ERASMUS UNIVERSITEIT ROTTERDAM

Deloitte.

Thesis MSc Econometrics and Management Science

Integrated modelling for non-life insurance using individual claim reserving

Author:
Robin Slakhorst (453775)

Supervisors:
dr. (Chen) C. Zhou (Erasmus University)
dr. (Erik) H.J.W.G. Kole (Erasmus University)
Yoeri Arnoldus (Deloitte)

October 22, 2018

Abstract

In this paper, we introduce an individual claim reserving model that uses policy characteristics and information on the claim process to estimate the reserves of the Incurred But Not Reported and Reported But Not Settled claims of a non-life insurer. First, we analyse the optimal model specification for the claim process. We use mixture distributions to model the level of the payments and introduce zero adjusted distributions to model claim payments of zero. Second, we study the effect of creating dependence between parts of our model on the respective fit. In addition, we examine whether including policy characteristics increases our model fit. We find that the inclusion of both policy characteristics and claim process characteristics improve the fit of our individual claim reserving model. Based on simulations, we examine the effect of incorporating covariates in the model on the prediction of IBNR and RBNS reserves. We conclude that the inclusion of covariates increases the accuracy of our individual claim reserving model. Our model can be used by insurance companies to coherently estimate their reserves.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Insurers face unknown future cash flows. As such, one of the most important tasks of a non-life insurer is managing its reserves. An insurer reserves the amount that is expected to be paid out to policyholders based on estimates of future claims. The frequency and severity of the insurance claims are not known with certainty. Moreover, the frequency and the value of expected future claims can differ for policies with different characteristics.

The widely accepted approaches for modelling insurance claims in non-life insurance make use of aggregated data. The models aggregate for each year the amount that is paid out due to claims in a specific year. An important drawback of such aggregated models is that the differences between the individual policies are lost due to the aggregation. This drawback can be overcome by using an individual claim reserving model. Individual claim reserving models aim to model the occurrence of a claim and its process to settlement on an individual level. By incorporating policy characteristics in the individual claim reserving model, the insurer can adjust the claim process per policy characteristic. Furthermore, an individual claim reserving model allows the insurer to model the claim process in parts. In such a way, the insurer can analyze each part separately. This gives insight into the dynamics of the claim process and the accuracy of the model for the different parts. Finally, with the individual claim reserving model, the insurer can allow for dependence across different parts of the model, such as the time of occurrence and the level and timing of the payments. With all the information of the individual policy characteristics and the different parts of the claim process, the estimation of the reserves of a non-life insurer could be improved.

To summarize, our main research question is: Does a structural model based on individual claim reserving techniques improve the estimation of reserves for a non-life insurer?

To answer this question, we first clarify the different stages in the claim process. First, the policyholder buys an insurance. In this stage, the future claims are unknown. The future claims are estimated in order to determine the premium applicable for that policyholder. Thereafter, the claim occurs and the policyholder reports the claim to the insurer. The time between the occurrence time and the time of reporting is defined as the reporting delay. If the claim is reported to the insurer, the claim is processed and there might be several payments before the claim is finally settled. The information of the insurer about the state of the claim differs throughout this process. To distinguish between claims based on this information, we identify the claims as being Incurred But Not Reported (IBNR) and Reported But Not Settled (RBNS), following the literature on claim reserve modelling. Additionally, we introduce an additional category for the claims in the first stage of the claim process. We name the claims that have been estimated in the premium calculation but have not yet incurred the Not Incurred Not Reported (NINR) claims. This terminology allows us to connect the premium reserve to the claim reserves. Namely, the unearned premium reserve is the reserve that an insurer holds based on the premium of a policyholder. This reserve is based on modelling NINR claims using the policy characteristics. In this paper, we estimate the IBNR and RBNS reserves based on policy characteristics as well. Therefore, estimating the reserves for both the NINR claims and the other two types of claims at an individual level can lead to more coherency within the insurance company, as all reserves are now based on the same characteristics and calculated with the same method. We focus on estimating the reserves from the claim process, which are the IBNR and RBNS reserves. However, the model used in this paper can be used to estimate the NINR reserve and links the NINR reserve to the IBNR and RBNS reserves, since every type of reserves is now estimated on an individual level.

We model the claim process in five parts. First, we model the occurrence of the claims by a Poisson Process. We will model the intensity of the Poisson Process to be dependent on covariates. In such a way, we incorporate policy characteristics and external risk characteristics into the process of claim occurrences. Second, we model the distribution of the reporting delay. Hesselager and Witting (1988) show that the number of claims is negatively correlated between early and late policy years. This indicates that the reporting delay of the claims depends on the time of occurrence of the claim.

We take this dependency into account in our model. We define the last part of the claim process as the development process. The development process consists of possible payments and a settlement.We divide the development process into three parts. The third part of our model consists of modelling the occurrence of events in the development process. The fourth part models the type of the events. The fifth and last part of our model is regarding the level of the payments. We will model the distribution of the level of the payments depending on the covariates. In particular, we incorporate policy information as well as information on the claim process as an indicator for the level of the payments.

First, we analyse the optimal model specification for each part of the claim process. Second, we study the effect of creating dependence between parts of our model and the inclusion of covariates on the fit of our model. Lastly, in order to evaluate our model, we perform simulations and compare our findings with the observed data. We examine the value of including covariates to each building block of our model separately.

We find that the time of occurrence of a claim and the reporting delay can be best specified in a piece-wise constant way. The time between events in the development process can be best modelled with a Weibull hazard rate as opposed to a piece-wise constant hazard rate. To capture the different shapes of the distributions of the payments in the development process, a mixture distribution of three Log-Normal distributions is optimal. For modelling the settlement payments, in addition to using a mixture distribution of three Log-Normal distributions, it is optimal to account for the probability of a zero payment at the moment of settlement.

We use the age of the car and the catalog value of the car as policy characteristics. We find that the inclusion of the age of the car as a covariate improves the fit of all parts of the model. Besides, the catalog value of the car is included as a covariate for the reporting delay and the level of the payments. Furthermore, the inclusion of the time of occurrence does not lead to a better fit in any of the other parts of the model. By contrast, we find that incorporating dependence between the other four parts of the model increases the fit of our model.

Lastly, we find that the mean of the simulations using our model with covariates is closer to the actual data than the simulations results using our model without covariates. We conclude that incorporating covariates increases the ability of our individual claim reserving model to accurately predict the IBNR and RBNS reserves.

Chapter 2

Literature Review

In this paper, we focus on an individual claim reserving model which differs from the literature on aggregated models. Within the current literature on individual claim reserving our model is positioned as follows. Our main framework is an extension of the Marked Poisson Process of Arjas (1989), which we use for modelling the occurrence of claims and the reporting delay. Furthermore, we use the framework for the payment process of Norberg (1993). In this section, we first discuss the literature on aggregated models, after which we will discuss the current literature on individual claim reserving and how this relates to our model.

England and Verrall (2002) gives an overview of different aggregated models. The models have several drawbacks, as discussed by Taylor et al. (2008). They argue that a large part of relevant information that can be found in the data of individual claims is lost by using aggregate data. Moreover, aggregating data is only valid for a portfolio that is homogeneous with respect to risk characteristics, as shown by Norberg and Sundt (1985). For a portfolio of insurances, this is a strong assumption. Furthermore, a change in the composition of the portfolio causes a change in the claims process, which is not captured by the aggregated models (Norberg, 1993). The use of run-off triangles has several other problems, such as the possibility of negative or zero-value cells (Kunkler, 2004), problems with robustness of the model and outliers in the data (Verdonck et al., 2009) and over-parametrization of the model due to lower amount of observations for recent accident years (Wright, 1990) and (Renshaw, 1994).

The above mentioned drawbacks can be overcome by using individual claim reserving methods. The individual claim reserving model was first introduced by Norberg (1986), Jewell (1989) and Arjas (1989). Arjas (1989) introduces the concept by using theory on point processes and martingales to model IBNR claims. They model the claims process as a Marked Point Process, which is a process that models points in time together with associated marks. As an example, in the case of individual claim reserving, these points in time can be seen as the occurrence times of claims, whereas the marks can be thought of as the development of those claims. Norberg (1993) uses this model specification whereby the points and their marks are represented by the occurrence times and their development, respectively. Furthermore, they propose that a Marked Poisson Process can be defined for groups with different observable risk characteristics. Compared with Norberg (1993), we implement the idea of using covariates. Instead of specifying a different model for every group, we make each part of the process dependent on covariates and allow for dependence across the different parts of the process. Furthermore, Norberg (1993) discusses the payment process. We adapt their framework of the payment process and adjust it to account for different events. Moreover, we compare the fit of different distributions for modelling the level of the payments and use zero adjusted distributions to account for the probability of a payment of zero.

In a follow-up paper, Norberg (1999) extends their work by discussing how the model proposed by Norberg (1993) can be applied in various situations. Haastrup and Arias (1997) introduces a model that is similar to that of Norberg (1993). They adjust it by using a non-parametric Bayesian approach for the estimation and prediction of claims, both for IBNR and RBNS claims. They argue that some parts of the model could be better estimated parametrically, such that structure is added and the computations are less time-consuming. For model estimation, the use of Bayesian methods has already been suggested by Arjas (1989), who states that the need for the conditional distribution of the occurrence process (reporting times) and the claim sizes on the available information at time t strongly support the use of Bayesian statistics. We compare the use of non-parametric and parametric approaches, such that we are able to allow for flexibility to some parts of the model and use parametric distributions that are known to perform well in the other parts.

In recent years, multiple papers have been written that extend the literature on individual claim reserving, for example by adding dependence between different types of claims by means of a copula (Maciak et al. (2018)), or by using the Marked Point Process framework for performing a case-study (Larsen (2007), and Plat and Antonio (2014)). In our model, we do not consider dependence between different types of claims. Alternatively, we add dependence between different parts of our model. For some parts of our model, we use survival theory, following Zhao et al. (2009) and Zhao and Zhou (2010).

Compared to the existing models for individual claim reserving, this paper extends these models by introducing dependence across different components in the model. Moreover, in previous literature, either a non-parametric or a parametric approach has been used for model estimation. We investigate in which part of the model a parametric or non-parametric approach fits better. Eventually, we use the appropriate methods in each part of the model. Third, we compare multiple ways of modelling the payment process by means of mixture distributions and introduce zero adjusted distributions to model claim payments of zero. Lastly, we introduce the link between the unearned premium reserve and the claim reserves, such that insurers are able to use our model as one coherent framework to estimate these reserves.

Chapter 3

Claim Reserving

In this chapter, we discuss the development of a claim from the moment it occurs until the moment at which the insurer handled the claim. Moreover, we discuss the types of reserves an insurer holds according to the different stages of a claim.

3.1 Claim Process

In this section, we discuss the process of a claim. We first introduce the characteristics of a claim. Then, we discuss how the characteristics of a claim fit into our model and how they are related. Every claim i consists of

(i) an occurrence time T_i: the moment at which the event that leads to the claim takes place;

(ii) a reporting delay U_i: the time between occurrence of the claim and the time of reporting;

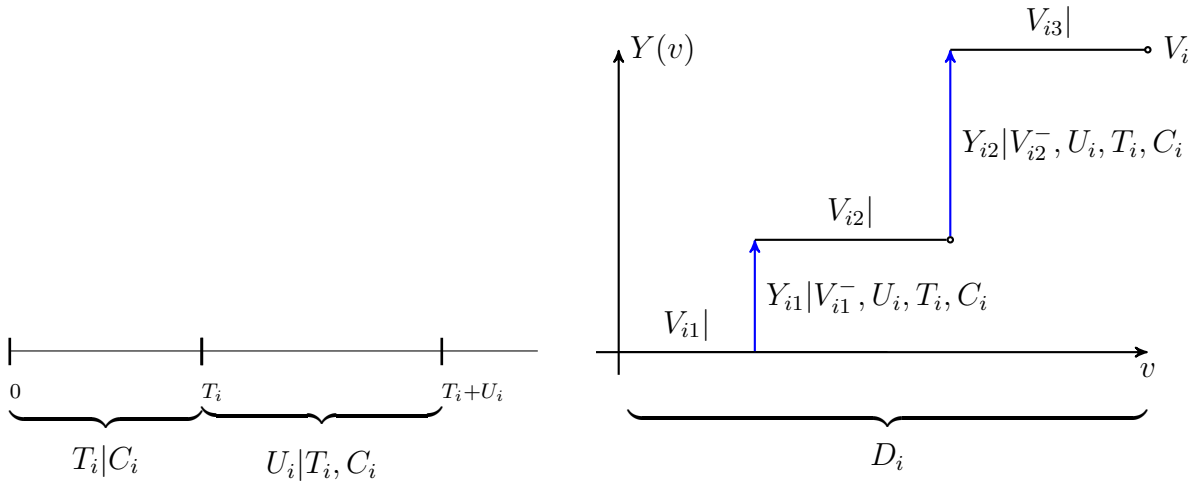
(iii) a development process D_i: the collection of events j, $j = 1, 2, \dots N_i$ with the total number of events for claim i given by N_i, between the moment of reporting and settlement, which is denoted as $D_i = \{(V_{ij}, K_{ij}, Y_{ij}) : j = 1, 2, \dots N_i\}$, where we define V_{ij} as the time between the $(j-1)$-th and the j-th event, K_{ij} as the type of event j, with $K_{ij} = 1$ indicating a payment and $K_{ij} = 0$ indicating a settlement, and where we define the payment process as

$$Y_i(v) = \begin{cases} Y_{ij} & \text{for } v = V_{i1} + \cdots + V_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $v \in [0, V_i]$ indicates the place in time in the development process and $V_i = \sum_{j=1}^{N_i} V_{ij}$ is the time it takes for claim i to settle from the moment of reporting;

(iv) and covariates C_i: the set of n policy characteristics known at the start of the policy, $C_i = (C_{i1}, C_{i2}, \ldots C_{in})$.

Figure 3.1 shows the process of a claim from occurrence until settlement. We will explain every part of the figure hereafter.



This figure displays the process of a claim from the start of the policy until the settlement of the claim. Note that $V_{ij}|$ indicates $V_{ij}|V_{ij}^-, U_i, T_i, C_i$.

Figure 3.1: A visual representation of the claim process

The first part of our model consists of modelling the occurrence times T_i, $i \geq 1$. We model the time at which the claim occurs as depending on characteristics of the policyholders C_i, i.e. $T_i|C_i$. We incorporate these characteristics into the claim occurrence process to account for groups with different characteristics. For example, we could argue that a man who has just obtained his drivers license is expected to be more likely to have an accident and to make a claim than someone who has had his drivers license for over ten years.

After the claim has occurred, there is a reporting delay before the claim is notified to the insurer, which we denote by U_i. We model the reporting delay as depending on the covariates C_i and the time of occurrence T_i, i.e. $U_i|T_i, C_i$. We do this, since there could be a difference in the time it takes for someone to report a claim between, for example, women and men.

The development process, D_i, is the third part of the claim process and starts when the claim is reported. As discussed before, the development process consists of different events. Each claim can have multiple payments during its development process. The claim process ends when the claim is settled, which can happen either with or without an associated payment.

We model the development process in two steps. First, we model the time until an event and the probability of the event being a payment or a settlement. The second step is determining the level of the payment. The level of the payment accompanying a settlement can take a value of zero, indicating that there was no payment. The timing of the payments and the levels of the payments could be different for claims with different characteristics. Moreover, the reporting delay and the time of occurrence of the claim could have an effect on the development process. For example, when it takes a while for the claim to be reported, it could indicate that it is a complicated claim. This could lead to larger gap times or higher payments due to the complexity of the claim. Therefore, we take the time of occurrence T_i, the reporting delay U_i and the covariates C_i associated with the claim into account for the development process part of our model. Furthermore, for the moment in time of the j-th event in the development process of claim i, we take into account the development process preceding the j-th event. The reason for this is that the number of previous payments or the total amount paid could have an impact on the timing and level of the payments. We denote the history of the development process of claim i up to the j-th event as V_{ij}^-, with $V_{ij}^- = (j-1, \sum_{b=1}^{j-1} V_{ib}, \sum_{b=1}^{j-1} Y_{ib})$. So, the history of the development period that we include consists of the number of previous payments $(j-1)$, the time the claim has been in the development process, $\sum_{b=1}^{j-1} V_{ib}$, and the total payout up to the j-th event $\sum_{b=1}^{j-1} Y_{ib}$. Information on the type of prior events K_{ij} is not included in the history of the development process, as they are always a payment. This is true since if the type of the previous event would have been a settlement, the development period would have ended.

Our model thus consists of five parts. We model

(i) time of occurrence $T_i|C_i$,

(ii) reporting delay $U_i|T_i, C_i$,

(iii) the time between events $V_{ij}|V_{ij}^-, U_i, T_i, C_i$,

(iv) the type of the events $K_{ij}|V_{ij}^-, U_i, T_i, C_i$,

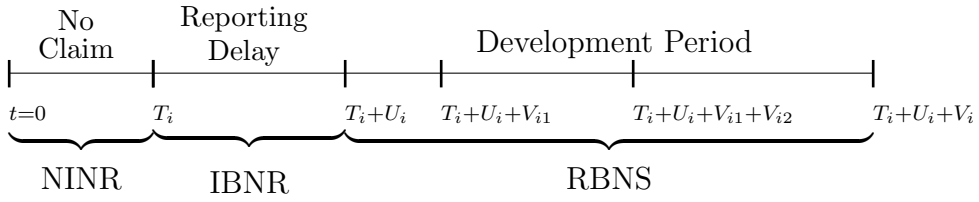
(v) and the level of the payments $Y_{ij}|V_{ij}^-, U_i, T_i, C_i$.

} in D_i

3.2 Reserves

The aim of our model is to accurately predict the reserves that are needed to account for the claims of an insurer. We have three types of claims for reserve modelling, namely NINR, IBNR and RBNS claims. In the introduction of this paper, we discussed that the reserve based on NINR claims is given by the unearned premium reserve. The unearned premium reserve is determined by the premium and is therefore based on the individual policy characteristics. The estimation of the NINR claim reserve is outside the scope of this paper. However, it is notable that this study provides for the link between the reserve based on NINR claims and the reserves based on IBNR and RBNS claims, since the latter are now also based on the individual claims. Our framework can therefore be used to estimate the NINR reserves. In this section, we examine the differences between NINR, IBNR and RBNS claims.

At time τ, the time for which the reserves have to be estimated, we can distinguish between NINR, IBNR and RBNS claims. Figure 3.2 illustrates this distinction. If the claim has not yet occurred, i.e. $\tau < T_i$, it is a NINR claim. If the claim has occurred, but it has not yet been reported, i.e., $T_i < \tau < T_i + U_i$, it is an IBNR claim. In the case of a RBNS claim, the claim has been reported and there has not been a settlement yet, i.e. $T_i + U_i < \tau < T_i + U_i + V_i$.



This figure illustrates the way in which the different types of claims are defined.

Figure 3.2: Types of claim reserves

Let $B_\tau = \{i|T_i < \tau < T_i + U_i\}$ and $S_\tau = \{i|T_i + U_i < \tau < T_i + U_i + V_i\}$ be the sets of indices corresponding to the IBNR and RBNS claims, respectively. Moreover, let $P_{i,\tau} = \{k|\tau < T_i + U_i + \sum_{j=1}^{k} V_{ij} < T_i + U_i + V_i\}$ be the set of indices corresponding to events in the development process from moment τ until the end of the development process for claim i. The total reserves of IBNR claims, X_{IBNR}, and RBNS claims, X_{RBNS}, at time τ are given by

$$X_{IBNR}(\tau) = \sum_{i \in B_\tau} \sum_{j \in P_{i,\tau}} E(Y_{ij}), \qquad X_{RBNS}(\tau) = \sum_{i \in S_\tau} \sum_{j \in P_{i,\tau}} E(Y_{ij}). \tag{3.2}$$

Chapter 4

Methodology

In this chapter, we introduce the methodology used to estimate our model. This chapter consists of three sections. First, we discuss the building blocks that are used to estimate the different parts of the claim process. In the section thereafter, we discuss the simulation procedure used to simulate the reserves. In the last section, we discuss the criteria that we use to choose an optimal model.

4.1 Building Blocks

In this section, we discuss the different building blocks of our model. As discussed before, our model consists of four building blocks. In Section 4.1.1, we discuss the model for the occurrence times of the claims conditional on the covariates, $T_i|C_i$. Thereafter, in Section 4.1.2, we introduce the model of the reporting delay of the claims conditional on the covariates and the time of occurrence of the claim, $U_i|T_i, C_i$. The third part of our model is modelling the event times in the payment process and the probability that an event is a payment or a settlement, which we discuss in Section 4.1.3. We introduce the model for our final building block, the level of the payments, in Section 4.1.4.

In each part of this section, we discuss the distribution of the building blocks. The building blocks of our model are conditionally independent. As a result, we are able to estimate the four parts separately. Therefore, in each part of this section, we discuss the likelihood of one specific building block and how we optimize the likelihood to find the right estimates.

4.1.1 Rate of Occurrence

The first part of our model is the occurrence of claims. The total number of claim occurrences $N(t)$ in time interval $[0, t)$, $t \in [0, \infty)$, follows a Poisson distribution. The intensity of the Poisson Process is modelled by a function $\lambda(t)$. In the case of a non-constant intensity, we call the process $\{N(t) : t \in [0, \infty)\}$ an inhomogeneous Poisson Process.

The number of occurrences in the time interval $[0, t)$ for an inhomogeneous Poisson Process follows a Poisson distribution as

$$N(t) \sim \text{Poisson}\bigg(\int_0^t \lambda(u)du\bigg), \qquad \text{where } N(0) = 0. \tag{4.1}$$

To estimate the intensity in our model, we need to take into account two different aspects. First, we incorporate information on the car insurance policies into the occurrence intensity. We consider intensity $\lambda(t|C_i)$, where C_i indicates the policy characteristics of claim i. The intensity is given by

$$\lambda(t|C_i) = \lambda_0(t) \exp(x_i'\beta) \tag{4.2}$$

where $x_i = C_i = (C_{i1}, \dots C_{in})$ are the covariates included in the model and β is a vector of length n indicating the effect of the covariates on the intensity. The density function of the occurrence time of a claim, following Cook and Lawless (2007), is given by

$$f_T(t|C_i) = \lambda_0(t) \exp(x_i'\beta) \exp\bigg(-\int_0^t \lambda_0(s) \exp(x_i'\beta)ds\bigg). \tag{4.3}$$

Second, we account for the number of policies that are in the portfolio at time t. Instead of solely using the policies that experienced a claim, we take all policies into account when estimating the rate of occurrence. We do this for two reasons. First, an insurer always takes the total number of active policies in his portfolio into consideration in order to get an overview of the total risk. Therefore, we consider it useful to use all available information to estimate the number of occurrences. Second, using this approach to estimate the rate of occurrence makes it possible to estimate the number of occurrences for NINR claims by using the estimated rate of occurrence and applying it to a specific (future) time period. Earlier we discussed that we do not estimate the reserves for the NINR claims in this paper. However, insurers now have the possibility to use this framework for estimating all three types of reserves.

An insurer has multiple policies in its portfolio, indexed by $p = 1, 2, \dots P$, such that P is the total number of policies of the insurer. To account for the number of active policies in the portfolio of an insurer, we define $I_p(t) = I(\text{Start date policy } p \leq t \leq \tau)$ as an indication of whether policy p was active at time t. As a result, $\int_0^\tau I_p(s)ds$ indicates the total amount of time that policy p was active from 0 to τ, which we define as the exposure of portfolio p from time 0 to τ. The total exposure of the insurer in the period $[0, \tau)$ is defined as $\sum_{p=1}^{P} \int_0^\tau I_p(s)ds$. For example, if we take $\tau = 30$, then if policy p was active for 20 days between $[0, 30)$, $\int_0^\tau I_p(s)ds$ is given by 20.

The likelihood of the occurrence process for the observed claims with intensity $\lambda(t|C_i)$ is given by

$$\begin{aligned} L(\lambda_0, \beta) = &\prod_{i \geq 1} \left(\lambda_0(t_i) \exp(x_i' \beta) \right) \\ &\times \prod_{p \geq 1} \exp\left(- \int_0^\tau I_p(s) \lambda_0(s) \exp(x_p' \beta) ds \right) \Bigg), \end{aligned} \tag{4.4}$$

where t_i is the time of occurrence of the i-th claim, and $x_i = C_i$ and $x_p = C_p$ are the vectors of covariates of the i-th claim and the p-th policy, respectively. The two terms are similar to (4.3), where the second term is corrected for the time that the policies were active. The derivation of the likelihood is given in the Appendix, Section 8.2.

In our model, we estimate $\lambda_0(t)$ in a piece-wise constant way. Hence, we model λ_0 as

$$\lambda_0(t) = \sum_{l \geq 1} I_{(a_{l-1} \leq t < a_l)} \lambda_{l,0}, \tag{4.5}$$

where $[a_{l-1}, a_l)$ for each $l = 1, 2, \dots$ indicates a monthly interval. This means that we define an occurrence rate for each month such that we are able to capture differences between months. Now, we adjust (4.4) to construct the likelihood of the claim occurrence process with a piece-wise constant intensity.

$$\begin{aligned} L(\lambda_{1,0}, \dots \lambda_{l,0}, \beta) = &\prod_{l \geq 0} \Bigg(\Bigg(\prod_{i \geq 1} \lambda_{l,0} \exp(x_i' \beta)^{I_{(a_{l-1} < t_i < a_l)}} \Bigg) \\ &\times \Bigg(\prod_{p \geq 1} \exp - \Bigg(\lambda_{l,0} \exp(x_p' \beta) \int_{a_{l-1}}^{a_l} I_p(s) ds \Bigg) \Bigg) \Bigg), \end{aligned} \tag{4.6}$$

where everything is as in (4.4) and $I_{(a_{l-1} \leq t_i < a_l)}$ indicates whether claim i occurred between a_{l-1} and a_l. $\int_{a_{l-1}}^{a_l} I_p(s)ds$ is similar as before, indicating the exposure of portfolio p in $[a_{l-1}, a_l)$.

We optimize the likelihood given in (4.6) with respect to $\lambda_{l,0}$ by taking

$$\hat{\lambda}_{l,0}(\beta) = \frac{\sum_{i\geq 1} I_{(a_{l-1}\leq t_i<a_l)}}{\sum_{p\geq 1}\exp(x_p'\beta)\int_{a_{l-1}}^{a_l} I_p(s)ds} \quad \text{for each } l = 1, 2, \ldots \tag{4.7}$$

We learn from this equation that the maximum likelihood estimator for $\lambda_{l,0}$ is the total number of occurrences from a_{l-1} to a_l divided by the exposure. Furthermore, we need to find estimates for β. We find the MLE for β by solving (4.8) for β. Then, we use it in the definition of $\hat{\lambda}_{l,0}(\beta)$ as in (4.7) to find $\hat{\lambda}_{l,0}$.

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i\geq 1}\left(x_i' - \frac{\sum_{l\geq 0} I_{(a_{l-1}\leq t_i<a_l)}\sum_{p\geq 1}\exp(x_p'\beta)x_p'\int_{a_{l-1}}^{a_l} I_p(s)ds}{\sum_{l\geq 0} I_{(a_{l-1}\leq t_i<a_l)}\sum_{p\geq 1}\exp(x_p'\beta)\int_{a_{l-1}}^{a_l} I_p(s)ds}\right) = 0. \tag{4.8}$$

The proof of the optimization is given in the Appendix, Section 8.3.1.

4.1.2 Reporting Delay

The second part of our model concerns the reporting delay. The reporting delay of a claim is defined as the time between the time of occurrence and the moment of reporting. The reporting of a claim occurs only once in a claim process. Therefore, we model the delays using survival theory.

The distribution of the reporting delay, $F_U(u|T_i, C_i)$, is characterized by a hazard rate $\gamma_U(u|T_i, C_i)$, which is defined as the instantaneous rate of occurrence of the event at time t. The hazard rate γ_U is formally defined as $\gamma_U = f_U(u)/(1 - F_U(u))$, where $F_U(u) = 1 - \exp(-\int_0^u \gamma_U(t)dt)$. Hence, we obtain the density by multiplying the hazard rate with one minus the distribution function. The distribution of the reporting delay is dependent on the time of occurrence of the claim T_i and the covariates C_i. To be able to incorporate the covariates in the hazard rate of the reporting delay, we write it as follows:

$$\gamma_U(u|T_i, C_i) = \gamma_0(u)\exp(x_i'\beta). \tag{4.9}$$

Here, $\gamma_0(u)$ is known as the baseline hazard rate, $x_i = (T_i, C_i)$, and β is a vector of length $n+1$ indicating the effect of the covariates on the hazard rate. The baseline hazard rate can be specified either in a parametric or in a non-parametric way. We use and compare the fit of the different approaches in our model. First, we discuss the piece-wise constant hazard rate, after which we discuss the parametric hazard rate.

Piece-wise Constant

A piece-wise constant baseline hazard rate can be seen as a non-parametric model. We consider a baseline hazard rate that is fixed for small intervals and define it by

$$\gamma_0(u) = \sum_{w \geq 1} 1_{(q_{w-1} \leq u < q_w)} \gamma_{w,0}. \tag{4.10}$$

For each w, $1 \leq w \leq W$, $[q_{w-1}, q_w)$ is the interval on which the hazard rates are assumed to be constant with W indicating the total number of intervals. We incorporate covariates in the hazard specification as in (4.9). The likelihood is given by

$$L(\gamma_{1,0}, \dots \gamma_{w,0}, \beta) = \prod_{w \geq 1} \left(\prod_{i \geq 1} (\gamma_{w,0} \exp(x_i' \beta))^{I_{(q_{w-1} \leq u_i < q_w)}} \exp\left(- \gamma_{w,0} \exp(x_i' \beta) \int_{q_{w-1}}^{q_w} I_i(s) ds \right) \right) \tag{4.11}$$

We can optimize the likelihood in (4.11) with respect to $\gamma_{w,0}$ given β by taking

$$\hat{\gamma}_{w,0}(\beta) = \frac{\sum_{i \geq 1} I_{(q_{w-1} \leq u_i < q_w)}}{\sum_{i \geq 1} \exp(x_i' \beta) \int_{q_{w-1}}^{q_w} I_i(s) ds} \quad \text{for each } w = 1, 2, \dots \tag{4.12}$$

We get the MLE's for β and $\gamma_{w,0}$ by solving (4.13) for β to obtain $\hat{\beta}$ and use this in (4.12) to obtain $\hat{\gamma}_{w,0}$.

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i \geq 1} \left(x_i' - \frac{\sum_{w \geq 1} I_{(q_{w-1} \leq u_i < q_w)} \sum_{i' \geq 1} \exp(x_{i'}' \beta) x_{i'}' \int_{q_{w-1}}^{q_w} I_{i'}(s) ds}{\sum_{w \geq 1} I_{(q_{w-1} \leq u_i < q_w)} \sum_{i' \geq 1} \exp(x_{i'}' \beta) \int_{q_{w-1}}^{q_w} I_{i'}(s) ds} \right) = 0. \tag{4.13}$$

The proof of the optimization is given in the Appendix, Section 8.3.2.

Weibull

In a parametric approach, we fit the distribution of u_i of a Weibull distribution. The specification of the Weibull distribution with covariates can be seen as both an accelerated failure time model, which assumes that the covariates accelerates or decelerates the time by some constant, and a proportional hazards model, which assumes that covariates have a multiplicative effect on the hazard. This is useful for our analysis, since now we do not have to make the distinction between those two types of models. Therefore, we model the reporting delay with a Weibull distribution. The Weibull baseline hazard rate is given by

$$\gamma_0(u) = \rho \zeta (\zeta u)^{\rho - 1}, \rho \geq 0, \zeta \geq 0 \tag{4.14}$$

where ζ and p are the parameters that should be estimated. By combining (4.9) and (4.14), we obtain the hazard rate of the waiting time

$$\gamma(u|T_i, C_i) = \rho\zeta(\zeta u)^{\rho-1} \exp(x_i'\beta) \tag{4.15}$$

$$= \rho\zeta^*(\zeta^* u)^{\rho-1}, \tag{4.16}$$

where $\zeta^* = \zeta \exp(x_i'\beta/\rho)$, with $x_i = (T_i, C_i)$, and β is a vector of length $n+1$ indicating the effect of the covariates on the hazard rate. The density of the reporting delay with a Weibull hazard rate is given by

$$f_U(u|T_i, C_i) = \rho\zeta^*(\zeta^* u)^{\rho-1} \exp\bigg(- (\zeta^* u)^{\rho}\bigg). \tag{4.17}$$

The likelihood is given by

$$L(\rho, \zeta, \beta) = \prod_{i \geq 1} \rho\zeta(\zeta u_i)^{\rho-1} \exp(x_i'\beta) \exp\bigg(- (\zeta u_i)^{\rho} \exp(x_i'\beta)\bigg). \tag{4.18}$$

The likelihood is optimized by a Newton Raphson algorithm. First, the Weibull distribution without covariates is fitted to the data. The ρ and ζ obtained from this optimization are iteratively updated by the Newton Raphson method. The algorithm stops when the likelihood has converged. The optimal ρ and ζ are stored. Then, we perform a linear regression of the u_i's on the covariates to find the starting values for β. The values of β, ρ and ζ are used as the starting values for the next Newton Raphson algorithm. After the algorithm converges, we find our optimal estimates for ρ, ζ and β. The optimization algorithm can be found in the Appendix, Section 8.3.4.

4.1.3 Event Rate and Event Probability

The third part of our model regards the events in the development process. First, we discuss how we model the event rate in the development period. Thereafter, we discuss how we model the probability of an event.

Event Rate

To model the events in the development process, we use theory on recurrent events. As discussed in Section 3, we denote the gap time between the $(j-1)$-th and the j-th event as V_{ij}. Theory on recurrent events assumes independence between the gap times of the events. In our model, this is a strong assumption, since the events correspond to the same claim with the same characteristics.

In other words, the claim process does not restart after a payment. Therefore, the gaps between the events can not be seen as independent, since they are influenced by the same claim and its characteristics. One possible solution to this issue is to add covariates of the claim and assume that the gaps are conditionally independent when taking into account those covariates. However, conditional on the covariates of the claims, the gaps between payments could still be dependent, since the payment structure can not be fully attributed to those covariates.

Another option is to include information on the prior events in the hazard rate specification. We use this method for controlling for the dependency between the gaps for two reasons. First, all the information that is needed is available, such that we can fully control for the dependency between the gaps. Furthermore, we want to include this information to study the relation between the events in a development process. We denote the information on the development process of a claim prior to an event with V_{ij}^-. The history of the development period that we include consists of the number of previous payments, total payout and the time the claim has been in the development process. In addition, we include the reporting delay U_i, time of occurrence T_i and the covariates C_i in the hazard rate specification.

We denote the hazard rate as $\phi(v|V_{ij}^-, U_i, T_i, C_i)$ and model it as

$$\phi(v|V_{ij}^-, U_i, T_i, C_i) = \phi_0(v) \exp(x_{ij}'\beta). \tag{4.19}$$

with $x_{ij} = (V_{ij}^-, U_i, T_i, C_i)$ and β is a vector of length $n + 5$ indicating the effect of the covariates on the hazard rate. The vector x_{ij} differs for the different gap times between events, but is by definition constant within a gap time.

The baseline hazard rate of the events in the payment process, ϕ_0, is modelled both parametrically and non-parametrically. As for the reporting delay, we consider two types of specifications for the hazard rate; the non-parametric approach and the parametric approach using the Weibull distribution.

The likelihoods of the non-parametric and parametric approach are similar to those in (4.11) and (4.18). The main difference is that some of the observations for v_{ij} are censored. This means that we have two types of observations: one where v_{ij} is the time between two events and one where v_{ij} is the time between the last event and τ, the end of the observation period. Therefore, we introduce the variable $\delta_{ij} = 1 - I(v_{ij} \text{ is censored})$, such that $\delta_{ij} = 1$ indicates a non-censored observation, whereas $\delta_{ij} = 0$ indicates a censored observation.

The likelihood of the non-parametric, i.e. the piece-wise constant, approach is given by

$$L(\phi_{1,0},\dots\phi_{z,0},\beta)=\prod_{z\geq 1}\prod_{i\geq 1}\left(\prod_{j\geq 1}(\phi_{z,0}\exp(x'_{ij}\beta))^{I_{(r_{z-1}\leq v_{ij}<r_z)}\delta_{ij}}\right. \tag{4.20}$$

$$\left.\times\exp\left(-\phi_{z,0}\exp(x'_{ij}\beta)\int_{r_{z-1}}^{r_z}I_i(s)ds\right)\right), \tag{4.21}$$

where $\int_{r_{z-1}}^{r_z}I_i(s)ds$ indicates the exposure of claim i in the development process between r_{z-1} and r_z, $x_{ij}=(V_{ij}^-,U_i,T_i,C_i)$, and β indicates the effect of the covariates on the hazard rate. Furthermore, for each z, $1\leq z\leq Z$, $[r_{z-1},r_z)$ is the interval on which the hazard rates are assumed to be constant, with Z indicating the total number of intervals. The likelihood in (4.21) can be optimized with respect to to $\phi_{z,0}$ given β by taking

$$\hat{\phi}_{z,0}(\beta)=\frac{\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij}I_{(r_{z-1}\leq v_{ij}<r_z)}}{\sum_{i\geq 1}\exp(x'_{ij}\beta)\int_{r_{z-1}}^{r_z}I_i(s)ds}\quad\text{for each } z=1,2,\dots \tag{4.22}$$

We get the MLE's for β and $\phi_{z,0}$ by solving (4.23) for β to obtain $\hat{\beta}$ and using this in (4.22) to obtain $\hat{\phi}_{z,0}$.

$$\frac{\partial\log L(\beta)}{\partial\beta}=\sum_{i\geq 1}\sum_{j\geq 1}\left(\delta_{ij}x'_{ij}-\frac{\sum_{z\geq 1}I_{(r_{z-1}\leq v_{ij}<r_z)}\sum_{i'\geq 1}\exp(x'_{i'j'}\beta)x'_{i'j'}\int_{r_{z-1}}^{r_z}I_{i'}(s)ds}{\sum_{z\geq 1}I_{(r_{z-1}\leq v_{ij}<r_z)}\sum_{i'\geq 1}\exp(x'_{i'j'}\beta)\int_{r_{z-1}}^{r_z}I_{i'}(s)ds}\right)=0. \tag{4.23}$$

The proof of the optimization is given in Section 8.3.3 of the Appendix.

The likelihood of the parametric approach is given by

$$L(\rho,\zeta,\beta)=\prod_{i\geq 1}\prod_{j\geq 1}\left(\rho\zeta(\zeta v_{ij})^{\rho-1}\exp(x'_{ij}\beta)\right)^{\delta_{ij}}\exp\left(-(\zeta v_{ij})^{\rho}\exp(x'_{ij}\beta)\right), \tag{4.24}$$

where ζ and ρ are parameters of the distribution to be estimated, $x_{ij}=(V_{ij}^-,U_i,T_i,C_i)$, δ_{ij} indicates whether an observation is censored, and β is a vector of length $n+5$ indicating the effect of the covariates on the hazard rate. The likelihood is optimized in the same way as the likelihood in (4.18).

Event Probability

After modelling the occurrence of an event in the development process, we need to determine whether it is a payment or a settlement. Let $\kappa_{ij} \in 0, 1$ indicate the type of event of the j-th event of claim i, with $\kappa_{ij} = 1$ indicating a payment and $\kappa_{ij} = 0$ indicating a settlement. As for the event hazard rates, we take into account the reporting delay U_i, time of occurrence T_i, covariates C_i, and the history of the development process V_{ij}^-. We use a logistic regression to compute the probability of the event being either a payment or a settlement. There are only two types of events, such that we have $P(\kappa = 0|V_{ij}^-, U_i, T_i, C_i) = 1 - P(\kappa = 1|V_{ij}^-, U_i, T_i, C_i)$. The probabilities obtained from the regression are given by

$$P(\kappa = 1|V_{ij}^-, U_i, T_i, C_i) = \frac{\exp(x'_{ij}\beta)}{1 + \exp(x'_{ij}\beta)}, \tag{4.25}$$

$$P(\kappa = 0|V_{ij}^-, U_i, T_i, C_i) = 1 - P(\kappa = 1|V_{ij}^-, U_i, T_i, C_i), \tag{4.26}$$

where $x_{ij} = (1, V_{ij}^-, U_i, T_i, C_i)$ and β is a vector of length $n + 6$ indicating the effect of the covariates on the event probability. Furthermore, we perform the regression without covariates, such that we obtain the probability of a payment without the effect of covariates. This way, we are able to compare the fit with and without covariates. The density of the type of events is given by

$$f_\kappa(\kappa|V_{ij}^-, U_i, T_i, C_i) = P(\kappa = 1|V_{ij}^-, U_i, T_i, C_i)^\kappa (1 - P(\kappa = 1|V_{ij}^-, U_i, T_i, C_i))^{1-\kappa}, \tag{4.27}$$

where $\kappa \in \{0, 1\}$. The likelihood is given by

$$L(\beta) = \prod_{i\geq 1} \prod_{j\geq 1} \left(\frac{\exp(x'_{ij}\beta)}{1 + \exp(x'_{ij}\beta)} \right)^{\kappa_{ij}} \left(\frac{1}{1 + \exp(x'_{ij}\beta)} \right)^{1-\kappa_{ij}}. \tag{4.28}$$

The values for β that maximize (4.28) are found by performing a logistic regression.

4.1.4 Payment Distribution

The last part of our model is modelling the level of the payments. Similarly to the previous parts of the claim process, we use covariates in the payment distribution. As before, the reporting delay U_i, time of occurrence T_i, covariates C_i, and the history of the development process V_{ij}^- are included. We denote the payment after gap time V_{ij} as Y_{ij}. First, we discuss the regular payments. In the second section, we discuss how we model the payments that are associated with settlements.

Payments

To account for the various shapes of our payment distributions, we use mixture distributions. Moreover, we add covariates to the payment distribution. We do this by introducing link functions, which map the parameters of a distribution to the covariates. We combine the use of the mixture distribution and the link function. In this section, we first shortly introduce the concept of mixture distributions, after which we discuss how to make the parameters of the distributions in the mixture distribution covariate dependent.

We start by introducing the mixture distribution. The two distributions that we use for the different combinations of distributions are the Log Normal distribution and the Gamma distribution. The distribution of a mixture distribution consisting of K different distributions is given by

$$F_{Y,MX}(y|\theta) = \sum_{k=1}^{K} \pi_k F_{Y,k}(y|\theta_k) \tag{4.29}$$

where π_k is the weight assigned to the k-th distribution in the mixture distribution, with $\sum_{k=1}^{K} \pi_k = 1$, and θ_k are the parameters of the k-th distribution.

We add covariates to the mixture distribution, i.e. we make the parameters of the density functions $f_{Y,k}$ dependent on x_{ij}, where x_{ij} are the covariates corresponding to the payment Y_{ij}. In our mixture model, we have K distributions, where each of the K distributions has its own parameters. Both of the distributions we use are two parameter distributions. We denote the parameters as $\theta_k = (\mu_k, \sigma_k)$, indicating the first and second parameter of the k-th distribution, respectively.

We use link functions to create the relation between the parameters and the covariates. Link functions are functions that map a parameter of the distribution to a function of the covariates. Let $g_{\mu,k}(\cdot)$ and $g_{\sigma,k}(\cdot)$ be link functions of parameters μ_k and σ_k of the k-th distribution in the mixture distribution. The link function of the distribution is linearly linked to the covariates, which means that the link function of the parameter is equal to product of the covariates and the coefficient, namely

$$g_{\mu,k}(\mu_{ij,k}) = x'_{ij}\beta_{\mu,k}, \tag{4.30}$$

$$g_{\sigma,k}(\sigma_{ij,k}) = x'_{ij}\beta_{\sigma,k}, \tag{4.31}$$

where $x_{ij} = (V_{ij}^{-}, U_i, T_i, C_i)$, the parameters $\mu_{ij,k}$ and $\sigma_{ij,k}$ are the first and second parameter of distribution k for claim i and event j, and $\beta_{\mu,k}$ and $\beta_{\sigma,k}$ are the vectors of length $n+5$ indicating the effect of the covariates on the link functions of $\mu_{ij,k}$ and $\sigma_{ij,k}$, respectively.

In (4.30) and (4.31) we observe that it is possible to specify a separate link function for each parameter and for each distribution. In Table 4.1, we give an overview of the link functions used. The link function for μ in the Log Normal distribution is the identity link function, whereas for the Gamma distribution it is a log link. The link function for σ is the same for both distributions. By inverting the link functions, we obtain the relation between the parameters of the distribution and the covariates, which are given in the last column of Table 4.1.

Table 4.1: Link Functions

This table shows for each parameter which type of link function is used to relate the covariates to that specific parameter. The last column indicates the direct relation between the covariates and the parameters.

	Parameter	Link Function	Inverse of Link Function
Log Normal	μ_k	$g(\mu_{ij,k}) = \mu_{ij,k} = x'_{ij}\beta_{\mu,k}$	$\mu_{ij,k} = x'_{ij}\beta_{\mu,k}$
	σ_k	$g(\sigma_{ij,k}) = \log(\sigma_{ij,k}) = x'_{ij}\beta_{\sigma,k}$	$\sigma_{ij,k} = \exp(x'_{ij}\beta_{\sigma,k})$
Gamma	μ_k	$g(\mu_{ij,k}) = \log(\mu_{ij,k}) = x'_{ij}\beta_{\mu,k}$	$\mu_{ij} = \exp(x'_{ij}\beta_{\mu,k})$
	σ_k	$g(\sigma_{ij,k}) = \log(\sigma_{ij,k}) = x'_{ij}\beta_{\sigma,k}$	$\sigma_{ij,k} = \exp(x'_{ij}\beta_{\sigma,k})$

Furthermore, since we use a mixture distribution, we need to estimate the values for π_k in (4.29) as well. To make it covariate dependent, we specify the following link function

$$g_{\pi,k}(\pi_{ij,k}) = x'_{ij}\beta_{\pi,k}, \tag{4.32}$$

where $x_{ij} = (V_{ij}^{-}, U_i, T_i, C_i)$, $\pi_{ij,k}$ is the weighting of distribution k for claim i and event j, and $\beta_{\pi,k}$ are vectors of length $n+5$ indicating the effects of the covariates on the link function of $\pi_{ij,k}$.

We use a logit link to relate the parameter π_k to the covariates. We need $\sum_{k=1}^{K} \pi_k = 1$. Therefore, we use a multinominal logistic regression, such that we have

$$\pi_k = \begin{cases} \dfrac{\exp(x'_{ij}\beta_{\pi,k})}{1 + \sum_{k'=1}^{K-1} \exp(x'_{ij}\beta_{\pi,k'})} \text{ for } k = 1, \ldots, K-1 \\ \dfrac{1}{1 + \sum_{k'=1}^{K-1} \exp(x'_{ij}\beta_{\pi,k'})} \text{ for } k = K \end{cases} \tag{4.33}$$

We can use the above mentioned link functions to construct the likelihood for the mixture distribution with covariates. Let $I_k(Y_{ij})$ indicate whether payment Y_{ij} belongs to distribution k for each k, $1 \leq k \leq K$. The likelihood for the payments in our model is given by

$$\begin{aligned} L(\mu, \sigma, \pi) = &\prod_{i \geq 1} \prod_{j \geq 1} (\pi_1 f_{Y,1}(y_{ij}|\mu_1, \sigma_1, x_{ij}))^{I_1(y_{ij})} \\ &\cdot (\pi_2 f_{Y,2}(y_{ij}|\mu_2, \sigma_2, x_{ij}))^{I_2(y_{ij})} \cdots (\pi_K f_{Y,K}(y_{ij}|\mu_K, \sigma_K, x_{ij}))^{I_K(y_{ij})} \end{aligned} \tag{4.34}$$

In order to determine the values for the parameters of the payment distribution, we need to find the values of $\beta_{\pi,k}, \beta_{\sigma,k}, \beta_{\mu,k}$ for each k that optimize the likelihood given in (4.34). We do this by performing an EM algorithm. The EM algorithm finds optimal values for the parameters of the k distributions and the weights that need to be assigned to each distribution in the mixed distribution by iteratively optimizing the likelihood. In our case, in each iteration, instead of using a normal maximization of the likelihood to find the parameters of the distribution, we perform a weighted likelihood optimization to find each of the $\beta_{k,j}$ by using a Newton Raphson algorithm. Using this approach, instead of finding the value of the parameters σ, μ, π, we find the $\beta_{k,j}$'s that link the covariates to the values of the parameters. A more detailed overview of the algorithm can be found in the Appendix, Section 8.4.1.

Settlements

The payments during the development process, i.e. not at the settlement date, are always positive. Contrarily, a settlement can occur either with or without a payment, which implies that there is a positive probability that the level of the payment is zero. In order to model the levels of the payments associated with settlements, we need to take into account the probability of a zero payment. The Log-Normal and Gamma distribution do not have a mass at zero. Therefore, to model the settlement payments, we add a point probability at zero. First, let us define ξ_0 as the probability that a payment Y_{ij} is zero, $0 \leq \xi_0 \leq 1$. The probability of a zero-payment is estimated as covariate dependent. Therefore, we use a logit link function to map ξ_0 to the covariates x_{ij},

$$g_{\xi_0}(\xi_0) = \text{logit}(\xi_0) = x'_{ij}\beta_{\xi_0}. \tag{4.35}$$

where $x_{ij} = (V_{ij}^-, U_i, T_i, C_i)$, ξ_0 is the probability of a zero payment, and β_{ξ_0} is the vector of length $n+5$ indicating the effect of the covariates on the link function of ξ_0. The inverse of the link function is given by

$$g_{\xi_0}^{-1}(x'_{ij}\beta_{\xi_0}) = \xi_0 = \frac{\exp(x'_{ij}\beta_{\xi_0})}{1+\exp(x'_{ij}\beta_{\xi_0})}. \tag{4.36}$$

Let $I_0(Y_{ij})$ indicate whether payment Y_{ij} is a zero payment. The likelihood of the zero adjusted model is given by

$$\begin{aligned} L(\mu,\sigma,\pi,\xi_0) = &\prod_{i\geq 1}\prod_{j\geq 1} \xi_0^{I_0(y_{ij})} \cdot \Big((1-\xi_0)(\pi_1 f_{Y,1}(y_{ij}|\mu_1,\sigma_1,x_{ij}))^{I_1(y_{ij})} \\ &\cdot (\pi_2 f_{Y,2}(y_{ij}|\mu_2,\sigma_2,x_{ij}))^{I_2(y_{ij})} \cdots (\pi_K f_{Y,K}(y_{ij}|\mu_K,\sigma_K,x_{ij}))^{I_K(y_{ij})} \Big)^{1-I_0(y_{ij})} \end{aligned} \tag{4.37}$$

where $x_{ij} = (V_{ij}^-, U_i, T_i, C_i)$ are the covariates corresponding to payment Y_{ij} and the additional parameter ξ_0 allows for the probability mass at zero.

To estimate the parameters of the payment distributions of settlements, we introduce a new step in the algorithm to calculate the value for β_{ξ_0} that optimizes the likelihood in (4.37). We first divide the payments into zero and non-zero payments. We estimate the value of ξ_0 based on the probability of the payment being a zero payment or a non-zero payment. Thereafter, as before, we use the Newton Raphson algorithm to find the values for β_{ξ_0}. The analysis continues as in 4.1.4, where the zero-payments are assigned a weight 0 for all iterations. The complete step can be found in the Appendix, Section 8.4.2.

4.2 Simulation

In the previous section, we discussed how the parameters of our model are estimated. In this section, we discuss how we simulate the IBNR and RBNS reserves at time τ using. We simulate the reserves for one year, using covariates for all parts of the model. The simulation procedure for the reserves without taking into account the covariates can be found in the Appendix.

First, since the time of occurrence, the covariates and the reporting delay of the IBNR claims are not known to the insurer at time τ, we need to simulate these variables. We will discuss how we simulate these first. Then, we discuss the way in which we simulate the development period for each IBNR and RBNS claim, as the simulation procedure is the same for both types of reserves. In the last section we discuss how the reserves are calculated based on the simulated claim processes for both types of reserves. Every simulation is performed M times, such that we have M reserves at the end of the simulation.

4.2.1 Time of Occurrence and Reporting Delay

We start the simulation of the IBNR claims by first simulating the time of occurrence, after which we simulate the corresponding covariates. Based on these covariates and time of occurrence, we simulate the reporting delay. The complete simulation is displayed in Algorithm 1.

We simulate the number of IBNR claims originated in month l. We first determine the exposure in month l for each combination of covariates x_i. We multiply this with $\lambda_{l,0} \exp(x_i'\beta)$, which is the intensity in month l with covariates x_i. We do this for every combination of covariates and then add it to get the total for month l. The resulting number is the expected number of occurrences for month l based on $\hat{\lambda}_{l,0}(\beta)$, which we indicate with N_l.

The rate of occurrence $\hat{\lambda}_{l,0}(\beta)$ that we estimated, is the rate based on the claims that have been reported. However, for each month l, there is a positive probability that there are claims that have occurred, but have not yet been reported. To obtain the number of claims that have occurred in month l, but have not yet been reported at time τ, denoted with N_l^{IBNR}, we need to correct N_l as

$$\begin{aligned} N_l^{IBNR} &= N_l \frac{\int_{a_{l-1}}^{a_l} (1 - F_U(\tau - s))ds}{\int_{a_{l-1}}^{a_l} F_U(\tau - s)ds} \\ &= N_l \left(\frac{\int_{a_{l-1}}^{a_l} 1ds}{\int_{a_{l-1}}^{a_l} F_U(\tau - s)ds} - 1 \right) \\ &= N_l \frac{\int_{a_{l-1}}^{a_l} 1ds}{\int_{a_{l-1}}^{a_l} F_U(\tau - s)ds} - N_l. \end{aligned} \tag{4.38}$$

Hence, we multiply the number of occurrences based on the observed occurrence rate with a factor to obtain the real number of occurrences in month l and subtract the observed number of occurrences to obtain the number of IBNR claims, i.e. the claims that occurred in month l but have not yet been reported at time τ. We need $F_U(\tau - t)$ to correct for the right amount of claims. Here, we use a simplistic version of the reporting delay distribution: we use the Weibull hazard rate without any covariates. This reporting delay distribution is only used for this calculation and is not consistent with the rest of the paper. We choose to do it this way, to make this part somewhat more simplistic.

From our simulation it follows that for each simulated claim, we also know the covariates of that claim, i.e. x_i. Hence, these do not have to be simulated separately. To determine the day of occurrence, we draw a random uniform distributed number between 1 and the number of days in that month for each claim. We use this number as the occurrence date of the IBNR claim in month l.

We continue our simulation by simulating the reporting delay for each of the simulated IBNR claims. To obtain the reporting delays, we simulate from the distribution function of the reporting delay, $F_U(u|T_i, C_i)$. First, we simulate a random variable from the Uniform distribution, which we use to compute the reporting delay. In Section 8.5, the simulation procedure using a hazard rate is explained in more detail.

After simulating the three variables, time of occurrence, reporting delay and the covariates, we have the same information of the IBNR claims as we do of the RBNS claims. Accordingly, in our next step, we continue with simulating the development process for both the IBNR and RBNS claims.

Algorithm 1 IBNR Claims With Covariates

for $m = 1 : M$ **do**
  procedure Occurrence Time and Covariates
    for Each Month l **do**
      Number of IBNR claims $\leftarrow$ RPoisson(N_l^{IBNR})
      Day in month $\leftarrow$ RUnif(1, Number of days in month l)
    $N \leftarrow$ Sum of simulated IBNR claims over all months $l = 1, 2, \ldots$
  for $i = 1 : N$ **do**
    procedure Reporting Delay
      $x_i \leftarrow (T_i, C_i)$
      $Y \leftarrow$ Draw a randomly distributed U(0,1) variable
      Weibull:
      $$U_i \leftarrow \frac{(-\zeta \log(Y))^{1/\rho}}{\exp(x_i'\beta)}$$
      Piece-wise Constant:
      $$U_i \leftarrow \frac{H^{-1}(-\log(Y))}{\exp(x_i'\beta)} \text{ with } H^{-1} \text{ as in (8.39) of the Appendix}$$

4.2.2 Development Process

The development process simulation is the same for IBNR claims as for RBNS claims. First, we determine the number of claims and simulate the event time and event type of the first event $j = 1$ for each claim i. Then, we simulate the payments Y_{i1} for that event for each claim i. The covariates are updated and the claims for which a settlement took place are closed. This simulation step is repeated until all claims have been settled. Algorithm 2 displays the simulation in more detail.

Algorithm 2 Development Process With Covariates

for $m = 1 : M$ **do**
  $j \leftarrow 0$
  Open Claims $\leftarrow$ All Claims i
  while There are Open Claims **do**
    for i in Open Claims **do**
      $j \leftarrow j + 1$
      procedure EVENT
        Y $\leftarrow$ Draw a randomly distributed U(0,1) variable
        $V_{ij} \leftarrow \dfrac{H^{-1}(-\log(Y))}{\exp(x'_{ij}\beta)}$ (See Appendix Section 8.5)
        $K_{ij} \leftarrow$ RBernoulli with probability $\dfrac{1}{1+\exp(x_i\beta)}$
      if $K_{ij} = 1$ **then**
        procedure PAYMENT
          Compute value for $\mu_{ij,k}$, $\sigma_{ij,k}$ and $\pi_{ij,k}$ for each k with covariates x_{ij} and $\beta_{\mu,k}$, $\beta_{\sigma,k}$ and $\beta_{\pi,k}$
          U $\leftarrow$ Draw a randomly distributed U(0,1) variable
          Payment $Y_{ij} \leftarrow F^{-1}_{Y,MX}(U|\mu_{ij}, \sigma_{ij}, \pi_{ij})$
      if $K_{ij} = 0$ **then**
        procedure SETTLEMENT
          Compute value for $\mu_{ij,k}$, $\sigma_{ij,k}$, $\pi_{ij,k}$, ξ_0 for each k with covariates x_{ij} and $\beta_{\mu,k}$, $\beta_{\sigma,k}$, $\beta_{\pi,k}$ and β_{ξ_0}
          U $\leftarrow$ Draw a randomly distributed U(0,1) variable
          Payment $Y_{ij} \leftarrow F^{-1}_{Y,0}(U|\mu_{ij}, \sigma_{ij}, \pi_{ij}, \xi_0)$
          Close claim i
      procedure UPDATE COVARIATES
        Total Payout = Total Payout + Y_{ij}
        Time in Development = Time in Development + V_{ij}

4.2.3 Reserves

To estimate the reserves, we add all the payments Y_{ij} for each month. Through this way, we get a prediction of the losses for each month. To obtain the total reserve amount, we add all the monthly losses, such that we have a total value for the losses of the coming year. We calculate it for one year ahead, as this is the period for which we have the actual payments. This way, we can compare the our simulated reserves with the actual values. The total simulation is performed M times, so that we have M estimates of the total reserve that should be held for both claims.

Algorithm 3 Reserve Calculation

1: **for** $m = 1 : M$ **do**
2: **procedure** RESERVES
3: **for** Each month l **do**
4: $X_{IBNR}(l) \leftarrow$ Sum of Y_{ij} of IBNR claims made in month l
5: $X_{RBNS}(l) \leftarrow$ Sum of Y_{ij} of RBNS claims made in month l
6: $X_{IBNR}^{(m)} \leftarrow \sum_{l \geq 1} X_{IBNR}(l)$
7: $X_{RBNS}^{(m)} \leftarrow \sum_{l \geq 1} X_{RBNS}(l)$

4.3 Methods of Comparison

In our analysis, we need to compare the different models. In this section, we introduce three methods of comparison that we use to compare (parts of) our models: the Kaplan-Meier estimate, the AIC and the VaR.

First, in order to compare the fit of the hazard rates, we use the Kaplan-Meier estimate. The Kaplan-Meier estimate is an estimate that can be used to check the fit of a hazard rate, which we use for both the reporting delay and the gap times of the events.

Second, we need to determine which covariates add value to the estimation of our parameters, i.e., we need to select the model that fits our data best. We have multiple covariates that we include in our analysis and therefore have many possible combinations of those covariates that we can include in our final model. One possibility is to consider the p-values of the covariates and select the significant covariates. A drawback of this approach is that p-values are only valid for comparing nested models. Therefore, we can not use this approach to compare non-nested models. To compare the different models, we will use the Akaike Information Criterion, or AIC.

Third, after performing our simulations, we want to determine whether incorporating covariates in all parts of the model, i.e. the occurrence process, reporting delay and development process, adds value. We want to compare the different models based on their accuracy of predicting the reserves. The reserves that an insurer holds are the expected total loss arising from payments of a specific type of claim. Therefore, we consider if including covariates improves the estimation of the reserves by considering the expected loss from our simulations, which is given by the mean of our simulations. Furthermore, we examine the standard deviation of the simulations for each model as an indication of the confidence interval of our simulations. Lastly, we examine the Value at Risk in order to have an indication of the tail of the distribution of our simulations. We will use the Monte Carlo method for estimating the Value at Risk.

4.3.1 Kaplan-Meier

The Kaplan-Meier estimator is a non-parametric estimator for the survival function (Kaplan and Meier, 1958). A survival function is a function $S(u)$, which is defined as $S(u) = \Pr[U > u]$, i.e. $S(u) = 1 - F(u)$, where $F(u)$ is the distribution function of u. The hazard rate $\gamma_U(u)$ can be linked to the survival function as $\gamma_U(u) = f_U(u)/(1 - F_U(u))$, where we now have $F_U(u) = 1 - S_U(u)$. The Kaplan-Meier estimator for the reporting delay is given by

$$\hat{S}(u) = \frac{\sum_{i\geq 1} I(u_i > u)}{\sum_{i\geq 1} 1}, \tag{4.39}$$

which is a step function that decreases each time a claim is reported.

A similar approach can be used to find the Kaplan-Meier estimate for the survival function of the gap times between events v_{ij}. A part of observations of the gap times are censored, which implies that we do not know exactly what the value of $I(v_{ij} > v)$ is. As before, we let $\delta_{ij} = (1 - I(v_{ij} \text{ is censored}))$ indicate whether event j of claim i is observed ($\delta_{ij} = 1$) or censored ($\delta_{ij} = 0$). The Kaplan-Meier estimate for the gap times is given by

$$\hat{S}(v) = \prod_{o:v_o^* \leq v} \left(1 - \frac{\sum_{i\geq 1}\sum_{j\geq 1} I(v_{ij} = v_o^*)\delta_{ij}}{\sum_{i\geq 1}\sum_{j\geq 1} I(v_{ij} \geq v_o^*)}\right) \tag{4.40}$$

where v_o^* are the distinct values among all v_{ij}, $i \geq 1$, $\sum_{i\geq 1}\sum_{j\geq 1} I(v_{ij} = v_o^*)\delta_{ij}$ is the number of uncensored events with a gap time equal to v_o^* and $\sum_{i\geq 1}\sum_{j\geq 1} I(v_{ij} \geq v_o^*)$ is the number of events with a gap time greater than or equal to v_o^* for both uncensored and censored observations. In the calculation of the Kaplan-Meier estimator, we allow for more than one value being equal to v_o^*.

We use the Kaplan-Meier estimate to check whether the Weibull distribution has a good fit to the data. To do this, we make use of a property of the Weibull distribution. The Weibull survival function for the reporting delay is given by

$$S(u) = \exp(-(\zeta u)^\rho). \tag{4.41}$$

By taking the log of the negative log of the survival function, we get

$$\log[-\log(S(u_i))] = \rho\log(\zeta) + \rho\log(u_i). \tag{4.42}$$

We observe that the log of the negative log plotted against the log of the time between events should have a linear relationship in order for the Weibull to have a good fit to the data.

Moreover, it is possible to compare the fit of the Weibull distribution with covariates for the reporting delay and the gap times by checking whether the linear relations in (4.43) and (4.44) hold.

$$\log[-\log(S(u_i|T_i, C_i))] = \rho \log(\zeta) + \rho \log(u_i) + x_i'\beta, \tag{4.43}$$

$$\log[-\log(S(v_{ij}|V_{ij}^-, U_i, T_i, C_i))] = \rho \log(\zeta) + \rho \log(u_i) + x_i'\beta, \tag{4.44}$$

where $\hat{S}(u_i|T_i, C_i))$ and $\hat{S}(v_{ij}|V_{ij}^-, U_i, T_i, C_i)$ are defined as the Kaplan-Meier estimate based on the covariates for the reporting delay and the gap times, respectively. The proof of this relation can be found in the Appendix, Section 8.6. However, as the number of covariates n becomes larger, the total number of combinations of covariates becomes larger. This implies that we need to check a large number of plots to verify the fit of the Weibull distribution. As this is not optimal, we choose to use the Kaplan-Meier estimate to check the fit of the Weibull distribution without covariates. Afterwards, we check the added value of the covariates, as explained in the next section.

4.3.2 Akaike Information Criterion

The Akaike Information Criterion, AIC, is often used for comparing models. One of its advantages is that it penalizes the use of too many parameters, such that overparameterization is avoided. This is especially useful for our analysis, as we want to examine whether including additional covariates improves our model. The AIC is given by

$$AIC = -2l(\hat{\theta}) + 2\eta, \tag{4.45}$$

where $l(\hat{\theta})$ denotes the value of the partial log likelihood at the maximum partial likelihood estimate for a model, $\hat{\theta}$, and η denotes the number of parameters in the model. The first term indicates the goodness of fit of the model, whereas the second term penalizes complexity due to a larger number of parameters. A low AIC indicates that a model fits the data well with few parameters. Therefore, we select the model that has the lowest AIC.

4.3.3 Mean, Standard Deviation and VaR

The values for the mean and standard deviation of our simulations are simply given by the average and the standard deviation of the M simulated losses for each type of reserve. We compare our expected value of the losses for the next year, which is given by the mean of our simulations, with the actual payments that have been made in that year. This way, we can examine whether incorporating covariates improves the reserve calculation.

Furthermore, we compute the Value at Risk. Value at Risk, VaR_α, is often used for measuring risk and is included in the Solvency framework used by insurance companies. It is defined as the minimum loss χ such that the probability of a larger loss than χ is smaller than $1-\alpha$, where $\alpha \in (0,1)$. The loss is estimated over a specified time period. In accordance with the Solvency II framework, we use $\alpha = 0.995(99.5\%)$ and estimate VaR_α over one year. Formally, we define VaR_α as

$$\text{VaR}_\alpha = \inf\{\chi \in \mathbb{R} : \Pr(X > \chi) \leq 1-\alpha\} \tag{4.46}$$

$$= \inf\{\chi \in \mathbb{R} : \text{F}(\chi) \geq \alpha\}. \tag{4.47}$$

The Monte Carlo method is an approach that estimates the VaR_α by simulating losses from their parametric distribution. First, we use the M simulated losses for each type of reserve. Then, we order the losses from small to large and take the α-th quantile of the ordered losses to obtain VaR_α. The computation of the expected value, standard deviation and the VaR is given in the Algorithm below.

Algorithm 4 Expected Value, Standard Deviation and VaR_α Calculation

procedure Mean and Standard Deviation
- Expected Loss IBNR $\leftarrow \sum_{m=1}^{M} X_{IBNR}^{(m)}$
- Standard Deviation IBNR $\leftarrow \sqrt{\frac{\sum_{m=1}^{M}(X_{IBNR}^{(m)} - \overline{X}_{IBNR})^2}{M-1}}$
- Expected Loss RBNS $\leftarrow \sum_{m=1}^{M} X_{RBNS}^{(m)}$
- Standard Deviation RBNS $\leftarrow \sqrt{\frac{\sum_{m=1}^{M}(X_{RBNS}^{(m)} - \overline{X}_{RBNS})^2}{M-1}}$

procedure VaR
- $X_{IBNR}^{(+)} \leftarrow$ Ordered $X_{IBNR}^{(m)}$
- $\text{VaR}_{99.5\%}$(IBNR) $\leftarrow$ 99.5% Quantile of $X_{IBNR}^{(+)}$
- $X_{RBNS}^{(+)} \leftarrow$ Ordered $X_{RBNS}^{(m)}$
- $\text{VaR}_{99.5\%}$(RBNS) $\leftarrow$ 99.5% Quantile of $X_{RBNS}^{(+)}$

Chapter 5

Data

In this paper, we use data from a car insurer from January 2011 until December 2016. We use the first five years of the data set, i.e. from 2011 until 2015, to estimate our model. We simulate the reserves based on payments for 2016 and compare the reserves with the actual payment data of 2016. Our data consists of two data sets: one data set with policy information of the 5,025,658 policies that were in the portfolio of the insurer and one data set with information on 319,640 claims that have been reported by the policyholders. Therefore, on average, we have 1 claim per 16 policies over the last five years.

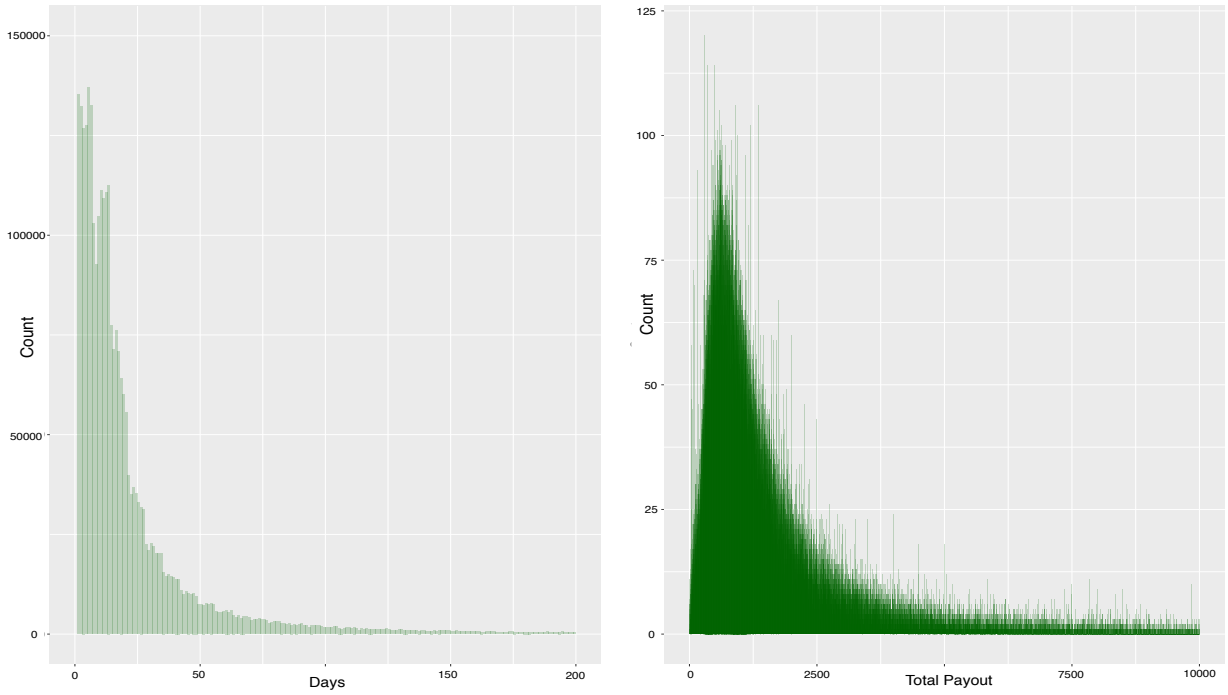
The first data set contains information on the development of a claim from the moment of occurrence until a settlement. Therefore, we have information on the date of occurrence, date of reporting, and dates and heights of the payments and settlement for each claim. There are 385,728 payments associated with the claims, of which 198,387 belong to regular payments and 187,341 are payments accompanying a settlement. This indicates that there are 121,297 settlements with a zero payment. Furthermore, we are able to calculate the exposure of the insurer for each month, which is defined as the total number of active policies weighted by the number of days that they were active in that month. In Table 5.1 we give a more detailed overview of the data.

Figure 5.1 displays histograms of the reporting delay and the total payout. We observe that the reporting delay differs for the first 15 days, after which it gradually declines. Moreover, the total payout has a long tail.

Table 5.1: Summary Statistics

This table reports summary statistics of the variables used in our analysis.

	Min	Mean	Max	Count
Time of Occurrence	0	161	524	319,640
Reporting Delay	0	25	1,475	319,640
Time between Events	0	24	2,037	868,330
Time to Settlement	0	70	2,120	6,831
Payments	0	1,582	901,233	198,387
Settlement Payments	0	423	77,850	308,638
Total Payout	0	890	1,040,159	308,638
Exposure per Month	35,790	67,595	86,480	-



(a) Reporting Delay

(b) Total Payout

Panel (a) and Panel (b) of this figure show the histograms of the reporting delay data and the total payout data for 2011-2015, respectively.

Figure 5.1: Histograms of the data

The second data set contains information on the policy for different versions of the insurance for each policyholder. For example, when a policyholder buys a different car, the policy is updated and a new version is reported in the data set. The start and end date of every version is known, so we link a claim to the right version based on the time of occurrence of the claim. Information about the policies consists of demographic statistics of the policyholder, information on the car that is insured and information on the type of insurance.

In our model, we only use the all-risk insurance car policies. We do this to avoid a bias for a specific type of insurance. The demographic statistics and information on the car can be found in Table 5.2. The information on the car consists of the catalog value of the car and an indicator that specifies whether the car is an old car. These two variables are available for the large majority of the policies. However, the data about the demographic statistics of the policyholder is not complete. Unfortunately, the majority of the claims, 88%, do not have information on the sex of the policyholder and from 20% of the claims it is not known what the age of the policyholder is. Therefore, we choose to not include these variables in our analysis.

We choose to include the catalog value of the car as a categorical variable for computational reasons. The out of sample simulations take a long time. Therefore, to make the simulation faster, we choose to use a categorical variable such that the number of possible distributions is limited.

Table 5.2: Policy Characteristics

This table provides descriptions for the four available policy characteristics in the data set. The number of NA's indicate the number of missing values of each variable.

Variable	Description	NA's
Catalog Value Car	Categorical variable for the catalog value of the car, *1 = Low, 2 = Intermediate and 3 = High*	1 (0%)
Old Car	Indicator variable on whether the car is an old car, *1 = Yes and 0 = No*	127 (0%)
Gender Policyholder	Indicator variable for the gender of the policyholder, *1 = Male and 0 = Female*	281,838 (88%)
Age Policyholder	Continuous variable for the age of the policyholder	63,992 (20%)

Chapter 6

Results

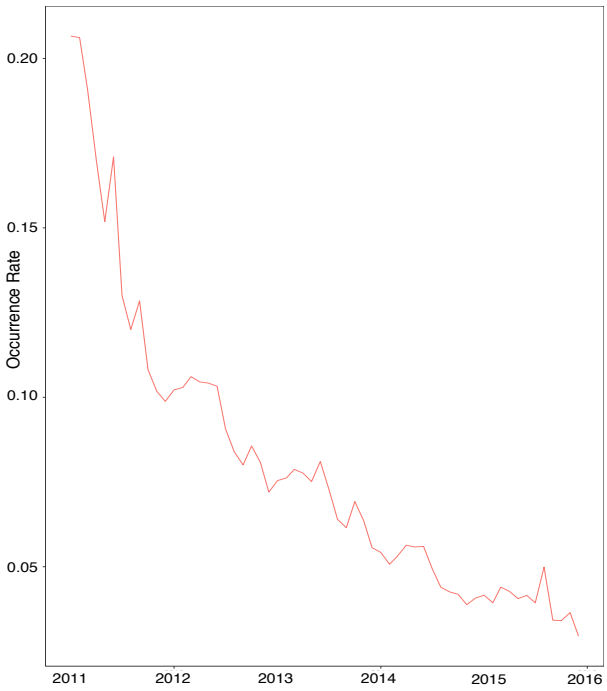
In this chapter, we discuss the results of our analysis. First, we discuss the empirical results of fitting our model. Thereafter, we compare the results of our simulations.

6.1 Empirical results

In this section, we discuss the empirical results of our analysis on the development of a claim and compare which model specification fits our data best.

6.1.1 Rate of Occurrence

The time of occurrence is modelled as a Poisson process with rate $\hat{\lambda}_{l,0}$. The values for $\hat{\lambda}_{l,0}$ for every month in the last years are displayed in Figure 6.1. We display the rate of occurrence by using calendar time instead of time passed since the start of the policy as this is standard practice for car insurers. One of the reasons for using calendar time is that the policies often change over time, due to changes in the terms and conditions of the policies. Therefore, there are many versions of the same policy of the same policyholder. It is critical to assess which change in the policy brings about a new policy in order to obtain the time that has passed since the start of the policy. It is more objective to use calendar time instead of time passed since the start of the policy, as this does not require an assessment of the terms and conditions. Therefore, in this part of our analysis, we obtain the occurrence rate per calendar month. We observe that the occurrence rate has declined over the past years. Furthermore, the occurrence rate differs for each month of each year, indicating that there is an added value for modelling the occurrence rate per month instead of per year or per season.



This figure displays the rate of occurrence for each month of the years 2011-2015.

Figure 6.1: Occurrence Rate

In order to check which covariates improve the fit of our model, we use a backwards step-wise analysis based on the AIC of the models. In this analysis, we first include all covariates and compute the AIC. Then, for each covariate, we compute the AIC of the model without that covariate. We select the model with the lowest AIC and perform the second step again. We do this iteratively until the AIC of the model without deleting a covariate is the lowest. The model specification of the rate of occurrence with the lowest AIC contains only the coefficient for the age of the car. The full model and the AIC of all models can be found in Table 8.2 and Table 8.3 of the Appendix, respectively. The coefficient of the covariate of the age of the car in Table 6.1 indicates that the rate of occurrence increases for old cars. An explanation is that an old car is more likely to have a failure of (a part of) the car, which causes a higher occurrence of a claim.

Table 6.1: Rate of Occurrence Coefficient

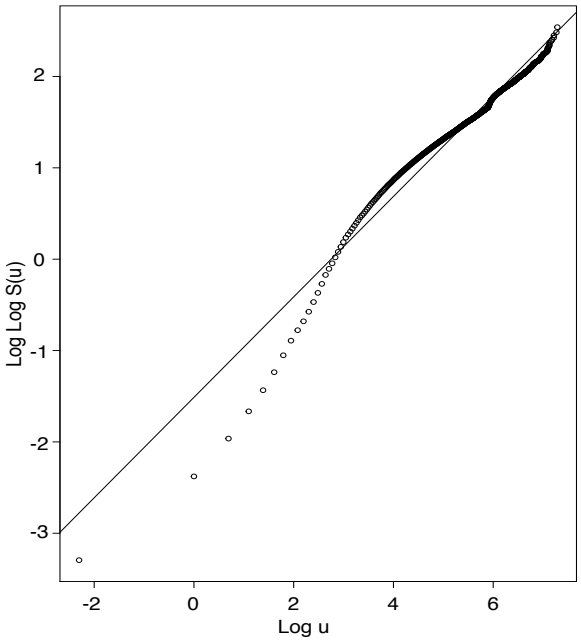
This table reports the coefficient estimate (standard deviation) for the variables in the optimal model for the rate of occurrence.

	Estimate
Old Car	0.051**
	(0.017)

*Note: ** p<.05*

6.1.2 Reporting Delay

The hazard rate of the reporting delay of the claims is modelled both parametrically and non-parametrically. To explore the fit of the Weibull hazard rate, we examine the plot of the log of the log of the survival function against the log of the reporting delay. If the Weibull hazard rate is a proper hazard for the data, the relation between these two variables should be linear. In Figure 6.2, we observe that the observations do not follow the straight line. As the Weibull distributions does not have a good fit with the data, we choose to model the reporting delay with a piece-wise constant baseline hazard. We inspect the histogram of the reporting delays to determine the intervals for which the hazard rates are assumed to be constant. The number of claims with a reporting delay within the first 15 days differ substantially. Therefore, we choose a daily interval for the first 15 days. Thereafter, we keep the hazard rate constant over a ten day interval.



This figure plots $\log[-\log[\hat{S}(u_i)]]$ *versus the log of the reporting delays* $\log(u_i)$*. A linear line indicates a good fit of the Weibull specification for the hazard rate.*

Figure 6.2: Visual inspection of the fit of the Weibull hazard rate for the reporting delay

On top of the piece-wise constant hazard rate, we add covariates to the hazard rate to check whether this improves the fit of our model. We do this with the backwards step-wise analysis. The full model, including the covariates for time of occurrence, the age of the car and the catalog value of the car, has the lowest AIC and is shown in Table 6.2. The AIC values of the different models considered in the backwards analysis can be found in Table 8.4 in the Appendix.

First, we observe that the effect of the time of occurrence on the reporting delay is significant, but small. An increase in the time of occurrence of 6 months increases the hazard rate with 0.01%. As the effect on the hazard rate is small, there does not seem to be a relation between the reporting delay and the time of occurrence. The coefficients of the policy characteristics are significant. This implies that the inclusion of policy characteristics contributes to the fit of the model, whereas incorporating dependency between the reporting delay and the time of occurrence has no added value for this part of the model.

As an example, the coefficient of the variable old car is given by -0.041, which indicates that the hazard is a factor $\exp(-0.041) = 0.960$ smaller. Therefore, the time to reporting is longer. This could be due to the fact that it takes longer to determine the damage to an old car, which leads to a longer time to reporting in case of an old car.

Table 6.2: Reporting Delay Coefficients

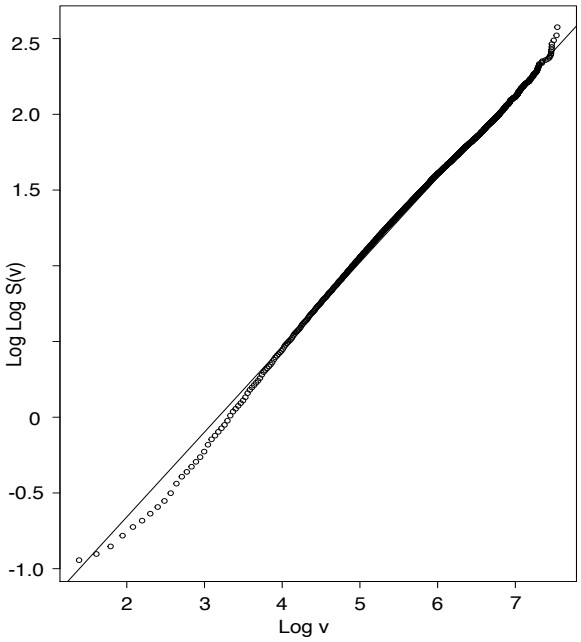
This table reports the coefficient estimates (standard deviation) for the variables in the optimal model for the reporting delay. The last column indicates the change in the hazard rate corresponding to the coefficient estimate of each variable.

	Estimate	Change in Hazard Rate
Old Car	-0.041*	-4.017%
	(0.017)	
Catalog Value "Intermediate"	-0.046***	-4.496%
	(0.004)	
Catalog Value "High"	-0.140***	-13.064%
	(0.007)	
Time of Occurrence	6.01E-7***	+6.01E-7%
	(0.000)	

*Note: * p<.10, ** p<.05, *** p<0.01*

6.1.3 Event Rate

The time between events in the development process is estimated with a piecewise constant and a Weibull hazard rate. First, we examine the fit of the Weibull model. Figure 6.3 shows that the observations closely follow the straight line of the double log of the Kaplan-Meier survival function. Therefore, the Weibull distribution is able to model the time between events in the development process accurately.



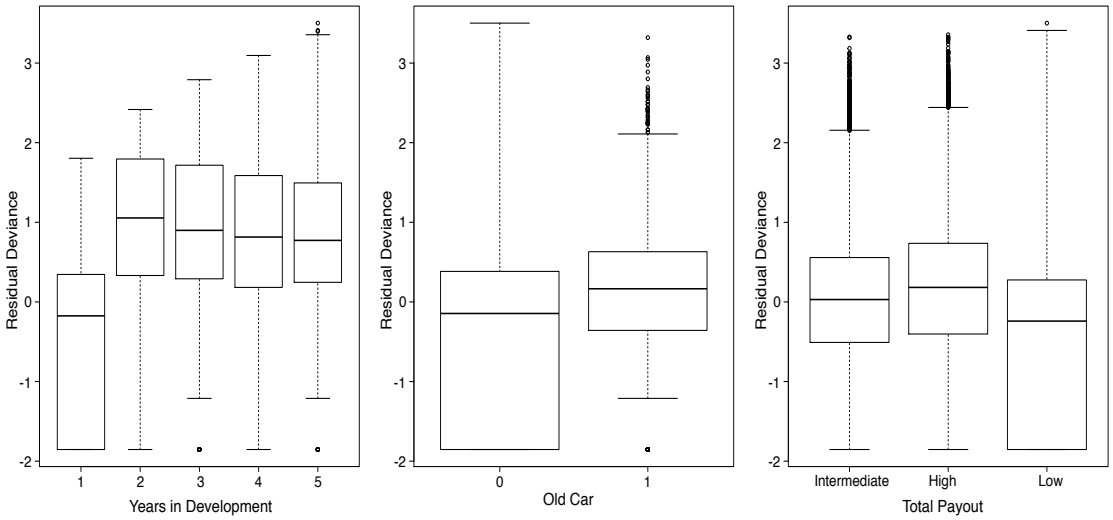
This figure plots $\log[-\log[\hat{S}(v_{ij})]]$ *versus the log of the event gap times* $\log(v_{ij})$*. A linear line indicates a good fit of the Weibull specification for the hazard rate.*

Figure 6.3: Visual inspection of the fit of the Weibull hazard rate for the event rate

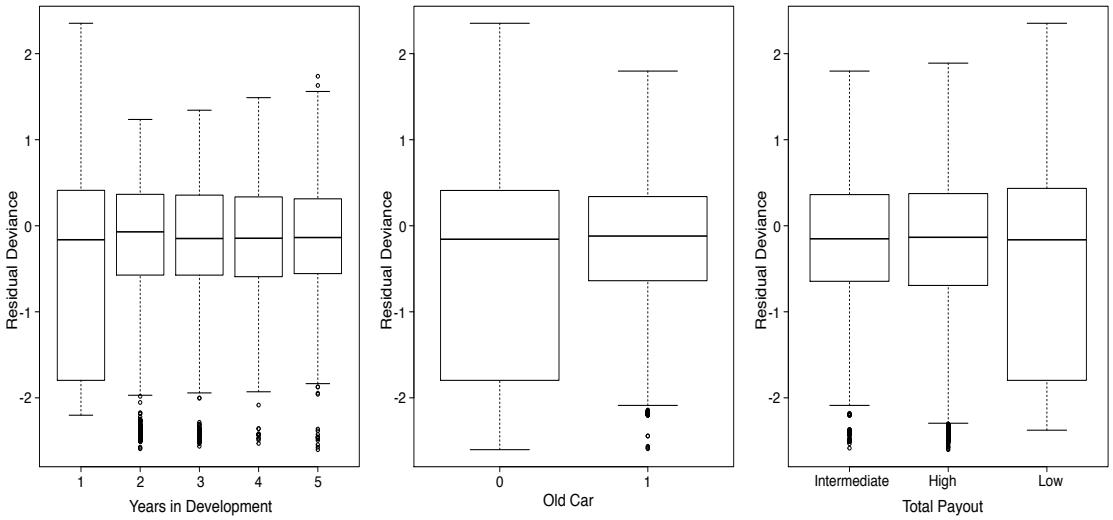
We estimate the Weibull hazard rate with and without covariates. The estimates of the model without covariates are given in the second column of Table 6.3. The shape and scale, ρ and ζ, of the Weibull distribution, are given by 0.625 and 27.33, respectively. Furthermore, we estimate the model with covariates. We again use the backwards step-wise analysis to check which covariates add value to our model. The results of the backwards step-wise analysis can be found in the Appendix, Table 8.5.

The optimal model following from the backwards step-wise analysis consists of the number of payments, the age of the car, the reporting delay, the time in the development process and the total payout as covariates. This implies that creating dependency between the model components, i.e. the reporting delay and the event hazard, as well as including the history of the development process improves the fit of our model. To further asses our model, we inspect the deviance residuals. Panel (a) in Figure 6.4 shows the deviance residuals of the data when we use the Weibull hazard rate without covariates. Each plot in the panel shows

the results for observations with a specific value of a covariate. The first, second, and third figure show the deviance results of the data with different values of the time in the development process, the age of the car, and the total payout, respectively. The lower panel, Panel (b), displays the deviance residuals of the data for different covariate values when we use a Weibull hazard rate with covariates. We observe that the deviance residuals are more evenly distributed in the lower plots than in the upper plots. This implies that the model with covariates has a better fit with our data.



(a) Residuals using a hazard rate without covariates



(b) Residuals using a hazard rate with covariates

Panel (a) and Panel (b) show the deviance residuals of the data after using a hazard rate without and with covariates, respectively. Each of the three plots in each panel displays the deviance residuals of the data for a specific covariate, where each bar in the plot respresents the deviance residuals of the data for a specific value of that covariate

Figure 6.4: Plots of the deviance residuals of the event gap times

Next, we examine the coefficients of the covariates, which are given in Table 6.3. First, we explore the relation of the event hazard rate with the other parts of the claim process, namely the time of occurrence and the reporting delay. The time of occurrence does not have a significant impact on the tap time, as it is not included in the optimal model.

Table 6.3: Event Gap Times Coefficients

This table reports the coefficient estimates (standard deviation) of the covariates included in the null model, the full model and the optimal model. The null model and the full model indicate the models with no and all covariates, respectively. The optimal model is the model with the lowest AIC.

	Null Model	Full Model	Optimal Model
2 Years in Development		-1.232***	-1.234***
		(0.014)	(0.014)
3 Years in Development		-1.377***	-1.380***
		(0.026)	(0.026)
4 Years in Development		-1.490***	-1.491***
		(0.041)	(0.041)
5 Years in Development		-1.709***	-1.707***
		(0.058)	(0.058)
Old Car		-0.291***	-0.297***
		(0.020)	(0.020)
Number of Payments		0.033***	0.036***
		(0.020)	(0.001)
Catalog Value "Intermediate"		0.053***	
		(0.004)	
Catalog Value "High"		0.007	
		(0.008)	
Total Payout $>$ €2000		0.033***	0.035***
		(0.008)	(0.008)
Total Payout $<$ €500		0.331***	0.330***
		(0.006)	(0.006)
Occurrence		0.000	
		(0.001)	
Reporting Delay		-0.001***	-0.001***
		(0.000)	(0.000)
Shape (ρ)	0.625	0.662	0.662
Scale (ζ)	27.331	33.115	31.817
AIC	4,510,262	4,448,090	4,436,905

*Note: * p<.10, ** p<.05, *** p<0.01*

By contrast, the effect of the reporting delay on the event hazard rate is significant. When the reporting delay becomes larger, the time to the next event also becomes larger. For each day the insurer takes to report a claim, the hazard rate is multiplied with factor $\exp(-0.001) = 0.999$, i.e. decreases with 0.1%. A one standard deviation increase in the reporting delay decreases the hazard rate with 4.8%, which is a notable effect. The decrease in the hazard rate for larger reporting delays could be due to the fact that claims that are complicated take a while to report, resulting in a larger reporting delay. These claims lead to larger gap times, since they are also more complicated to handle by the insurer.

Among the other covariates, we observe that the history of the development process has a significant effect on the time until the next event. For claims that have been longer in the development process, the time to the next event is longer. This effect is large: a claim that has been in the development process for 5 years has a hazard rate that is 82% smaller than the hazard rate of a claim that has been in the development process for less than a year. A large proportion of the claim that have just been reported can be handled quickly, leading to a short time to the next event. The fact that the claims have been in the development period for a long time implies that some aspects of the claim make it a difficult claim to handle. Therefore, the expected time to the next event is longer than claims that have been in the development process for a short period of time.

Lastly, the coefficient of the old car indicator indicates that the time to the next event in the development process for an old car is longer than for a new car. The effect on the hazard rate for an old car is a decrease of 26%. Claims of regular cars are more likely to be similar than claims of old cars, as the repairs of old cars are more difficult. This makes handling a claim of an old car more complicated, which can lead to longer gap times between events.

6.1.4 Event Probability

The probability that an event is a payment or a settlement is determined in a specification with and without covariates. Table 6.4 shows the results of three different logistic regressions. The first model estimates the probability of a payment without covariates. The coefficient of our first model of -0.276 indicates that the probability of a payment is 43%. The second model contains all our covariates, whereas the third model shows the optimal model after using the backwards step-wise procedure. A comparison of the AIC of the first model, 657,364, and the AIC of the third model, 657,328, indicates that the inclusion of

covariates is useful for predicting the type of event.

The time of occurrence is not included in the optimal model, indicating that there is no significant relation between the time of occurrence and the probability of an event. By contrast, the reporting delay does have a significant impact. The probability of a settlement increases with 4% as a result of a one standard deviation increase in the reporting delay.

Moreover, the dependency of the event probability on the history of the development process is significant, since the coefficients for the years in the development period and the total payout are significant. Thus, including dependence between the parts of our model improves the fit of our model. A previous total payout which is larger than €2000 or smaller than €500 leads to an increase of the probability of a payment from 24% to 45% and 44%, respectively. Thus, the effect of the previous payout on the probability of a payment is economically significant. Furthermore, incorporating the policy characteristic of the age of the car improves the fit of our model as well. Claims arising from policies with an old car have a lower probability of a payment, indicating that there are on average less payments for those claims. This could be due to the fact that claims of old cars are not standard claims, such that someone needs to review the claim in person and thereby minimizes the number of payments. However, the effect on the probability is small, as it decreases the probability of a payment with 2 percentage points.

Table 6.4: Event Probability Coefficients

This table reports the coefficient estimates (standard deviation) of the covariates included in the null model, the full model and the optimal model. The null model and the full model indicate the models with no and all covariates, respectively. The optimal model is the model with the lowest AIC.

	Null Model	Full Model	Optimal Model
(Intercept)	-0.276***	-1.142***	-1.142***
	(0.002)	(0.009)	(0.009)
2 Years in Development		0.744***	0.744***
		(0.019)	(0.019)
3 Years in Development		1.208***	1.208***
		(0.036)	(0.036)
4 Years in Development		1.531***	1.532***
		(0.062)	(0.062)
5 Years in Development		1.449***	1.450***
		(0.081)	(0.081)
Old Car		-0.092***	-0.089***
		(0.011)	(0.011)
Number of Payments		0.001	
		(0.001)	
Catalog Value "Intermediate"		-0.023**	
		(0.006)	
Catalog Value "High"		0.010	
		(0.011)	
Total Payout $> €2000$		0.960***	0.960***
		(0.012)	(0.011)
Total Payout $< €500$		0.898***	0.898***
		(0.009)	(0.008)
Time of Occurrence		0.000	
		(0.001)	
Reporting Delay		-0.005***	-0.005***
		(0.000)	(0.000)
AIC	678,722	657,364	657,328

*Note: * p<.10, ** p<.05, *** p<0.01*

6.1.5 Payment Distribution

To model the payments during the development period, we make a distinction between payments that have been paid out during the development period and payments that have been paid out together with a settlement. In the first part of this section, we first discuss the payments model. Thereafter we discuss the settlement payments. For both types of payments, we first find the optimal model for modelling the payments without covariates. After we have found the optimal model without covariates, we will compare the fit of models with different combinations of covariates.

Payments

We compare the fit of multiple combinations of distributions to find the optimal model for the payments without covariates. In order to find the optimal model, we need to find the optimal number of distributions K and the optimal mix of the type of the K distributions. To do this, we compare the AIC of the different models after running the EM algorithm. We compare the AIC of a the mixture distributions for $K = 2, 3, 4$ and use different combinations of the Normal and Gamma distributions. The optimal model is given by a mixture of three Log Normal distributions, with an AIC of 533,686. The AIC of the optimal model for each K can be found in Table 8.6 of the Appendix.

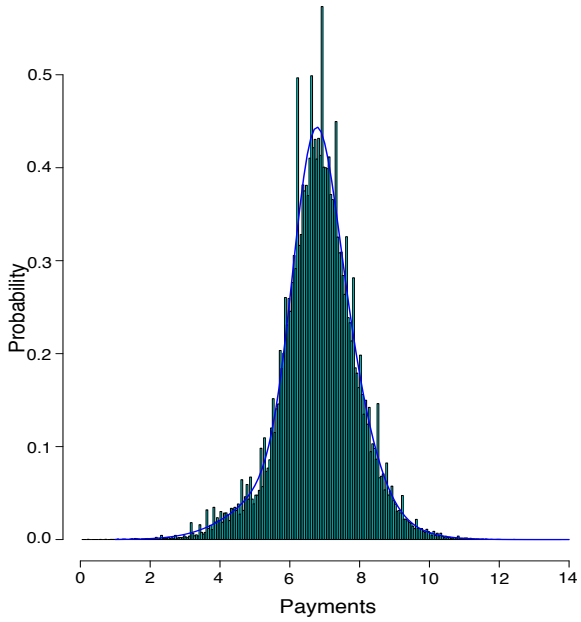
The parameter estimates of the model can be found in Table 6.5. To check whether our EM algorithm has found a local or a global maximum, we run the EM algorithm ten times with different starting values. We find the normal mixture model as displayed in Table 6.5 for each iteration. Hence, without taking covariates into account, we can best model the payments with the mixture distribution displayed in Table 6.5.

Table 6.5: Payments Mixture Model Parameter Estimates

This table reports the parameter estimates for the optimal mixture model, which consists of three Log-Normal distributions. The π_k's indicate the probability weights that are given to the distributions in the mixed distribution.

	μ_k	σ_k	π_k
$k = 1$	6.648	0.603	0.363
$k = 2$	6.555	1.545	0.312
$k = 3$	7.278	0.837	0.325

We can construct the mixture distribution arising from the coefficients in Table 6.5. In Figure 6.5, we observe that the mixture distribution closely follows the observed payments.



This figure displays the histogram of the payments during the development period together with the fitted mixture distribution consisting of three Log-Normal distributions given by the blue line.

Figure 6.5: Histogram and fitted mixture distribution of the payments

Now that we have found the optimal model without covariates, we examine whether incorporating covariates improves the fit of our model. For each distribution in the mixture model, we have three parameters to estimate, i.e. μ_k, σ_k and π_k, $k = 1, 2, 3$. We consider eight different specifications of the model: one model for each combination of μ_k, σ_k and π_k modelled either with or without covariates. An optimal solution would be to compare all specifications with all different combinations of covariates. However, due to the large amount of options that results, we choose to make two separate selections. We choose to first select which parameters should be modelled with covariates and afterwards determine the right covariates. We do this because the optimal set of covariates for one specification does not need to be the optimal set of covariates for the other specification, which leaves us with the problem on which model we should select our optimal set of covariates. To avoid

this problem, we choose to first select parameters. The model where only μ_k is estimated with covariates for each $k = 1, 2, 3$ has the lowest AIC. The AIC of all the models are given in Table 8.7.

Next, we use the backwards step-wise algorithm to determine our optimal set of covariates. The final parameter estimates for our model can be found in Table 6.6. The covariates that are in the model with the lowest AIC are the years in development, the total payout, the age of the car and the reporting delay. The full model can be found in the Appendix, Table 8.9.

Table 6.6: Payments Mixture Model with Covariates Coefficients

This table reports the coefficient estimates (standard deviation) obtained from the EM algorithm. The μ_k's are modelled with covariates, whereas σ_k and π_k are modelled without covariates for each k. The values for π_k are given in the lowest row as 'Probability'.

	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
(Intercept)	5.481***	7.353***	6.133***	0.315***	0.406***	0.281***
	(0.023)	(0.010)	(0.012)	(0.003)	(0.003)	(0.003)
2 Years in Development	0.065***	0.301***	0.491***			
	(0.033)	(0.015)	(0.018)			
3 Years in Development	0.496***	2.119***	-0.829***			
	(0.049)	(0.024)	(0.025)			
4 Years in Development	0.740***	0.783***	1.552***			
	(0.074)	(0.031)	(0.044)			
5 Years in Development	1.161***	0.603***	1.527***			
	(0.098)	(0.043)	(0.055)			
Total Payout $>$ €2000	-0.020	0.436***	2.054***			
	(0.028)	(0.012)	(0.015)			
Total Payout $<$ €500	1.09***	0.606***	0.942***			
	(0.023)	(0.010)	(0.012)			
Old Car	0.593***	-0.032	0.489***			
	(0.054)	(0.025)	(0.029)			
Reporting Delay	-0.002***	-0.001***	-0.006***			
	(0.000)	(0.000)	(0.000)			
Catalog Value "Intermediate"	0.154***	-0.011*	0.065***			
	(0.013)	(0.005)	(0.006)			
Catalog Value "High"	0.577***	-0.020*	0.310***			
	(0.022)	(0.010)	(0.011)			
Probability (π_k)	0.288	0.380	0.332			

*Note: * p<.10, ** p<.05, *** p<0.01*

First, the time the claim has spend in the development process has a positive effect on the mean of the distribution. So, a payment that takes place later in the development period is, on average, a larger payment. This could be explained by the fact that an insurer takes some time to verify large claims, therefore taking longer to pay out a large amount. For example, a claim reported in 2012 that is still open in 2015 indicates that there is a high probability that it is a complicated or large claim. Therefore, the average payments for that claim are expected to be higher. The mean of the mixture distribution for claims that have been in the development period for four years is 1,607 higher than the mean of the mixture distributions for claims that have been in the development process for less than one year. The effect of the covariates is calculated by taking the coefficients of the covariates into account when determining the mean of the mixture distribution, which is given by the weighted average of the means of the three Log-Normal distribution. The average payment is given by €1,582, which implies that the effect of the time the claim has spend in the development process has a large effect on the level of the payment.

Second, an old car has a higher average payment of €153 than a newer car. This could be caused by the fact that damage on an old car is more expensive to repair because of expensive parts. Third, the higher the catalog value of the car, the higher the average payment of a claim, which can be explained similarly. The effect of a car with an intermediate and high catalog value is small, increasing the mean of the mixture distribution with 17 and 113, respectively. Fourth, the total amount paid up to the moment of payment also has a significant impact on the average payment. If the total amount paid is above €2,000 or below €500, the average payment is €1592 and €1040 higher, respectively. This effect is large and could be explained by the fact that when that total amount paid is above €2,000, it is a large claim compared to the average total payout, €890. Therefore, following payments could on average also be higher. Furthermore, if the total payment is lower than €500, the payout is lower than average, indicating that the average next payment is probably higher than when the total payout is between €500 - 2000.

Lastly, the reporting delay is significant in the model. If the claim is reported early, the average payment is higher than when it is reported late. The effect is relatively small, as a one standard deviation decrease in the reporting delay increases the mean of the mixture distribution with 90. This could imply that if the claim concerns a larger amount of money, the policyholder is more likely to report sooner. Since he is willing to put in more effort in

order to get the money sooner compared to when it is a small claim. Based on these results, we conclude that incorporating dependency between the parts of our model, including the dependency within the development process, leads to a better fit of the mixture model for the payments.

Settlement Payments

For the settlement payments, we have the same approach as for the payments. However, we use the parameter ξ_0 to account for the probability that a payment is zero. We compare the fit of different values of K and different combinations of distributions for each $k = 1, 2, \ldots K$ together with the parameter ξ_0. Based on the AIC of these models, the optimal model for the settlement payments is given by the parameter ξ_0 for the zero probability together with a mixture of three Log-Normal distributions for the non-zero payments, similar to the regular payments. The AIC of the different models can be found in the Appendix, Table 8.8. The parameter estimates and probabilities of the different distributions are displayed in Table 6.7.

Table 6.7: Settlement Payments Mixture Model Parameter Estimates

This table reports the parameter estimates for the optimal mixture model, which consists of three Log-Normal distributions and an extra parameter for the probability at zero. The π_k's indicate the probability weights that are given to the distributions in the mixed distribution.

	μ_k	σ_k	π_k
ξ_0	-	-	0.393
$k = 1$	4.109	0.200	0.150
$k = 2$	6.401	1.139	0.318
$k = 3$	6.040	0.432	0.138

We compare the fit of the model without covariates to a model with covariates. In our mixture distribution, we have 10 parameters consisting of the μ_k, σ_k and π_k for each of the three normal distributions and ξ_0 to account for the probability of a zero payment. Each of the parameters can be chosen to be estimated with- or without covariates. Comparing the AIC of the different models, we find that the model that estimates μ_k for each k and ξ_0 all based on covariates has the best fit to our data. Hence, we take this model for modelling the settlement payments.

The model parameters of the optimal model can be found in Table 6.8. First, we analyse the effects of the history of the development period on the height of the settlement payment. The time in the development process and the total payout in the development period have a positive effect on the probability of a zero settlement, whereas the number of previous payments has a diminishing effect. The probability of a payment during settlement increases from 18% to 74% if the total payout is less than €500 compared to a total payout of €500-2000, which is a significant increase. A lower total payout before the settlement could indicate that there still needs to be paid out an amount of money, which leads to a higher probability of a payment during the settlement. Furthermore, the longer a claim has been in the development process, the higher the probability of a settlement without payment. Specifically, for claims that have been in the development process for 5 years, the probability of a zero payment is very high (99%). This indicates that these claims have been fully paid out before the settlement occurs. One explanation for this might be that the insurer wants to make sure everything is handled correctly and is fully paid before giving the order to settle the claim. The policyholder has already waited a long time for the claim to be settled, such that the insurer might want to make sure that it does not have to be reopened after settlement. The last result of the development process history is that the probability of a payment at the settlement date increases with the number of previous payments. A one standard deviation increase in the total number of previous payments at the moment of settlement increases the probability of a payment from 18% to 38%, which is a significant increase. This indicates that if a claim has experienced more payments, this process is expected to continue.

Second, the results of the reporting delay are consistent with our previous findings. We found that claims with a short reporting delay have a higher average height of the payments, which could be explained by the willingness of a policyholder to put in extra effort to report a claim with a large payout sooner. For a small claim, there might be less urgency for the policyholder to get his money back soon. This is consistent with the findings in this section, which indicate that claims with a short reporting delay have a lower probability of a zero payment at the settlement date and a higher average payment. The probability of a zero payment decreases with 1 percentage point as a result of a one standard deviation decrease in the reporting delay. Furthermore, the mean of the mixture distribution, given by the weighted average of the means of the three Log Normal distributions, increases with €34 in the case of a reporting delay that is one standard deviation smaller. As the average settlement payment is €432, these results indicate that the effect of the reporting delay on the settlement payment of a claim is small.

Third, an old car and a car with a higher catalog value have a higher average payment at settlement than a regular car and a car with a low catalog value, respectively. These results are consistent with our findings on the height of the regular payments. The effect of the old car is large, as the mean of the mixture distribution increases with €228 for an old car, which is an increase of more than 50% of the average settlement payment.

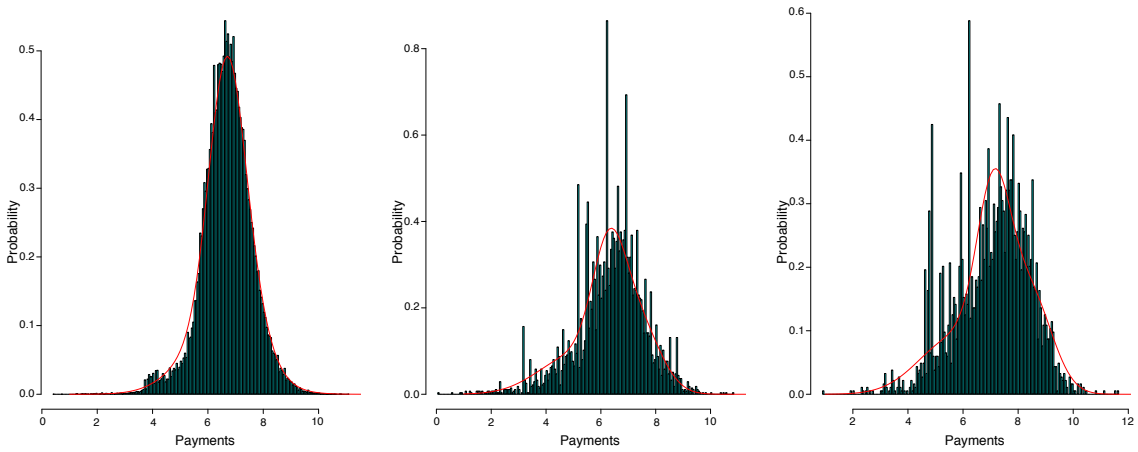
Table 6.8: Settlement Payments Mixture Model with Covariates Coefficients

This table reports the coefficient estimates (standard deviation) obtained from the EM algorithm. The μ_k's and ξ_0 are modelled with covariates, whereas σ_k and π_k are modelled without covariates for each k. The values for π_k are given in the lowest row as 'Probability'.

	ξ_0	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
(Intercept)	1.522***	4.287***	5.598***	7.102***	2.159***	0.2393***	2.405***
	(0.014)	(0.003)	(0.014)	(0.003)	(0.004)	(0.002)	(0.005)
2 Years in Development	1.198***	0.033***	0.047***	0.007			
	(0.036)	(0.008)	(0.031)	(0.008)			
3 Years in Development	3.235***	2.066	1.237***	0.526***			
	(0.109)	(0.020)	(0.069)	(0.020)			
4 Years in Development	5.847***	1.222***	1.759***	-0.887***			
	(0.257)	(0.033)	(0.132)	(0.030)			
5 Years in Development	12.480***	1.385***	2.635***	4.373***			
	(0.425)	(0.053)	(0.183)	(0.087)			
Old Car	-0.014*	0.705***	0.582***	0.039***			
	(0.007)	(0.007)	(0.031)	(0.008)			
Number of Payments	-1.217***	-0.003	-0.102***	-0.329***			
	(0.104)	(0.002)	(0.006)	(0.002)			
Catalog Value "Intermediate"	-0.137***	0.086***	0.470***	-3.111***			
	(0.009)	(0.001)	(0.007)	(0.002)			
Catalog Value "High"	-0.251***	0.666***	0.081***	-0.942***			
	(0.016)	(0.003)	(0.012)	(0.003)			
Total Payout > €2000	0.563***	0.005	0.514***	0.008			
	(0.017)	(0.004)	(0.073)	(0.004)			
Total Payout < €500	-2.579***	1.542***	0.094***	-0.006*			
	(0.012)	(0.003)	(0.013)	(0.003)			
Reporting Delay	0.001***	-0.001***	-0.001***	-0.001****			
	(0.000)	(0.000)	(0.000)	(0.000)			
Probability (π_k)		0.146	0.740	0.114			

*Note: * p<.10, ** p<.05, *** p<0.01*

Figure 6.6 shows the fit of different distributions based on covariates. We use three covariate groups for this figure, which are explained in the table below the figure. The value for the age of the car and the catalog value are kept constant. We choose these three covariate groups in order to examine the effects of small changes in the mean of the distribution. For example, the diminishing effect of the number of payments is one of the smaller effects compared to the other covariates. However, Figure 6.6(a) and 6.6(b) make it clear that there is a substantial difference between the two distributions and that it has added value to use a different distribution for these two groups. We learn from the three plots in Figure 6.6 that the distribution is able to adjust to the different payment patterns for the different covariate groups.



(a) Group 1 (b) Group 2 (c) Group 3

This figure displays the histograms of the settlement payments for three different groups of payments with specific covariate values, as explained in the table below, together with the fitted mixture distribution.

Figure 6.6: Histogram and fitted mixture distribution of the settlement payments

Table 6.9: Covariate Groups

This table explains the three groups of payments used in Figure (6.6)

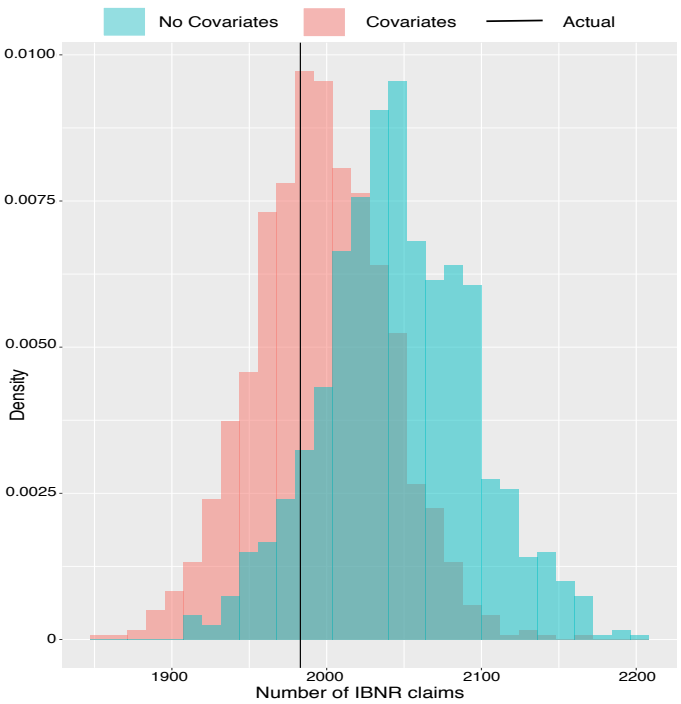
	Number of Payments	Development Process	Total Payout	Reporting Delay
Group 1	1	1	<€ 500	>5
Group 2	2	1	<€ 500	>5
Group 3	2	1	>€ 2000	<5

6.2 Simulation

In this section, we discuss the simulation results for the reserves. First, we analyze the results for the time of occurrence and the reporting delay for the IBNR claims. In the previous section, we concluded that covariates have an added value to the fit of our model. In this section, we explore the effect of the covariates on the simulations. First, we determine whether covariates add value to the time of occurrence and reporting delay simulations for IBNR claims. Thereafter, we inspect the simulation results of the development period for both IBNR and RBNS claims. We compare the mean, standard deviation and VaR of all simulations to check which parts of the model should be modelled with covariates.

6.2.1 Time of Occurrence

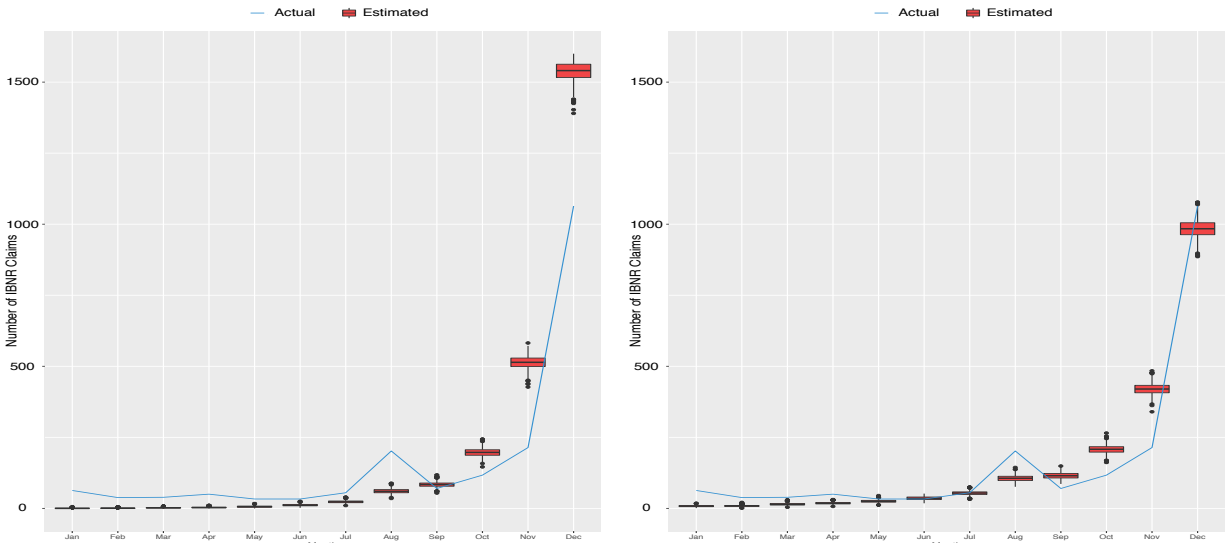
We simulate the number of IBNR claims that have occurred up until 2016. In Figure 6.7, the histograms of the simulated number of IBNR claims are shown. We observe that the model with covariates predicts the number of IBNR claims more accurately, as the simulations are more centered around the actual number of IBNR claims, given by the black line.



This figure displays the simulated number of IBNR claims with (pink) and without (blue) covariates.

Figure 6.7: Histograms of the simulated number of IBNR claims

Furthermore, the number IBNR claims can be determined for each month in which they occurred. In Figure 6.8, we observe that the model with covariates is more accurate in determining in which months the IBNR claims have occurred.



(a) Without Covariates (b)With Covariates

Panel (a) and Panel (b) of this figure show the estimated number of IBNR claims (red blocks) that originated in a specific month for each month of 2015 without and with covariates, respectively, together with the true number of IBNR claims incurred in each month (blue line).

Figure 6.8: Simulated monthly number of IBNR claims

Reporting Delay

The results of the simulation of the reporting delay with and without covariates are displayed in Table 6.10. We observe that both models underestimate the length of the reporting delay. The means of the models with and without covariates are respectively 7% and 26% less than the observed reporting delay. Moreover, the maximum observed reporting delay is significantly larger than that predicted by either model. Furthermore, the minimum reporting delay is underestimated as well. One potential explanation is that our model does not take into account New Year's Eve and New Year's Day. In reality, the probability of reporting a claim on these days is small, which leads to a longer reporting delay for the claims incurred on the last days of a year. Therefore, this might be the reason why the

observed minimum reporting delay is three days, whereas our model's is lower as it does not take into account these two days. Though both models underestimate the reporting delay, the model with covariates is more accurate than the model without covariates.

Table 6.10: Simulation Results Reporting Delay

This table provides the simulation results of the reporting delay with and without covariates. The columns 'Min', 'Mean', and 'Max' indicate the average of the minimum, mean and maximum simulated reporting delay over all simulations, respectively.

	Min	Mean	Max
Without Covariates	1.2	97	818
With Covariates	1.5	118	1195
Observed	3	131	1475

Development Process

We simulate the different elements of the development process with and without covariates to check whether incorporating covariates in all parts of the model adds value. Moreover, we compare the effects of including covariates for the different parts. First, we discuss the IBNR claims, followed by discussing the RBNS results.

The simulation results of the total payout from IBNR claims of the different models is displayed in Table 6.11. The first column reports the means of the different models. Compared to the observed IBNR payout in 2016, the worst performing model is the model without covariates, whereas the best performing model is the model with all parts of the model estimated and simulated with covariates. Besides the full model, the next three best models are models in which the event probability and the event hazard are estimated with covariates. This indicates that for IBNR claims, it is important to add covariates to the estimation of the probability of the event being a payment or a settlement.

The results of the standard deviation are similar across all models. The standard deviation seems to decline with the number of covariates added. For the models in which payments are modelled with covariates, the standard deviation is relatively lower than the other models. One explanation for this as follows. There is a large difference between the average payout of claims that have been in the development process for multiple years and claims that have just been reported. IBNR claims are by construction less than a year in the development process if we only consider payments in 2016. Therefore, there is a smaller variety in the

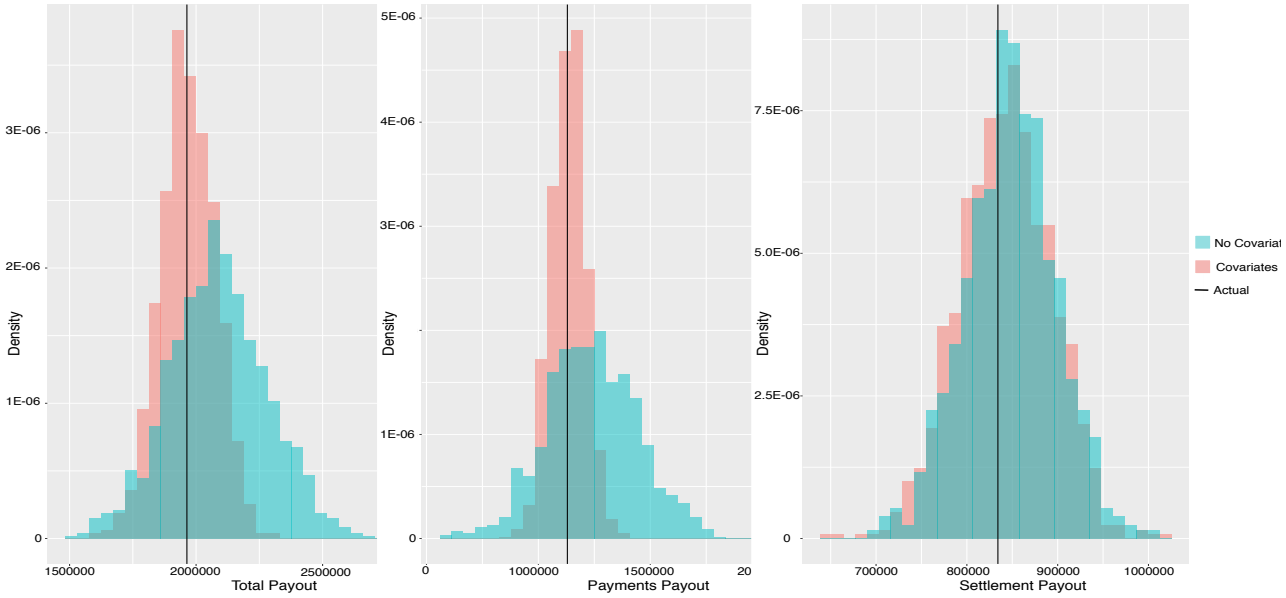
payment heights, as they all come from the same time in the development process. This leads to a smaller standard deviation for the IBNR reserves when the payments are modelled with covariates. Lastly, we observe that the VaR declines as more covariates are added to the model.

Table 6.11: Simulation Results for IBNR Reserve

This table reports the mean, standard deviation and $VaR_{0.995}$ of the sum of the simulated payments arising from IBNR claims in 2016. Each row represents a different combination of parts of the model that are simulated with covariates, where the last row indicates the actual total loss arising from IBNR claims in 2016.

	Mean	St. Dev.	$\text{VaR}_{0.995}$
No Covariates	2,087,156	196,322	3,834,995
Event Hazard	2,081,919	191,595	3,878,373
Settlement	2,068,084	199,043	3,777,222
Payments + Settlement	2,062,134	127,806	2,881,739
Event Probability	2,051,659	161,074	2,861,443
Payments	2,050,803	111,793	2,727,802
Event Hazard + Probability	2,044,655	133,556	2,681,421
Settlement + Event	1,995,968	134,943	2,638,207
Payment + Event	1,984,435	89,617	2,141,827
All Covariates	1,968,537	108,603	2,111,473
Actual	1,963,686		

To make the distinction between models without and with covariates more clear, we plot the simulation results of these two models in Figure 6.9. We observe that for the total reserve and the reserve coming from regular payments, incorporating covariates improves the mean estimation of the reserves as well as the standard deviation. For settlements, the added value is less significant. This coincides with the standard deviation results in Table 6.11.



(a) Total (b) Payments (c) Settlement

The first, second and third plot in this figure exhibit the histograms of simulations of the sum of all the payments, the sum of the regular payments, and the sum of the settlement payments in 2016 arising from IBNR claims, respectively. The pink histogram in each plot is given by simulations in which all parts of the claim process are modelled with covariates, whereas for the blue histogram every part is modelled without covariates.

Figure 6.9: Simulated total payout for IBNR claims

Next, we examine the results of the simulation of the RBNS reserve. First, we observe from the simulation results in Table 6.12 that the covariates add value in the estimation of reserves when considering the mean. Namely, the mean of the simulations of the model with no covariates is the furthest away from the actual RBNS reserve, whereas the model with all covariates has a mean closest to the actual loss in 2016. All four models in which the payments are estimated with covariates is closer to the mean than the models in which payments are not simulated with covariates. An explanation can be found when we consider the time the claim spend in the development period. For the RBNS claims, the amount of claims that have been in the development process for longer than a year is significantly larger than for IBNR claims. The average payment for IBNR payments in 2016 is 1,209, whereas the average payment of all payments in 2016 is 1,937. The last is significantly higher than the first, indicating that claims that are paid out in 2016 which are reported before

2016, i.e. the RBNS claims, have a higher average payout. The models without covariates for payments underestimate the RBNS payout, since they do not take into account the significantly higher payment level for claims that have been longer in the development period.

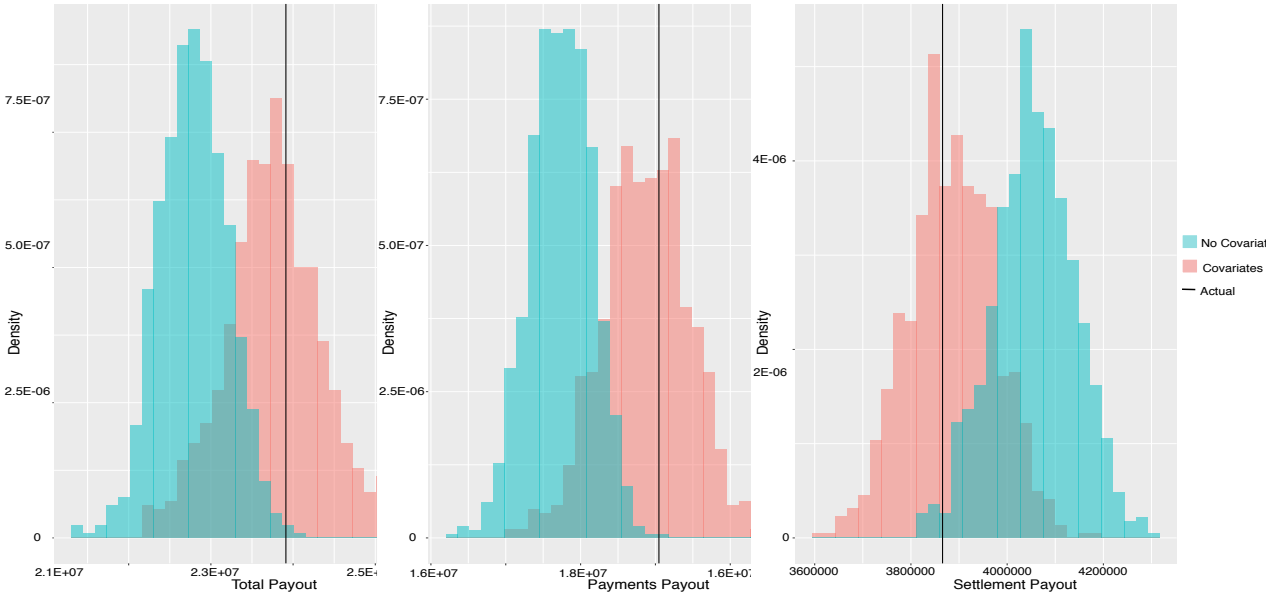
Second, the models in which the events are simulated with covariates have the lowest standard deviation. RBNS claims differ in the histories of their development processes: some claims have had multiple payments and have been in the development process a long time, whereas other claims await their first payment. These differences have an important impact on the probability of a settlement, since it might be that the former group of claims has a higher probability of a settlement than the latter. Therefore, incorporating covariates will give a higher probability of a settlement to the first group. By not including covariates, all claims are treated as equal and are thereby treated as if they just entered the development process. If covariates are incorporated, there thus is a larger variety in the events of the claims and thereby a larger standard deviation of the reserves. As opposed to the IBNR simulation results, the VaR increases as more covariates are added to the model.

Table 6.12: Simulation Results for RBNS Reserve

This table reports the mean, standard deviation and $VaR_{0.995}$ of the sum of the simulated payments in 2016 arising from RBNS claims. Each row represents a different combination of parts of the model that are simulated with covariates, where the last row indicates the actual total loss arising from RBNS claims in 2016.

	Mean	St. Dev.	$\text{VaR}_{0.995}$
No	21,789,969	417,191	22,694,731
Event Hazard	21,881,263	352,592	22,831,306
Event Probability	21,906,653	305,067	22,915,875
Event Hazard + Probability	21,945,462	387,995	23,071,588
Settlement + Event	22,075,386	400,512	23,176,598
Settlement	22,244,367	381,428	23,243,362
Payments	22,462,854	472,017	23,640,038
Payments + Settlement	22,525,988	446,815	23,872,620
Payments + Event	22,551,426	537,804	24,038,638
All	22,760,946	583,188	24,257,041
Actual	22,912,495		

In Figure 6.10, we observe the difference between the model with and without covariates clearly. Although the standard deviation of the model with covariates is larger than that of the model without covariates, the actual payout is located in the center of the distribution as opposed to the tail. Therefore, we conclude that including covariates adds value to our RBNS reserve estimation.

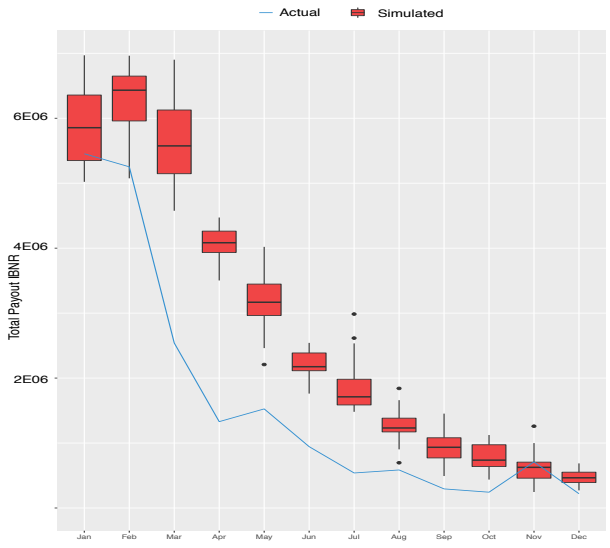


(a) Total (b) Payments (c) Settlement

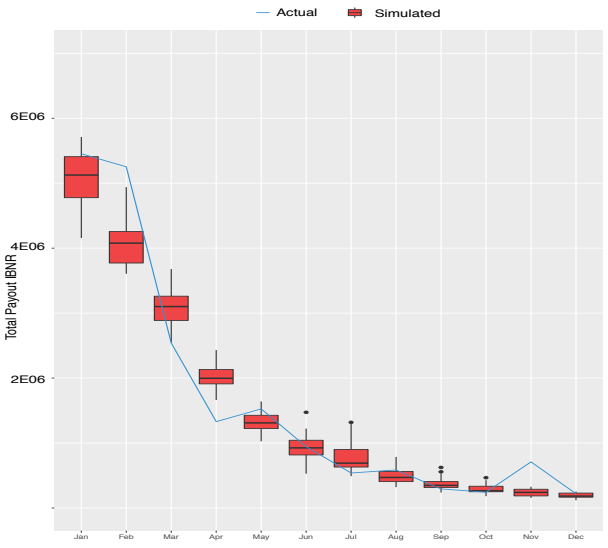
The first, second and third plot in this figure exhibit the histograms of simulations of the sum of all the payments, the sum of the regular payments, and the sum of the settlement payments arising from RBNS claims in 2016, respectively. The pink histogram in each plot is given by simulations in which all parts of the claim process are modelled with covariates, whereas for the blue histogram every part is modelled without covariates.

Figure 6.10: Simulated total payout for RBNS claims

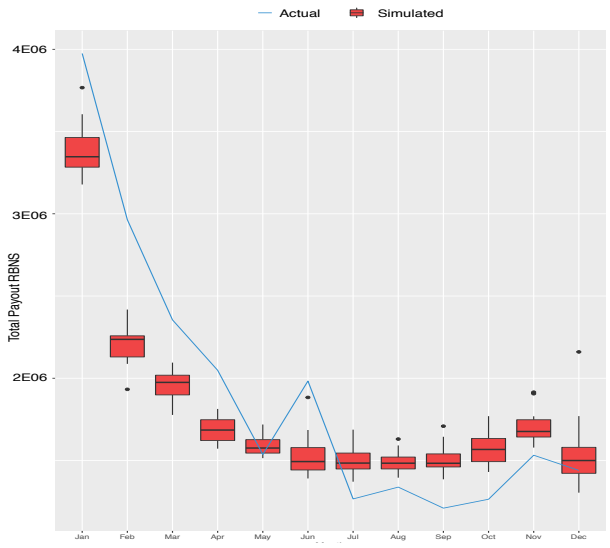
A main advantage of our model over the models currently used by insurance companies is the possibility of estimating the reserves for each month of the year as opposed to one reserve for the entire year. Figure 6.11 shows the cash flow for 2016 based on our simulations for both the RBNS and IBNR claims. For the IBNR reserves, we observe the added value of covariates for the reserve estimation, even on a monthly level. However, for the RBNS reserves, there is no added value to the monthly distribution of the payments by incorporating covariates.



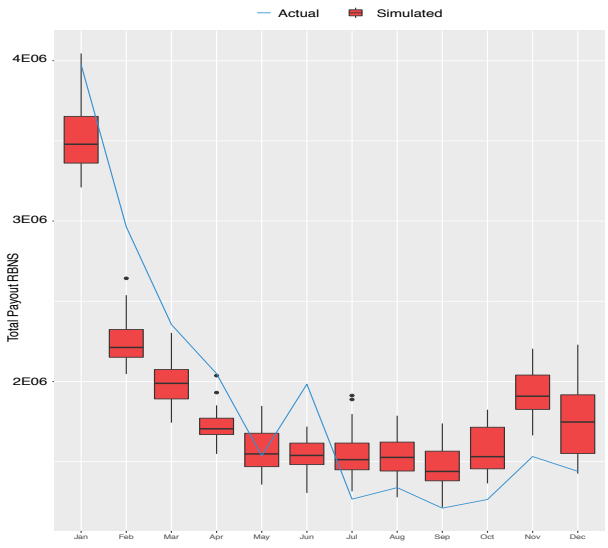
(a) IBNR without Covariates



(b) IBNR with Covariates



(c) RBNS without Covariates



(d) RBNS with Covariates

Panel (a) (Panel (c)) and Panel (b) (Panel(d)) display the sum of all payments arising from IBNR (RBNS) claims simulated without and with covariates, respectively, for each month in 2016 as indicated by the red blocks. The blue line in each plot indicates the observed total monthly payout arising from IBNR (RBNS) claims in 2016.

Figure 6.11: Comparison of the total monthly payout per claim type

Chapter 7

Final Remarks

7.1 Conclusion

In this paper, we compare different model specifications for an individual claim reserving model. First, we analyse the optimal model for the different parts of the claim process. We find that the time of occurrence of a claim and the reporting delay can be best specified in a piece-wise constant way. The time between events in the development process can be best modelled with a Weibull hazard rate as opposed to a piece-wise constant hazard rate. This allows for more simplicity in the model. To capture the different shapes of the distributions of the payments in the development process, a mixture distribution of three Log-Normal distributions is optimal. For modelling the settlement payments, in addition to using a mixture distribution of three Log-Normal distributions, it is optimal to account for the probability of a zero payment at the moment of settlement.

Second, we examine the impact of creating dependence between the different parts of our model. We create this dependence by including information about the claim process as covariates. Using this approach, more information of the claim is used in all parts of the process. First, we find that the inclusion of the time of occurrence does not lead to a better fit in any of the other parts of the model. Second, incorporating the reporting delay as a covariate in all following parts of the model leads to a better fit. The time between events increases for a larger reporting delay, whereas the average level of a payment, both with and without a settlement, decreases as the reporting delay increases. Third, the history of the development process has a significant effect on the timing and probability of events and the level of the accompanying payments. Therefore, we conclude that introducing dependence between different parts of the model improves its respective fit.

Lastly, we analyse whether incorporating policy characteristics leads to a better fit of the individual claim reserving model. We use two policy characteristics: the age of the car and the catalog value of the car. We find that the inclusion of the age of the car as a covariate improves the fit of all parts of the model. Besides, the catalog value of the car is included as a covariate for the reporting delay and the level of the payments.

We conclude that the inclusion of both policy characteristics and claim process characteristics improves the fit of our individual claim reserving model.

After analysing the optimal model specification for each part of the claim process, we perform simulations in order to examine whether incorporating covariates in each part of the model results in a higher accuracy of the prediction of IBNR and RBNS reserves. We find that incorporating covariates increases the ability of our individual claim reserving model to accurately predict the IBNR and RBNS reserves. Additionally, for IBNR claims we observe that the standard deviations of the models with covariates are lower and that accuracy of the monthly estimation of the reserves improves when using covariates. We conclude that the inclusion of covariates to all parts of our model increases the ability of our individual claim reserving model to accurately predict the IBNR and RBNS reserves.

One application of our model and findings is the application on current literature on individual claim reserving in a Bayesian setting. Arjas (1989) shows that a Bayesian approach can be used for individual claim reserving. However, a serious limitation of their approach is the usage of a non-parametric estimator for each part of their model, as this increases the computation time. They argue that some parts of the model could be better modelled using a parametric specification. Our results provide information on which parts of the model can be modelled parametrically in order to reduce the computation time of the Bayesian model. Furthermore, our paper can be extended in multiple ways. One example is to use more complicated structures for modelling the dependencies between the different parts of the model, such as copulas.

For insurance companies, the applications of this paper are threefold. First, the models that are currently used by insurance companies, which are based on aggregate data, are outdated as they are not aligned with the current data rich environment. Insurance companies do not use the enormous amount of data that is available to them for estimating

their reserves. Our paper provides evidence that the inclusion of policy characteristics and information on the claim process is useful for the reserve calculation by insurance companies. Second, our model introduces the link between the premium reserve and the claim reserve. An insurance company can use our findings to estimate both reserves in one model. This increases the coherency within their reserve calculations. Third, by modelling the reserves on an individual level, the reserve calculation becomes variable over time: the added value to the reserve can be estimated for each new policy in the portfolio of the insurer. This way, the reserves change over time as opposed to being constant for the time for which the reserve is estimated.

7.2 Limitations

Besides the advantages of our model discussed in the previous section, there are three important limitations that apply to our results.

The first limitation of our model is the absence of characteristics of the policy holders. Ideally, these covariates would have been included in our model. However, it is not feasible due to the limited amount of policies with such covariates. In practice, incorporating these characteristics can easily be implemented by including them in the original set of covariates.

A different limitation concerns selecting the optimal model among all possible specifications of the different parts of our model. For example, the time between events can be modelled with a parametric or a non-parametric specification, and both these specifications can be modelled with and without covariates. This leads to multiple model specifications. The number of possible models further increases as we incorporate more covariates, since there are multiple combinations of included covariates. In order to compare all possible options, we need to compute all models and afterwards compare their results. Performing these comparisons for all parts of our model would take a very long time. Therefore, we chose to make some of the decisions sequentially, such as first determining whether to use a parametric or non-parametric approach before checking which covariates improve the fit of the model. It would be optimal to compare all possible models instead of having multiple selection procedures sequentially. Due to the large amount of possible models, this is not feasible.

A final limitation of our model is the computation time of the simulations. By incorporating covariates in a part of the model, we increase the computation time of our simulations increases significantly. An insurance company needs to take the computation time into account when using this model.

Chapter 8

Appendix

8.1 Weibull Hazard Rate Proof

The Weibull hazard rate with covariates has the same form as the Weibull hazard rate without covariates,

$$\gamma(u|T_i, C_i) = \rho\zeta(\zeta u)^{\rho-1}\exp\{x_i'\beta\} \tag{8.1}$$

$$= \rho\zeta(\zeta u)^{\rho-1}\exp\{x_i'\beta(\rho-1)/(\rho)\}\cdot\exp\{x_i'\beta/\rho\} \tag{8.2}$$

$$= \rho\zeta\exp\{x_i'\beta/\rho\}(\zeta\exp\{x_i'\beta/p\}u)^{\rho-1} \tag{8.3}$$

$$= \rho\zeta^*(\zeta^* u)^{\rho-1}, \tag{8.4}$$

where $\zeta^* = \zeta\exp\{x_i'\beta/\rho\}$.

8.2 Occurrence Likelihood

In this section, we derive the likelihood of the claim occurrence process. Let the first P^* policies be policies that experience a claim and let the policies $P^* + 1, \ldots, P$ be policies that did not experience a claim from 0 to τ. For each $i = 1, \ldots P^*$, the probability density of policy i experiencing a claim at time t_i is given by

$$\lambda_0(t_i) \exp(x_i'\beta) \exp\bigg(- \int_0^{t_i} \exp(x_i'\beta) I_i(s) ds \bigg), \tag{8.5}$$

where $I_i(s)$ indicates whether policy i was active at time s. The complete derivation of this density can be found in Cook and Lawless (2007). Additional to policy i observing a claim at time t_i, it does not observe a claim from t_i to τ. The probability of policy i not observing a claim from t_i until τ is

$$P(N(\tau) - N(t_i) = 0 | C_i) = \exp\bigg(- \int_{t_i}^{\tau} \exp(x_i'\beta) I_i(s) ds \bigg). \tag{8.6}$$

We consider this probability since if policy i would experience a claim after t_i, it is considered as a new policy. In car insurance, this is considered standard practice, as a claim alters the terms of the policy. Therefore, if there is a claim with occurrence time t_i, there is no claim between t_i and τ for policy i. Furthermore, for policies $p = P^* + 1, \ldots, P$, the probability of not experiencing a claim from 0 to τ is given by

$$P(N(\tau) - N(0) = 0 | C_p) = \exp\bigg(- \int_0^{\tau} \exp(x_p'\beta) I_p(s) ds \bigg), \tag{8.7}$$

where $x_p = C_p$ indicate the covariates of policy p.

The joint likelihood of (1) P^* claims occurring at times t_i, $i = 1, \ldots, P^*$, (2) the same P^* claims having no claim between t_i until τ and (3) the other policies $P^*+1, \ldots, P$ experiencing no claim between 0 and τ, is given by

$$
\begin{aligned}
L(\lambda_0, \beta) &= \prod_{i=1}^{P^*} \left(\lambda_0(t_i) \exp(x_i'\beta) \exp\left(- \int_0^{t_i} \lambda_0(s) \exp(x_i'\beta) I_i(s) ds \right) \right. \\
&\quad \left. \times \exp\left(- \int_{t_i}^{\tau} \lambda_0(s) \exp(x_i'\beta) I_i(s) ds \right) \right) \prod_{p=P^*+1}^{P} \left(\exp\left(- \int_0^{\tau} \lambda_0(s) \exp(x_p'\beta) I_p(s) ds \right) \right) \\
&= \prod_{i=1}^{P^*} \left(\lambda_0(t_i) \exp(x_i'\beta) \exp\left(- \int_0^{\tau} \lambda_0(s) \exp(x_i'\beta) I_i(s) ds \right) \right) \\
&\quad \times \prod_{p=P^*+1}^{P} \left(\exp\left(- \int_0^{\tau} \lambda_0(s) \exp(x_p'\beta) I_p(s) ds \right) \right) \\
&= \prod_{i=1}^{P^*} \left(\lambda_0(t_i) \exp(x_i'\beta) \right) \\
&\quad \times \left(\prod_{i=1}^{P^*} \exp\left(- \int_0^{\tau} \lambda_0(s) \exp(x_i'\beta) I_i(s) ds \right) \right) \left(\prod_{p=P^*+1}^{P} \exp\left(- \int_0^{\tau} \lambda_0(s) \exp(x_p'\beta) I_p(s) ds \right) \right) \\
&= \prod_{i \geq 1} \left(\lambda_0(t_i) \exp(x_i'\beta) \right) \prod_{p=1}^{P} \exp\left(- \int_0^{\tau} I_p(s) \lambda_0(s) \exp(x_p'\beta) ds \right), \qquad (8.8)
\end{aligned}
$$

where $p = 1, \ldots P$ indicate all policies in the portfolio of the insurer.

8.3 Optimization Likelihood

8.3.1 Occurrence Rate

The likelihood of the time of occurrence is given by

$$\begin{aligned} L(\lambda_{1,0}, \dots \lambda_{l,0}, \beta) = \prod_{l \geq 1} \Bigg(\Bigg(\prod_{i \geq 1} \left(\lambda_{l,0} \exp(x_i' \beta)\right)^{I_{(a_{l-1} \leq < t_i < a_l)}} \Bigg) \\ \times \Bigg(\prod_{p \geq 1} \exp\left(- \lambda_{l,0} \int_{a_{l-1}}^{a_l} I_p(s) \exp(x_p' \beta) ds \right) \Bigg) \Bigg) \end{aligned} \tag{8.9}$$

In order to find the values of $\lambda_{l,0}$ and β that optimize (8.9), we first rewrite (8.9) to a log likelihood as

$$\begin{aligned} \log L(\lambda_{1,0}, \dots \lambda_{l,0}, \beta) &= \sum_{l \geq 1} \sum_{i \geq 1} I_{(a_{l-1} \leq t_i < a_l)} \log(\lambda_{l,0}) + \sum_{l \geq 1} \sum_{i \geq 1} I_{(a_{l-1} \leq t_i < a_l)} x_i' \beta \\ &\quad + \sum_{l \geq 1} \sum_{p \geq 1} \left(- \lambda_{l,0} \int_{a_{l-1}}^{a_l} I_p(s) \exp(x_p' \beta) ds \right) \\ &= \sum_{l \geq 1} \sum_{i \geq 1} I_{(a_{l-1} \leq t_i < a_l)} \log(\lambda_{l,0}) + \sum_{i \geq 1} x_i' \beta \\ &\quad - \sum_{l \geq 1} \lambda_{l,0} \sum_{p \geq 1} \left(\int_{a_{l-1}}^{a_l} I_p(s) \exp(x_p' \beta) ds \right). \end{aligned} \tag{8.10}$$

For each l, we take the derivative with respect to $\lambda_{l,0}$ and set it equal to zero. We get

$$\frac{\partial \log L(\lambda_{1,0}, \dots \lambda_{l,0}, \beta)}{\partial \lambda_{l,0}} = \frac{\sum_{i \geq 1} I_{(a_{l-1} \leq t_i < a_l)}}{\lambda_{l,0}} - \sum_{p \geq 1} \int_{a_{l-1}}^{a_l} I_p(s) \exp(x_p' \beta) ds = 0. \tag{8.11}$$

Solving (8.11) for $\lambda_{l,0}$ gives the MLE of $\lambda_{l,0}$ given β, which is given by

$$\hat{\lambda}_{l,0}(\beta) = \frac{\sum_{i \geq 1} I_{(a_{l-1} \leq t_i < a_l)}}{\sum_{p \geq 1} \int_{a_{l-1}}^{a_l} I_p(s) \exp(x_p' \beta) ds} \quad \text{for each } l = 1, 2, \dots \tag{8.12}$$

Now, we use the value for $\hat{\lambda}_{l,0}(\beta)$ in the log likelihood of (8.10) such that it is not a function of $\lambda_{l,0}$. We use i' instead of i to indicate the claims in $\hat{\lambda}_{l,0}(\beta)$ to avoid confusion of the summations. This gives

$$
\begin{aligned}
\log L(\beta) &= \sum_{l\geq 1}\sum_{i\geq 1} I_{(a_{l-1}\leq t_i<a_l)} \log\left(\frac{\sum_{i'\geq 1} I_{(a_{l-1}\leq t_{i'}<a_l)}}{\sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds}\right) \\
&+ \sum_{i\geq 1} x_i'\beta \\
&- \sum_{l\geq 1} \frac{\sum_{i'\geq 1} I_{(a_{l-1}\leq t_{i'}<a_l)}}{\sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds} \sum_{p\geq 1}\left(\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds\right) \\
&= \sum_{l\geq 1}\sum_{i\geq 1} I_{(a_{l-1}\leq t_i<a_l)} \log \sum_{i'\geq 1} I_{(a_{l-1}<t_{i'}<a_l)} \\
&- \sum_{l\geq 1}\sum_{i\geq 1} I_{(a_{l-1}\leq t_i<a_l)} \log\left(\sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds\right) \\
&+ \sum_{i\geq 1} x_i'\beta - \sum_{l\geq 1}\sum_{i'\geq 1} I_{(a_{l-1}\leq t_{i'}<a_l)} \\
&\propto \sum_{i\geq 1} x_i'\beta - \sum_{l\geq 1}\sum_{i\geq 1} I_{(a_{l-1}\leq t_i<a_l)} \log\left(\sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds\right) \\
&= \sum_{i\geq 1} x_i'\beta - \sum_{i\geq 1} \log\left(\sum_{l\geq 0} I_{(a_{l-1}\leq t_i<a_l)} \sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds\right) \qquad (8.13)
\end{aligned}
$$

Taking the first derivative with respect to β and setting it equal to zero yields

$$
\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i\geq 1}\left(x_i' - \frac{\sum_{l\geq 1} I_{(a_{l-1}\leq t_i<a_l)} \sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)x_p'ds}{\sum_{l\geq 1} I_{(a_{l-1}\leq t_i<a_l)} \sum_{p\geq 1}\int_{a_{l-1}}^{a_l} I_p(s)\exp(x_p'\beta)ds}\right) = 0. \qquad (8.14)
$$

We find the MLE for β and $\lambda_{l,0}$ by solving (8.14) for β and using it in the definition of $\hat{\lambda}_{l,0}(\beta)$ as in (8.11) to find $\hat{\lambda}_{l,0}$.

8.3.2 Reporting Delay Hazard

The likelihood of the reporting delay is given by

$$
\begin{aligned}
L(\gamma_{1,0},\ldots\gamma_{w,0},\beta) = \prod_{w\geq 1}\prod_{i\geq 1}\Bigg(&\left(\gamma_{w,0}\exp(x_i'\beta)\right)^{I_{(q_{w-1}\leq u_i<q_w)}} \\
&\times \exp\left(-\gamma_{w,0}\int_{q_{w-1}}^{q_w} I_i(s)\exp(x_i'\beta)ds\right)\Bigg)
\end{aligned} \qquad (8.15)
$$

In order to find the values of $\lambda_{l,0}$ and β that optimize (8.9), we first rewrite (8.15) to a log likelihood as

$$\begin{aligned}\log L(\theta) &= \sum_{w\geq 1}\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}\log(\gamma_{w,0}) + \sum_{w\geq 1}\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}x_i'\beta \\ &\quad - \sum_{w\geq 1}\sum_{i\geq 1}\left(\gamma_{w,0}\int_{q_{w-1}}^{q_w} I_i(s)\exp(x_i'\beta)ds\right) \\ &= \sum_{w\geq 1}\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}\log(\gamma_{w,0}) + \sum_{i\geq 1} x_i'\beta \\ &\quad - \sum_{w\geq 1}\gamma_{w,0}\sum_{i\geq 1}\left(\int_{q_{w-1}}^{q_w} I_i(s)\exp(x_i'\beta)ds\right)\end{aligned} \tag{8.16}$$

For each w, we take the derivative with respect to $\gamma_{w,0}$,

$$\frac{\partial \log L(\gamma_{1,0},\ldots\gamma_{w,0},\beta)}{\partial\gamma_{w,0}} = \frac{\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}}{\gamma_{w,0}} - \sum_{i\geq 1}\exp(x_i'\beta)\int_{q_{w-1}}^{q_w} I_i(s)ds = 0. \tag{8.17}$$

Solving (8.17) for $\gamma_{w,0}$ gives

$$\hat{\gamma}_{w,0}(\beta) = \frac{\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}}{\sum_{i\geq 1}\exp(x_i'\beta)\int_{q_{w-1}}^{q_w} I_i(s)ds} \quad \text{for each } w = 1,2,\ldots \tag{8.18}$$

We plug the value of $\hat{\gamma}_{w,0}$ as in (8.18), with i' instead of i to indicate the claims to avoid confusion in the summations, in the log likelihood of (8.16) to obtain $\log L(\beta)$

$$\begin{aligned}\log L(\beta) &= \sum_{w\geq 1}\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}\log\left(\frac{\sum_{i'\geq 1} I_{(q_{w-1}\leq u_{i'}<q_w)}}{\sum_{i'\geq 1}\exp(x_{i'}'\beta)\int_{q_{w-1}}^{q_w} I_{i'}(s)ds}\right) + \sum_{i\geq 1} x_i'\beta \\ &\quad - \sum_{w\geq 1}\frac{\sum_{i'\geq 1} I_{(q_{w-1}\leq u_{i'}<q_w)}}{\sum_{i'\geq 1}\exp(x_{i'}'\beta)\int_{q_{w-1}}^{q_w} I_{i'}(s)ds}\sum_{i\geq 1}\exp(x_i'\beta)\int_{q_{w-1}}^{q_w} I_i(s)ds \\ &\propto \sum_{i\geq 1} x_i'\beta - \sum_{w\geq 1}\sum_{i\geq 1} I_{(q_{w-1}\leq u_i<q_w)}\log\left(\sum_{i'\geq 1}\exp(x_{i'}'\beta)\int_{q_{w-1}}^{q_w} I_{i'}(s)ds\right) \\ &= \sum_{i\geq 1}\left(x_i'\beta - \log\left(\sum_{w\geq 1} I_{(q_{w-1}\leq u_i<q_w)}\sum_{i'\geq 1}\exp(x_{i'}'\beta)\int_{q_{w-1}}^{q_w} I_{i'}(s)ds\right)\right)\end{aligned} \tag{8.19}$$

Taking the derivative with respect to β and setting it equal to zero gives

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i\geq 1}\left(x_i' - \frac{\sum_{w\geq 1} I_{(q_{w-1}\leq u_i<q_w)} \sum_{i'\geq 1}\int_{q_{w-1}}^{q_w} I_{i'}(s)\exp(x_{i'}'\beta)x_{i'}' ds}{\sum_{w\geq 1} I_{(q_{w-1}\leq u_i<q_w)} \sum_{i'\geq 1}\int_{q_{w-1}}^{q_w} I_{i'}(s)\exp(x_{i'}'\beta) ds}\right) = 0. \quad (8.20)$$

We get the MLE's for β and $\gamma_{w,0}$ by solving (8.20) for β to obtain $\hat{\beta}$ and use $\hat{\beta}$ in (8.18) to obtain $\hat{\gamma}_{w,0}$.

8.3.3 Event Hazard

The likelihood of the event gap times is defined as

$$\begin{aligned} L(\phi_{1,0},\dots\phi_{z,0},\beta) = \prod_{z\geq 1}\prod_{i\geq 1}\Bigg(\prod_{j\geq 1}(\phi_{z,0}\exp(x_{ij}'\beta))^{I_{(r_{z-1}\leq v_{ij}<r_z)}\delta_{ij}} \\ \times \exp\Bigg(-\phi_{z,0}\exp(x_{ij}'\beta)\int_{r_{z-1}}^{r_z} I_i(s)ds\Bigg)\Bigg). \end{aligned} \quad (8.21)$$

Then, we can write the log likelihood as

$$\begin{aligned} \log L(\phi_{1,0},\dots\phi_{z,0},\beta) &= \sum_{z\geq 1}\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij} I_{(r_{z-1}\leq v_{ij}<r_z)}\log(\phi_{z,0}) \\ &+ \sum_{z\geq 1}\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij} I_{(r_{z-1}\leq v_{ij}<r_z)} x_{ij}'\beta \\ &- \sum_{z\geq 1}\sum_{i\geq 1}\sum_{j\geq 1}\left(\phi_{z,0}\exp(x_{ij}'\beta)\int_{r_{z-1}}^{r_z} I_i(s)ds\right) \\ &= \sum_{z\geq 1}\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij} I_{(r_{z-1}\leq v_{ij}<r_z)}\log(\phi_{z,0}) + \sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij}x_{ij}'\beta \\ &- \sum_{z\geq 1}\phi_{z,0}\Bigg(\sum_{i\geq 1}\exp(x_{ij}'\beta)\int_{r_{z-1}}^{r_z} I_i(s)ds\Bigg)\Bigg). \end{aligned} \quad (8.22)$$

Then, for each l, we take the derivative with respect to $\phi_{z,0}$,

$$\frac{\partial \log L(\phi_{1,0},\dots\phi_{z,0},\beta)}{\partial \phi_{z,0}} = \frac{\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij} I_{(r_{z-1}\leq v_{ij}<r_z)}}{\phi_{z,0}} - \sum_{i\geq 1}\exp(x_{ij}'\beta)\int_{r_{z-1}}^{r_z} I_i(s)ds = 0. \quad (8.23)$$

Solving (8.11) for $\phi_{z,0}$ gives

$$\hat{\phi}_{z,0}(\beta) = \frac{\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij} I_{(r_{z-1}\leq v_{ij}<r_z)}}{\sum_{i\geq 1}\exp(x_{ij}'\beta)\int_{r_{z-1}}^{r_z} I_i(s)ds} \quad \text{for each } z = 1, 2, \dots \quad (8.24)$$

Then, we plug in the value of $\hat{\phi}_{z,0}$ as in (8.24), with i' instead of i and j' instead of j to indicate the claims and events to avoid confusion in the summations, in the log likelihood in (8.22) to obtain $\log L(\beta)$,

$$\begin{aligned}\log L(\beta) &= \sum_{z\geq 1}\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij}I_{(r_{z-1}\leq v_{ij}<r_z)}\log\left(\frac{\sum_{i'\geq 1}\sum_{j'\geq 1}\delta_{i'j'}I_{(r_{z-1}\leq v_{ij'}<r_z)}}{\sum_{i'\geq 1}\exp(x'_{i'j'}\beta)\int_{r_{z-1}}^{r_z}I_{i'}(s)ds}\right)+\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij}x'_{ij}\beta \\ &\quad -\sum_{z\geq 1}\frac{\sum_{i'\geq 1}\sum_{j'\geq 1}\delta_{i'j'}I_{(r_{z-1}\leq v_{i'j'}<r_z)}}{\sum_{i'\geq 1}\exp(x'_{i'j'}\beta)\int_{r_{z-1}}^{r_z}I_{i'}(s)ds}\left(\sum_{i\geq 1}\exp(x'_{ij}\beta)\int_{r_{z-1}}^{r_z}I_i(s)ds\right)\Bigg) \\ &\propto \sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij}x'_{ij}\beta-\sum_{z\geq 1}\sum_{i\geq 1}\sum_{j\geq 1}\delta_{ij}I_{(r_{z-1}\leq v_{ij}<r_z)}\log\left(\sum_{i'\geq 1}\exp(x'_{i'j'}\beta)\int_{r_{z-1}}^{r_z}I_{i'}(s)ds\right) \\ &= \sum_{i\geq 1}\sum_{j\geq 1}\left(\delta_{ij}x'_{ij}\beta-\log\left(\sum_{z\geq 1}\delta_{ij}I_{(r_{z-1}\leq v_{ij}<r_z)}\sum_{i'\geq 1}\exp(x'_{i'j'}\beta)\int_{r_{z-1}}^{r_z}I_{i'}(s)ds\right)\right).\end{aligned} \tag{8.25}$$

Taking the derivative with respect to β and setting it equal to zero gives

$$\frac{\partial\log L(\beta)}{\partial\beta}=\sum_{i\geq 1}\sum_{j\geq 1}\left(\delta_{ij}x'_{ij}-\frac{\sum_{z\geq 1}I_{(r_{z-1}\leq v_{ij}<r_z)}\sum_{i'\geq 1}\int_{r_{z-1}}^{r_z}I_{i'}(s)\exp(x'_{i'j'}\beta)x'_{i'j'}ds}{\sum_{z\geq 1}I_{(r_{z-1}\leq v_{ij}<r_z)}\sum_{i'\geq 1}\int_{r_{z-1}}^{r_z}I_{i'}(s)\exp(x'_{i'j'}\beta)ds}\right)=0. \tag{8.26}$$

We get the MLE's for β and $\phi_{z,0}$ by solving (8.26) for β to obtain $\hat{\beta}$ and using this in (8.24) to obtain $\hat{\phi}_{z,0}$.

8.3.4 Weibull Hazard Rate

Algorithm 5 Weibull Hazard Rate

procedure INITIALIZATION
    $\rho^{(0)}, \zeta^{(0)} \leftarrow$ MLE of Weibull distribution
    Old Likelihood $\leftarrow$ Log Likelihood of Weibull distribution with $\rho^{(0)}$ and $\zeta^{(0)}$
    $\rho^{(1)}, \zeta^{(1)} \leftarrow$ Update $\rho^{(0)}, \zeta^{(0)}$ with Newton Raphson
    New Likelihood $\leftarrow$ Log Likelihood of Weibull distribution with $\rho^{(1)}$ and $\zeta^{(1)}$
    $b \leftarrow 1$
procedure NEWTON RAPHSON ALGORITHM 1
    while |New Likelihood - Old Likelihood| $> \epsilon$ **do**
        $b = b + 1$
        Old Likelihood $\leftarrow$ New Likelihood
        $\rho^{(b)}, \zeta^{(b)} \leftarrow$Update $\rho^{(b-1)}, \zeta^{(b-1)}$ with Newton Raphson
        New Likelihood $\leftarrow$ Log Likelihood evaluated with $\rho^{(b)}$ and $\zeta^{(b)}$
procedure INITIALIZATION 2
    $\rho^{(0)} \leftarrow \rho^{(b)}$
    $\beta^{(0)} \leftarrow$ Coefficients of linear regression of dependent variable on x_i
    $\zeta^{(0)} \leftarrow$ Intercept of linear regression of dependent variable on x_i
    Old Likelihood $\leftarrow$ Log of the Likelihood of Weibull distribution with covariates as in (4.18) with $\rho^{(0)}, \zeta^{(0)}$ and $\beta^{(0)}$
    Update $\rho^{(0)}, \zeta^{(0)}$ and $\beta^{(0)}$ with Newton Raphson
    New Likelihood $\leftarrow$ Log Likelihood of Weibull distribution with $\rho^{(1)}, \zeta^{(1)}$ and $\beta^{(1)}$
    $b \leftarrow 1$
procedure NEWTON RAPHSON ALGORITHM 2
    while |New Likelihood - Old Likelihood| $> \epsilon$ **do**
        $b = b + 1$
        Old Likelihood $\leftarrow$ New Likelihood
        Update $\rho^{(b)}, \zeta^{(b)}$ and $\beta^{(b)}$ with Newton Raphson
        New Likelihood $\leftarrow$ Likelihood evaluated $\rho^{(b)}, \zeta^{(b)}$ and $\beta^{(b)}$
Final values are given by $\rho^{(b)}, \zeta^{(b)}$ and $\beta^{(b)}$

8.3.5 Event Probability

The values for β in that optimize the likelihood in (4.28) are found by iteratively reweighted least squares. The algorihtm is shown below.

Algorithm 6 Iteratively Reweighted Least Squares

procedure INITIALIZATION
    $n \leftarrow$ Total number of events
    $\psi^{(0)} \leftarrow$ Vector with weights of length n with 1's
    $\beta^{(0)} \leftarrow$ Coefficients from weighted regression with $\psi^{(0)}$ as weights
    Old Likelihood $\leftarrow$ Log Likelihood of (4.28) with $\beta^{(0)}$
    $\psi^{(1)} \leftarrow$ Update $\psi^{(0)}$ with Newton Raphson
    $\beta^{(1)} \leftarrow$ Coefficients from weighted regression with $\psi^{(1)}$ as weights
    New Likelihood $\leftarrow$ Log Likelihood of (4.28) with $\beta^{(1)}$
    $b \leftarrow 1$
procedure NEWTON RAPHSON ALGORITHM 1
    while |New Likelihood - Old Likelihood| $> \epsilon$ **do**
        $b = b + 1$
        Old Likelihood $\leftarrow$ New Likelihood
        $\psi^{(b)} \leftarrow$ Update $\psi^{(b-1)}$ with Newton Raphson
        $\beta^{(b)} \leftarrow$ Coefficients from weighted regression with $\psi^{(b)}$ as weights
        New Likelihood $\leftarrow$ Log Likelihood of (4.28) with $\beta^{(b)}$
Final values are given by $\beta^{(b)}$

8.4 EM Algorithm Payments

8.4.1 Payments

The algorithm used for payments is given by

Algorithm 7 Payment Distribution

The following EM algorithm is used to find $\beta_{\mu,k}, \beta_{\sigma,k} and \beta_{\pi,k}$ for each distribution k.

procedure INITIALIZATION
  $\pi_k^{(0)} \leftarrow \frac{1}{K}$ for each k
  $N \leftarrow \sum_{i,j} Y_{ij}$ (total number of payments)
  $\psi_k^{(0)} \leftarrow$ RBinominal(N, $\pi_k^{(1)}$, $1 - \pi_k^{(1)}$) for each k
  $b \leftarrow 0$
while |New Likelihood - Old Likelihood| $> \epsilon$ **do**
  $b = b + 1$
  Old Likelihood $\leftarrow$ New Likelihood
  procedure ESTIMATION OF DISTRIBUTION PARAMETERS
    for $k = 1 : K$ **do**
      Set starting values:
      $\mu_{k,NR} \leftarrow$ MLE for μ_k of distribution k with weights $\psi_k^{(b-1)}$
      $\sigma_{k,NR} \leftarrow$ MLE for σ_k of distribution k with weights $\psi_k^{(b-1)}$
      NL $\leftarrow$ Log Likelihood of distribution k evaluated with $\mu_{k,NR}$ and $\sigma_{k,NR}$
      OL $\leftarrow 0$
      while |NL - OL| $> \epsilon$ **do**
        OL $\leftarrow$ NL
        Update $\mu_{k,NR}$ and $\sigma_{k,NR}$ with Newton Raphson
        Evaluate $\beta_{\mu,k}$ and $\beta_{\sigma,k}$ in (4.31) by a GAM model with weights $\psi_k^{(b-1)}$
        Store fitted values $\hat{\mu}_{k,NR} = g^{-1}(x'_{ij}\beta_{\mu,k})$ and $\hat{\sigma}_{k,NR} = g^{-1}(x'_{ij}\beta_{\sigma,k})$
        NL $\leftarrow$ Log Likelihood of distribution k evaluated with $\hat{\mu}_{k,NR}$ and $\hat{\sigma}_{k,NR}$
      $\hat{\mu}_k^{(b)} \leftarrow \hat{\mu}_{k,NR}$
      $\hat{\sigma}_k^{(b)} \leftarrow \hat{\sigma}_{k,NR}$
      $\beta_{\mu,k}^{(b)} \leftarrow \beta_{\mu,k}$
      $\beta_{\sigma,k}^{(b)} \leftarrow \beta_{\sigma,k}$
      $\log L_k \leftarrow$ NL
  procedure MAXIMIZE LIKELIHOOD FOR PROBABILITIES
    $\psi_k^{(b)} \leftarrow \left(\frac{\pi_k f_k(x_1)}{\sum_{k'=1}^{K} \pi'_k f_{k'}(x_1)}, \ldots \frac{\pi_k f_k(x_N)}{\sum_{k'=1}^{K} \pi'_k f_{k'}(x_N)} \right)$ for each k
  Evaluate $\beta_{\pi,k}^{(b)}$ in (4.32) by performing a multinominal logistic regression with weights $\psi^{(b)} = (\psi_1^{(b)}, \ldots, \psi_k^{(b)})$
  Store fitted values $\hat{\pi}_k^{(b+1)}$ from the multinominal logistic regression
  New Likelihood $\leftarrow$ Log of Likelihood in (4.34) evaluated with $\hat{\mu}_k^{(b)}$, $\hat{\sigma}_k^{(b)}$, and $\hat{\pi}_k^{(b)}$
$\beta_{\mu,k} \leftarrow \beta_{\mu,k}^{(b)}$
$\beta_{\sigma,k} \leftarrow \beta_{\sigma,k}^{(b)}$
$\beta_{\pi,k} \leftarrow \beta_{\pi,k}^{(b)}$

8.4.2 Settlement

The following EM algorithm is used to find $\beta_{\mu,k}, \beta_{\sigma,k}$ and $\beta_{\pi,k}$ for each distribution k.

Algorithm 8 Settlement Distribution

procedure INITIALIZATION
    $N \leftarrow \sum_{i,j} Y_{ij}$ (total number of payments)
    $\psi_0 \leftarrow$ Vector of length N
    $\psi_0 \leftarrow$ ifelse$(Y_{ij} = 0, 1, 0)$ for each i, j
    $\xi_0^{(1)} \leftarrow$ MLE of Binominal Distribution with weights ψ_0
    $b \leftarrow 0$
while |New Likelihood - Old Likelihood| $> \epsilon$ **do**
    $b = b + 1$
    Old Likelihood $\leftarrow$ New Likelihood
    Update $\xi_0^{(b)}$ with Newton Raphson
    Evaluate β_{ξ_0} in (4.35) by a GAM model
    New Likelihood $\leftarrow$ Log Likelihood evaluated with fitted values $\hat{\xi}_0^{(b)} = g^{-1}(x'_{ij}\beta_{\xi_0})$
$\beta_{\xi_0} \leftarrow \beta_{\xi_0}^{(b)}$
$\psi_k^{(b)}[\psi_0 = 1] \leftarrow 0$ for all b
Continue with Algorithm 7

8.5 Simulation of Hazard Rate

In this section, we discuss how we can simulate the reporting delay with a piece-wise constant and a Weibull hazard rate. We will only discuss simulating the reporting delay here, however, the same methods are applied for simulating the time until a next event. First, we discuss how we can simulate from a general hazard rate. Thereafter, we will apply the found simulation equations to the piece-wise constant and Weibull hazard rate.

The hazard rate $\gamma(u)$ is related to F(u) as

$$F(u) = 1 - \exp(-\int_0^u \gamma(s)ds) \tag{8.27}$$

Let $Y \sim$ Uni(0,1) be a randomly drawn variable. We have

$$\begin{aligned} Y &= F(u) && (8.28)\\ &= 1 - \exp\big(-\int_0^u \gamma(s)ds\big) && (8.29)\\ &= 1 - \exp(-H(u)). && (8.30) \end{aligned}$$

Inversing the equation gives

$$u = H^{-1}(-\log(1-Y)). \tag{8.31}$$

Hence, by using Y in (8.31), we obtain the value for the reporting delay u. Equivalently, we can draw a random Uniform variable and use this in $H^{-1}(-\log(Y))$ to obtain an estimate for the reporting delay.

In the case of a model with covariates, we need to incorporate the covariates into the simulation. The hazard rate is then written as

$$\begin{aligned} F(u) &= 1 - \exp\left(\int_0^u \gamma(s)\exp(x_i'\beta)ds\right) && (8.32)\\ &= 1 - \exp\left(\exp(x_i'\beta)\int_0^u h(s)ds\right) && (8.33)\\ &= 1 - \exp\left(\exp(x_i'\beta)H(u)\right) && (8.34) \end{aligned}$$

We can rewrite $F(u) = Y$ with the above found relation to obtain

$$u = \frac{H^{-1}(-\log(Y))}{\exp(x_i'\beta)}. \tag{8.35}$$

We have found the general simulation equations for a hazard rate $\gamma(u)$. Next, we consider the two hazard rates that we apply: the piece-wise constant hazard rate and the Weibull hazard rate. To obtain the simulation equations for these two hazard rates, we need to find H^{-1}, such that we can use it in the simulation equations (8.31) and (8.35).

For the Weibull distribution, the hazard function without covariates is given by $\gamma(u) = \rho\zeta(\zeta u)^{\rho-1}$. The cumulative hazard and its inverse are computed as

$$H(u) = \int_0^u \rho\zeta(\zeta s)^{\rho-1} ds \tag{8.36}$$

$$= (\zeta u)^{\rho} \tag{8.37}$$

$$H^{-1} = (\zeta u)^{1/\rho} \tag{8.38}$$

The hazard rate and the cumulative hazard with a piece-wise constant specification are given by

$$\gamma(u) = \begin{cases} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_W \end{cases} \qquad H(u) = \begin{cases} \gamma_1 u & \text{for } 0 \le u < q_1 \\ \gamma_1 q_1 + \gamma_2(u - q_1) & \text{for } q_1 \le u < q_2 \\ \vdots & \\ \gamma_1 q_1 + \gamma_2(q_2 - q_1) + \cdots + \gamma_W(u - q_W) & \text{for } q_{W-1} \le u < q_W) \end{cases}$$

By inverting the cumulative hazard function, we get

$$H^{-1}(u) = \begin{cases} u/\gamma_1 & \text{for } 0 \le u < \gamma_1 q_1 \\ q_1 + (u - \gamma_1 q_1)/\gamma_2 & \text{for } \gamma_1 q_1 \le u < \gamma_1 q_1 + \gamma_2(q_2 - q_1) \\ \vdots & \\ q_k + (u - \gamma_1 q_1 - \sum_{w=2}^{W-1} \gamma_w(q_w - q_{w-1}))/\gamma_W & \text{for } \gamma_{W-1} q_{W-1} \le u < \ldots \\ & \ldots \gamma_{W-1} q_{W-1} + \gamma_k(q_W - q_{W-1}) \end{cases} \tag{8.39}$$

We use the values of the inverse hazard in the equations (8.31) and (8.35) to obtain the simulation equations given in Table 8.1.

Table 8.1: Simulation Equations

	Without Covariates	With Covariates
Weibull	$(-\zeta \log(Y))^{1/\rho}$	$\frac{(-\zeta \log(Y))^{1/\rho}}{\exp(x_i'\beta)}$
Piece-wise Constant	$H^{-1}(-\log(Y))$	$\frac{H^{-1}(-\log(Y)}{\exp(x_i'\beta)}$

Note: $Y \sim U(0,1)$ *and* H^{-1} *is as in* (8.39)

8.6 Kaplan-Meier with Covariates

The survival function of the Weibull distribution with covariates is given by

$$S(u|T_i, C_i) = S_0(u)^{\exp(x_i'\beta)} \tag{8.40}$$

where $S_0(u)$ is the baseline survival function, $x_i = (T_i, C_i)$ and β indicates the effect of the covariates on the hazard rate. We rewrite this by using (4.41) as

$$\log[-\log(\hat{S}(u_i|T_i, C_i))] = \log[-\log(S_0(u_i))] + x_i'\beta, \tag{8.41}$$

$$= \rho \log(\zeta) + \rho \log(u_i) + x_i'\beta. \tag{8.42}$$

This indicates that for groups for which $x_i'\beta$ is constant, we again have a linear relation between the log of the reporting delay and the double log of the Kaplan-Meier survival function. To check whether the linear relation holds, we first compute the Kaplan-Meier estimate of the survival function, $\hat{S}(u_i)$ for each group of claims that have the same covariates. Then, we compute $y_i = \log[-\log[\hat{S}(u_i)]]$ for each i and plot it against u_i. We fit a straight line through these points, $y = b + m \log(u)$. Here, we should approximately have $b = \rho \log(\zeta)$ and $m = \rho$.

The same procedure can be performed for the gap times v_{ij}, where $\hat{S}(u_i|T_i, C_i)$ in (8.42) is replaced by $\hat{S}(v_{ij}|V_{ij}^-, U_i, T_i, C_i)$, the Kaplan-Meier estimate of the gap times as given in (4.40).

8.7 Tables

Here, the tables of the results are given.

	Estimate		Estimate
Old Car	0.051**	Catalog Value 3	0.410***
	(0.017)		(0.007)
Catalog Value 2	0.010*		
	(0.004)		

Table 8.2: Covariate Estimates of Full Model Rate of Occurrence

Delete Variable	AIC
None	1886453
Old Car	1887000
Catalog Value	1886000
Both	1887000

Table 8.3: AIC of Different Models Rate of Occurrence

Delete Covariate	AIC
None	2,141,013
Old Car	2,141,908
Catalog Value	2,141,512
Time of Occurrence	2,141,113

Table 8.4: AIC Different Models Reporting Delay

Delete Covariate	AIC
None	4,436,905
Number of Payments	4,436,907
Old Car	4,439,044
Reporting Delay	4,440,105
Total Payout	4,458,003
Years in Development	4,470,850

Table 8.5: Backwards Analysis: AIC after deleting Covariates

	AIC	Optimal Model
K = 2	534,414	Log Normal, Log Normal
K = 3	533,686	Log Normal, Log Normal, Log Normal
K = 4	533,694	Gamma, Log Normal, Log Normal, Log Normal

Table 8.6: AIC of Payment Mixture Models

Parameters with Covariates	AIC
None	533,686
μ_k	518,608
μ_k, σ_k	519,674
μ_k, σ_k, π_k	519,673
σ_k	518,900
σ_k, π_k	519,886
μ_k, π_k	518,711
pi_k	523,670

Table 8.7: AIC of Different Covariate Models for Payments

	AIC	Optimal Model
K = 2	559,384	ξ_0, Log Normal, Log Normal
K = 3	553,666	ξ_0, Log Normal, Log Normal, Log Normal
K = 4	553,700	ξ_0, Gamma, Gamma, Log Normal, Log Normal

Table 8.8: AIC of Settlement Mixture Models

	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
(Intercept)	6.940***	5.292***	7.464***	0.230***	0.232***	0.282***
	(0.010)	(0.022)	(0.024)	(0.002)	(0.003)	(0.003)
2 Years in Development	0.334***	0.106**	3.676***			
	(0.015)	(0.030)	(0.043)			
3 Years in Development	0.670***	0.343***	0.1759***			
	(0.023)	(0.046)	(0.182)			
4 Years in Development	0.816***	3.341***	-0.458***			
	(0.036)	(0.066)	(0.138)			
5 Years in Development	0.761***	3.449***	0.465***			
	(0.049)	(0.090)	(0.869)			
Old Car	0.148***	0.331***	2.373***			
	(0.023)	(0.048)	(0.047)			
Total Payout $>$ €2000	0.701***	0.335***	-0.096***			
	(0.012)	(0.024)	(0.027)			
Total Payout $<$ €500	0.279***	1.286***	-0.569***			
	(0.009)	(0.020)	(0.022)			
Reporting Delay	-0.002***	-0.002***	-0.019***			
	(0.000)	(0.000)	(0.000)			
Number of Payments	0.004*	0.002	-0.612*			
	(0.001)	(0.002)	(0.007)			
Time of Occurrence	-0.000	0.000*	-0.000**			
	(0.000)	(0.000)	(0.000)			
Catalog Value "Intermediate"	0.026***	0.139***	0.044***			
	(0.005)	(0.011)	(0.011)			
Catalog Value "High"	0.112***	0.266***	2.657***			
	(0.009)	(0.019)	(0.022)			
Probability	0.630	0.322	0.047			

*Note: * p<.10, ** p<.05, *** p<0.01*

Table 8.9: Parameter Estimates Payments Mixture Model with Covariates

	ξ_0	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
(Intercept)	1.522***	4.287***	5.598***	7.102***	2.159***	0.2393***	2.405***
	(0.014)	(0.003)	(0.014)	(0.003)	(0.004)	(0.002)	(0.005)
2 Years in Development	1.198***	0.033***	0.047***	0.007			
	(0.036)	(0.008)	(0.031)	(0.008)			
3 Years in Development	3.235***	2.066	1.237***	0.526***			
	(0.109)	(0.020)	(0.069)	(0.020)			
4 Years in Development	5.847***	1.222***	1.759***	-0.887***			
	(0.257)	(0.033)	(0.132)	(0.030)			
5 Years in Development	12.480***	1.385***	2.635***	4.373***			
	(0.425)	(0.053)	(0.183)	(0.087)			
Old Car	-0.014*	0.705***	0.582***	0.039***			
	(0.007)	(0.007)	(0.031)	(0.008)			
Number of Payments	-1.217***	-0.003	-0.102***	-0.329***			
	(0.104)	(0.002)	(0.006)	(0.002)			
Catalog Value "Intermediate"	-0.137***	0.086***	0.470***	-3.111***			
	(0.009)	(0.001)	(0.007)	(0.002)			
Catalog Value "High"	-0.251***	0.666***	0.081***	-0.942***			
	(0.016)	(0.003)	(0.012)	(0.003)			
Total Payout $> €\,2000$	0.563***	0.005	0.514***	0.008			
	(0.017)	(0.004)	(0.073)	(0.004)			
Total Payout $< €\,500$	-2.579***	1.542***	0.094***	-0.006*			
	(0.012)	(0.003)	(0.013)	(0.003)			
Time of Occurrence	0.000	0.000	0.000***	-0.000**			
	(0.000)	(0.000)	(0.000)	(0.000)			
Reporting Delay	0.001***	-0.001***	-0.001***	-0.001****			
	(0.000)	(0.000)	(0.000)	(0.000)			
Probability		0.146	0.740	0.114			

Table 8.10: Parameter Estimates Model

Bibliography

Arjas, E. (1989). The claims reserving problem in non-life insurance: some structural ideas. *Astin Bulletin*, 19(2):139–152.

Cook, R. J. and Lawless, J. (2007). *The statistical analysis of recurrent events*. Springer Science & Business Media.

England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8(3):443–518.

Haastrup, S. and Arias, E. (1997). Claims reserving in continuous time; a nonparametric bayesian approach. *Insurance Mathematics and Economics*, 2(19):153.

Hesselager, O. and Witting, T. (1988). A credibility model with random fluctuations in delay probabilities for the prediction of ibnr claims. *ASTIN Bulletin: The Journal of the IAA*, 18(1):79–90.

Jewell, W. S. (1989). Predicting ibnyr events and delays: I. continuous time. *ASTIN Bulletin: The Journal of the IAA*, 19(1):25–55.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kunkler, M. (2004). Modelling zeros in stochastic reserving models. *Insurance: Mathematics and Economics*, 34(1):23–35.

Larsen, C. R. (2007). An individual claims reserving model. *ASTIN Bulletin: The Journal of the IAA*, 37(1):113–132.

Maciak, M., Okhrin, O., and Pešta, M. (2018). Dynamic and granular loss reserving with copulae. *arXiv preprint arXiv:1801.01792*.

Norberg, R. (1986). A contribution to modelling of ibnr claims. *Scandinavian Actuarial Journal*, 1986(3-4):155–203.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance1. *ASTIN Bulletin: The Journal of the IAA*, 23(1):95–115.

Norberg, R. (1999). Prediction of outstanding liabilities ii. model variations and extensions. *ASTIN Bulletin: The Journal of the IAA*, 29(1):5–25.

Norberg, R. and Sundt, B. (1985). Draft of a system for solvency control in non-life insurance. *ASTIN Bulletin: The Journal of the IAA*, 15(2):149–169.

Plat, R. and Antonio, K. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 2014(7):649–669.

Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin: The Journal of the IAA*, 24(2):265–285.

Taylor, G., McGuire, G., and Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. *Annals of Actuarial Science*, 3(1-2):215–256.

Verdonck, T., Van Wouwe, M., and Dhaene, J. (2009). A robustification of the chain-ladder method. *North American Actuarial Journal*, 13(2):280–298.

Wright, T. S. (1990). A stochastic method for claims reserving in general insurance. *Journal of the Institute of Actuaries*, 117(3):677–731.

Zhao, X. and Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. *Insurance: Mathematics and Economics*, 46(2):290–299.

Zhao, X. B., Zhou, X., and Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics*, 45(1):1–8.