

Figure 5.10: *Deviations of ten simulations with high variance, for data generated using inverse Gaussian distributed frailty. Rows, 1 and 2, containing results using lognormal and gamma frailty respectively to fit the data. Columns A, B and C contain sums of fixed effects, $\hat{\Lambda}_i$, frailties, $\hat{z}_i$, and expected number of claims, $\hat{z}_i\hat{\Lambda}_i$. Colour and shape indicating the distribution used for the frailty when fitting the model to the data. The line indicate the result of using a generalized additive smoothing function on the deviations.*

expected number of claims. This makes it more important to choose the right distribution for data with a high degree of heterogeneity than for those with a low degree of heterogeneity. The differences are also more prominent when the number of claims is higher. Seeing as the policyholders with many claims are often also the customers that will cost an insurance company the most, it is important to get the estimated number of claims for these policyholders as close to the true value as possible.

6

Concluding remarks

We have seen how both Poisson regression and recurrent events models can be used to model the number of claims to expect on a car insurance policy. We have also seen how the same is true of their extensions, i.e., including a random effect or frailty.

As we have established this interchangeability between models, the focus is now only on the recurrent events model and the distribution of the frailty. The results of fitting a model to the example data set using lognormal, gamma and inverse Gaussian frailties show that there are some differences in the predicted frailties and the number of claims to expect between the three. The lognormal distribution seems to be the one that predicts the most correct number of claims. Although we do not know the distribution that was used to generate this set of data, the fact that it was made for use in an example on generalized linear mixed models with random intercept, it is highly likely that a lognormal distribution was used for the random effect (i.e., a normal distribution on the random intercept).

To alleviate the problem of not knowing the true distribution of the frailty in the data, we simulate our own data. This way we are also able to look at possible different effects the choice of frailty distribution, when fitting the data, has for different "true" distributions of frailty. We chose to keep the covariates the same through all the data sets we generated and only changed the number of claims, by setting different distributions and variances on the "true" frailty.

Similarly to the results fitting models to the example data set, fitting models using the different distributions of frailty on our simulated data indicates that there is an effect of the choice of distribution to use on the frailty when fitting a model. The deviations between choices of distribution to use when fitting the model were larger when the "true" variation was higher, making the choice of distribution to use on the frailty when fitting a model more important for data with a higher degree of heterogeneity than for data with a lower degree of heterogeneity.

The differences were also more notable when the number of "observed" claims were higher. Since the policyholders with the most claims are also often the most costly customers to an insurance company, it will be in their interest to have a model that also does well for higher number of expected claims to allocate enough for future compensations.

Regarding possible further improvements the deviation of the fitted values from the "true" ones could also be considered, to check if the distribution used to generate the data also actually produces the best fitting model (like we have assumed it did). Another aspect to consider is the number of replications used. It could also be possible to improve the computational times by using other optimization methods.

The results of the examples in chapter 4 and the simulations in chapter 5 do not give quite the same results, but due to the time limitations of this thesis we have not had time to look at this further. One possible explanation for the different results may be that the variance is higher in the examples than in the simulations. Another possibility could be in the composition of the data set used in the examples. The reduced data set used for the examples keeps all the policyholders from the highest value categories from the full data set, possibly making this data set somewhat less realistic.

We have also only looked at three different distributions on the frailty, but there are of course many other feasible distributions that can be used, and that may yield different results. Some potential distributions are the other possible choices of distribution to use on the frailty in `parfm`, the positive stable and the loglogistic. Another possibility could be to use any of the other distributions in the family of power variance function distributions (Aalen et al. 2008, ch. 6.2.3).

For a real-world data set one does not know the true distribution of the frailty, and as our results have shown the choice of distribution of the frailty does have an effect on a prediction model. So it might be wise to try other distributions than your standard go-to choice of distribution for the frailty, and use some model selection criteria to choose the one that fits the data best.

References

Aalen, O., Borgan, Ø., and Gjessing, H. 2008. *Survival and Event History Analysis: A Process Point of View.* 1st ed. 2008. Statistics for Biology and Health. New York, NY: Springer New York : Imprint: Springer.

Agresti, A. 2015. *Foundations of Linear and Generalized Linear Models.* Wiley series in probability and statistics. Hoboken, N.J: Wiley.

Bates, D., Mächler, M., Bolker, B., and Walker, S. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (no. 1): 1–48.

Broström, G. 2019. "glmmML: Generalized Linear Models with Clustering." R package version 1.1.0.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. 2019. "xtable: Export Tables to LaTeX or HTML." R package version 1.8-4.

Davison, A. C. 2003. *Statistical Models.* X, 726. Cambridge series in statistical and probabilistic mathematics. Cambridge: University Press.

De Jong, P. and Heller, G. Z. 2008. *Generalized Linear Models for Insurance Data.* International series on actuarial science. Cambridge: University Press.

Garnier, S. 2018. "viridis: Default Color Maps from 'matplotlib'." R package version 0.5.1.

Günther, C.-C., Tvete, I. F., Aas, K., Hagen, J. A., Kvifte, L., and Borgan, Ø. 2014. "Predicting Future Claims Among High Risk Policyholders Using Random Effects." In *Modern Problems in Insurance Mathematics,* edited by Silvestrov, D. and Martin-Löf, A., 171–185. EAA Series. Cham: Springer International Publishing.

Hougaard, P. 2000. *Analysis of Multivariate Survival Data.* XVII, 542. Statistics for biology and health. New York: Springer.

Liu, Q. and Pierce, D. A. 1994. "A note on Gauss—Hermite quadrature." *Biometrika* 81, no. 3 (September 1, 1994): 624–629.

Munda, M., Rotolo, F., and Legrand, C. 2012. "parfm: Parametric Frailty Models in R." *Journal of Statistical Software* 51, no. 1 (1 2012): 1–20.

Munda, M., Rotolo, F., and Legrand, C. 2017. “parfm: Parametric Frailty Models in R.” Package vignette, V. 1.4.

Pedersen, T. L. 2019. “patchwork: The Composer of Plots.” R package version 1.0.0.

R Core Team. 2019. “R: A Language and Environment for Statistical Computing,” R Foundation for Statistical Computing.

Statistisk sentralbyrå. 2020a. “07459: Alders- og kjønnsfordeling i kommuner, fylker og hele landets befolkning (K) 1986 - 2020.” Accessed May 16, 2020. http://www.ssb.no/statbank/table/07459/.

———. 2020b. “08581: Alder og bilmerke på person- og varebiler 2008 - 2019.” Accessed May 16, 2020. http://www.ssb.no/statbank/table/08581/.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, H., François, R., Henry, L., and Müller, K. 2020. “dplyr: A Grammar of Data Manipulation.” R package version 0.8.5.

Appendices

A

Results for GLMMs

Results from running `glmmML` and `glmer` on the full data set of chapter 2, using Laplace approximation and Gauss–Hermite with different number of quadrature points.

A.1 Results using `glmmML`

Table A.1 shows the result when using Gauss–Hermite approximation with 8 quadrature points, while results in tables A.2 and A.3 was produced using 15 and 20 quadrature points respectively. All fitted using `glmmML`.

A.2 Results using `glmer`

Table A.4 shows the result when using Laplace approximation Gauss–Hermite approximation with 8 quadrature points, while results in tables A.5, A.6 and A.7 was produced using 8, 15 and 20 quadrature points respectively with a Gauss–Hermite approximation. All fitted using `glmer`.

Table A.1: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Gauss–Hermite approximation with 8 quadrature points with `glmmML` on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-2.9299	0.0309	0.0534	0
Age category				
Group 1	0.2648	0.0482	1.3031	3.919e-08
Group 2	0.0417	0.0383	1.0426	0.2766
Group 3	0.0463	0.0367	1.0474	0.207
Group 4	0.0000		1.0000	
Group 5	-0.1867	0.0418	0.8297	7.993e-06
Group 6	-0.1385	0.0489	0.8707	0.00462
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1989	0.0329	1.2201	1.461e-09
$50 - 75$	0.0768	0.0939	1.0799	0.4134
$75 - 100$	-0.6207	0.3733	0.5376	0.09641
$100 - 125$	-0.4435	0.5710	0.6418	0.4373
> 125	-1.2719	0.6963	0.2803	0.06775
Period				
1	0.0000		1.0000	
2	0.1062	0.0149	1.1121	8.514e-13
3	0.2344	0.0144	1.2641	0
Standard error of random intercept				
$\hat{\sigma}_u$	1.6456	0.0140		0

Table A.2: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Gauss–Hermite approximation with 15 quadrature points with `glmmML` on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-2.9453	0.0313	0.0526	0
Age category				
Group 1	0.2652	0.0486	1.3037	4.744e-08
Group 2	0.0422	0.0386	1.0431	0.2744
Group 3	0.0464	0.0370	1.0475	0.2097
Group 4	0.0000		1.0000	
Group 5	-0.1865	0.0422	0.8299	9.662e-06
Group 6	-0.1383	0.0493	0.8709	0.005025
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1995	0.0331	1.2208	1.743e-09
$50 - 75$	0.0770	0.0947	1.0800	0.4161
$75 - 100$	-0.6268	0.3774	0.5343	0.09678
$100 - 125$	-0.3842	0.5716	0.6810	0.5015
> 125	-1.1872	0.6919	0.3051	0.08618
Period				
1	0.0000		1.0000	
2	0.1061	0.0149	1.1119	9.073e-13
3	0.2343	0.0144	1.2640	0
Standard error of random intercept				
$\hat{\sigma}_u$	1.6653	0.0147		0

Table A.3: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Gauss–Hermite approximation with 20 quadrature points with `glmmML` on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-2.9447	0.0312	0.0526	0
Age category				
Group 1	0.2640	0.0486	1.3022	5.442e-08
Group 2	0.0421	0.0386	1.0430	0.2751
Group 3	0.0467	0.0370	1.0479	0.206
Group 4	0.0000		1.0000	
Group 5	-0.1867	0.0421	0.8297	9.467e-06
Group 6	-0.1380	0.0493	0.8711	0.005103
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1995	0.0331	1.2208	1.748e-09
$50 - 75$	0.0781	0.0946	1.0812	0.4093
$75 - 100$	-0.6177	0.3769	0.5392	0.1012
$100 - 125$	-0.4458	0.5766	0.6403	0.4394
> 125	-1.2616	0.7018	0.2832	0.07223
Period				
1	0.0000		1.0000	
2	0.1063	0.0149	1.1121	8.26e-13
3	0.2343	0.0144	1.2640	0
Standard error of random intercept				
$\hat{\sigma}_u$	1.6646	0.0147		0

Table A.4: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Laplace approximation with* `glmer` *on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-3.0973	0.0333	0.0452	0
Age category				
Group 1	0.2643	0.0504	1.3025	1.583e-07
Group 2	0.0418	0.0399	1.0427	0.2949
Group 3	0.0456	0.0382	1.0467	0.2322
Group 4	0.0000		1.0000	
Group 5	-0.1830	0.0433	0.8327	2.393e-05
Group 6	-0.1369	0.0507	0.8721	0.006929
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1977	0.0343	1.2185	8.613e-09
$50 - 75$	0.0750	0.0978	1.0779	0.443
$75 - 100$	-0.5950	0.3804	0.5516	0.1178
$100 - 125$	-0.4487	0.5852	0.6385	0.4433
> 125	-1.1964	0.7000	0.3023	0.08742
Period				
1	0.0000		1.0000	
2	0.1062	0.0148	1.1121	8.266e-13
3	0.2344	0.0144	1.2641	1.946e-59
Standard error of random intercept				
$\hat{\sigma}_u$	1.7934			

Table A.5: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Gauss–Hermite approximation with 8 quadrature points with `glmer` on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-3.0973	0.0333	0.0452	0
Age category				
Group 1	0.2642	0.0504	1.3024	1.582e-07
Group 2	0.0418	0.0399	1.0427	0.2944
Group 3	0.0456	0.0382	1.0467	0.2321
Group 4	0.0000		1.0000	
Group 5	-0.1830	0.0433	0.8327	2.364e-05
Group 6	-0.1368	0.0506	0.8721	0.006859
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1976	0.0343	1.2185	8.646e-09
$50 - 75$	0.0749	0.0978	1.0778	0.4434
$75 - 100$	-0.5948	0.3808	0.5517	0.1183
$100 - 125$	-0.4487	0.5854	0.6385	0.4434
> 125	-1.1961	0.7004	0.3024	0.08767
Period				
1	0.0000		1.0000	
2	0.1062	0.0148	1.1121	8.24e-13
3	0.2344	0.0144	1.2641	1.962e-59
Standard error of random intercept				
$\hat{\sigma}_u$	1.7934			

Table A.6: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Gauss–Hermite approximation with 15 quadrature points with* `glmer` *on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-3.0973	0.0333	0.0452	0
Age category				
Group 1	0.2642	0.0504	1.3024	1.573e-07
Group 2	0.0418	0.0398	1.0427	0.2939
Group 3	0.0456	0.0381	1.0467	0.2314
Group 4	0.0000		1.0000	
Group 5	-0.1830	0.0433	0.8327	2.354e-05
Group 6	-0.1368	0.0506	0.8722	0.006857
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1976	0.0343	1.2185	8.601e-09
$50 - 75$	0.0749	0.0978	1.0778	0.4436
$75 - 100$	-0.5948	0.3807	0.5517	0.1182
$100 - 125$	-0.4485	0.5852	0.6386	0.4434
> 125	-1.1963	0.7000	0.3023	0.08743
Period				
1	0.0000		1.0000	
2	0.1062	0.0148	1.1121	8.256e-13
3	0.2344	0.0144	1.2641	1.884e-59
Standard error of random intercept				
$\hat{\sigma}_u$	1.7934			

Table A.7: *Estimate of coefficients for the vehicle insurance data, with standard errors and claim rates for the mixed model using Gauss–Hermite approximation with 20 quadrature points with `glmer` on the full data set.*

	$\hat{\beta}$	se	$e^{\hat{\beta}}$	p-value
Intercept	-3.0973	0.0333	0.0452	0
Age category				
Group 1	0.2643	0.0504	1.3025	1.59e-07
Group 2	0.0418	0.0399	1.0427	0.2944
Group 3	0.0456	0.0382	1.0467	0.2317
Group 4	0.0000		1.0000	
Group 5	-0.1830	0.0433	0.8328	2.392e-05
Group 6	-0.1368	0.0506	0.8721	0.006883
Value category($ 000's)				
< 25	0.0000		1.0000	
$25 - 50$	0.1976	0.0343	1.2185	8.595e-09
$50 - 75$	0.0749	0.0978	1.0778	0.4434
$75 - 100$	-0.5948	0.3805	0.5517	0.118
$100 - 125$	-0.4488	0.5853	0.6384	0.4432
> 125	-1.1961	0.7000	0.3024	0.08748
Period				
1	0.0000		1.0000	
2	0.1062	0.0148	1.1121	8.407e-13
3	0.2344	0.0144	1.2641	2.008e-59
Standard error of random intercept				
$\hat{\sigma}_u$	1.7934			

B

Code

Some code used in the thesis. Listing B.1 covers the conversion of a dataset into one that is compatible with a clustered survival setting as described in section 3.2.1, and is used in chapters 3, 4 and 5 to prepare the data for use with `parfm`. Listing B.2 is a function to generate the covariates, while listing B.3 covers the simulation of a dataset and the fitting of this with the different distributions used for the frailty. The listings B.2 and B.3 are used for the simulations in chapter 5.

Listing B.1: Function to make the data survival compatible

```r
survivalData <- function(claims_data){
  idx = seq_along(claims_data[,1])
  claims_surv = data.frame()
  for (i in idx){
    if(claims_data$numclaims[i]>1){
      numcl = claims_data$numclaims[i]
      for (j in seq_len(numcl)){
        numclaims.surv = 1
        exposure = 1/numcl
        claims_surv = rbind(claims_surv,cbind(claims_data[i,],numclaims.surv,
                                              exposure))
      }
    }
    else{
      numclaims.surv = claims_data$numclaims[i]
      exposure = 1
      claims_surv = rbind(claims_surv,cbind(claims_data[i,],numclaims.surv,
                                            exposure))
    }
  }
  return(claims_surv)
}
```

Listing B.2: Function to generate the covariates

```r
simCov <- function(n=100,K=3){
  pid <- rep(1:n,each=K) # policy id
  period <- rep(1:K,n)#no fixed effect from this

  # Based on population numbers from table 07459 from SSB, for 2020
  Age.p.seq <- seq(18,100,1)
  Age.p.prob <- c(0.015,0.0157,0.0157,0.0156,0.016,0.0165,0.0165,0.0167,0.017,
                  0.0175,0.0179,0.0183,0.0182,0.018,0.0174,0.0174,0.0171,0.0169,
                  0.0168,0.0169,0.0165,0.0167,0.0165,0.0164,0.0159,0.0163,0.0167,
                  0.0174,0.0174,0.018,0.018,0.0177,0.0181,0.0179,0.0174,0.0173,
                  0.0169,0.0166,0.0159,0.0155,0.0152,0.015,0.0149,0.0147,0.0143,
                  0.0144,0.014,0.0136,0.0133,0.013,0.0124,0.0125,0.0124,0.0124,
                  0.0125,0.0127,0.0111,0.0105,0.0091,0.0082,0.0069,0.0069,0.0064,
                  0.0059,0.0053,0.0048,0.0043,0.004,0.0036,0.0035,0.0031,0.0028,
                  0.0023,0.0019,0.0016,0.0012,0.0009,0.0007,0.0005,0.0004,0.0003,
                  0.0002,0.0001)
  Age.p <- as.vector(apply(data.frame(rep(sample(Age.p.seq,n,prob=Age.p.prob,
                                                 replace=T),each=K)),2,
                           function(x) x+period-1))

  # Based on table 08581 on age of cars from SSB, numbers for 2019
  # Grouped ages in table -> divided probabilities equally among ages in same group
  Age.v.seq <- seq(0,21,1) # 21=20+ i.e. includes all cars above 20 years
  Age.v.prob <- c(0.042,0.042,0.042,0.042,0.057,0.057,0.057,0.057,0.051,0.051,
                  0.051,0.051,0.046,0.046,0.046,0.046,0.025,0.025,0.025,0.025,
                  0.025,0.091)
  Age.v <- as.vector(apply(data.frame(rep(sample(Age.v.seq,n,prob=Age.v.prob,
                                                 replace=T),each=K)),2,
                           function(x) x+period-1))

  mileage.cat <- c("1","2","3","4") # 1: few kilometers 4: many kilometers
  mileage.v <- rep(sample(mileage.cat, n, replace=T),each=K)

  return(data.frame(pid=pid,period=period,Age.p=Age.p,Age.v=Age.v,
               mileage.v=factor(mileage.v)))
}
```

Listing B.3: Functions to simulate and fit data with different frailties

```
library(statmod)
library(parfm)
library(tidyverse)

# Importing function to make the data survival compatible
source("codefiles/survivalData.R")

simModel <- function(covMat,frailty="lognorm",sigma2){
  n = length(unique(covMat$pid))
  K = length(unique(covMat$period))
  #intercept,period,Age.p,Age.v,mileage.v2,mileage.v3,mileage.v4
  beta = c(-2.7,0.005,0.009,log(1.2),log(1.4),log(1.6))
  X <- model.matrix(~covMat$Age.p+covMat$Age.v+covMat$mileage.v)

  fixed = exp(X%*%beta)

  if(frailty=="lognorm"){
    z <- rep(rlnorm(n,-(log(sigma2+1)/2),sqrt(log(sigma2+1))),each=K)
  }
  else if(frailty=="gamma"){
    z <- rep(rgamma(n,1/sigma2,scale=sigma2),each=K)
  }
  else if(frailty=="ingau"){
    z <- rep(rinvgauss(n,1,dispersion=sigma2),each=K)
  }

  numclaims = rpois(n*K,fixed*z)

  data = cbind(covMat,numclaims=numclaims)
  data_surv = survivalData(data)

  frailties=c("lognorm","gamma","ingau")
  fit.parfm = list()
  z = list()
  Lambda = list()
  for(f in frailties){
    fit.parfm[[f]] = parfm(Surv(exposure,numclaims.surv) ~ Age.p + Age.v +
                             mileage.v, cluster = "pid",
                           dist= "exponential", frailty=f,
                           data=data_surv,
                           showtime=T, method="nlminb")
    if(f=="lognorm"){
      z.lognorm = predict(fit.parfm[[f]])
      z[[f]] = z.lognorm/exp(fit.parfm[[f]][1,1]/2)
      Lambda.lognorm = attr(fit.parfm[[f]],"cumhaz")
      Lambda[[f]] = Lambda.lognorm*exp(fit.parfm[[f]][1,1]/2)
    }
    else{
      z[[f]] = predict(fit.parfm[[f]])
      Lambda[[f]] = attr(fit.parfm[[f]],"cumhaz")
    }
  }
  results = getResults(data,z,Lambda,frailty)
  return(list(results=results))
}

getResults <- function(data,z,Lambda,frailty){
  res = data %>%
      group_by(pid)%>%
      summarise(tot_claims=sum(numclaims))%>%
      bind_cols(z.lognorm=as.matrix(z$lognorm),
                z.gamma=as.matrix(z$gamma),
                z.ingau=as.matrix(z$ingau),
                Lambda.lognorm=Lambda$lognorm,
                Lambda.gamma=Lambda$gamma,
                Lambda.ingau=Lambda$ingau)%>%
      mutate(claims.lognorm = z.lognorm*Lambda.lognorm,
             claims.gamma = z.gamma*Lambda.gamma,
             claims.ingau = z.ingau*Lambda.ingau)
  if(frailty=="lognorm"){
    results = res %>%
        mutate(dev.Lambda.gamma = Lambda.gamma-Lambda.lognorm,
               dev.Lambda.ingau = Lambda.ingau-Lambda.lognorm,
               dev.z.gamma = z.gamma-z.lognorm,
               dev.z.ingau = z.ingau-z.lognorm,
               dev.claims.gamma = claims.gamma-claims.lognorm,
               dev.claims.ingau = claims.ingau-claims.lognorm)
  }
  else if(frailty=="gamma"){
    results = res %>%
      mutate(dev.Lambda.lognorm = Lambda.lognorm-Lambda.gamma,
             dev.Lambda.ingau = Lambda.ingau-Lambda.gamma,
             dev.z.lognorm = z.lognorm-z.gamma,
             dev.z.ingau = z.ingau-z.gamma,
             dev.claims.lognorm = claims.lognorm-claims.gamma,
             dev.claims.ingau = claims.ingau-claims.gamma)
  }
  else if(frailty=="ingau"){
    results = res %>%
      mutate(dev.Lambda.lognorm = Lambda.lognorm-Lambda.ingau,
             dev.Lambda.gamma = Lambda.gamma-Lambda.ingau,
             dev.z.lognorm = z.lognorm-z.ingau,
             dev.z.gamma = z.gamma-z.ingau,
             dev.claims.lognorm = claims.lognorm-claims.ingau,
             dev.claims.gamma = claims.gamma-claims.ingau)
  }
  return(results)
})
```