University of Economics, Prague

Faculty of Informatics and Statistics

Study program: Quantitative Methods in Economics

Field of study: Statistics

MICRO-MODEL FOR RESERVING IN NON-LIFE INSURANCE WITH THE USE OF GENERALIZED LINEAR REGRESSION MODEL EXTENSIONS

Master thesis

Author: Bc. Marcel Vrobel

Supervisor: Ing. Pavel Zimmermann, Ph. D.

Prague, **May 2018**

Declaration

I declare that I carried out this thesis independently and cited all used sources and literature.

In Prague date 14.05.2018

……………………………….
Signature

Acknowledgement

I would like to express my sincere gratitude to Ing. Pavel Zimmermann, Ph.D. for his valuable help, guidance, patience and support throughout the whole writing process of this thesis.

Abstract

MICRO-MODEL FOR RESERVING IN NON-LIFE INSURANCE WITH THE USE OF GENERALIZED LINEAR REGRESSION MODEL EXTENSIONS

This paper develops a deterministic model for individual claim level reserving. The proposed model is based on collected literature and its structure is decomposed and presented as a list of variables. These variables are firstly defined and then predicted with the usage of generalized linear models (GLM) and its extensions in form of Hurdle and Zero-inflated models. These variables are then combined and as a result, a deterministic model for individual claim level reserving model is obtained. For the practical implementation and evaluation, a dataset of MTPL claims, originating from non-insured cars, is used and presented. Based on these claims the estimated reserve is obtained. The model is then compared with a traditional Chain-Ladder model reserve estimate. In the end the proposed model proved to have very accurate predictions but is biased in comparison with the real claim development. The last section suggests how the model can be improved and further developed.

Keywords: Micro model, Individual claim level model, GLM, Hurdle model, Zero-inflated model.

MICRO RESERVING MODEL V NEŽIVOTNÍM POJIŠTĚNÍ S VYUŽITÍM ROZŠIŘENÝCH ZOBECNĚNÝCH LINEÁRNÍCH REGRESNÍCH MODELŮ.

V této práci byl navržen deterministický model určených, pro tvorbu rezerv na základě individuálního vývoje škodných událostí. Navržený model je vytvořen na základě představené literatury a jeho struktura je představena jako seznam proměnných. Tyto proměnné jsou napřed definované, a poté predikované s pomocí zobecněných lineárních regresních modelů (GLM) a jeho rozšíření ve formě Hurdle a Zero-inflated modelů. Na základě kombinace těchto proměnných se získá model určený pro tvorbu rezerv na základě individuálního vývoje škodných událostí. Praktická ukázka a zhodnocení modelu je provedeno na datovém souboru MTPL škodných událostí původem z nepojištěných vozidel. Pro tyto události se vytvoří model a jeho výsledky jsou poté porovnány s výsledky z tradičního Chain-Ladder modelu. Po porovnaní bylo zjištěno, že navržený model má velmi přesné odhady, ale je vychýlený v porovnání s reálným vývojem škod. V závěru práce se navrhuje, jak lze tento model vylepšit a dále rozšířit.

Klíčová slova: Micro model, Individual claim level model, GLM, Hurdle model, Zero-inflated model.

Contents

1. Introduction

In this section, the objectives of this work will be laid out and the main terms will be briefly introduced in the context of the insurance industry.

1.1 Overview

The structure of this work is as follows:

- Chapter 1 introduces the basic terms in the context of the insurance industry and objectives of this work are laid out.
- Chapter 2 focuses on the difference between aggregate and individual claim level models.
- Chapters 3 – 7 present the proposed model and study it.
- Chapter 3 presents the proposed individual claim level.
- Chapter 4 presents the framework of the GLM models, hurdle models and zero inflated models.
- Chapter 5 describes the provided dataset and model variables in detail.
- Chapter 6 presents the practical implementation of the chapter 3 model, chapter contains descriptions of the model components in detail.
- Chapter 7 presents model results and compares them with the real claim development and a traditional chain ladder model and proposes how the model can be improved.
- Chapter 8 provides conclusions.

1.2 Introduction to Insurance

Insurance is a service that provides coverage, in the form of compensation resulting from loss. Loss can be described as damage, injury, treatment or loss. To provide this coverage the premium is collected. When determining the premium, the risk is calculated as the probability of loss occurrence and the cost to replace the associated loss.

Based on this description the following questions needs to be discussed.

- Who can provide coverage?
- How to define loss?
- How to determine the premium?
- How to calculate the risk?

These questions will be further discussed below.

Who can provide coverage?

To be permitted to provide coverage as an insurance company in Czech Republic it is necessary to obtain license from the Czech National Bank. In addition, the insurance company needs to follow the following laws:

- Act No. 277/2009 Coll. on Insurance
- Act No. 38/2004 Coll. on Insurance Intermediaries and Independent Loss Adjusters.

Finally, the insurance company behavior is also subjected to the European Parliament Directive Solvency II.

How to define loss?

The insurance service is always provided based on the contract (insurance policy) and the loss definition is presented in this contract. The insurance contract should contain following sections.

- Declarations – contract summarization and introduction of contact parties
- Definitions – definition of insurance terms and phases
- Terms of Insurance – defines what is the loss, can be presented in two forms:
    - Named Perils Coverage – coverage for only the named perils
    - All-Risk Coverage – coverage for all losses except for exclusions
- Exclusions – named list of losses that are not covered by the policy, three types:
    - Excluded causes of loss – specific scenarios that are not covered
    - Excluded losses – types of loss that are not covered
    - Excluded property – exclusion of property from the coverage
- Conditions – list of conditions that apply to the contact
- Endorsements – contract modifications

The insurance policy should always contain clear definition what a loss is. When provided the insurance customer (insured) can be able to ask for coverage when the loss occurs.

How to determine the premium?

The pricing actuary is responsible for determining the premium. The following factors are commonly determining the premium value:

- type of coverage – the higher number of possible risk to cover, the larger the premium
- coverage amount – the larger amount to replace, the larger the premium
- costs – every premium should also pay for the insurance company costs
- score – increases or decreases in the premium based on the personal history of insured
- competitiveness – premium can be affected by other insurance company behavior

How these factors are evaluated and how are they combined is mostly dependent on the pricing actuary.

How to calculate the risk?

The risk evaluation process is done by the reserving actuary. The insurance company does not know what size the risk would be. Therefore, it is necessary to create a sufficient reserve. The term sufficient means that the reserve will not be too large (unnecessary costs for insurance company) or too small (risk of insolvency). The final reserve is obtained when proper reserving process is used. This reserving process is the focus of this work and will be further discussed and developed in the following chapters. The following sections will present how the reserving process is concluded and what are the requirements of this process as presented by the Solvency II Directive.

1.3 Reserving Process

The risk reserving process is trying to predict future liabilities of the insurance company. This is also the biggest problem of the insurance business where the liabilities resulting from insurance contracts are not known for a long time to the insurance company (insurer). This means that technical results of the insurance company for a given period (mostly a year) are not known even after several years. As a result, the insurer needs to hold reserves (technical provisions) until all claims are paid out.

This work will focus on these reserves.

- The reserves on claims that occurred and were reported, but are not yet settled (RBNS)
- The reserves on claims that occurred, but have not yet been reported (IBNR)

The reserving process for each individual claim can be described as follows.

When a claim is reported to the insurer, a claim handler will be selected by the insurer to collect all available information and create a reserve. This reserve is an estimate of the ultimate loss paid to the customer (insured). This reserve will be updated when more information about the claim will become available to the claim handler. Then after a set period or when enough information is obtained the claim is settled and will be paid out. If any additional information about the loss are found afterwards there still exists the possibility of claim being reopened and its reserve adjusted and paid out.

The previous description of the reserving process presents following variables:

- Reserve variable – will be paid out in the future
- Paid out variable – have been paid out in the past

When these variables are summed for given claim at one point the **incurred** value is obtained. This variable represents the insurers **liability** for given claim. The main problem is that final (ultimate) value of this variable is only known when the claim is settled.

To be able to predict this incurred value the reserving models were developed by the reserving actuaries. These models, based on the way how the incurred value prediction is obtained, can be categorized as follows.

- Aggregate claim level reserving models.
- Individual claim level reserving models.

These models provide estimate of the insurer liability which may differ from the reality. The ultimate liability can be lower (resulting in additional costs of capital) or be much higher (resulting in solvency problems). For this purpose, the volatility of the previous estimate is also evaluated. This will provide additional information about the estimate.

Descriptions of these models will be provided in the following chapter. Apart from the model description the Solvency II Directive set up a list of requirements the insurance company needs to fulfill.

1.4 Solvency II

Solvency II is a European Parliament directive (framework) for insurers and reinsurers which requires them to meet certain solvency requirements. These requirements are based on analyzing risk profile of each individual insurance company to promote comparability, transparency and competitiveness. Its main goals are to:

- Ensure the financial health of insurance undertakings
- Protect policyholders (consumers, businesses) and the stability of the financial system

The main reason for creating a new directive for an insurance market was the financial crisis, which led to severe shocks in the insurance market. The crisis has drawn attention to the weaknesses of the current system and developed a need for an improved risk management by updating the Solvency framework. The new requirements moved away from a gauche "one-model-fits-all" way of estimating capital requirements to more entity-specific requirements.

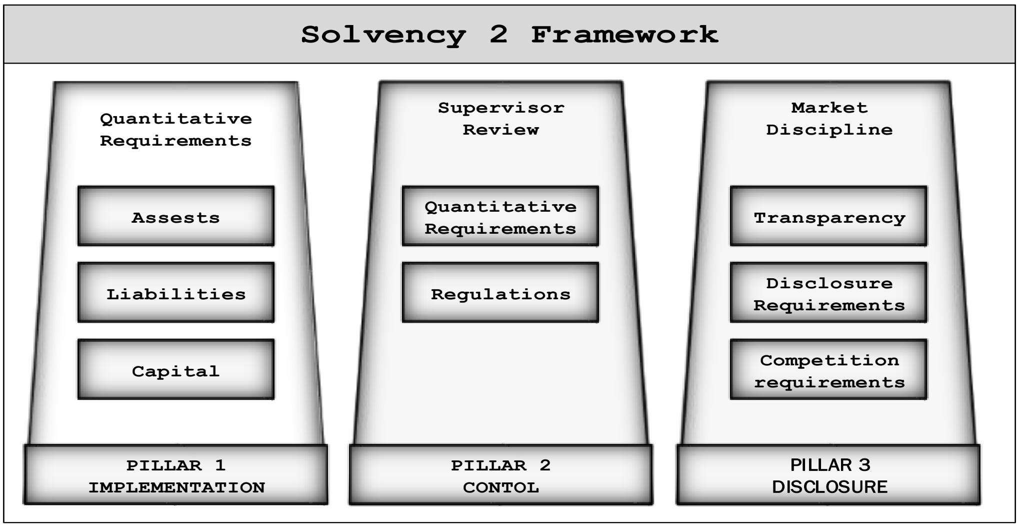
The Solvency II regulatory framework is structured into three pillars:

- Pillar 1 represents the quantitative requirements for evaluating the technical provisions, Solvency Capital Requirement (SCR) and Minimum Capital Requirement (MCR), measurement of assets and liabilities and determining the required data quality.

- Pillar 2 sets out the qualitative requirements for internal controls, risk management and governance. The main part of Pillar II is the Own Risk and Solvency Assessment (ORSA) which defines the overall solvency needs related to the specific risk profile of the insurance company.

- Pillar 3 focuses on disclosure, reporting and transparency requirements around these risks and capital requirements.

Figure 1*: Three-pillar structure of Solvency II regulatory framework*



The directive also states how to determine the MCR and SCR.

The Solvency Capital Requirement (SCR) corresponds to the economic capital a (re)insurance undertaking needs to hold to limit the probability of ruin over a one-year period to 0,5 % (1 in 200 years). SCR is estimated through the Value at Risk (VaR) measure, commonly used in financial services to assess the risk associated with a portfolio of assets and liabilities. VaR enables to quantify how much money would be lost, if events developed in an adverse and unexpected way. In other words, it measures the worst expected loss under normal conditions over a specific time interval at a given confidence level. Specifically, for Solvency II framework, the VaR is measured over a one-year period at a confidence level of 99.5 %.

The Minimum Capital Requirement (MCR) represents the absolute minimum level of capital below which policyholders' interests would be seriously endangered if the undertakings could continue to operate. In the case that the Minimum Capital Requirement is breached ultimate supervisory action is triggered, i.e. license is withdrawn. Undertakings are therefore required to hold eligible basic own funds to cover the Minimum Capital Requirement.

There are two approaches to determine the MCR:

- As a fixed percentage of SCR * 1/3
- As a lower confidence for VaR (e.g. 90 %).

When determining the SCR and MCR the directive also states that internal model or standard formula approach should be used. Internal model is a model that was created to forecast the probability distribution of risks to which (re)insurance undertakings are exposed. This model needs to be well documented and approved by the regulatory authorities. The SCR and MCR should be obtained from this model. An alternative to the internal model is the standard formula approach where SCR and MCR are obtained from formula provided by the directive. The internal model and standard formula can be combined to create the partial internal model approach.

Reserve models under Solvency II

The Directive defines the following principles for the evaluation of the reserves:

1) Technical provisions shall be calculated in a prudent, reliable and objective manner.
2) Technical provisions calculation shall be based on their current exit value.
3) Technical provisions calculation shall make use of and be consistent with information provided by the financial markets and generally should have available data on insurance and reinsurance technical risks.

The second principle is meant for case of buying the insurance liabilities where the current exit value is higher than the expected value of the future cash flows. Buyer will need to hold an appropriate amount of the solvency capital to continue running the business. Holding an extra capital relates to additional costs of capital.

The last principle speaks about the market price of liabilities which is not directly observable for insurance liabilities. The estimate of the market price of the insurance liabilities is assessed by splitting the insurance liabilities into hedgeable and non-hedgeable obligations. The hedgeable obligations are those obligations for which the associated future cash flows can be replicated using such financial instruments that their market value is directly observable. In those cases, the value of technical provisions shall be determined by their market value. In case of non-hedgeable obligations, the sum of best estimate and risk margin can be used.

The best estimate is defined as:

'The best estimate shall be equal to the probability-weighted average of future cash-flows, taking account of the time value of money (expected present value of future cash-flows), using the relevant risk-free interest rate term structure.'

The risk margin is defined as:

'The risk margin shall be such as to ensure that the value of the technical provisions is equivalent to the amount insurance and reinsurance undertakings would be expected to require taking over and meet the insurance and reinsurance obligations.'
'The risk margin is defined as the expected cost of future capital required for non-hedgeable risks necessary to support the insurance liabilities. Therefore, the risk margin is the probability weighted average of future cash flows stemming from the cost of future capital, considering the time value of money.'

The directive requires separate calculation of these two components (best estimate and risk margin) of the value of the technical provisions for non-hedgeable risk.

1.5 Objectives of This Work

With the provided introduction of risk reserving modeling the following objectives will be laid out and solved in this work:

1) Research of the existing literature on the topic of the reserve risk models based on aggregate claim level and individual claim level.
2) Define an individual claim level model with the usage of hurdle models and zero inflated models.
3) Practically implement the model on given dataset.
4) Describe the model results and compare them with alternative approach.
5) Propose how the model can be improved.

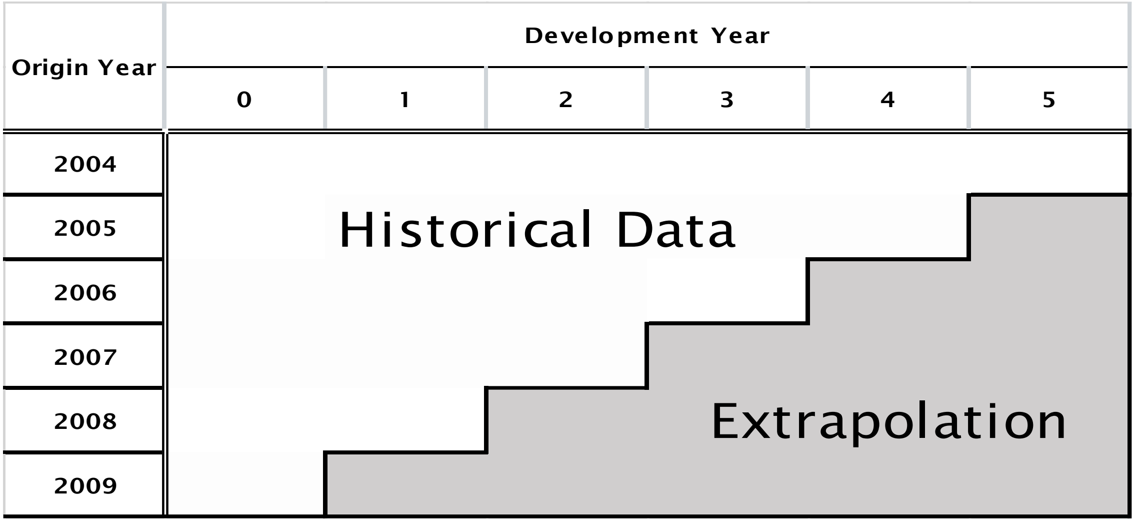
2. Literature Overview - Reserve Risk Models

In this chapter the collected literature on the topic of reserve risk modelling will be presented. Firstly, the aggregate claim level models will be presented. These models are often described as the 'traditional' models by the actuarial community. An alternative to aggregate claim level models are the individual claim level models. These will also be introduced in the second part of this chapter. In the final part of this chapter these models will be compared. This section was inspired by the work of Zimmermann 2010.

2.1 Aggregate Claim Level Models

The aggregate claim level models are commonly based on the so-called triangle schemes. (see figure 2). The triangle scheme is basically a contingency table that contains in the rows cumulative value of claims that occurred in an occurrence year, and in columns value of claims that appeared in a certain development period (so called incremental triangle) or values that occurred up to a certain development period (so called cumulative triangle).

Figure 2: *Triangle scheme*.



Source: Zimmerman 2010

The historical observations can be found only in the upper left triangle part of the contingency table. When considering the incremental claim data (increments) of the total value of the claims of the *i*-th occurrence year in the j-th development period:

$$\{C_{i,j}: i = 1,2, \dots ,n; j = 1,2, \dots ,n - i + 1\}. \qquad (2.1)$$

The table containing $C_{i,j}$ is the incremental triangle scheme. Commonly this scheme is used to represent the paid, incurred or the reserve value. From the incremental triangle scheme the cumulative triangle scheme can be obtained as follows:

$$D_{i,j} = \sum_{k=1}^{j} C_{i,k}. \tag{2.2}$$

Cumulative triangle scheme presents the total incurred or paid value up to a certain development period for each occurrence year.

The reserving problem can then be understood as the problem of extrapolating the historical development to the future to estimate the ultimate value of the claims that have already occurred.

2.2 The Chain-Ladder Method

The Chain-Ladder method is the prime example of the aggregate claim level reserving method. The description of this method provided here was taken from the England and Verrall 2002. Chain-ladder method has the following assumptions.

1) Increments of the claims value are independent in the development periods

$$\{\, C_{i,1}, \dots, C_{i,n}\}, \{\, C_{j,1}, \dots, C_{j,n}\}, i \neq j \text{ are indenpendent} \tag{2.3}$$

2) Development is stable in time. The development has the same characteristics in each occurrence period.

The chain-ladder method estimates the so-called development factors for each development period (common for each occurrence period):

$$\hat{\lambda}_{i,j} = \frac{\sum_{i=1}^{n-j+1} D_{i,j}}{\sum_{i=1}^{n-j+1} D_{i,j\,-\,1}}. \tag{2.4}$$

These are then applied to the latest cumulative claims in each row ($D_{i,n-i+1}$) to produce forecasts of future values of cumulative claims:

$$\widehat{D}_{i,n-i+2} = D_{i,n-i+1}\,\hat{\lambda}_{n-i+2}, \tag{2.5}$$

and the reserve yearly estimate:

$$\hat{R}_i = D_{i,n-i+1} * \left(\, \hat{\lambda}_{n-i+1} \dots \hat{\lambda}_{n-1} - 1\,\right). \tag{2.6}$$

The overall reserve estimate $\hat{R}$ is the sum of reserve yearly estimates $\widehat{R_\iota}$.
This method is reliable for early development years, but for late development years (years where no observations are available) the tail factor estimates are often used. Tail factor estimate is a curve extrapolation of the development in late years. Several authors have set up model with known statistical method. These models are presented and compared in the England and Verrall (2002).

2.3 Mack's Model

Thomas Mack presented model based on statistical theory in Mack (1993). This model defines how the standard error for chain ladder reserve estimate can be obtained. Mack's model is built on the following assumptions:

1) Developing factors $\lambda_1, \dots, \lambda_{n-1}$> 0 exist and the expected cumulative claim value can be obtained as

$$E\left[D_{i,j}\middle|D_{i,j-1}\right] = \lambda_j D_{i,j-1}. \tag{2.7}$$

2) Cumulative claims variables of different accident years are independent.

$$\left\{ D_{i,1}, \dots, D_{i,n}\right\}, \left\{ D_{j,1}, \dots, D_{j,n}\right\}, i \neq j \; are \; indenpendent \tag{2.8}$$

3) The variance can be obtained as

$$Var\left[D_{i,j}\middle|D_{i,j-1}\right] = \sigma^2{}_j D_{i,j-1}. \tag{2.9}$$

No assumptions about the variable distributions are required therefore this model is often called „distribution free “. The estimation of the parameter λ_j can be obtained from (2.4) and the parameter $\hat{\sigma}_j^2$ is obtained as follows.

$$\hat{\sigma}_j^2 = \frac{1}{n-j} \sum_{i=1}^{n-j+1} D_{i,j} \left(\frac{D_{i,j+1}}{D_{i,j}} - \hat{\lambda}_j \right)^2 . \tag{2.10}$$

Based on the model the mean squared error for the chain ladder reserve yearly estimate as

$$\widehat{mse(R_\iota)} = \hat{D}_{i,n}^2 \sum_{k=n-i+1}^{n-1} \frac{\hat{\sigma}_k^2}{\hat{\lambda}_k^2} \left(\frac{1}{D_{i,k}} + \frac{1}{\sum_{j=1}^{n-k} D_{j,k}} \right). \tag{2.11}$$

The standard error ($s.e\,(\hat{R}_i\,)$) is obtained as the root of the mean squared error. It is not possible to obtain the total reserve estimate as the sum of yearly estimates because of correlation via the common estimators $\hat{\lambda}_{i,j}$ and $\hat{\sigma}_j^2$ (proof in Mack (1993)). The mean squared error for the chain ladder total reserve estimate then obtained as:

$$\widehat{mse(R)} = \sum_{i=2}^{n} \{\, (s.e\,(\hat{R}_i\,))^2 + \hat{D}_{i,n} \left(\sum_{j=i+1}^{n} \hat{D}_{j,n} \right) \sum_{k=n+1-i}^{n-1} \frac{\frac{2\hat{\sigma}_k^2}{\hat{\lambda}_k^2}}{\sum_{j=1}^{n-k} D_{j,k}} \} . \tag{2.12}$$

2.4 Over-Dispersed Poisson Model

The reserving model developed based on the over-dispersed Poisson distribution assumes that the variance of the Poisson distribution is not equal to the mean, rather it is proportional to the mean. Namely this model assumes that the incremental claims $C_{i,j}$ are distributed as independent random variables from the Poisson Distribution with mean and variance:

$$\begin{aligned} E\left[C_{i,j}\right] &= \mu_{i,j} = x_i\, y_j, \\ Var\left[C_{i,j}\right] &= \varphi\mu_{i,j} = \varphi x_i\, y_j, \end{aligned} \tag{2.13}$$

where $\sum_{j=1}^{n} y_j = 1$.
The mean $\mu_{i,j}$ has a multiplicative structure and is a product of two factors. The x_i as the expected ultimate claims (observed in the triangle) and y_i the proportion of ultimate claims to emerge each development year. Over-dispersion is introduced in the parameter φ. Allowing for over-dispersion does not affect estimation of parameters, but increases their standart error by proportion of y_i.

2.5 Predictive Distribution

The previous models focused on providing the estimate of the expected value and estimation error as a measure of the precision of the estimate. By estimating the predictive distribution of the expected value additional information are obtained. These are necessary for estimating the extreme quantiles (VaR) of the company's predicted liability as required by the Solvency II directive. For these models the prediction error consists of the estimation error and the process error.

Simple technique to estimate the predictive distribution is the bootstrapping model which was presented in England and Verrall (2002) and developed in England and Verrall (2006). The bootstrapping model is created by repeating two steps. Firstly, the estimation error is obtained using the bootstrapping method, where residuals are calculated and resampled. Then the process error is estimated from an assumed distribution. This is achieved by generating a random value from the underlying distribution of the provided residual sample. These steps are then repeated substantial number of times. Result of this model is a predictive distribution with estimated parameters and a precision based on the estimation error and process error.

2.6 Individual Claim Level Models

Individual claim reserve risk modelling (Micro modelling) is the new reserve risk approach in non-life insurance. The point of this approach is to evaluate each claim and to predict the future claim development on per claim basis. This approach was fundamentally developed by work by Norberg (1993), Haastrup and Arjas (1996) and Norberg (1999), Antonio and Plat (2013) model the development of individual claims in continuous time. Drieskens et al. (2012), Rosenlund (2012) and Pigeon et al. (2013) work in discrete time and the framework for the complete micro model was presented in Pigeon, et. al (2014). This section will present the few possible approaches how to work with individual claim reserve risk modelling.

2.7 Marked Poisson Process

Marked Poisson process as presented in Norberg (1993) is an individual claim level model. The individual claims are presented as a random variable:

$$C = (O, Z), \tag{2.14}$$

where O is the time of claim occurrence (defined as O = [0, ∞)) and Z is the mark describing the claim development from the time of occurrence to the final settlement. The mark Z can be presented in the following form:

$$Z = (I, T, U, \{ U'(v'), 0 < v' < U\}), \tag{2.15}$$

where I is the waiting time from occurrence until the notification, T is the waiting time from notification until the final settlement, U is the eventual total claim amount (ultimate liability) and $U'(v')$ is the amount paid within the v' time after the notification. The individual claims can be then classified (at a present time t) based on their stage of development as:

- Settled → $t \geq T$
- Reported but not settled (RBNS) → $t < T$ and $U'(t) < U$
- Incurred but not reported (IBNR) → $t < I$
- Covered but not incurred. → $t < O$

The Norberg (1993) then describes the marked Poisson process as:
The claim process of an insurance business can be described as a random collection of points in a claim space $\{(O_t, Z_t)\}_{t=1,\dots,N}$ where t is the indicator of chronological order from 0 to N ≤ ∞. This model assumes that the O_t are generated by an inhomogeneous Poisson process with intensity $w(t)$ at time t > 0 and that the marks are of the form Z_t = Z_T where $\{Z_t\}_{\ t>0}$ is a family of random elements in Z that are mutually independent and independent of the Poisson process, and Z_t ~ $P_{Z|t}$. We then speak about a marked Poisson process with intensity of $\{w_t\}_{\ t>0}$ and Poisson-dependent marking by $\{P_{Z|t}\}_{\ t>0}$, and write

$$\{(O_t, Z_t)\}_{t=1,\dots,N} \sim Po\big(w_t, P_{Z|t};\ t > 0\big), \tag{2.16}$$

where the w_t is the intensity of the O_t and $P_{Z|t}$ is the join distribution of the $\{(O_t, Z_t)\}_{t=1,\dots,N}$.

Then the claim process outcome can be viewed as a Poisson number of independent and identically distributed occurrences and development marks. Firstly, when creating a claim process the number of N claims from the Poisson distribution should be generated and a random sample of N pairs from the join distribution of a random pair $\{(O_t, Z_t)\}_{t=1,\dots,N}$ be taken. The total loss (sum of all individual final claim amounts U) of these random pairs has a generalized or compound Poisson distribution.

2.8 Dynamic Claims Reserving Model

Dynamic claims reserving model was presented in Larsen 2007. The author also considers a discrete marked Poisson process with the following adjustment done to the claim development mark Z:

$$Z_l = \left(J_l, Y_{J,l}, Y_{J+1,l}, \dots, Y_{D,l}, G_l\right), \tag{2.17}$$

where J_l is the stochastic reporting delay in years (J_l ∈ {1, …, D}), $Y_{k,l}$ is the stochastic incremental incurred value in the development period k (k ∈ {J_l, …, D}) and $G_l$is a discrete stochastic characteristic of the claim (claim-type, loss description, etc.). This model assumes that claims are settled after D development years. The claim process can be then decomposed to the following components:

- The intensity of the Poisson process w_t (claim generating process).
- The distribution of G_l (claim characteristics based on a business mix).
- The conditional distribution of J_l, given the value of G_l (reporting delay for provided claim characteristics).
- The conditional distribution of $(Y_{J,l}, Y_{J+1,l}, \dots, Y_{D,l})$ given the occurrence year, development year, claim characteristics (claim incurred value at given point).
    - Assumptions of this distribution:
        - No additional time dependency apart from given variables.
        - Conditional distribution depends on its history based on an existing h functions as follows:

$$P_{Z:t}(Y_k | Y_{k-1}, \dots, Y_j) \sim P_{Z:t}(Y_k | h(Y_{k-1}, \dots, Y_j)), \tag{2.18}$$

Larsen 2007 also provides an example how to model the discrete claim process based on one-year periods using the generalized linear models (GLM). The initial incurred value Y_0 is obtained from the unconditional distribution and then the incremental incurred values Y_k are obtained from the conditional distributions. The initial incurred value distribution is mixture of distributions, where the components are a mixture of disjoint events (no initial incurred value Y_0 = 0, large initial incurred value Y_0 ≥ L (threshold), small initial incurred value Y_0 < L). The incremental incurred values are then modelled separately based on the initial incurred value (S_0 = 0 and S_0 ≠ 0). For the positive initial incurred value $S_0 > 0$ or previous incurred values S_{k-1}, the results can be presented in form of disjoint events. (no change in $Y_k = S_{k-1}$, increase in $Y_k > S_{k-1}$, increase in $Y_k > S_{k-1}$ over $Y_k > L$, decrease in $Y_k < S_{k-1}$ and reserve dismissal $Y_k = -S_{k-1} < 0$).

2.9 Individual Claim Loss Reserving

Taylor, McGuire and Sullivan (2008) focused describing the difference between aggregate models and individual models. In their work they presented an individual claim level model predicting the total loss reserve as.

$$U = h(X_1, X_2, \dots, X_n, \theta), \tag{2.19}$$

where h is a function, X_i is a vector of claim-dependent covariates and θ is a vector of parameters that are independent of the claims. The author then distinguishes the claim-dependent covariates into the following categories.

- Static covariates → do not change over life of a claim
- Time covariates → can change in time, but are predictable.
- Unpredictable dynamic covariates → can change in time, but are unpredictable

Their work then focuses on these covariates and how the individual model performance changes based on which covariates are to be included.

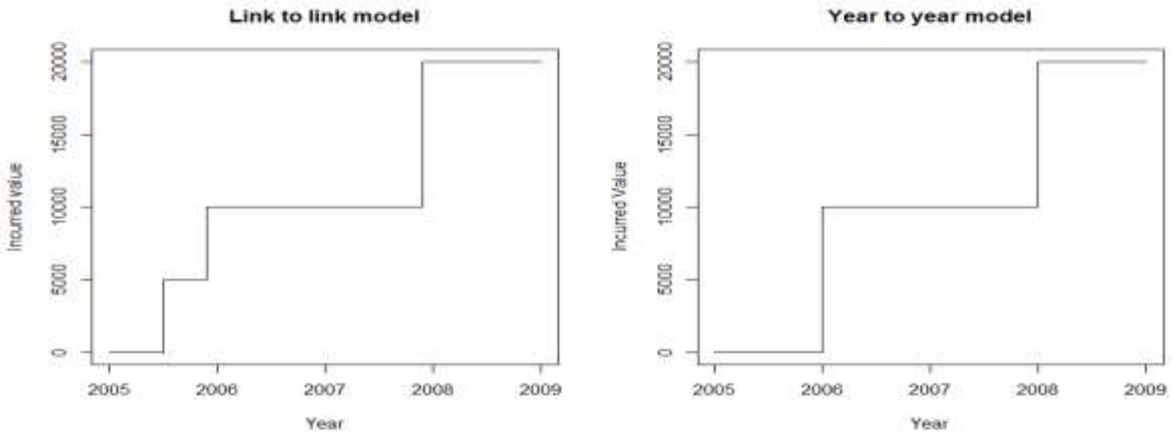
2.10 Individual Chain Reserving Model

Individual chain reserving model was presented by Pigeon *et al*. (2014). This model focuses on predicting the ultimate value as a combination of first incurred value and link factors as.

$$U_k = [\,Y_1 \;\; \lambda_1 \;\; \lambda_2 \ldots \;\; \lambda_N\,]\,, \tag{2.20}$$

where *k* is a defined period. The final model can be based on link to link approach (modelling changes as they occur) or year to year approach (modelling claim states as the end of each year). The difference between these approaches is presented in the following figure:

Figure 3: *Comparison of link to link and year to year models.*



The first model predicts how the claims will develop and presents λ_k link factors where k is in 1, …, N number of changes that will be modelled. (Chain of changes that have occurred) This model requires severity and time components to predict when the claim should change. The λ_k link factors represent the claim change after a time internal. For the year to year model the change is assumed to be fixed at the length of one year. Therefore, no time component is required for this model. Pigeon et al. (2014) focuses on the link to link model and presents the year to year model as an alternative approach. This work will try to further develop this approach.

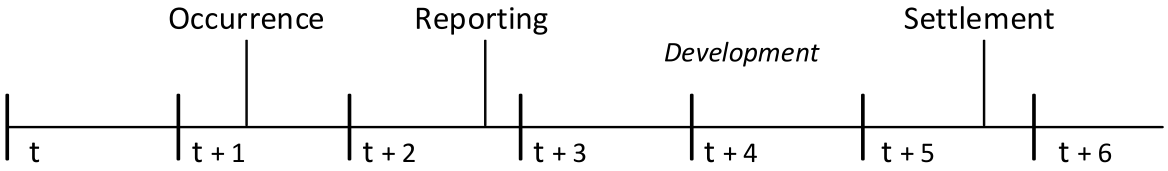
3. Model Overview

This chapter and the following chapters will provide theoretical introduction of the proposed model. The model will be an individual claim level model. This model will focus on developing a claim development process. This process will be modelled with a multiple sub models created by using generalized linear models (GLM). The goal of this work is to use hurdle (HM) and zero-inflated models (ZIM) to obtain better performance on a link factor sub model rather than using standard GLM model. For the model evaluation the results will be compared based on the estimate of the ultimate incurred value (liability) $Y_{i,\infty}$.

3.1 Claim Development Process

Before the model can be presented it is necessary to present the claim development process that is assumed. The claim development process is assumed based on the work of Pigeon et al. (2014) and its year to year model. This year to year model and its development process is presented in the following figure.

Figure 4: *Claim development process.*



where *t* represents the year when the policy was sold and the following years $t + k$ contains the events of claim occurrence, delay until reporting d ($t + 2$ in **figure 2**), claim development and the settlement delay s ($t + 6$ in **figure 6**). The year to year model focuses on modelling the individual claim development process Z at a fixed time points $t + j$ which represents the end of a year. The *j* represents the development year of the individual claim after the claim have occurred. Under this claim development process, the incurred value $Y_{i,j}$ can develop as follows:

$$\begin{aligned} Y_{i,j} &= 0, & 0 < j \leq t + d \\ Y_{i,j} &\geq 0, & t + d < j < t + s \\ Y_{i,j} &= Y_{i,\infty}, & j \geq t + s \end{aligned} \tag{3.1}$$

where the first state of $Y_{i,j} = 0$, represents claims in the incurred but not reported (IBNR). The last state $Y_{i,j} = Y_{i,\infty}$ is for all settled claim where the incurred value will not change. This model does not assume the possibility of claim being reopened. The second state of $Y_{i,j} \geq 0$ represents the reported but not settled (RBNS) and assumes following equation $Y_{i,j} \geq 0 = \lambda_{i,j} Y_{i,j-1}$ where the incremental value $Y_{i,j}$ can be obtained as a combination of the previous incremental value $Y_{i,j-1}$ and the link factor $\lambda_{i,j}$. This assumtions leads to the multiplicative specification.

$$Y_{i,t+s} = Y_{i,t+d+1}(\lambda_{i,t+d+2} \dots \lambda_{i,t+s}), \tag{3.2}$$

where the $Y_{i,t+d+1}$ is the first incurred value in the assumed claim development process. Other way how to describe this process can be based on the adjustments of $X_{i,t}$ that have occurred as

$$X_{i,t} = Y_{i,t} - Y_{i,t-1} \tag{3.3}$$

when using this form the final incurred value $Y_{i,t}$ can be obtained as a sum of all adjustments $X_{i,1}, X_{i,2}, \dots, X_{i,S}$. For this claim development process to work the following variables are needed to be defined.

- Reporting delay → probability that the claim will reported in the development year *t.*
- First incurred value → initial incurred value after the claim has been reported.
- Claim development → link factor or the claim adjustment in the development year *t.*
- Claim settlement → probability that the claim will be closed in the development year *t*
- Ultimate incurred value → final incurred value after the claim has been settled.

This claim development process and its variables will be further developed.

3.2 Individual Claim Level Model

The individual claim development process Z_i as presented in the previous section can be presented as a list of following variables:

$$Z_i = \left(Y_{i,1}, Y_{i,\infty}, \lambda_{i,1}, \dots, \lambda_{i,S}, Pr\left(O_{i,j}\right), Pr\left(S_{i,j}\right)\right), \tag{3.4}$$

where $Y_{i,1}$ is the first positive incurred value, $Y_{i,\infty}$ is the ultimate incurred value, $\lambda_{i,1}, \lambda_{i,2}, \dots, \lambda_{i,S}$ are the individual link factors, $Pr\left(O_{i,j}\right)$ is the probability of *i*-th claim being reported in the *j*-th development year and the $Pr(S_{i,j})$ is the probability of *i*-th claim being settled in the *j*-th development year. Each variable in the claim development process will be modelled (in chapter 6). Before the model practical implementation can be provided the generalized linear model (GLM), hurdle models (HM) and zero-inflated models (ZIM) needs to be introduced (chapter 4). The following section will focus on model variables and their definition.

3.3 Individual Claim Level Model Variables

This section will introduce the individual claim level model variables. The variables here are those presented in the beginning of this chapter and additional variables that are the sub results of the claim development process.

Ultimate Incurred Value

The ultimate incurred value Y_N represents the final liability that is associated with the claim. The ultimate incurred value can be defined as

$$Y_{i,\infty} = \sum_{j=1}^{S_i} X_{i,j} + Y_{i,1} = \prod_{j=1}^{S_i} \lambda_{i,j} \, Y_{i,1}, \tag{3.5}$$

where the $X_{i,j}$ is the yearly incremental adjustment of the i-th claim incurred value after the first incurred value $Y_{i,1}$. Each claim can be open $1, 2, \ldots, S_i$ years before it is finally settled (S_i may differ for each individual claim). The claim incremental adjustment $X_{i,j}$ or link factor $\lambda_{i,j}$ can result in increase or decrease of the incurred value Y_i. This work assumes that the $X_{i,j}$ are from the Gamma distribution.

$$X_i \sim (k_i, \theta_i), \tag{3.6}$$

where k_i is the shape parameter and the θ_i is the scale parameter.

The incremental adjustment X_i can be of any other continuous distribution which allows positive and negative occurences. The incremental adjustments will not be modelled in this work. The link factor approach was selected.

First Incurred Value

The first incurred value is often a result of an expert estimate or the insurance company standard for creating the first reserve. This value is obtained after the claim is reported and enough information about the claim is collected. The first incurred value needs to be a positive continuous variable which can be obtained from the Log-normal distribution as

$$Y_{i,1} \sim LN(\mu_i,\ \sigma_i^2), \tag{3.7}$$

where μ_i is the mean and σ_i is the standart deviation. This distribution is proposed based on the data exploration of the provided dataset. For more information see chapter 4.

Link Factors

The link factor represents the relative change between incurred value at the end of the previous year and the incurred value at the end of the actual year as

$$\lambda_i = Y_i / Y_{i-1,}, \tag{3.8}$$

The link factor is assumed to be a positive variable, where the following claim states can be observed.

- $\lambda_i = 0$ → The claim has not been reported yet.
- $0 < \lambda_i < 1$ → The incurred value has decreased.
- $\lambda_i = 1$ → The incurred value has not changed or has been closed
- $\lambda_i > 1$ → The incurred value has increased.

The link factor will be assumed to be from the gamma distribution in this work and will be subject to the modelling by the GLM, HM and ZIM models (see chapter 3).

$$\lambda_i \sim \Gamma(k_i, \theta_i). \tag{3.9}$$

Based on the link factor the following variables can be obtained.

Yearly Incurred, Paid and Reserve Value

Yearly incurred value $Y_{i,j}$ describes the total liability associated with i-th claim by the j-th development year (cumulative incurred value). The yearly incurred value can be decomposed as

$$Y_{i,j} = P_{i,j} + R_{i,j}, \tag{3.10}$$

where $P_{i,j}$ is the yearly paid value (cumulative paid value) from the i-th claim by the j-th development year and $R_{i,j}$ is the yearly reserve value for the i-th claim in the j-th development year. The yearly incurred value can be obtained from the equation (3.15).

Paid Ratio

The cumulative paid value from the i-th claim by the j-th development year can be also obtained with the use of incurred value to paid value ratio as.

$$\varsigma_{i,j} = \frac{P_{i,j}}{Y_{i,j}}, \tag{3.11}$$

where $\varsigma_{i,j}$ is the paid ratio. Based on this ratio a model will be introducted in the chapter 6.

Probability of Claim Being Reported

The probability of i-th claim being reported in j-th development year can be presented as.

$$Pr(O_{i,j}) = Pr(Y_{i,j} > 0 \mid Y_{i,j-1} = 0) = Pr(R_{i,j} > 0 \mid R_{i,j-1} = 0), \tag{3.12}$$

where $Y_{i,j}$ is the incurred value of the i-th claim in j-th year and $R_{i,j}$ is the reserve value of the i-th claim in j-th development year. Similar works (as presented in chapter 2) focused often on the time aspect the difference here is that claim is reported only when its first incurred value was observed. This probability will be observed from binary variable $I(O_{i,j}) \sim B(1, \pi)$.

Probability of Claim Being Settled

The probability of i-th claim being settled (closed) in j-th development year can be presented as.

$$Pr(S_{i,j}) = Pr(Y_{i,j} = Y_{i,\infty}) = \Pr(R_{i,j} = 0 \mid R_{i,j} > 0), \tag{3.13}$$

where $Y_{i,j}$ is the incurred value of the i-th claim in j-th development year and $R_{i,j}$ is the reserve value of the i-th claim in j-th year**.** The probability predicts the first year when the reserve value should be equal to 0. This assumes that there is no possibility of claim being reopened in the following years. This probability will be observed from binary variable $I(S_{i,j}) \sim B(1, \pi)$.

4. Model Framework - Generalized Linear Models and Its Extensions

Generalized linear models (GLM) were introduced by Nelder and Wedderburn (1972) and extended McCullagh and Nelder (1989). Generalized linear model describe the dependence of scalar variable y_i (i = 1, …, n) on a vector of regressors x_i. The conditional distribution of $y_i|x_i$ is a linear exponential family with probability density function of

$$f(y;\lambda;\phi) = \exp\left(\frac{y *\ \lambda - b(\lambda)}{\phi} + c(y,\phi)\right), \qquad (4.1)$$

where λ is the canonical parameter that depends on the regressors via a linear predictor and ϕ is a dispersion parameter that is often known. The function b(.) and c(.) are known and determine which member of the family is used. Family can be of **Normal**, **Binomial**, **Poisson** or other distribution. The generalized linear model is made up of a linear predictor.

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \qquad (4.2)$$

where the β_0 is the model intercept and β_i $(\mathrm{i} = 1, \ldots, \mathrm{p})$ are regressor coeficients associated the vector of regressors x_i when combined they are often a vector of regression coefficients B. For coefficients estimation the maximum likelihood (ML) using the iterative weighted least squares (IWLS) algorithm is used. At each stage of the iterative algorithm the model is increasing goodness-of-fit to the current set of data against increasing complexity of the model. The fitting of the parameters at each stage is done by maximizing the likelihood for the current model and the matching of the model to the data will be measured quantitatively by the quantity -2* L max which is called the deviance. The deviance is measured from that of the complete model, so that terms involving constants, the data alone, or the scale factor alone are omitted

Apart from the linear predictor the generalized linear model utilizes two functions. The mean function of y_i is given by $E[y_i|x_i] = \mu_i$ is defined as follows

$$g(\mu_i) = \eta_i, \qquad (4.3)$$

where g(.) is a known link function and the variance function of y_i is given by $VAR[y_i|x_i]$ as follows

$$VAR(y_i) = \phi V(\mu_i), \qquad (4.4)$$

is also called variance function where ϕ is the dispersion parameter. The selection of the link and variance function depends on the selected family. This model will be utilized multiple times during the modelling process in the following sections.

4.1 Probability Models

Probability models are used in predicting a binary (0|1) categorical variable. In case of micro claim level modelling these models are often used to predict the probability of a specific event occurrence. The events could be for example the claim settlement (probability of claim being settled in selected year), claim reopening (probability of claim being reopened after closure), claim development change (probability that the claim incurred value the total loss will change in any year). For the purposes of predicting the probability of these events the logistic regression model is used.

Logistic Regression

The logistic regression (logit regression) is a regression model where the dependent variables is binary. This model was developed by David Cox in 1958. Logistic regression can be defined as follows.

$$f(y;\pi) = logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta \tag{4.5}$$

Where π is the log odds of the mean (the probability of event occurrence), the η is a linear predictor as defined in (3.1). The link function in case of the logistic regression model is the logit link as described above.

4.2 Frequency Models

The frequency models are used in micro claim level modelling for predicting discrete variables (count variables). A typical example of discrete variable could be the number of occurred claims or the number of policies sold in the future (to predict how the portfolio will change, therefore how will it affect the possible number of occurred claims). These models are the Poisson model, Quasi-Poisson model and the Negative binomial model.

Poisson Model

The simplest distribution used for modeling count data is the Poisson distribution with probability density function

$$f(y;\mu) = \frac{\exp(-\mu) * \mu^{y}}{y!}, \tag{4.6}$$

where the probability density function is a special case of GLM model. The canonical link is g(μ) = log(μ) resulting in a log-linear relationship between mean and linear predictor. The variance in the Poisson model is identical to the mean, thus the dispersion is fixed at ϕ = 1 and the variance function is V (μ) = μ. In practice, the Poisson model is often useful for describing the mean μ_i but underestimates the variance in the data.

Quasi-Poisson Model

Another way of dealing with over-dispersion is to use the mean regression function and the variance function from the Poisson GLM but to leave the dispersion parameter ϕ unrestricted. Thus, φ is not assumed to be fixed at 1 but is estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but inference is adjusted for over-dispersion.

Negative Binomial Models.

A third way of modeling over-dispersed count data is to assume a negative binomial (NB) distribution for $y_i|x_i$ which can arise as a gamma mixture of Poisson distributions. Parameterization of its probability density function is

$$f(y; \mu;\ \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta) \cdot y!} * \frac{\mu^{y} * \theta^{\theta}}{(\mu + \theta)^{y+\theta}}, \tag{4.7}$$

with mean μ and shape parameter θ. Γ(·) is the gamma function. It also has φ = 1 but with variance function V (μ) = μ + μ·2·θ. Negative binomial probability density function is also a special case of GLM model.

4.3 Severity Models

The severity models are used in micro claim level modelling for predicting continuous variables. These continuous variables are often positive and right-skewed. As an example of these variables the claim incurred value, claim paid value or reserve could be presented. For these purposes the Gamma, Log-Normal or Poisson models are used.

Log-normal Model

The log-normal model is used for predicting positive continuous variables which are right-skewed. The log-normal model is estimated based on the maximal likelihood function which formula is as follows.

$$f(y; \mu; \sigma\ ) = \frac{1}{y}\frac{1}{\sqrt{2\pi}\sigma} exp\left[- \frac{(ln(y) - \mu)^2}{2\sigma^2} \right], \tag{4.8}$$

where $\mu = \ln\left(\beta_0 + \beta_1 x_{i1} + \cdots + \ \beta_p x_{ip}\right) - \frac{\sigma^2}{2}$. The estimates of β_i and σ are obtained during algorithm iteration.

Gamma Model

The gamma model is used for predicting positive continuous variables as an alternative to the log-normal model. The probability density function of the gamma distributed model can be described as follows.

$$f(y_i;\ \alpha_i;\ \beta_i) = \frac{1}{\beta_i^{\alpha_i}\Gamma(\alpha_i)}\ y_i^{(\alpha_i - 1)} e^{-(y_i/\beta_i)}, \tag{4.9}$$

where $\Gamma(.)$ is a gamma function, α_i is the shape parameter and β_i is the scale parameter of the gamma distribution. The mean and variance functions of y_i are as follows:

$$
\begin{aligned}
E(y_i) &= \alpha_i \beta_i, \\
Var(y_i) &= \alpha_i \beta_i^2,
\end{aligned} \tag{4.10}
$$

The gamma regression is focused on the scale parameter β_i (which is the source of variantion). For the shape parameter α_i it is assumed that this parameter is the same for all observations therefore the shape parameter is only a multiplier. The inverse value of the shape parameter is equal to the inverse of the dispersion parameter ϕ.

Beta Model

The beta regression model is used for predicting rates and proportional variables (between 0 and 1). The beta density function can be expressed as

$$
f(y_i;\, p_i;\, q_i) = \frac{\Gamma(p_i + q_i)}{\Gamma(p_i)\Gamma(q_i)} y_i^{p_i - 1}(1 - y_i)^{q_i - 1}, \tag{4.11}
$$

where 0 < $y_i < 1$ and p_i, q_i > 0 and $\Gamma(.)$ is a gamma function. The mean and variance functions of y_i are as follows:

$$
\begin{aligned}
E(y_i) &= p_i/(p_i + q_i), \\
Var(y_i) &= (p_i * q_i)/(p_i + q_i)^2(p_i + q_i + 1)\,,
\end{aligned} \tag{4.12}
$$

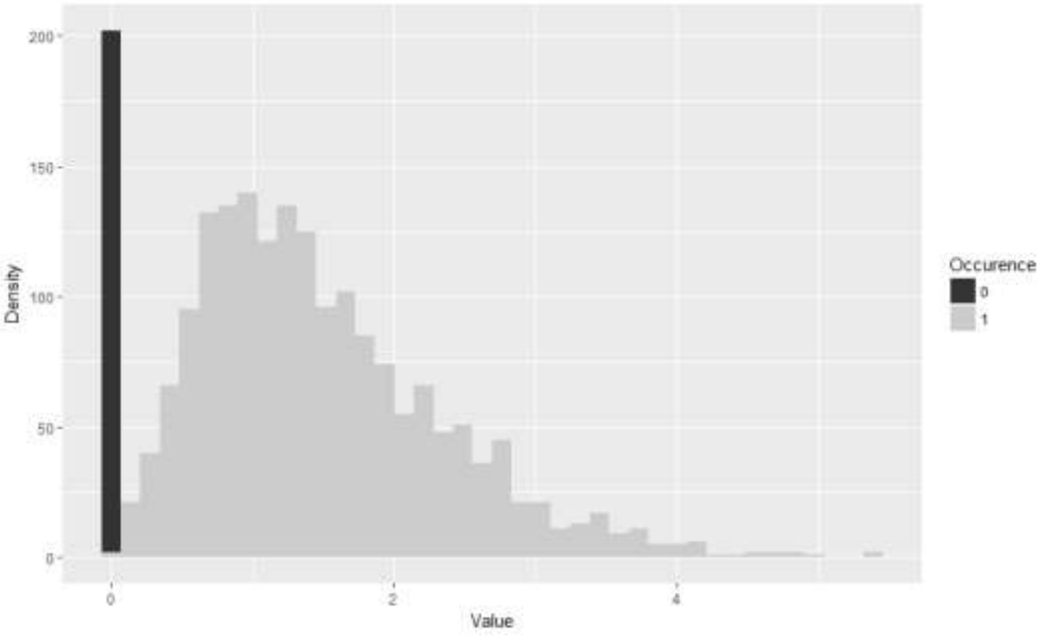
4.4 Hurdle Models

Hurdle model is an extension of the GLM models that was firstly introduced for count data to handle excess zeros and overdispersion. The hurdle model is a multiple-part model that specifies one process for no occurrence and another process for positive occurrences. The idea is that positive occurrences occur once a threshold is crossed, or put another way, a hurdle is cleared. The threshold is not crossed with probability $f_1(0)$ in which case there are no occurences. If the threshold is crossed, we observe positive occurrences, with probabilities coming from the density $f_2(y)$ with the associated truncated density $f_2(y)/\,(1 - f_2(0))$ that needs to be multiplied by $(1 - f_1(0))$ to ensure probabilities sum to one. The resulting hurdle is then defined as

$$
P(y_i) = \begin{cases} f_1(0) & y_i = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y_i) & y_i > 0. \end{cases} \tag{4.13}
$$

The $f_1(0)$ is typically a binary Logistic regression model. This process predicts whether an observation takes a positive occurrence or not with the probability of π_1. The $f_2(y)$ is usually a truncated/censored Poisson or Negative Binomial (for discrete) or Gamma (for continuous) model. The truncated/censored means that model was fitted only of positive occurrences. For difference between truncated/censored see Cameron and Trivedi 2013. The model can have multiple hurdles, for example hurdle equal to 0 occurrences, hurdle equal to 1 occurrence and truncated model for all other positive occurrences. The graphical description of Gamma hurdle model is provided in the following figure.

Figure 5: *Gamma Hurdle model*



The black area is the hurdle to be cleared by the binary Logistic regression model f_1 and the grey is the truncated/censored Gamma model f_2 variable distribution.

The mean and variance functions of a general hurdle model (in case of two components) can be presented as

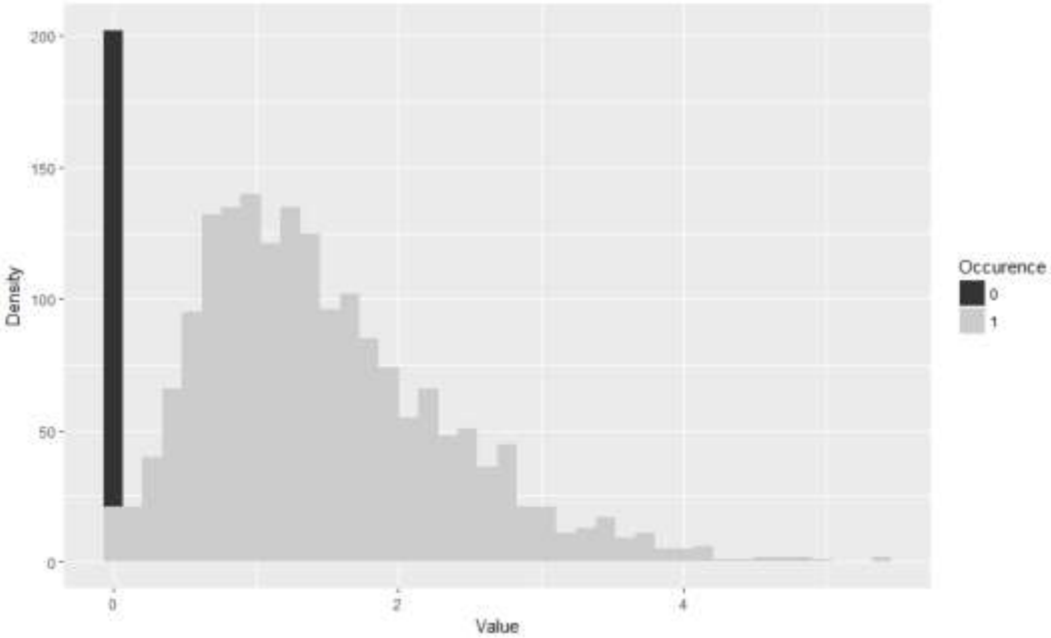
$$E(y) = (1 - \pi_1)\, E_2(y|y > 0)\,,$$
$$Var(y) = (1 - \pi_1) Var_2(y|y > 0) + \pi_1(1 - \pi_1) E_2(y|y > 0) \qquad (4.14)$$

where π_1 is the probability of positive occurrence, $E_2(y|y \neq 0)$ is the mean for positive occurrences from the truncated/censored model and $Var_2(y|y \neq 0)$ is the variance of the truncated/censored model. For more details on this expression and hurdle models in general, see Cameron and Trivedi (2013).

4.5 Zero-inflated Models

Zero-inflated model is an also extension of the GLM model. The zero-inflated model differs from the hurdle model in definition what is the source of the variable no occurrence. In the Hurdle model the no occurrences are assumed to be a from an independent source (structural origin). The Zero-inflated model assumes than there are two sources of no occurrence. The structural origin and the sampling origin. The sampling origin represents that no occurrences are the taken from the variable distribution. As a result, the no occurrences are generated by two sources. First source is the Logistic regression model (structural) and the variable distribution (sampling). The same variable distributions as in case of the hurdle model can be assumed. The graphical description of the Zero-inflated model is presented in the following figure.

Figure 6: *Gamma Zero-inflated model*



In comparison with the hurdle model the black area is smaller because the no occurrences (= 0) are not only generated by the Logistic regression model f_1 but also by the Gamma model f_2. The probability for the Zero-inflated model can be presented as

$$P(y_i) = \begin{cases} \pi + (1-\pi) f_2(0) & y_i = 0 \\ (1-\pi) f_2(y_i) & y_i > 0, \end{cases} \tag{4.15}$$

where π is the proportion of no occurences and this in increased by the probability of no occurrences from the second model as $(1-\pi)f_2(0)$ to ensure that probabilities sum to one.

The mean and variance functions of a general zero-inflated model (in case of two components) can be presented as

$$\begin{aligned} E(y) &= (1-\pi)\mu_2, \\ Var(y) &= (1-\pi)(\sigma_2^2 + \pi\mu_2^2), \end{aligned} \tag{4.16}$$

where μ_2 is the mean and σ_2^2 is the variance of the full model f_2. For more details on the zero-inflated models see Mills (2013) or Cameron and Trivedi (2013).

4.6 Framework Conclusion

This section provided and overview of the Generalized Linear Models (GLM) and its extensions. The GLM framework is used in the practical model implementation (chapter 6). The proposed model is comparing the GLM, Hurdle and Zero-inflated models (chapter 7). From the presented models the frequency models are not used because the proposed model focuses on the severity modelling and therefore frequency modelling is out of scope of this work. In the following chapter the provided dataset is prepared and explored.

5. Dataset

We have a dataset at our disposal with detailed information on the development of individual claims from Motor Third Party Liability (MTPL) line of business which evolves from occurrence of the accident until the settlement or censoring of the claim. This dataset was obtained from the **Czech Insurers‘ Bureau** (www.ckp.cz). Dataset contains records from year 2004 till 2016. The year 2004 till 2010 will be used for model implementation (chapter 6) while the years 2011 till 2016 will be used for model evaluation. (chapter 7). The table 1 illustrates the structure of individual claims. Each record stores static columns Claim ID, Occurrence Date and Registration Date and dynamical columns Change Date, Amount in CZK and Log Type columns. Static columns are associated with the Claim ID and provide a basic claim description. Static columns value does not change when the claim is modelled while the dynamical columns describe the individual claim change in time.

Table 1: Claim Occurrence History Log Data.

Claim ID	Occurrence Date	Registration Date	Change Date	Amount CZK	Log Type
78002	14.03.2005	26.04.2005	02.05.2005	540 000	Reserve
85134	25.05.2005	28.05.2005	07.05.2005	6 000	Reserve
78002	14.03.2005	26.04.2005	15.05.2005	-540 000	Reserve
86514	05.05.2005	17.05.2005	20.05.2005	254 560	Reserve
78002	14.03.2005	26.04.2005	25.05.2005	540 000	Paid
85134	25.06.2005	28.06.2005	04.06.2005	12 000	Reserve
85134	25.06.2005	28.06.2005	15.06.2005	-12 000	Paid
…	…	…	…	…	…

5.1 Data Preparation

For the purposes of analysis, it is necessary to adjust the dataset by changing it from continuous time framework to the discrete time framework as described in Pigeon et al. (2014). One-year period discrete time framework will be used. All amount changes for each claim will be summed by the Log Type. The results for paid changes are as follows.

Table 2: Cumulative paid value at the end of the year for each claim.

Claim ID	Occurrence Date	2005	2006	2007	2008	…
78002	14.03.2005	540 000	540 000	540 000	540 000	…
85134	25.05.2005	6 000	25 000	72 000	160 000	…
86514	05.05.2005	0	0	0	0	…
…	…	…	…	…	…	…

By summing the cumulative paid value and the reserve value at the end of the year the incurred value is obtained. The cumulative incurred value development can be described as.

Table 3: Cumulative incurred value at the end of the year for each claim.

Claim ID	Occurrence Date	2005	2006	2007	2008	…
78002	14.03.2005	540 000	540 000	540 000	540 000	…
85134	25.05.2005	18 000	56 000	84 000	160 000	…
86514	05.05.2005	254 560	124 000	80 200	0	…
…	…	…	…	…	…	…

For the modelling purposes, it is necessary to prepare data structure describing each claim and its variables at the given year. This work proposes the following structure.

Table 4: Analytical table used in model

Claim ID	Loss Year	Dev Year	Month	Prev. Incurr.	Prev. Paid.	Incurred	Paid	…
78002	2005	2005	3	0	0	540 000	540 000	…
78002	2005	2006	3	540 000	540 000	540 000	540 000	…
78002	2005	2007	3	540 000	540 000	540 000	540 000	…
…	…	…	…	…	…	…	…	…

Previous incurred and paid describes the claim state at the start of the year. This year incurred and paid describes the claim state at the end of year. The table 5 describes all variables assumed to be needed for modelling.

Table 5: Analytical table variables description

Variable	Description
Claim ID	ID of the Claim
Loss Year	Loss Year from the Occurrence Date
Dev Year	Development Year
Development year	Difference between Development Year and Loss Year
Previous Incurred	Cumulative Incurred Value at the end of previous Year
Previous Paid	Cumulative Paid Value at the end of previous Year
Incurred	Cumulative Incurred Value at the end of current Year
Paid	Cumulative Paid Value at the end of current Year
Reserve	Difference between Incurred and Paid
Ultimate	Final cumulative Incurred Value found for the claim
Reported	Binary variable if the Incurred > 0 for the first time.
Closed	Binary variable if the Incurred = Paid and claim settled.
Order	Rank variable based on order of the Occurrence Dates
Quarter	Quarter when the Loss Occurred
Month	Month when the Loss Occurred
Link Factor	Incurred /Previous Incurred
Paid/Incurred Ratio	Paid/Incurred
Change in Incurred	Binary variable if Incurred value changed from Previous Incurred
Change in Payment	Binary variable if Paid value changed from Previous Paid

5.2 Variable Description

This section describes available variables of the provided dataset (see **table 5)**. These variables can be used to create following variable groups.

- **Description** variables (Claim ID, Ultimate, Loss Year, Quarter, Month, Order)
- **State** variables (Incurred, Paid, Reserve, Closed)
- **Time** variables (Dev Year, Development year)
- **Change** variables (Link factors variables, Paid to Incurred ratio, Change in variables)

Description variables provide basic information about the individual claim. These variables are static and therefore their value does not change during the claim development. Other variable groups change its values based on the development year. The state variables present the incurred value, paid value and reserve value of the individual claim. The time variable holds information about the development year and the difference between loss year and dev year. Lastly the change variables hold information about the link factor variable, incurred to paid value ratio and the change in incurred value and paid value variables.

Description Variables

Claim ID is numerical variable to distinguish between individual claims. There exists 18016 unique claim IDs in the provided dataset. Each claim ID is represented by multiple rows to presents the development history of this claim. The claim ID variable will be used to create the training and testing dataset. When the claim ID is randomly selected its complete history is moved into the training/testing dataset.

Loss year, quarter, month and order variables are discrete number variables derived from the claim original occurrence date. Together with the claim count the claim frequency per year, quarter and month. In addition, the occurrence date was used to determine the claim order. These variables will be tested in the modelling if they hold additional information that can improve the individual claim level model. Finally, the order is obtained as the chronological order in which the claims have occurred.

Ultimate is the ultimate incurred value $Y_{i,\infty}$ obtained as the last available incurred value of the i-th claim as of year 2016 for the provided claim. This variable will be used in model evaluation.

State Variables

Each row of the presented dataset contains the incurred, paid and reserve variable. For the previous incurred value $Y_{i,j-1}$and the incurred value $Y_{i,j}$. Both the previous incurred value and incurred value represents the cumulative incurred value of i-th claim at the end of j-1 and j-th development year. The difference between these values can be interpreted as the incremental adjustment $X_{i,j}$ of i-th claim in j-th development year. This description can also apply to the variables previous paid value $P_{i,j-1}$ and the paid value $P_{i,j}$. The reserve variable $R_{i,j}$ is obtained by applying the formula (3.10). These prediction of variables is the result of the claim development process. The closed binary variable $I(S_i)$ represents if the claim has already been settled. The claim is settled when the $Y_{i,j}$ = $Y_{i,N}$ and $Y_{i,j}$ > 0. This variable does not consider fake claim which ends with 0 ultimate claim value. These claims were removed from the dataset. This variable will be predicted with the probability of closure $Pr(S_i)$ model.

Time Variables

These variables are containing the time aspect of the state variables. The dev year represents the actual year of the state variables. The development year represents the j-th year of i-th claim after occurrence. These variables will be tested in the modelling if they hold addititonal information that can improve the individual claim level model.

Change Variables

These variables represent the change in the state variables that occurred in the j-th development year. The link factor $\lambda_{i,j}$ presents the relative difference between the incurred value $Y_{i,j}$ and the previous incurred value $Y_{i,j-1}$. This variable will be predicted with link factor GLM, HM and ZIM models. The paid to incurred ratio $\varsigma_{i,j}$ represents the relative difference between the incurred value $Y_{i,j}$ and $P_{i,j}$. This variable will be predicted with the beta regression model. The final variable change in incurred value is the binary variable for the HM and ZIM model to distinguish development years when the incurred value has not changed. This variable is equal to 0 when no change occurred ($\lambda_{i,j} = 1$) and 1 when change occurred ($\lambda_{i,j} \neq 1$).

5.3 Describing Claim Development

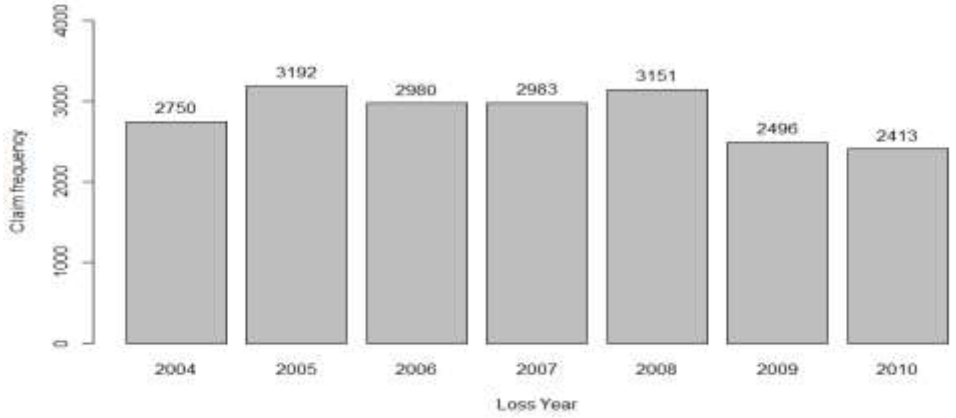
This section will provide basic introduction to the modelled variables and explore their development. The proposed variables are presented as they change during the development years. Firstly, the number of occurred claims per year will be standardized with the total exposition for each year to provide the claim frequency. Then the individual claim development of not reported, open and closed claims is considered. Then the development of incurred, paid and reserve variables for each development year is presented. Lastly the development of the link factor is provided.

Claim Frequency

When describing the overall development of the claims for the provided dataset it is important to look at how the years differ in their exposure. The dataset contains information about claims arising from the Motor Third Party Liability (MTPL) line of business. These claims are representing only material damages done by non-insured cars. The total number of claims is presented in figure 7. Firstly, it is important to look at the number of occurred claims for each year. Number of occurred material claims tend to be very similar with on average 2 976 material claims occurred each year with only exception of the years 2009 and 2010 where the number of claims occurred has decreased to 2 537 and 2413. The main factor which led to this decrease was the overall decrease in number of car accidents occurred. This information is collected by the Czech police car accidents statistics[1]. In year 2008 the total number of car accidents registered by Czech police was 160 376 and in the following year 2009 the total number decreased to 74 815. The reason for this change is that there is new law enforced which changes the rules when the police must attend to the car accident. The main change is that the car accident loses must now be more than 100 000 rather than 50 000 before year 2009. Overall the number of car accidents was lower but not so much as would the official statistics would describe.

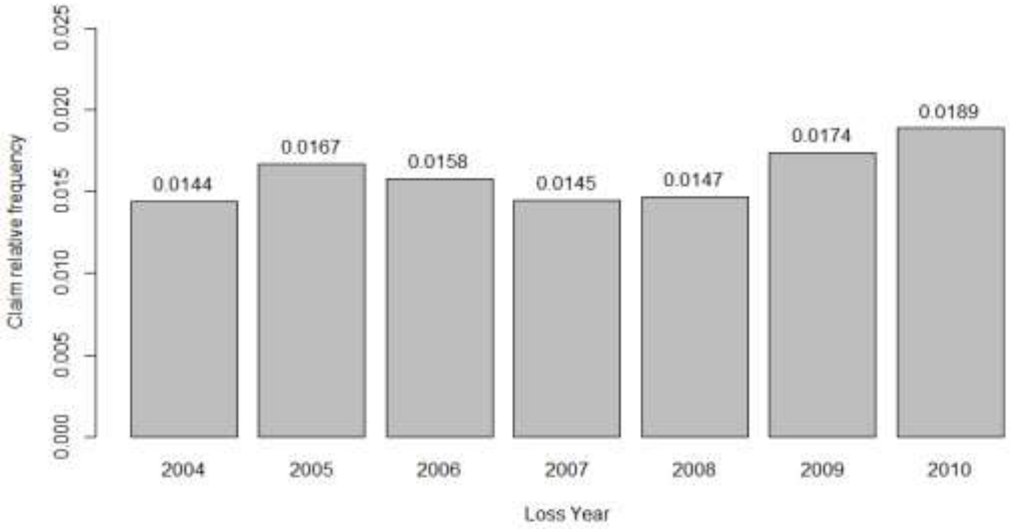
[1] http://www.policie.cz/clanek/statistika-nehodovosti-900835.aspx?q=Y2hudW09OA%3d%3d

Figure 7: *Claim counts per year.*



To complete the picture the following figure was created where the total number of occurred claims are compared with the estimate total number of non-insured cars. This information is presented in the figure 8.

Figure 8: *Claim frequency per year.*



Source: Supin s. r. o.

The relative number of claims was around 0.015 every year except for the years 2009 and 2010 where this value was increased to 0.018. This may in fact because of decrease in the exposition between year 2008 and 2009. The table 6 presents how the figures 7 and 8 were obtained.

Table 6: *Claim frequency per year.*

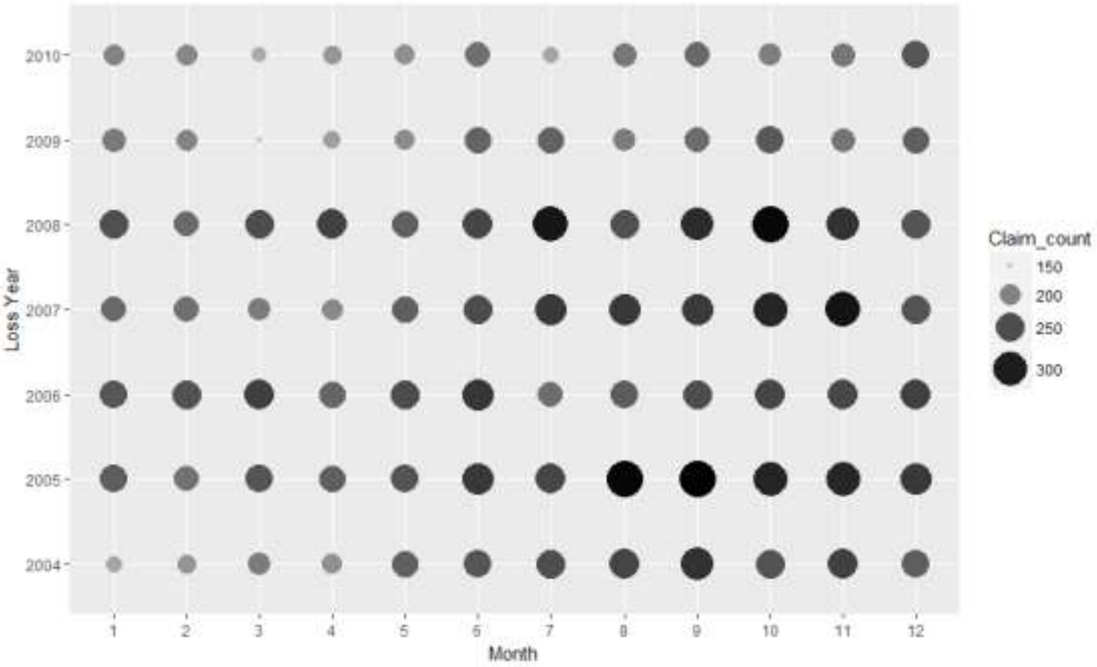
Loss Year	2004	2005	2006	2007	2008	2009	2010
Claim count	2 750	3 192	2 980	2 983	3 151	2 496	2 413
Exposure	191 281	191 379	188 389	205 621	214 029	143 725	127 780
Frequency	0.0144	0.0167	0.0158	0.0145	0.0147	0.0174	0.0189

Source: Supin s. r. o

Claim Frequency Development

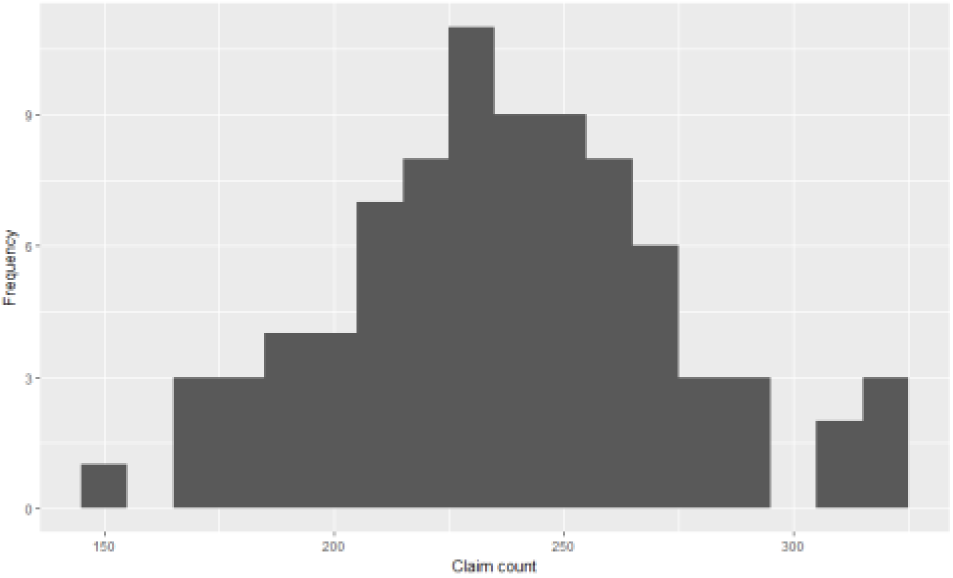
Claim frequency in the loss year will be presented in the number of occurred claim per month in the loss year. The claim frequency is not uniformly distributed as displayed in the figure 10. The highest number of claims was observed in 2005, 2007 and 2008 from July to November with the maximum of 325 claims in October of 2008. On the contrary, the lowest number of 150 claims was observed in March of 2009. The average number of observed claims is 220 from January to June of the year and 254 from July to December.

Figure 9: *Number of material claims occurred per month for each loss year.*



In addition to the relation between month and loss year the overall claim frequency distribution is presented in figure 10. The claim frequency histogram is relatively normally distributed with majority of claim counts occurring between 180 and 280 per month. The claim frequency is not subject of this work model but the information was included to present complete picture about the provided dataset.

Figure 10: *Claim frequency distribution.*



Claim Development by Claim State

In the previous section the number of occurred claims for each loss year and month was presented. This section will focus on individual claim development, i.e. how claims are developing during each year. Claim states can be described as:

1) Not Reported $I(S_{i,j}) = 0$
   → Claims that were not reported by the end of year.
2) Open $I(S_{i,j}) = 1$ & $I(O_{i,j}) = 0$
   → Claims that were reported and remained open at the end of year
3) Closed $I(S_{i,j}) = 1$ & $I(O_{i,j}) = 1$
   → Claims that were reported and were closed by the end of year

Not reported is the first state that the claim has until it is reported to the insurance company. The not reported claims can remain in this state for several years before they are reported. An alternative name for these claims in the literature is the incurred but not reported claims (IBNR). The second state the claim can belong to is the open state. The claim is marked as open when it is reported to the insurance company but not all payments have been done. An alternative name for these claims in the literature is the reported but not settled claims (RBNS). The final state for each claim is the closed claim. For these claims all payments are paid out and no further development is done to these claims. The proposed dataset was explored and number of claims per claim state have been collected and presented in table 7.

Table 7: *Claim development by claim state.*

Loss Year	Claim State	Development Year					
		0	1	2	3	4	5
2004	Not Reported	650	22	5	1	1	1
	Open	1392	483	289	213	168	144
	Closed	760	2297	2508	2588	2633	2657
2005	Not Reported	655	39	12	1	0	0
	Open	1710	640	406	295	259	44
	Closed	875	2561	2822	2944	2981	3196
2006	Not Reported	667	42	12	2	2	0
	Open	1410	476	288	218	24	8
	Closed	964	2523	2741	2821	3015	3033
2007	Not Reported	643	56	14	7	0	0
	Open	1243	535	274	203	50	21
	Closed	1144	2439	2742	2820	2980	3009
2008	Not Reported	633	62	15	5	1	1
	Open	1361	480	334	260	141	77
	Closed	1214	2666	2859	2943	3066	3130
2009	Not Reported	475	49	11	4	1	0
	Open	1031	368	299	208	77	0
	Closed	1031	2120	2227	2325	2459	2537
2009	Not Reported	475	49	11	4	1	0
	Open	1031	368	299	208	77	0
	Closed	1031	2120	2227	2325	2459	2537
2010	Not Reported	482	78	19	4	1	0
	Open	870	382	285	245	28	4
	Closed	1061	1953	2109	2164	2384	2409

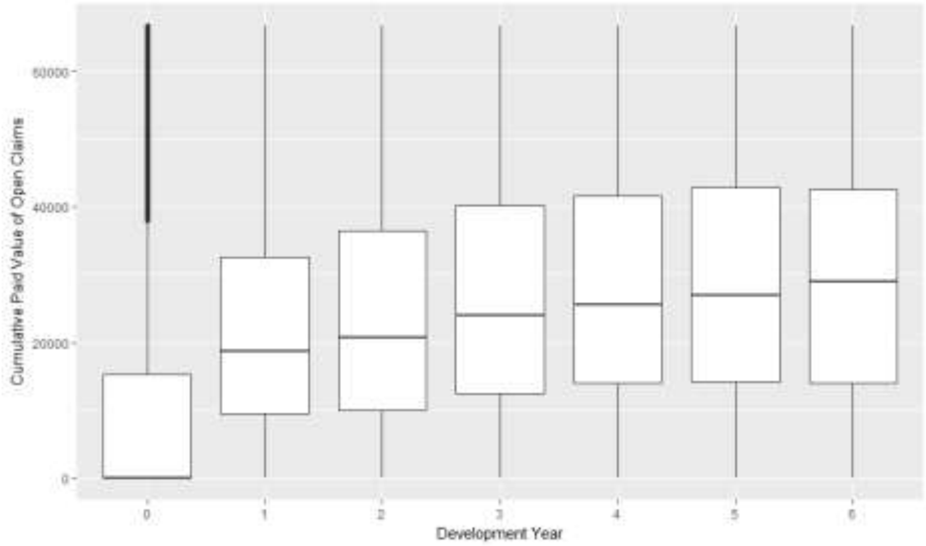
At the end of the loss year (0 development year) about 20-25 % of claims remain not reported and the 75-80 % reported claims are in 30-40 % of cases closed and in 40-45% of cases remained open. Not reported claims are by the end of first development year nearly all reported. From all the claims about 20 % of claims are long running claims which need more than 2 years to be processed while the majority will be closed by the end of first development year. At the end of fifth development year only small fraction of claims remain open. There exists a possibility of claim being reopened and this occurs in the provided dataset in 1 % of cases but the proposed model in chapter 6 is neglecting this option.

By this section the claim frequency exploration is concluded and the individual claim level model implementation will focus on the individual claim level development rather than modelling the claim frequency.

Paid Value Development for Open Claims

This section will focus on open claims paid value development. The cumulative paid value for all the open claims by development year is presented in **figure 11.** In the first year, there exist only few material claims, where payment has occurred and therefore the paid value median is equal to 0 and the most of paid values are less than 50 000. In the second year the claims starting to be paid out and the median will increase 23 720 and the most of paid values are less than 100 000. From the development year the increase in median is slower and the median will reach 26 752 by the sixth development year and the most of paid values will be less than 150 000.

Figure 11: *Open claims paid value development.*

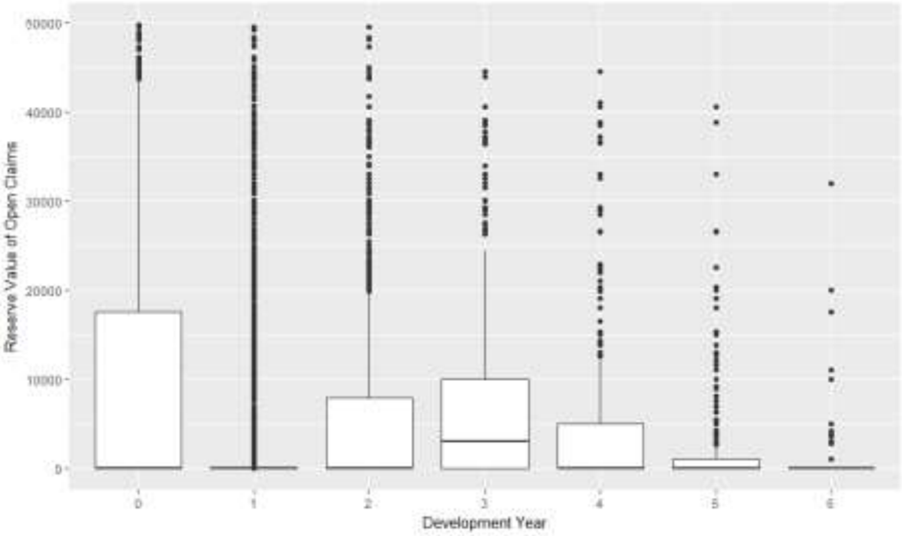


The paid value distribution is positively skewed and by each development year the number of still open claims is very small (see table 7) therefore the variability is very high in the latter development years.

Reserve Value Development for Open Claims

This section will focus on reserve value development of open claims. For closed claims, the reserve value is always equal to 0. No additional payments are expected arising from these claims once they are settled (Reopening is not considered). The reserve value differs based on the moment the claim is reported. For claims reported in the first year the initial reserve is created. This reserve will be paid out in the following year therefore in the first development year there are very small reserves left. In the following years the number of still open claims is very small and therefore the reserve value is visible in comparison with the first development year. The reserve value is obtained as a difference between the incurred value and the paid value in the modelled year.

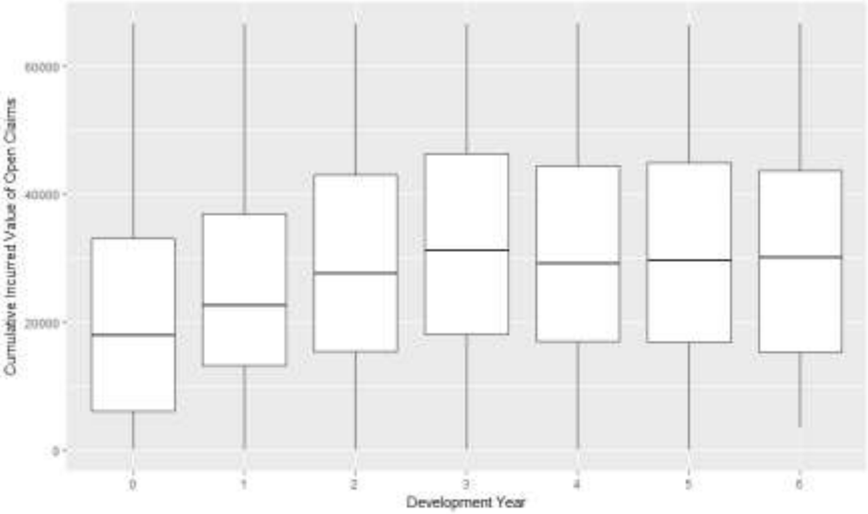
Figure 12: *Open claims reserve value development.*



Incurred Value Development for Open Claims

The development of incurred value for open claims is the mixture of the previous developments. The incurred value development is presented **figure 13**. The median of the incurred value is increasing till the third development year. After the third yeas the median of the incurred value is relatively the same. In numbers the median is equal to 22500 in the zeroth development year and 26875 by the sixth development year.

Figure 13: *Open claims incurred value development.*

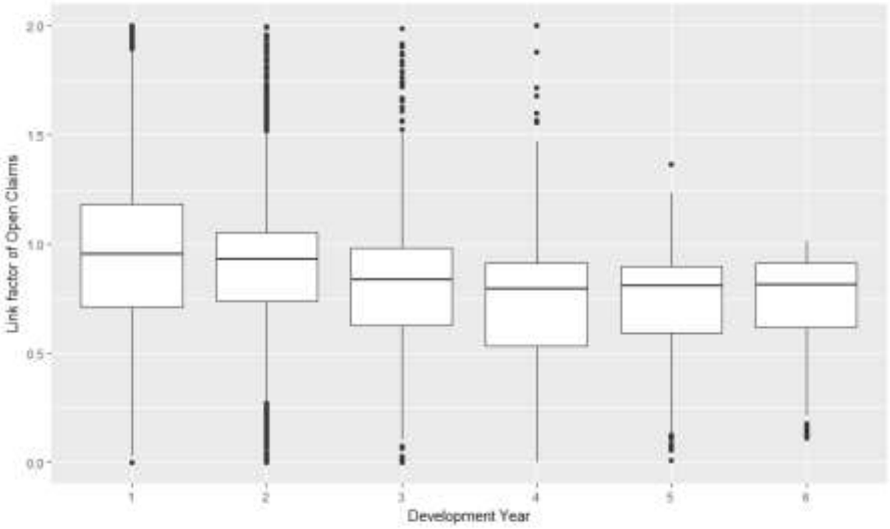


The incurred value will not be modelled directly rather it will be result of the previous incurred value and a link factor.

Link Factor Development for Open Claims

The link factor represents the incurred value change that have occurred in the development year. The link factor variable development is presented in the following figure. The link factor median for all development yeas is less than 1 which indicates that the change is more often a decrease in the incurred value from the previous year rather than increase as displayed in figure 14.

Figure 14: *Open claims link factor development.*



The figure 14 was created from link factors that were not equal to 1 therefore, only link factors that represented a change in the incurred value. In the first development year the median close to 1. In the following years the link factor median is decreasing and ending close to 0.8. Based on the figure 14 the link factor is assumed to decrease in the following years and therefore, the incurred value is expected to decrease if the incurred value should change in later years.

Concluding the Claim Development

This section focused on presenting the provided dataset and presenting how claims develop in the dataset. The section presented how from the standard claim structure the model input table was prepared. Then from the model input table the variable development was observed and presented in this chapter. The following chapter will present how the proposed model will work to predict the future individual claim development.

6. Model Implementation

The individual claim level model implementation will be presented in this chapter. Firstly, the model structure is presented. The individual claim level model handles separately based on the claim state. For claims in not reported state (IBNR) the **incurred but not reported models** are used. For claims in open state (RBNS) the **development models** are used. These models are based on variables presented in the chapter 4. The presented models are not considering the claim frequency as this not the focus of this work. This chapter assumes that the claim counts are fixed and only the individual claim level development is assumed and modelled.

6.1 Individual Claim Level Model

The practical implementation of the individual claim level model is proposed in this section. When modelling the individual claim level model, it is assumed that the available dataset will be split to a training and testing part. In the training part the individual models are fitted and during the testing part the models are used to predict the individual claim development and save the result to the analytical table for years that are unknown. For example, when the model contains history for years 2004 and 2016 the model can be used to evaluate claims from year 2004 to 2010 and train models based on their history in these years. Then in the testing part the model will focus on modelling the development of claims originating from years 2004 to 2010 (claims that were not used in training the model) in the years 2011 to 2016. Then the model results are collected and saved to the analytical table. One analytical table is needed for predicting the model and the second will be used for model testing. The table 8 contains information about the variables that the model uses and how they are obtained when testing the model.

Table 8: *Analytical table variables source*

Variable	Description
Claim ID	Available at the start of modelling.
Loss Year	Available at the start of modelling.
Dev Year	Updated based on the model.
Dev	Updated based on the model.
Previous Incurred	Calculated with usage of link factor.
Previous Paid	Calculated with usage of paid ratio.
Incurred	Calculated with usage of link factor.
Paid	Calculated with usage of paid ratio.
Reserve	Incurred - Paid
Ultimate	Available at the start of modelling.
Reported	**Obtained from probability of claim being reported model.**
Closed	**Obtained from probability of claim closure.**
Order	Available at the start of modelling.
Quarter	Available at the start of modelling.
Month	Available at the start of modelling.
Link Factor	**Obtained from the link factor model.**
Paid/Incurred Ratio	**Obtained from the beta regression model.**
Change in Incurred	**Obtained from probability that incurred value change during the year.**
Change in Paid	**Obtained from probability of payment occurrence model**

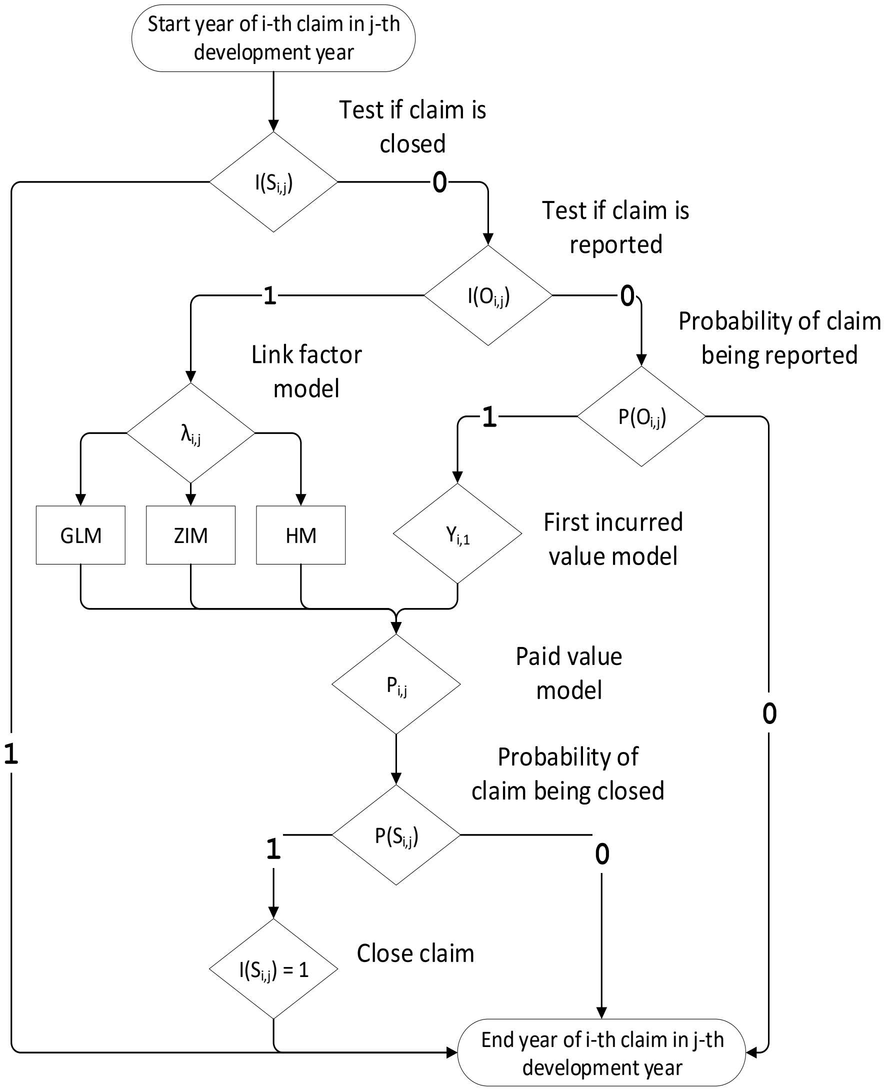
The variables known at the start of the year are as follows: **Loss Year, Development Year, Order, Quarter, Month, Previous Year Incurred Value, Previous Year Paid Value.**

Predicted variables are as follows: This Year Incurred Value, This Year Paid Value, Reserve, Closed, Grossing Factor Incurred, Grossing Factor Paid, Change in Paid and Change in Incurred value.

6.2 Modelling Claim Change

For modelling claim change two support variables $I(O_{i,j})$ and $I(S_{i,j})$ be used as introducted in chapter 2. These variables are used to determine if claim have been reported, is open or has been already closed. The $I(O_{i,j})$ is an indicator of i-th claim being reported in the j-th development year. When the $I(O_{i,j})$ = 0 the claim is not reported and therefore only the incurred but not reported models are used. On the other hand, in case of $I(O_{i,j}) = 1$ the claim has been reported and the claim development models are used. After the claim has been reported $I(O_{i,j}) = 1$ the closure model will determine if the claim has been settled as $I(S_{i,j}) = 1$ or remained open $I(S_{i,j}) = 0$ in the j-th development year. When the claim has been settled no additional modelling is done in the following years and all reserves are paid out. The following figure represents this model.

Figure 15: *Modelling claim change in one year.*


Start year of i-th claim in j-th development year
Test if claim is closed
I(S_{i,j})
0
1
Test if claim is reported
I(O_{i,j})
1
0
Probability of claim being reported
Link factor model
λ_{i,j}
P(O_{i,j})
1
0
GLM
ZIM
HM
Y_{i,1}
First incurred value model
P_{i,j}
Paid value model
Probability of claim being closed
P(S_{i,j})
1
0
I(S_{i,j}) = 1
Close claim
End year of i-th claim in j-th development year

The proposed model parts will be further discussed and presented in detail in this section. Examples presented chapter are results of only one model run on the training dataset (on 90 % of the dataset) and contains an evaluation of each submodel on the testing dataset (on 10 % of the dataset). The complete model is expected to be run multiple times and reevaluated. The complete model results and evaluation is available in chapter 7.

6.3 Incurred but Not Reported Models

The incurred but not reported models (IBNR) focus on handling individual claims that have not yet been reported. The individual claim is firstly subjected to the probability of claim being reported $P\left(O_{i,j}\right)$ model and based on this model output the claim is either reported or remain unreported. For reported claims the first incurred value is predicted as $Y_{i,j}$ and in the following years the claim is processed by the reported but not settled model until the claim is closed. The IBNR model can be viewed as a hurdle model based on how it was implemented.

Probability of Claim Being Reported

The probability of claims being reported is predicting the outcome of binary variable $I\left(O_{i,j}\right)$ with a logistic regression model (as presented in the section 4.1). The model was created based on the formula (6.1)

$$I\left(\hat{O}_{i,j}\right) \sim Loss\ Year_i + Month_i + \log(Order_i) + Development\ year_{i,j}, \qquad (6.1)$$

where the loss year is the year that the individual claim has occurred, the month is the month of the loss year, the order is the chronological id of the provided claim and the development year is the difference between the j-th year and the loss year when the claim has occurred. When predicting the probability of claim being reported the incurred, reserve and paid value are equal to zero and therefore cannot improve the model. The formula (6.1) variables were selected based on ANOVA. The variables known at the start of the year were assumed for this model and only those that proved to be significant (Pr(>Chi) < 0.001) were kept in the model. The ANOVA results are presented in table 9.

Table 9: *Probability of claim being reported model variable significance.*

Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			38 529	53 337	
Loss Year	6	69.5	38 523	53 267	0.000000
Development Year	6	17 248.0	38 517	36 019	0.000000
Month	11	388.6	38 506	35 631	0.000000
Order	1	118.9	38 505	35 512	0.000000

The development year variable proved to be the most significant from the selected variables but even with all these variables the residual deviance is still very large. The model was fitted and its results are enclosed in the Appendix.

The model was fitted on the training dataset (90 %) and for the testing dataset (10 %) the model will be evaluated. For this purpose, a Confusion matrix was created and presented in table 10.

Table 10: *Probability of claim being reported confusion matrix.*

Confusion Matrix		Reported	
		0	1
Predicted	0	1 735	429
	1	442	1 568

Accuracy : 0.7913

95% CI : (0.7787, 0.8036)

Sensitivity : 0.7970

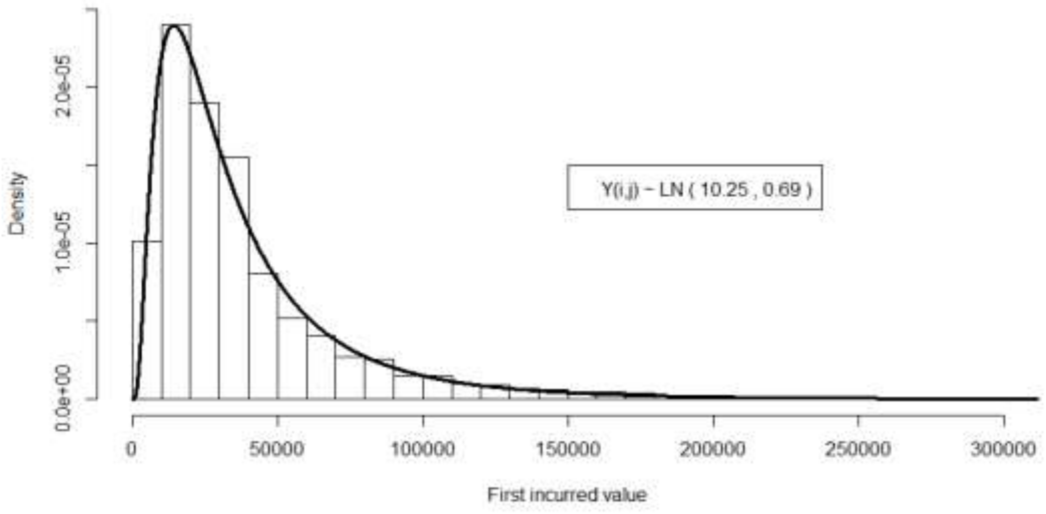
Specificity : 0.7852

The model predicts the $I\big(\hat{O}_{i,j}\big)$ correctly on average in 0.7913 number of cases. When the $I\big(\hat{O}_{i,j}\big) = 1$ the first incurred value model is used to determine the initial $Y_{i,j}$.

First Incurred Value Model

The first incurred value model is used to determine the initial $Y_{i,j}$ as an outcome of a random variable from the Log-normal distribution. The modelled variable from the training dataset is presented in the figure 16. The histogram represents the observations of the first incurred value and the line represents the estimated Log-normal distribution to fit this variable. The assumed Log-normal distribution would have mean on the log scale equal to 10.25 and variance on the log scale equal to 0.69. (These values might change depending on the model run).

Figure 16: *First incurred value distribution*



To predict the first incurred value distribution a GLM gaussian model with log link was created. The proposed model was fitted based on formula (6.2).

$$\hat{Y}_{i,j} \sim Development\ year_{i,j} + \log(Order_i), \tag{6.2}$$

where the development year is the difference between the j-th year and the loss year when the claim has occurred and the order is the chronological id of the provided claim

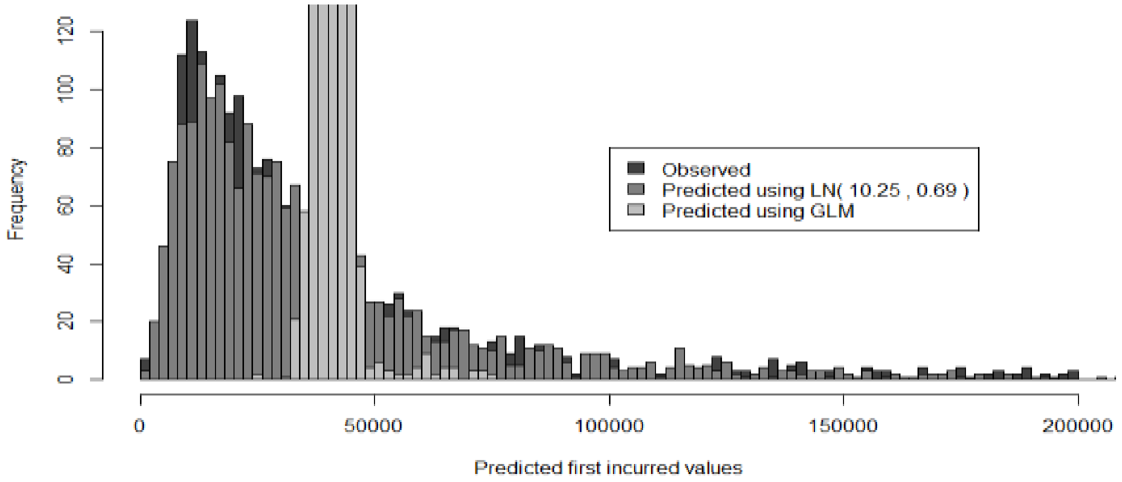
The model was fitted based on the subset of training data (truncated) that has changed its incurred value from 0 to higher than 0 in a development year. The formula (6.2) variables were selected based on ANOVA. The variables known at the start of the year were assumed for this model and only those that proved to be significant (Pr(>Chi) < 0.1) were kept in the model. The lower boundary was set because the assumed variables proved to be an insignificant except for the development year. The ANOVA results are presented in table 11.

Table 11: *First incurred value model variable significance*

Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			18402	87 172 907 892 599	
Development Year	6	151 542 294 217	18396	87 021 365 598 382	0.00001
Order	1	16 730 785 642	18395	87 004 634 812 740	0.06000

The development year can explain a small proportion of the variable variance but this it is a very poor fit. The order variable was kept in the model to allow for small variable variations. The fitted model is provided in the appendix. As expected, the model is not very good at estimating the first incurred value. Assumed variables are not enough to explain the variability in the first incurred value. The figure 17 represents the observed problem with this model.

Figure 17: *First incurred value model performance.*



Based on the mean value the both models are very similar, but the GLM model is not able to provide necessary variability based on the input variables a therefore its performance is not very good. (Nearly all predictions are small adjustments of the variable mean). The final model will use this model, but this information will be taken into consideration when evaluating the complete model. When this model will be used apart from the prediction the standard error will be collected and marked as s. e. $\hat{Y}_i$. The standart error will be used in the model evaluation chapter 7.

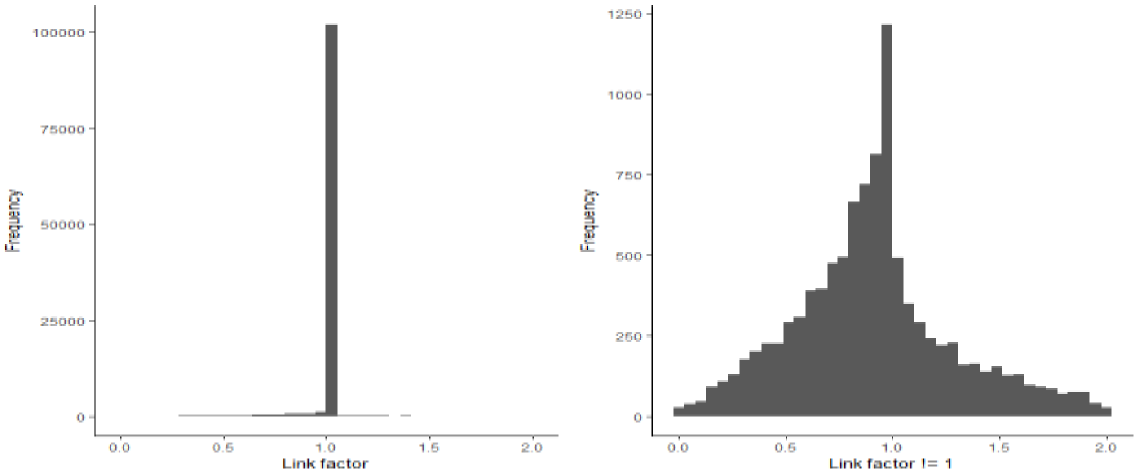
6.4 Reported but Not Settled Models

Reported but not settled models predicts how the claim incurred, paid and reserve variable changes during the development year after the claim have been reported. The incurred value obtained from the link factor model where the link factor is predicted with usage of GLM/HM/ZIM models. The paid value is obtained from the paid ratio model where the paid ratio is obtained from the HM model.

Link Factor Models

The link factor model is presented in 3 scenarios as the generalized linear model (GLM), hurdle model (HM) and zero inflated model (ZIM) based on the gamma distribution. These scenarios will be compared in the model evaluation chapter. Before the models will be presented in more detail the link factor $\lambda_{i,j}$ variable is presented in the figure 18.

Figure 18: *Link factor distribution for open claims.*



The left figure presents all link factors observed for all open claims in the training dataset. The right figure contains only a subset of link factors that are not equal to 1 (only claims that have changed in the development year). The ratio between $\lambda_{i,j} = 1$ and $\lambda_{i,j}\ ! = 1$ is 11 to 1.

The left figure will be a base for the full gamma model. This model will be fitted based on all available open claims. The right figure will be a base for the subset gamma model. This model will leave out all open claims based on the condition $\lambda_{i,j}\ ! = 1$. Both models will be fitted based on the formula (6.3).

$$\hat{\lambda}_{i,j} \sim Loss\ year_i + \ Development\ year_{i,j} + \log(Y_{i,j-1}), \qquad (6.3)$$

where the loss year is the year that the individual claim has occurred, the development year is the difference between the j-th year and the loss year when the claim has occurred. Finally, the $Y_{i,j-1}$ is incurred value at the start of the year. The formula (6.3) variables were selected based on ANOVA for the full and truncated gamma model. The variables known at the start of the year were assumed for this model and only those that proved to be significant (Pr(>Chi) < 0.001) were kept in the model. The ANOVA results for the full model are presented in table 12 and for the truncated model are presented in table 13.

Table 12: *Full gamma link factor model variable significance.*

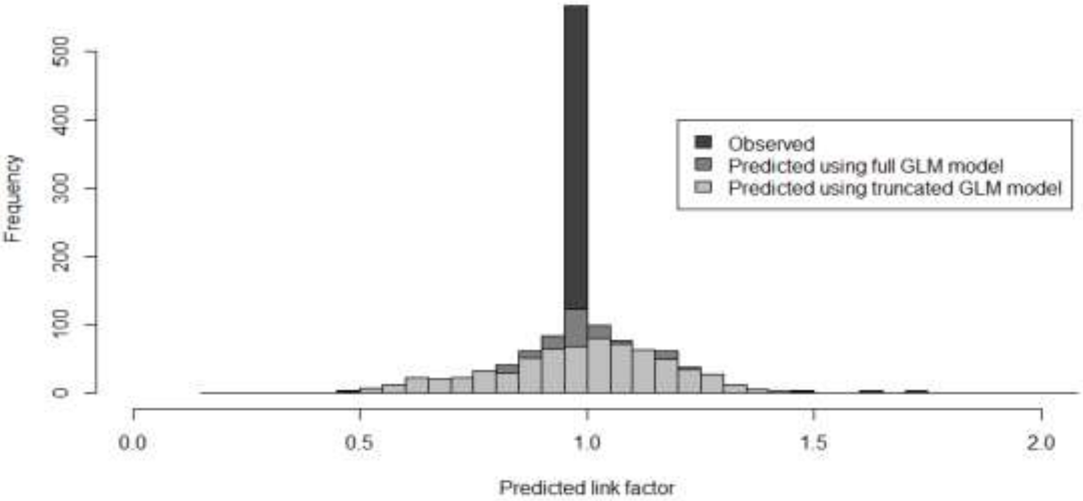
Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			20 123	13005.4	
Loss Year	6	40.9	20 117	12964.4	0.00000
Development Year	6	10538.4	20 111	2425.9	0.00000
Previous Incurred Value $Y_{i,j-1}$	1	2048.7	20 110	377.1	0.00000

Table 13: *Subset gamma link factor model variable significance.*

Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			14 906	5682.4	
Loss Year	6	15.2	14 900	5667.2	0.00000
Development Year	6	3522.5	14 894	2144.7	0.00000
Previous Incurred Value $Y_{i,j-1}$	1	1869.9	14 893	274.8	0.00000

While at the start the models have different residual deviance when the assumed variables are included the models are comparable with the subset model having lower residual deviance with less degrees of freedom. Both fitted models are enclosed in the Appendix section of this work the full model is marked as $\hat{\lambda}_{i,j}$ and the subset is marked as $\hat{\lambda}_{i,j}|\hat{\lambda}_{i,j} \,!= 1$. When testing these models based on the testing dataset the following results can be obtained. The GLM models predicts reasonably close but they are not able to handle the mass of $\lambda_{i,j} = 1$ in the assumed variable.

Figure 19: *Predicting link factor for open claims.*



These models will serve as a base for the following scenarios. Results of these scenarios will be compared in the model evaluation chapter 7.

Gamma Link Factor Model

The gamma link factor model scenario will use the full model $\hat{\lambda}_{i,j}$ to predict the change in the incurred value. The model scenario will use only the severity model.

- **Binary model: None**
- **Severity model: Full Gamma Model $\hat{\lambda}_{i,j}$**

Zero-inflated Gamma Link Factor Model

The Zero-inflated gamma link factor model scenario will use a binary variable $I(\hat{\lambda}_{i,j})$ to determine if change in incurred value has occurred and then the full model $\hat{\lambda}_{i,j}$ to predict the change in the incurred value. The model scenario will use the following binary and severity model.

- **Binary model: Probability of Incurred Value Change Model $I(\hat{\lambda}_{i,j})$**
- **Severity model: Full Gamma Model $\hat{\lambda}_{i,j}$**

Hurdle Gamma Link Factor Model

The Hurdle gamma link factor model scenario will use a binary variable $I(\hat{\lambda}_{i,j})$ to determine if change in incurred value has occurred and then the subset model $\hat{\lambda}_{i,j}|I(\hat{\lambda}_{i,j}) = 1$ to predict the change in the incurred value. The model scenario will use the following binary and severity model.

- **Binary model: Probability of Incurred Value Change Model $I(\hat{\lambda}_{i,j})$**
- **Severity model: Subset Gamma Model $\hat{\lambda}_{i,j}|I(\hat{\lambda}_{i,j}) = 1$**

Probability of Incurred Value Change

This section will introduce the binary variable for the HM and ZIM models as $I(\lambda_{i,j})$. The modelled link factors $\lambda_{i,j}$ contains a mass of occurrences where the link factor is equal to 1. (as presented in the figure 19). This model will try to determine if change in the incurred value have occurred or not. The incurred value change can represent a change in payments or reserves. The model is fitted based on formula (6.4).

$$I(\hat{\lambda}_{i,j}) \sim Loss\ Year_i + Month_i + \log(Order_i) + Development\ year_{i,j} + \log(Y_{i,j-1}), \tag{6.4}$$

where the loss year is the year that the individual claim has occurred, the month is the month of the loss year, the order is the chronological id of the provided claim and the development year is the difference between the *j*-th year and the loss year when the claim has occurred. Finally, the $Y_{i,j-1}$ is incurred value at the start of the year. When $I(\hat{\lambda}_{i,j})$ = 1 it is assumed that change has occured in the incurred value and $I(\hat{\lambda}_{i,j})$ = 0 that change has not occurred. The ratio between these two states is 1:11 in favor of change has not occurred. The training dataset needs to be readjusted. If the model would be fitted for this dataset the model would predict all results as not occurred and predict correctly in more than 90 % of cases. For this purpose, the training dataset would be randomly resampled to provide a smaller sample with the same number of has occurred and has not occurred. For this purpose, the sample balancing method is introduced.

Sample Balancing Method

The sample balancing method will ensure that both categories for binary category will have the same relative frequency and thus the model will try to predict the difference between category rather than giving favor to one category above the other. For this method the first step is to estimate the sample size from the following formula.

$$count = MIN(count\_zero, count\_one) * training\_ratio \qquad (6.5)$$

where count_zero and count_one is the absolute category count for the binary variable categories and the training ratio as selected (here 0.9) for the model. This count changes will be randomly sampled without returning for each of the count_zero and count_one datasets. As a result, the new training dataset would equal frequencies for both categories.

The formula (6.5) variables were selected based on ANOVA on this adjusted dataset. The variables known at the start of the year were assumed for this model and only those that proved to be significant (Pr(>Chi) < 0.001) were kept in the model. The lower boundary was set because the assumed variables proved to be an insignificant except for the development year. The F-Test results are presented in table 14.

Table 14: *Probability of incurred value change model variable significance.*

Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			38529	31916.2	
Loss Year	6	80.2	38523	31835.9	0.00000
Development Year	6	12789.6	38517	19046.3	0.00000
Previous Incurred Value $Y_{i,j-1}$	1	1832.7	38516	17213.5	0.00000
Month	11	275.7	38505	16937.8	0.00000
Order	1	12.4	38504	16925.3	0.00042

The most significant variables proved to be the development year and the previous incurred value $Y_{i,j-1}$. Based on this resampled dataset the model was fitted and presented in the Appendix section of this work. The evaluation of this model is provided in the table 15.

Table 15: *Probability of incurred value change confusion matrix.*

Confusion Matrix		Incurred value change	
		0	1
Predicted	0	332	188
	1	253	921

Accuracy : 0.7397

95% CI : (0.7181, 0.7604)

Sensitivity : 0.5675

Specificity : 0.8305

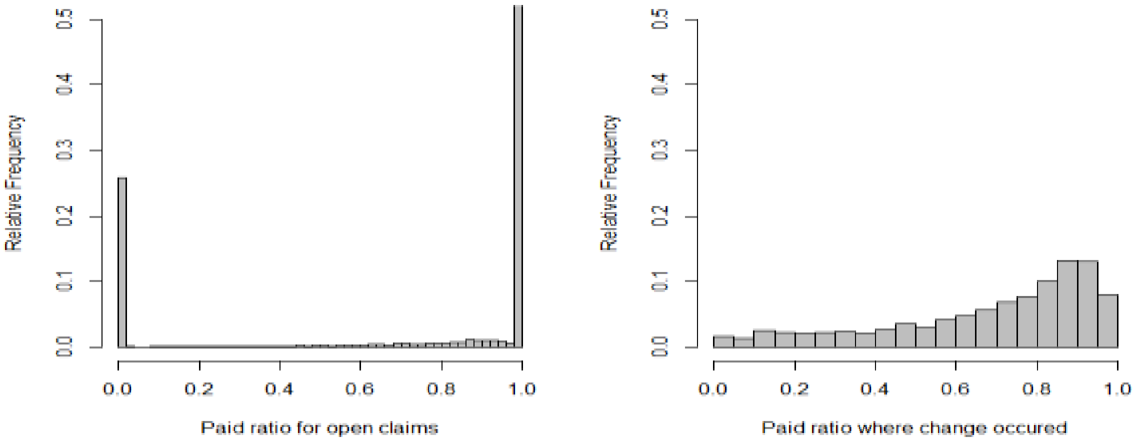
The model predicts the $I(\hat{\lambda}_{i,j})$ correctly on average in 0.7397 number of cases. When the $I(\hat{\lambda}_{i,j}) = 1$ the link factor HM or ZIM model is used to determine the change in $Y_{i,j}$.

When $I(\hat{\lambda}_{i,j}) = 0$ there is no change in $Y_{i,j}$ in the development year. This component will only be used with the HM and ZIM models for the standard GLM model this component is not considered.

Paid Value Models

The paid value models predict the paid ratio $\varsigma_{i,j}$ between the incurred value and the paid value at the end of the development year. This variable is presented in the figure 20.

Figure 20: *Paid ratio variable.*



Based on the left figure the paid value contains two masses. First is the mass of claims where no payment has been done yet and the second is the mass of claims where incurred value is equal to paid value therefore all losses were paid out. The right figure represents only claims where the change has occurred in the development year. This variable will be modelled by a hurdle beta regression model with a binary variable $I(\varsigma_{i,j})$ to determine if change from the previous year has occurred and $\varsigma_{i,j}$ to model the actual variable ratio in case change has occurred. The binary variable $I(\varsigma_{i,j})$ is predicted with the probability of payments being done model in a year and the ratio $\varsigma_{i,j}$ is predicted with the paid beta regression model. These models are presented in the following sections.

Probability of Payment Occurrence Model

The probability of payment occurrence model is based on predicting the binary variable $I(\varsigma_{i,j})$ outcome. The $I(\varsigma_{i,j})$ represents if the payment ratio from the previous year have changed ($I(\varsigma_{i,j}) = 1$) or remained the same ($I(\varsigma_{i,j}) = 0$). For this purpose, a Logistic regression model will be used. The model was fitted based on formula (6.6).

$$I(\hat{\varsigma}_{i,j}) \sim Loss\ year_i + \ Month_i + Development\ year_{i,j} + \log(Y_{i,j}), \qquad (6.6)$$

where the loss year is the year that the individual claim has occurred, the month is the month of the loss year, the development year is the difference between the j-th year and the loss year when the claim has occurred. Finally, the $Y_{i,j}$ is incurred value at the end of the year (This year incurred value). These variables proved to be significant in the assumed model. Before the model was fitted, the sample balancing method was used. (As described in the link factor model). The method needed to be used because the ratio between change has occurred and change has not occurred was 1:20.

The formula (6.6) variables were selected based on ANOVA on this adjusted dataset. The variables known at the end of the year were assumed (after the link factor model is used) for this model and only those that proved to be significant (Pr(>Chi) < 0.001) were kept in the model.

Table 16: *Probability of payment occurrence model variable significance.*

Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			1489	2065.5	
Development Year	6	714.2	1483	1351.3	0.00000
This Year Incurred Value $Y_{i,j}$	1	86.0	1482	1265.3	0.00000
Month	11	33.9	1471	1231.3	0.00037
Order	1	20.3	1470	1211.0	0.00000

The fitted model on the training dataset is enclosed in the Appendix. The main variable affecting this probability is the development year and the incurred value at the end of the year. Confusion matrix obtained from the testing dataset is presented in the following table.

Table 17: *Probability of claim being reported confusion matrix.*

Confusion Matrix		Reported	
		0	1
Predicted	0	1 938	473
	1	59	1 221

Accuracy : 0.8559

95% CI : (0.8441, 0.8670)

Sensitivity : 0.9705

Specificity : 0.7208

The model predicts the $I(\hat{\varsigma}_{i,j})$ correctly on average in 0.8559 number of cases. When the $I(\hat{\varsigma}_{i,j}) = 1$ the paid beta model is used to determine the change in $P_{i,j}$. When $I(\hat{\varsigma}_{i,j}) = 0$ there is no change in $P_{i,j}$ in the development year.

Paid Beta Regression Model

Paid beta model focuses on modelling the change $\varsigma_{i,j}$ in the cumulative paid value $P_{i,j}$. Their relation was described in the (3.13). For modelling the $\varsigma_{i,j}$ following formula will be used.

$$\hat{\varsigma}_{i,j}|(I(\hat{\varsigma}_{i,j}) = 1) \sim Development\ Year_{i,j} + \log(Y_{i,j}) + Month_i, \tag{6.7}$$

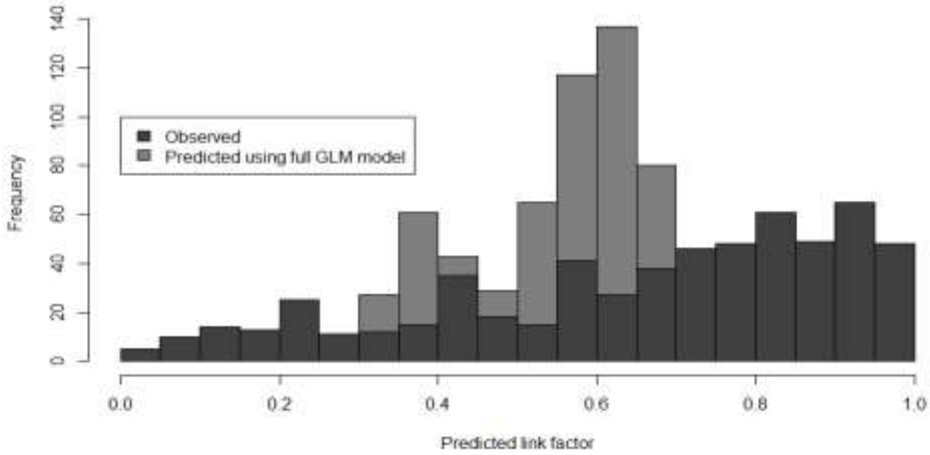
where the month is the month of the loss year, the development year is the difference between the j-th year and the loss year when the claim has occurred and $Y_{i,j}$ is incurred value at the end of the year. The formula (6.7) variables were selected based on ANOVA on this adjusted dataset. The variables known at the end of the year were assumed (after the link factor model is used) for this model and only those that proved to be significant (Pr(>Chi) < 0.001) were kept in the model. The results are presented in table 18.

Table 18: *Paid beta regression model variable significance.*

Variable	Df	Chisq	Pr(>Chi)
Development Year	6	6248.0	0.00000
This Year Incurred Value $Y_{i,j}$	1	509.6	0.00000
Month	11	211.1	0.00000

The final model was fitted with a beta regression approach and the model is presented in the Appendix section of this work. The model was used to create prediction on the testing dataset and the results are presented in the figure 21.

Figure 21: *Predicting paid ratio variable.*



The same problem as in case of the first incurred value have occurred here. The model is not able to represent the variability in the testing dataset. The predicted occurrences are forming two groups. The first one is near the mean of 0.6 and the second one is near 0.4. This model will be used in the final model. When claim is closed it is assumed that all reserves are paid out and therefore the paid value will be set equal to the incurred value. The closure model will be presented in the following section.

Probability of Claim Closure Model

The probability of claim closure is based on predicting the outcome of binary variable $I(S_{i,j})$. The $I(S_{i,j})$ determines if the individual claim development has been closed (claim have been settled) and all losses are paid out or the claim remain open and modelled in the following year. When $I(S_{i,j}) = 1$ it is assumed that claim have been settled and $I(S_{i,j}) = 0$ claim has remained open. For this purpose, a Logistic regression was fitted with formula (6.8).

$$I(\hat{S}_{i,j}) \sim Loss\ year_i + Development\ year_{i,j} + \log(P_{i,j}) + Month_i, \tag{6.8}$$

where the loss year is the year when the claim occurred, the month is the month of the loss year, the development year is the difference between the j-th year, the loss year when the claim has occurred and $P_{i,j}$ is paid value at the end of the year. The formula (6.8) variables were selected based on ANOVA on this adjusted dataset. The variables known at the end of the year were assumed (after the link factor model and beta regression model is used) for this model and only those that proved to be significant (Pr(>Chi) < 0.001) were kept in the model. The results are presented in table 19.

Table 19: *Probability of Claim Closure model variable significance.*

Variable	Df	Deviance	Residual Df	Residual Deviance	Pr(>Chi)
NULL			29477	40865.1	
Loss Year	6	74.4	29471	40790.7	0.000000
Development Year	6	3780.8	29465	37009.8	0.000000
This Year Loss Paid $P_{i,j}$	1	15822.3	29464	21187.4	0.000000
Month	11	82.1	29453	21105.3	0.000000

The $P_{i,j}$ proved to be the most significant variable. Before this model was selected and alternative test with the $Y_{i,j}$ was used. But the model led to claim being closed in the next year and there was nearly no claim development observed. The version with the $P_{i,j}$ offers an alternative where the claims develop before they are closed. The model was fitted on the training dataset and the resulting model is enclosed in the Appendix section of this work. The evaluation of this model on the testing dataset is provided in the following table.

Table 20: *Probability of claim being reported confusion matrix.*

Confusion Matrix		Reported	
		0	1
Predicted	0	1 602	571
	1	574	1 427

Accuracy : 0.7257

95% CI : (0.7119, 0.7392)

Sensitivity : 0.7362

Specificity : 0.7142

The model predicts the $I(\hat{S}_{i,j})$ correctly on average in 0.7257 number of cases. When the $I(\hat{S}_{i,j}) = 1$, the claim is closed, and all reserves are paid out. When $I(\hat{S}_{i,j}) = 0$ the claim will remain, open and will be modelled in the following years.

6.5 Collecting Individual Claim Level Model Results

This section will present these outputs and describe them how they could benefit. These outputs can be described as follows.

- Analytical table
- Individual model outputs
- Triangulation scheme output
- Estimate of the insurance liability and standard error

Analytical table (table 8) provides an overview of the individual claim development. Individual model outputs are the sub model's summary (provided in the Appendix) and the model variable significance tests. The year to year model allow to quite easily create output in the form of a triangulation schema for the training and testing dataset. The table 21 contains the model output for the testing dataset.

Table 21: *Testing dataset triangulation table.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Cumulative incurred value	2004	75 048 485	93 436 674	97 936 130	92 358 384	90 707 263	89 697 918	75 048 485
	2005	95 504 282	120 651 986	119 818 582	119 199 470	115 216 864	111 546 110	95 504 282
	2006	85 008 321	108 952 885	109 079 466	108 598 103	104427 284	104 167 233	85 008 321
	2007	85 009 501	111 630 245	112 187 038	111 326 410	107 509 755	107 258 823	85 009 501
	2008	89 317 790	115 114 926	115 913 322	115 534 218	112 426 730	111 375 194	89 317 790
	2009	73 467 219	85 888 628	86 748 608	86 541 949	84 508 011	83 585 402	73 467 219
	2010	69 746 519	82 709 977	83 158 862	83 049 213	79 504 456	78 886 006	69 746 519

The table contain incurred value development as was observed on the testing dataset. The observed means that this is the claim development without the need of modelling. This information will be used in the model evaluation part of the work to compare the predicted results with the actual results. The results if presented in this form can also be used to compare with a traditional aggregate claim level model. The final individual claim level model output is the prediction of the insurer liability and the standard error of the prediction. The individual model analytical table can be used to create the following output table.

Table 22: *Model output table.*

Claim ID	Loss Year	Incurred Y_i	Predicted Ultimate Incurred $\hat{Y}_i$	Observed Ultimate	Reserve $\hat{R}_i$	Bias
3100	2008	187 047	470 933	432 685	283 886	38 248
3115	2007	5 668	5 668	5 668	0	0.
3120	2008	20 175	20 175	20 175	0	0
3127	2009	46 145	41 856	9 450	- 4288	32 406

The incurred value represents the observed incurred value in 2010, the predicted incurred represent the model output as predicted by the individual claim level model in 2016. The ultimate is the observed incurred value which was not used in modelling and collected from the real claim development. The reserve is obtained as a difference between the predicted incurred and the ultimate value. The bias is obtained as the difference between the predicted incurred and the ultimate incurred. Based on this the total reserve estimate and the MSE of this estimate is obtained as.

$$\hat{R} = \sum_{i=1}^{N} \hat{R}_i, \tag{6.9}$$

where $\hat{R}_i$ is the claim reserve estimate obtained from the table 22 and N is total number of claims (rows) in the table 22. For the MSE of the reserve estimate the traditional form of variable variance and bias is used with the addition of the model variance. The MSE estimate from the individual claim level model is obtained as follows

$$MSE(\hat{R}) = Var(\hat{R}_i) + [E(B_i)]^2, \tag{6.10}$$

where $Var(\hat{R}_i)$ is the variability of the predicted reserve value, the B_i is the bias between the predicted reserve value and the real reserve value.

6.6 Aggregate Claim Level Model

This section will present an aggregate claim level model in form of average cost chain ladder model. The average cost chain ladder model is based on separating the claim frequency and the claim severity modelling as presented in the following formula.

$$Y_{i,j} = N_{i,j} + E(Y_{i,j}), \tag{6.11}$$

where $N_{i,j}$ is the claim frequency represented with the number of reported claims from the i-th origin year by the end of the j-th development year and $E(Y_{i,j})$ is the average incurred value for claims originating in the i-th year by the end of the j-th development year. The claim frequency development will be assumed to be **known** to keep the same assumption as in case of the individual claim level model and the only the severity modelling will be done in this work. This is assumed for the sake of comparing the model outputs in the model evaluation chapter. All examples presented in this section provide an example from one model run (see chapter 7), therefore the examples are just for references. The table 23 contain the known claim frequency.

Table 23: *Number of reported claims in the training dataset.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Cumulative Claim Count	2004	1 967	2 542	2 558	2 563	2 563	2 563	2 563
	2005	2 365	2 919	2 943	2 955	2 956	2 956	2 956
	2006	2 138	2 712	2 740	2 751	2 752	2 754	2 754
	2007	2 177	2 700	2 736	2 742	2 749	2 749	2 749
	2008	2 313	2 842	2 886	2 894	2 897	2 897	2 897
	2009	1 858	2 237	2 274	2 280	2 283	2 284	2 284
	2010	1 761	2 132	2 184	2 200	2 200	2 200	2 200

On the contrary the claim severity will be modelled based on the average incurred value as presented in table 24. For the average incurred value, the development factor and variance estimates will be obtained as presented in the chapter 2. As presented the testing dataset is equal to the 90 % of the original dataset.

Table 24: *Training average incurred value.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Average Incurred Value	2004	44 812	43 875	45 690	43 621	42 743	42 281	41 407
	2005	45 886	47 091	46 519	46 429	45 052	43 906	
	2006	44 151	46 364	46 146	45 946	44 400		
	2007	44 834	48 138	47 573	47 108			
	2008	43 682	46 699	46 623				
	2009	44 709	44 405					
	2010	44 837						

The resulting development factors are presented in table 25. The development factors represent a decreasing trend in the average incurred value on average from the first year the average will decrease by 5 % by the end of the sixth development year.

Table 25: *Developing factors and variance***.**

Metric	1	2	3	4	5	6
Development factor	1.0317	1.0024	0.9848	0.9720	0.9816	0.9793
LDF	0.9518	0.9226	0.9203	0.9345	0.9614	0.9793
Variance	69.3	22.3	18.8	2.1	4.7	2.1
Variance (dev. factor)	8.3	4.7	4.3	1.5	2.2	1.5

These developing factors will be used with the testing dataset to obtain the predicted average incurred value. For this purpose, table 26 was created with the usage of testing dataset and the grey areas represent the predicted average incurred value.

Table 26: *Predicting the average incurred value for the testing dataset.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Average Incurred Value	2004	44 920	38 287	38 365	37 969	37 558	37 483	36 314
	2005	46 808	47 421	48 929	48 327	48 194	47 674	46 689
	2006	45 064	47 868	48 121	47 654	45 126	44 300	43 384
	2007	45 189	47 926	47 189	46 555	45 254	44 425	43 507
	2008	42 708	51 182	51 429	50 647	49 232	48 330	47 331
	2009	43 718	42 587	42 693	42 044	40 869	40 121	39 292
	2010	44 148	45 548	45 661	44 967	43 710	42 910	42 023

In addition, the assumption about the known claim frequency is still valid and table 27 holds number of reported claim in the testing dataset. As presented the testing dataset is equal to the 10 % of the original dataset.

Table 27: *Number of reported claims in the testing dataset*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Cumulative Claim Count	2004	214	291	292	293	293	293	293
	2005	260	320	323	324	324	324	324
	2006	237	297	300	302	302	303	303
	2007	250	303	303	303	304	304	304
	2008	257	320	324	325	325	325	325
	2009	209	252	256	256	256	256	256
	2010	200	236	238	239	239	239	239

When these two tables are combined based on the formula (6.11) the cumulative incurred value can be obtained. The results are presented in the table 28. This table will be used for reserve estimation and the model evaluation.

Table 28: Predicted cumulative incurred value for the testing dataset.

Origin Year		Development Year						
		0	1	2	3	4	5	6
Cumulative Incurred Value	2004	9 612 880	11 141 517	11 202 580	11 124 917	11 004 494	10 982 519	10 640 002
	2005	12 170 080	15 174 720	15 804 067	15 657 948	15 614 856	15 446 376	15 127 236
	2006	10 680 168	14 216 796	14 436 300	14 391 508	13 628 052	13 422 900	13 145 352
	2007	11 297 250	14 521 578	14 298 267	14 106 165	13 757 216	13 505 200	13 226 128
	2008	10 975 956	16 378 240	16 662 996	16 460 275	16 000 400	15 707 250	15 382 575
	2009	9 137 062	10 731 924	10 929 408	10 763 264	10 462 464	10 270 976	10 058 752
	2010	8 829 600	10 749 328	10 867 318	10 747 113	10 446 690	10 255 490	10 043 497

The difference between last known value and the grey estimate for each origin year will be used to create a reserve estimate. In addition to the reserve estimate it is needed to obtain the estimation MSE. The Mack's model offers and estimate of variability of based on the process and parameter variance. In Murphy 2007 the original Mack's formula (2.12) was adjusted and proposed as.

$$MSE\left(\hat{R}\right) = Var(R_i) + \ Var\left(\hat{R}_i\right) + [E(B_i)]^2, \tag{6.12}$$

where $Var(R_i)$ represents the process varaince, $Var\left(\hat{R}_i\right)$ represents the estimation variance and $E(B_i)$ is the bias. The variance parameters $Var(R_i)$ and $Var\left(\hat{R}_i\right)$ are obtained as in the formula (2.12). The bias is obtained as the difference between the predicted mean and the observed mean. The table 29 presents the aggregate claim level model results.

Table 29: *Mack's model variables obtained from the testing dataset for incurred value.*

Origin Year	Predicted Reserve	Predicted Ultimate	Real Ultimate	MSE$(\hat{\boldsymbol{R}})$	RMSE$(\hat{\boldsymbol{R}})$
2005	-319 140	15 127 236	15 354 095	113 609 082 650	337 059
2006	-482 700	13 145 352	13 620 520	196 096 229 364	442 827
2007	-880 037	13 226 128	13 693 414	357 641 448 561	598 031
2008	-1 280 421	15 382 575	15 270 836	877 138 140 295	936 556
2009	-673 172	10 058 752	10 563 771	948 565 824 343	973 943
2010	1 213 897	10 043 497	10 701 815	3 170 453 021 813	1 780 576
Total	-2 421 573	-	-	5 663 503 747 025	2 379 811

From the aggregate claim level model, the results contained in the total row will be collected and used in the model evaluation chapter for comparison with the individual claim level model. The aggregate claim level was proposed to be created based on the same assumptions as the individual claim level model to ensure that these scenarios are comparable in the next chapter.

7. Model Evaluation

The previous chapter presented how the individual and aggregate claim level models were created. This section will provide an overview of the model performance. The model performance is evaluated based on the model MSE and the reserve estimate $R_{i,j}$.

7.1 Model Scenarios

The following scenarios were assumed a will be compared in this chapter:

- Real claim development
- Aggregate claim level model
- Individual claim level model with usage of GLM link factor model
- Individual claim level model with usage of Hurdle link factor model
- Individual claim level model with usage of Zero-inflated link factor model

The real claim development scenario will be used as the ideal state that the other scenarios needs to achieve. The aggregate claim level model will use the chain ladder method as presented in the chapter 6.6. The individual claim level models were introduced in the previous section and the evaluation will also compare how these approaches differ in the reserve estimate $R_{i,j}$. The following subsections present on what bases the conclusive results were obtained.

Training, Testing and Validation Dataset

The dataset provided contains claim data from year 2004 till 2016, the claim development in the years 2004 to 2010 will be used for model training and 2011 to 2016 will be used for model testing, but only for claims originating from years 2004 to 2010. The testing dataset complete history is named the real claim development and will be used for validation.

Handling Large Claims

Before using the obtained dataset described in table 8. It is also important to identify outliers and remove them from the modelling. For the purposes of this work it is assumed that outliers are the identified as follows.

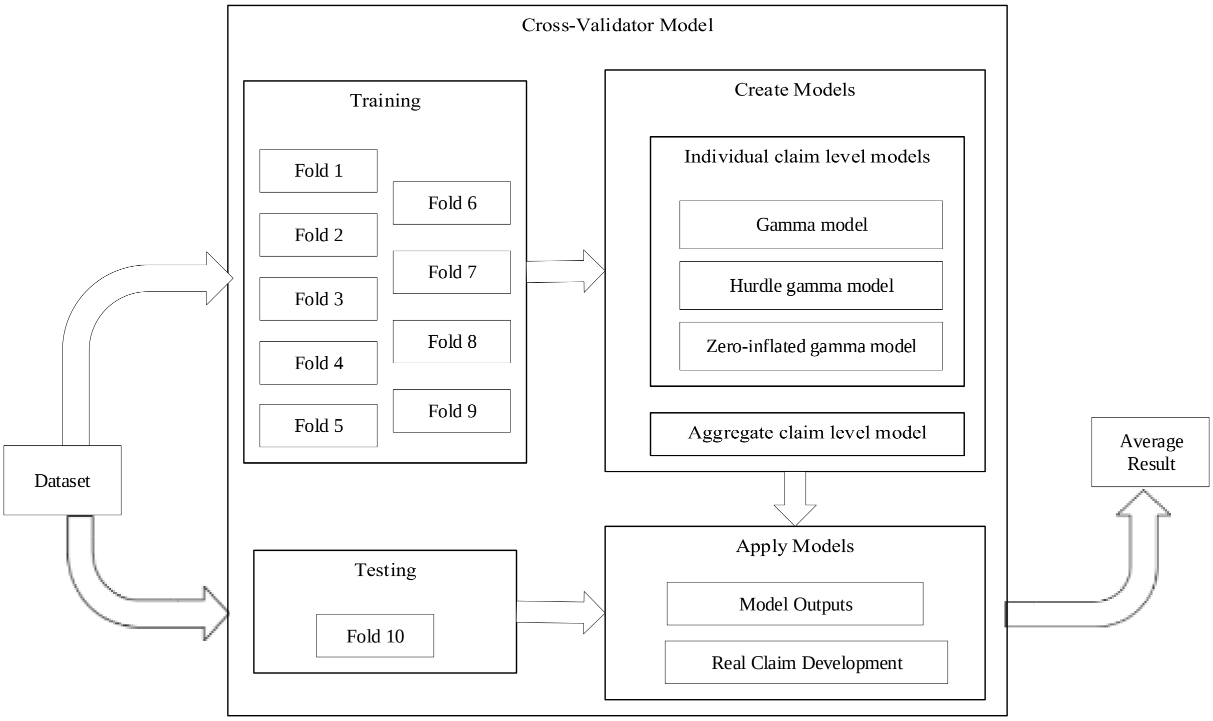
- ➢ Ultimate value above 300 000
- ➢ Link factor above 2

The ultimate threshold is selected as the approximate 99 % quantile of the ultimate incurred value. The link factor threshold is selected as the approximate 95 % quantile of the link factor distribution (figure 18). This assumption was created based on data exploration (chapter 5) and modelling done (chapter 6). Without these assumptions the model scenarios tend to be biased.

Cross-Validator Technique

For model validation the k-fold cross-validation technique is used. From the available dataset the claim identification numbers were collected and randomly split into 10 folds (subsets) of equal sizes. One of these folds is used for testing (10 % of the dataset) while the rest is used for training (90 % of the dataset) when the model runs. The model will be run 10 times and results from each run will be collected and averaged. The following figure 22 represents how the Cross-Validator model is used in context with the models proposed in the chapter 6.

Figure 22: *Cross-Validator model schema*



For each Cross-Validator model run the assumed dataset is randomly split into 10 equal folds. The proposed models (individual and aggregate) are fitted on the 9 folds of the dataset and then used on the 1 testing fold to obtain the model outputs. In addition, the 1 testing fold real claim development is saved. These results are then averaged with previously saved results. In the next run the testing fold is replaced with another fold that has not yet been used for testing. Therefore, each fold is used 9 times for training and once for testing.

7.2 Model Outputs

For each scenario the results are collected after the cross-validator model is used. The table 30 contains the real claim cumulative incurred values that have occurred in the testing dataset.

Table 30: *Real claim development cumulative incurred value.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Cumulative Incurred Value	2004	10 013 117	12 162 379	12 405 710	12 171 836	11 979 058	11 842 755	11 561 307
	2005	12 208 533	15 106 886	15 474 011	15 390 073	15 061 192	14 739 202	14 643 054
	2006	10 653 298	14 116 260	14 257 084	14 301 143	13 812 341	13 688 150	13 687 959
	2007	10 875 158	14 366 687	14 247 702	14 080 010	13 689 056	13 664 087	13 530 505
	2008	11 209 994	15 504 833	15 812 278	15 833 735	15 652 287	15 467 671	14 911 026
	2009	9 439 013	11 227 606	11 538 964	11 485 594	11 316 661	11 224 439	11 224 439
	2010	8 699 292	10 595 126	10 814 164	10 755 689	10 415 215	10 290 598	10 290 598

The table 31 contain the difference (deviations) between the real cumulative incurred value and the predicted cumulative incurred value from the chain ladder model. The Chain-Ladder method overestimates the cumulative incurred value in the years 2005, 2008 and 2010 while it underestimates in the years 2006, 2007 and 2009.

Table 31: *Aggregate claim level model cumulative incurred value deviations.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Deviation	2004	0	0	0	0	0	0	0
	2005	0	0	0	0	0	0	4 848
	2006	0	0	0	0	0	-30 799	-115 774
	2007	0	0	0	0	-28 114	-156 469	-106 990
	2008	0	0	0	-135 909	-422 092	-408 688	54 742
	2009	0	0	-195 988	-225 393	-391 267	-421 416	-489 053
	2010	0	491 101	385 910	363 108	-372 554	375 990	309 695

Tables 32, 33 and 34 contain the deviation for the individual claim level model scenarios. The link factor models tend to heavily overestimate the cumulative incurred value (possible bias). The largest overestimation was observed for claims originating from year 2008.

Table 32: *Individual claim level model cumulative incurred value deviations.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Deviation	2004	0	0	0	0	0	0	0
	2005	0	0	0	0	0	0	326 991
	2006	0	0	0	0	0	172 378	218 517
	2007	0	0	0	0	590 789	603 454	728 803
	2008	0	0	0	680 848	1 081 722	1 224 483	1 750 640
	2009	0	0	408 733	423 310	551 356	605 238	584 541
	2010	0	444 325	159 368	129 457	377 262	444 645	418 360

Table 33: *Hurdle individual claim level model cumulative incurred value deviations.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Deviation	2004	0	0	0	0	0	0	0
	2005	0	0	0	0	0	0	164 248
	2006	0	0	0	0	0	110 475	114 936
	2007	0	0	0	0	243 323	228 265	346 228
	2008	0	0	0	68 905	-61 808	5 691	503 709
	2009	0	0	28 602	79 503	45 131	34 178	-9 493
	2010	0	731 404	494 901	550 963	125 427	36 447	-66 835

Table 34: *Zero-inflated individual claim level model cumulative incurred value deviations.*

Origin Year		Development Year						
		0	1	2	3	4	5	6
Deviation	2004	0	0	0	0	0	0	0
	2005	0	0	0	0	0	0	326 991
	2006	0	0	0	0	0	141 341	169 985
	2007	0	0	0	0	551 753	564 942	690 584
	2008	0	0	0	68 912	327 773	472 483	1 000 651
	2009	0	0	23 800	75 007	206 908	268 926	256 093
	2010	0	490 378	214 710	270 720	458 550	538 173	512 889

The source of the overestimation for the link factor models is the first incurred value model, because the first incurred value plays the significant role in the claim development and in the following years only small adjustments are done to the incurred value (as presented in the chapter 5). The following section presents the reserve estimate and MSE of the estimate.

7.3 Scenario Evaluation

This section will compare the assumed scenarios based on their reserve estimate and MSE estimate. (see chapter 6 how these values were obtained). The conclusive results 10 model runs are presented in the table 35. The RMSE is the root of the MSE.

Table 35: *Final model output comparison.*

Model	Reserve	MSE	RMSE
Real claim development	248 498	-	-
Aggregate claim level model	-408 871	5 544 337 129 982	2 354 642
Gamma individual claim level model	3 819 217	254 191 833	15 943
Hurdle Gamma individual claim level model	1 038 327	78 423 089	8 856
Zero-inflated Gamma individual claim level model	2 953 073	104 435 716	10 219

Firstly, the real claim development average reserve is equal to **248 498**. When comparing the real claim development with the aggregate claim level model the expected reserve from this model is negative therefore, the model expects that the incurred value will decrease in the following years. This would lead to an underestimation and may result in possible loss for the insurance company. The MSE obtained from the Mack's model is large in comparison with the individual claim level model. The individual claim level model tends to heavy overestimate the expected reserve. The smallest reserve is the hurdle gamma individual claim level model where the expected reserve is equal to 1 038 327 while the MSE is very small with 78 423 089. The individual claim level model MSE is very small because all predictions done by the previously proposed sub models are very close to the mean and on the overall these models have very high accuracy. The following chapter describes how the individual claim level model can be improved.

7.4 Model Improvements

As of now the individual claim level model estimations are biased in comparison with the real claim development. When creating the individual claim level models, the following issues were found and if assessed can improve the model.

- Not enough variables usable for prediction.
    - The current dataset needs to be adjusted and additional external information about the claim is needed to improve the prediction power of the severity models. (Link factor, Beta Regression, First incurred value models)

- Predicting first incurred value need to be improved.
    - The first incurred value prediction is critical for the final reserve estimation.
    - The choice of distribution was not explored.
    - Another possibility is to categorize incoming claims into groups based on the possibility → fast claim, large claim, other, then predict the incurred value based on the development of these subcategories

- Handling large claims and outliers.
    - As of now these were left out, but should be also handled by the model.
    - Possible the claims could be split to more groups as there seem to be different patterns of claim development.

- Combining models from multiple runs.
    - As of now the models are being recalculatd each time the Cross-Validator model is run and only the results are combined, but not the created models.
    - This could be achieved with the usage of Bayesian methods or Assembling approach.

Apart from directly improving the actual model, it is also important to consider the other aspects of the reserving process. There are two additional aspects of the claim reserving process which, were not addressed in this work that should be addressed in the following works.

- Claim frequency modelling
    - The proposed individual claim level model assumed that claim frequency was fixed. Which is not the case and the individual level model needs to provide claim frequency estimation for IBNR and RBNS claims.

- Reserve variable predictive distribution
    - As presented in this work the reserve estimate and a MSE is not enough to be usable in claim reserving. The next step is to create predictive distribution of the reserve variable to be able to obtain the Value at Risk (VaR)
    - The VaR is a requirement for the Solvency Capital Requirement (SCR) and the Minimal Capital Requirement (MCR) by the Solvency II.

8. Conclusion

The work proposed an individual claim level model based on an example proposed by Pigeon et al. (2014). In their work they developed the link to link model and proposed the year to year model. This work tried to develop the year to year model which predicts the individual claim value at the end of year development year (very similar to chain ladder). This approach is assuming that time factor of the individual claim level models is fixed (in comparison with link to link model where the time factor is also modelled). This work focused on creating the year to year model and presenting it in three different scenarios. The simple model with GLM model usage for modelling the individual claim value change and the more complex Hurdle and Zero-inflated models. The Hurdle and Zero-inflated models were introduced to handle cases when there is no development in the individual claim.

Before these models could be presented it was important to provide overview of the insurance business and the regulatory framework that exist (**chapter 1)**. The regulatory framework is the reason why these models were developed in the first place to provide a better and more robust estimate of the insurer liability. In the **chapter 2** the work focused on presenting the existing literature. These two chapter concludes the first work objective to research the existing literature on the topic of the reserve risk models based on aggregate claim level and individual claim level. The **first objective was partially completed** because apart from the presented models there exist more source on this topic that could be put into this work therefore, only the most relevant sources were included. Based on the existing literature the individual claim level model was proposed as a combination of variables in **chapter 3**. These variables were suggested based on the provided literature and own research done when creating this work. These variables were modelled with the usage of Generalized linear models (GLM) and their extensions in form of Hurdle and Zero-inflated model (definition in **chapter 4**). These two chapters presented the second objective of this work to define an individual claim level model with the usage of hurdle models and zero inflated models. The **second objective was completed** and a model was presented in these chapters. The **chapter 5** proposed how the dataset should be prepared before the variables can be obtained. This approach is based on the own research done when creating this work. These variables were explored from the claim development point of view and findings were presented. In the **chapter 6** the model structure was presented and for each variable a sub model was proposed and fitted (examples are provided in the Appendix) and presented based on its prediction power. The three scenarios were equal in all sub models except for the link factor model where the change between the insurer liability at the start and of the year and at the end of the year is modelled. For these scenarios the GLM, Hurdle and Zero-inflated models were used. As a result, the model provides an estimate of the insurer liability. These two chapters presented the third objective of this work and proposed how to practically implement the model. The **third objective was completed** and steps by step described how the model would be implemented. For comparison purposes a chain-ladder model was presented at the end of **chapter 6**. The chain-ladder model, and individual claim level models were then compared with the real liability development in **chapter 7**. The comparison proved that the individual claim level model as proposed is not a very good fit of the insurer liability (the model overestimates the insurer liability). The **chapter 7** ended with presenting the fundamental issues that were found when creating the model and what should be improved in the following studies. The final two objectives were presented in this chapter. The **final two objectives were completed** and presented a way how to compare these models and how to improve the individual claim level model.

Sources

Aggregate Claim Reserving

MACK T. (1993) Distribution-free calculation of the standard error of chain-ladder reserve estimates. ASTIN Bulletin, 23:213–225, Available at: http://www.actuaries.org/LIBRARY/ASTIN/vol23no2/213.pdf.

CLAIM RESERVING MANUAL (1997), Faculty and Institute of Actuaries, Volume 1 & 2 ISBN 0901066281, 9780901066282. Available at: https://www.actuaries.org.uk/documents/claims-reserving-manual-vol1-contents

ENGLAND P. D., VERRALL R. J. (2002). STOCHASTIC CLAIMS RESERVING IN GENERAL INSURANCE, British Actuarial Journal., 8(3):443–518, Available at: https://www.actuaries.org.uk/documents/stochastic-claims-reserving-general-insurance.

ENGLAND P. D., VERRALL R. J. (2006). Predictive distributions of outstanding liabilities in general insurance. Annals of Actuarial Science., 1(2):221-270, 2006.
Available at: https://www.cambridge.org/core/journals/annals-of-actuarial-science/article/predictive-distributions-of-outstanding-liabilities-in-general-insurance/B00511D17FEF780E903A6C4B9E729ADF

MEYERS G. G. (2007) Estimating Predictive Distributions for Loss Reserve Models CASUALTY ACTUARIAL SOCIETY, Volume 1:2, 248-272, Available at: http://www.variancejournal.org/issues/?fa=article&abstrID=6417

MURPHY D. M., FCAS & MAAA (2007) Chain Ladder Reserve Risk Estimators, CAS E-Forum Summer, Available at: www.casact.org

Individual Claim Reserving

NORBERG R. (1993a) Prediction of outstanding liabilities in non-life insurance. ASTIN Bull. Vol. 23. no. I, 95–115
NORBERG R. (1993b) Prediction of outstanding liabilities: model variations and extensions
NORBERG R. (1993c) Prediction of outstanding liabilities: parameter
estimation. Proceedings of the XXIV ASTIN Coll., 255–266

NORBERG, R. (1999a). Prediction of Outstanding Liabilities in Non-Life Insurance. ASTIN Bulletin Volume 23, No. 1.
NORBERG, R. (1999b). Prediction of Outstanding Liabilities. II. Model Variations and Extensions. ASTIN Bulletin Volume 29, No. 1.

HAASTRUP, S., ARJAS, E. (1996). Claims Reserving in Continuous Time; A Nonparametric Bayesian Approach. *ASTIN Bulletin, 26*(2), 139-164. doi:10.2143/AST.26.2.563216

TAYLOR, G. C., MCGUIRE, G. (2004). Loss Reserving GLMs: A Case Study.
Centre for Actuarial Studies, Department of Economics, University of Melbourne.

LARSEN C.R. (2007). An individual claim reserving model. ASTIN Bulletin International Actuarial Association, 37(1):113-132. Available at: http://www.casact.org/library/astin/vol37no1/113.pdf.

TAYLOR G., MCGUIRE G., SULLIVAN J., (2008) Individual claim loss reserving conditioned by case estimates. Annals of Actuarial Science, 3:215-256, 2008. Available at: http://www.actuaries.org.uk/data/assets/pdffile/0016/24442/taylorreserving.pdf.

ZIMMERMAN P. (2010), General Insurance Reserve Risk Modeling Based on Unaggregated Data. University of Economics Prague, Available at: https://insis.vse.cz/auth/lide/clovek.pl?zalozka=7;id=38305;studium=85448;zp=14606;download_prace=1

DRIESKENS, D., HENRY, M., WALHIN, J.-F., and WIELANDTS, J. (2012). Stochastic projection for large individual losses. Scandinavian Actuarial Journal, 1:1–39.

ROSELUND, S. (2012). Bootstrapping individual claim histories. ASTIN Bulletin, 42:291–324.

ANTONIO K., PLAT R. (2013), Micro-level stochastic loss reserving for general insurance, Scandinavian Actuarial Journal 2014, N 7, 649-669 DOI 10.1080/03461238.2012.755938, Available at https://doi.org/10.1080/03461238.2012.755938

PIGEON. M., ANTONIO. K., DENUIT. M. (2013) Individual Loss Reserving with the Multivariate Skew Normal Framework (May 21. 2013). ASTIN Bulletin. 43(3). 399-428 (2013). Available at: https://ssrn.com/abstract=1996455 or http://dx.doi.org/10.2139/ssrn.1996455

PIGEON. M., ANTONIO. K., DENUIT. M. (2014). Individual loss reserving using paid–incurred data. In Insurance: Mathematics and Economics. Volume 58. 2014. Pages 121-131. ISSN 0167-6687. Available at: http://www.sciencedirect.com/science/article/pii/S0167668714000845

Generalized Linear Models

NELDER J. A., WEDDERBURN R. W. M. (1972) Generalized Linear Models, Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3 (1972), pp. 370-384 Published by: Blackwell Publishing for the Royal Statistical Society Available at: http://www.jstor.org/stable/2344614

MCCULLAGH P, NELDER J. A. (1989). Generalized Linear Models. 2nd edition. Chapman & Hall, London.

CAMERON A. C., TRIVEDI P. K. (1990), Regression-based tests for overdispersion in the Poisson model, In Journal of Econometrics, Volume 46, Issue 3, 1990, Pages 347-364, ISSN 0304-4076, https://doi.org/10.1016/0304-4076(90)90014-K. Available at: http://www.sciencedirect.com/science/article/pii/030440769090014K

Hurdle Model and Zero Inflated Models

CAMERON TRIVEDI (2013) Regression Analysis of Count Data, Econometric Society Monographs. Cambridge University Press. doi:10.1017/CBO0971139013567

MILLS, E. D. (2013). Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinous data. University of Iowa.

Available at: http://ir.uiowa.edu/cgi/viewcontent.cgi?article=4712&context=etd

List of Variables

$C_{i,j}$	incremental claim value
$\hat{C}_{i,j}$	incremental claim value estimation
$D_{i,j}$	cumulative claim value
$\widehat{D}_{i,j}$	cumulative claim value estimation
$R_{i,j}$	reserve value
$\hat{R}_{i,j}$	reserve value estimation
$Y_{i,j}$	incurred value
$\hat{Y}_{i,j}$	incurred value estimation
$P_{i,j}$	paid value
$\widehat{P}_{i,j}$	paid value estimation
$\lambda_{i,j}$	development factor (year to year)
$\hat{\lambda}_{i,j}$	development factor estimation (year to year)
$\varsigma_{i,j}$	paid to incurred ratio
$\hat{\varsigma}_{i,j}$	paid to incurred ratio estimation
$\sigma^2{}_j$	variance
$\hat{\sigma}^2{}_j$	variance estimation
$I(S_{i,j})$	claim closure indicator binary variable
$I(\hat{S}_{i,j})$	claim closure indicator binary variable estimation
$I(O_{i,j})$	claim reporting indicator binary variable
$I(\widehat{O}_{i,j})$	claim reporting indicator binary variable estimation
$I(\lambda_{i,j})$	incurred value change indicator binary variable
$I(\hat{\lambda}_{i,j})$	incurred value change indicator binary variable estimation
$I(\varsigma_{i,j})$	paid to incurred ratio change indicator binary variable
$I(\hat{\varsigma}_{i,j})$	paid to incurred ratio change indicator binary variable estimation
$P(S_{i,j})$	probability of claim being closed in a year
$P(O_{i,j})$	probability of claim being reported in a year
$P(\lambda_{i,j})$	probability of incurred value change in a year
$P(\varsigma_{i,j})$	probability of paid to incurred ratio change in a year

List of Tables. Figures. Equations

Figures

Tables

Equations

Appendix

Probability of Claim Being Reported Model

Formula: **Model:** Logistic Regression

$$I\left(\hat{O}_{i,j}\right) \sim Loss\ Year_i + Month_i + \log(Order_i) + Development\ year_{i,j},$$

Deviance Residual

Min	1Q	Median	3Q	Max
-2.150	-0.793	-0.105	0.686	3.354

Regressor	Coefficients	Std. Error	t-value	Pr(>|t|])	
Intercept	3.31562	0.15976	20.754	0.00000	***
Loss Year 2005	0.04010	0.04813	0.833	0.40473	
Loss Year 2006	0.06791	0.04930	1.378	0.16831	
Loss Year 2007	0.13841	0.04987	2.775	0.00552	**
Loss Year 2008	0.15217	0.04943	3.079	0.00208	**
Loss Year 2009	0.11642	0.05248	2.218	0.02653	*
Loss Year 2010	0.13530	0.05338	2.535	0.01125	*
Development Year 1	-2.12351	0.02778	-76.444	0.00000	***
Development Year 2	-3.90196	0.07157	-54.517	0.00000	***
Development Year 3	-4.81175	0.13288	-36.212	0.00000	***
Development Year 4	-6.12380	0.29069	-21.067	0.00000	***
Development Year 5	-6.65133	0.57868	-11.494	0.00000	***
Development Year 6	-6.13665	0.71060	-8.636	0.00000	***
February	0.12922	0.08823	1.465	0.14302	
March	0.18525	0.09050	2.047	0.04066	*
April	0.18525	0.09294	1.970	0.04879	*
May	0.16731	0.09356	1.788	0.07374	.
June	0.15168	0.09480	1.600	0.10959	
July	0.09069	0.09646	0.940	0.34711	
August	0.06093	0.09820	0.620	0.53499	
September	-0.01534	0.09877	-0.155	0.87659	
October	-0.06246	0.10071	-0.620	0.53509	
November	-0.05161	0.10227	-0.505	0.61379	
December	-0.04327	0.10419	-0.415	0.67792	
Order	-0.30015	0.02808	-10.688	0.00000	***

Null deviance:	53337 on 38529 degrees of freedom
Residual deviance:	35512 on 38505 degrees of freedom
AIC:	35562

First Incurred Value Model

Formula: **Model:** Log-normal

$$\log(\hat{Y}_{i,j}) \sim Development\ year_{i,j} + \log(Order_i),$$

Deviance Residual

Min	1Q	Median	3Q	Max
-89 822	-29 211	-16 704	5 051	4 679 915

Regressor	Coefficient	Std. Error	t-value	Pr(>|t|])	
Intercept	10.87548	0.08037	135.313	0.00000	***
Development Year 1	-0.00003	0.03185	-0.001	0.99919	
Development Year 2	0.37126	0.07147	5.194	0.00000	***
Development Year 3	0.46839	0.12150	3.854	0.00011	***
Development Year 4	0.33265	0.29853	1.114	0.26516	
Development Year 5	-2.57843	11.81942	-0.218	0.82731	
Development Year 6	1.01198	0.39973	2.531	0.01136	*
Order	-0.02300	0.01181	-1.946	0.05163	.

Null deviance:	87 172 907 892 599 on 18402 degrees of freedom
Residual deviance:	87 004 634 812 740 on 18395 degrees of freedom
AIC:	462 201

Probability of Claim Incurred Value Change

Formula: **Model:** Logistic Regression

$$I\left(\hat{\lambda}_{i,j}\right) \sim Loss\ Year_i + Month_i + \log(Order_i) + Development\ year_{i,j} + \log(Y_{i,j-1}),$$

Deviance Residual

Min	1Q	Median	3Q	Max
-2.4542	0.0000	0.0000	0.4258	2.3007

Regressor	Coefficient	Std. Error	t-value	Pr(>|t|])	
Intercept	19.4533	131.0762	0.148	0.00001	***
Loss Year 2005	0.1521	0.0663	2.293	0.02182	*
Loss Year 2006	0.4543	0.0738	6.155	0.00000	***
Loss Year 2007	0.2555	0.0733	3.483	0.00049	***
Loss Year 2008	0.1414	0.0703	2.011	0.04433	*
Loss Year 2009	0.1630	0.0757	2.153	0.03133	*
Loss Year 2010	0.1145	0.0770	1.487	0.13693	
Development Year 1	-15.1480	131.0762	-0.116	0.90799	
Development Year 2	-16.9912	131.0762	-0.130	0.89686	
Development Year 3	-17.6590	131.0762	-0.135	0.89283	
Development Year 4	-16.3879	131.0762	-0.125	0.90050	
Development Year 5	-16.2553	131.0762	-0.124	0.90130	
Development Year 6	-15.0911	131.0763	-0.115	0.90834	
Previous Year Loss Incurred $Y_{i,j-1}$	-0.3692	0.0156	-23.653	0.00000	***
February	-0.0089	0.1230	-0.073	0.94217	
March	0.0031	0.1257	0.025	0.98008	
April	0.0980	0.1296	0.756	0.44946	
May	0.0637	0.1286	0.496	0.62017	
June	0.1949	0.1318	1.479	0.13922	
July	0.2818	0.1345	2.095	0.03620	*
August	0.4939	0.1374	3.594	0.00032	***
September	0.6325	0.1391	4.547	0.00032	***
October	0.5729	0.1403	4.083	0.00004	***
November	0.6969	0.1441	4.837	0.00000	***
December	0.6738	0.1497	4.500	0.00000	***
Order	0.0928	0.0388	2.393	0.01669	*

Null deviance:	31 916 on 38 529 degrees of freedom
Residual deviance:	16 925 on 38 504 degrees of freedom
AIC:	16 977

Full Gamma Link Factor Model

Formula: **Model:** Gamma (Log)

$$\hat{\lambda}_{i,j} \sim Loss\ Year_i + Development\ Year_{i,j} + \log(Y_{i,j-1})$$

Deviance Residual

Min	1Q	Median	3Q	Max
-1.6175	-0.0048	-0.0026	0.0027	1.4038

Regressor	Coefficients	Std. Error	t-value	Pr(>|t|])	
Intercept	-1.6121	0.0026	-606.4890	0.00000	***
Loss Year 2005	0.0057	0.0034	1.6651	0.09590	.
Loss Year 2006	0.0011	0.0036	0.3051	0.76027	
Loss Year 2007	-0.0035	0.0036	-0.9567	0.33872	
Loss Year 2008	0.0053	0.0035	1.4930	0.13543	
Loss Year 2009	0.0040	0.0038	1.0421	0.29736	
Loss Year 2010	0.0075	0.0039	1.9314	0.05344	.
Development Year 1	0.0613	0.0037	16.2518	0.00000	***
Development Year 2	0.0167	0.0051	3.2225	0.00127	***
Development Year 3	0.0015	0.0056	0.2787	0.78040	
Development Year 4	-0.0056	0.0070	-0.8100	0.41791	
Development Year 5	-0.0045	0.0096	-0.4722	0.63678	
Development Year 6	-0.0064	0.0197	-0.3251	0.74510	
Previous Year Loss Incurred $Y_{i,j-1}$	0.1463	0.0003	370.4295	0.00000	***

Null deviance:	13 005 on 19 196 degrees of freedom
Residual deviance:	377 on 19 172 degrees of freedom
AIC:	-66 579

(Dispersion parameter for Gamma family taken to be 0.01964572369)

Subset Gamma Link Factor Model

Formula: **Model:** Gamma (Log)

$$\hat{\lambda}_{i,j}|(I(\hat{\lambda}_{i,j}) = 1) \sim Loss\ Year_i + Development\ Year_{i,j} + Month_i + \log(Y_{i,j-1})$$

Deviance Residual

Min	1Q	Median	3Q	Max
-1.6588	-0.0040	-0.0000	0.0009	1.3695

Regressor	Coefficients	Std. Error	t-value	Pr(>|t|])	
Intercept	-1.6158	0.0028	-559.8186	0.00000	***
Loss Year 2005	0.0072	0.0038	1.8663	0.06201	.
Loss Year 2006	0.0054	0.0040	1.3611	0.17348	
Loss Year 2007	0.0056	0.0040	1.3769	0.16855	
Loss Year 2008	0.0104	0.0040	2.5985	0.00937	**
Loss Year 2009	0.0064	0.0043	1.4846	0.13766	
Loss Year 2010	0.0113	0.0044	2.5709	0.01015	*
Development Year 1	0.0285	0.0037	7.5836	0.00000	***
Development Year 2	-0.0316	0.0074	-4.2500	0.00002	***
Development Year 3	-0.1100	0.0137	-7.9859	0.00000	***
Development Year 4	-0.3675	0.0248	-14.8146	0.00000	***
Development Year 5	-0.3218	0.0514	-6.2563	0.00000	***
Development Year 6	-0.2504	0.0481	-5.1977	0.00000	***
Previous Year Loss Incurred $Y_{i,j-1}$	0.1537	0.0004	331.3332	0.00000	***

Null deviance:	5 682 on 14 906 degrees of freedom
Residual deviance:	274 on 14 893 degrees of freedom
AIC:	-60 900

(Dispersion parameter for quasipoisson family taken to be 0.01841842095)

Probability of Payment Occurrence

Formula: **Model:** Logistic Regression

$$I(\hat{\varsigma}_{i,j}) \sim Development\ Year_{i,j} + \log\left(Y_{i,j}\right) + Month_i + \log(Order_i)$$

Deviance Residual

Min	1Q	Median	3Q	Max
-2.7261	-0.0001	0.0390	0.7573	1.7677

Regressors	Coefficients	Std. Error	t-value	Pr(>|t|])	
Intercept	-17.4850	513.3713	-0.034	0.97283	
Development Year 1	20.2389	513.3704	0.039	0.96855	
Development Year 2	21.2090	513.3704	0.041	0.96704	
Development Year 3	22.0783	513.3704	0.043	0.96569	
Development Year 4	20.0769	513.3704	0.040	0.96776	
Development Year 5	21.0753	513.3705	0.041	0.96725	
Development Year 6	22.8006	513.3710	0.044	0.96457	
This Year Loss Incurred $Y_{i,j}$	0.4430	0.0826	5.358	0.00000	***
February	0.3625	0.5662	0.640	0.52203	.
March	1.6397	0.6476	2.532	0.01134	*
April	1.6465	0.5850	2.814	0.00489	**
May	1.7470	0.5850	2.986	0.00282	**
June	2.0309	0.6094	3.332	0.00086	***
July	2.1936	0.6138	3.574	0.00035	***
August	2.2392	0.6187	3.654	0.00025	***
September	2.3664	0.6187	3.825	0.00013	***
October	2.2302	0.6362	3.506	0.00045	***
November	1.8247	0.6427	2.839	0.00452	**
December	1.7994	0.6634	2.712	0.00667	**
Order	-1.2752	0.1926	-6.620	0.00000	***

Null deviance:	2 065 on 1 489 degrees of freedom
Residual deviance:	1 211 on 1 470 degrees of freedom
AIC:	1 251

Paid Beta Regression Model

Formula: **Model:** Beta

$$\hat{\varsigma}_{i,j}|(I(\hat{\varsigma}_{i,j}) = 1) \sim Development\ Year_{i,j} + \log(Y_{i,j}) + Month_i$$

Deviance Residual

Min	1Q	Median	3Q	Max
-2.8914	-0.7180	-0.6162	0.6641	3.8846

Regressor	Coefficients	Std. Error	t-value	Pr(>|t|])	
Intercept	-3.3729	0.1271	-26.520	0.00000	***
Development Year 1	0.9429	0.0258	36.474	0.00000	***
Development Year 2	1.6966	0.0318	53.292	0.00000	***
Development Year 3	1.9844	0.0357	55.440	0.00000	***
Development Year 4	2.1144	0.0498	42.450	0.00000	***
Development Year 5	2.1302	0.0747	28.505	0.00000	***
Development Year 6	1.9934	0.1598	12.473	0.00000	***
This Year Loss Incurred $Y_{i,j}$	0.2102	0.0113	18.555	0.00000	***
February	-0.2212	0.0605	-3.656	0.00025	***
March	-0.2026	0.0590	-3.430	0.00060	***
April	-0.2745	0.0578	-4.742	0.00000	***
May	-0.2617	0.0549	-4.764	0.00000	***
June	-0.3144	0.0540	-5.823	0.00000	***
July	-0.3926	0.0533	-7.359	0.00000	***
August	-0.4182	0.0527	-7.929	0.00000	***
September	-0.4378	0.0514	-8.511	0.00000	***
October	-0.4618	0.0511	-9.029	0.00000	***
November	-0.5097	0.0518	-9.836	0.00000	***
December	-0.5815	0.0566	-10.259	0.00000	***

Probability of Claim Closure

Formula: **Model:** Logistic Regression

$$I(\hat{S}_{i,j}) \sim Loss\ year_i +\ Month_i + Development\ year_{i,j} + \log(P_{i,j})$$

Deviance Residual

Min	1Q	Median	3Q	Max
-3.2474	-0.9347	-0.0807	0.9289	2.2149

Regressor	Coefficients	Std. Error	t-value	Pr(>|t|])	
Intercept	1.23252	0.18525	6.653	0.00000	***
Loss Year 2005	0.04075	0.04806	0.848	0.39641	
Loss Year 2006	0.31092	0.05012	6.204	0.00000	***
Loss Year 2007	0.40896	0.05030	8.129	0.00000	***
Loss Year 2008	0.35108	0.04931	7.119	0.00000	***
Loss Year 2009	0.35223	0.05231	6.733	0.00000	***
Loss Year 2010	0.39361	0.05302	7.423	0.00000	
Development Year 1	1.78050	0.03415	52.137	0.00000	***
Development Year 2	0.20974	0.04791	4.377	0.00000	***
Development Year 3	-0.51746	0.06387	-8.101	0.00000	***
Development Year 4	0.77407	0.06683	11.582	0.00000	***
Development Year 5	1.12523	0.09806	11.475	0.00000	***
Development Year 6	2.27334	0.19388	11.725	0.00000	***
This Year Loss Paid $P_{i,j}$	0.16503	0.00629	26.225	0.00000	***
February	0.21008	0.08070	2.603	0.00923	**
March	0.28963	0.08364	3.462	0.00053	***
April	0.34547	0.08728	3.958	0.00000	***
May	0.19399	0.08835	2.196	0.02811	*
June	0.24846	0.09063	2.741	0.00611	**
July	0.11198	0.09288	1.206	0.22793	*
August	0.11632	0.09507	1.223	0.22114	
September	-0.01205	0.09617	-0.125	0.90026	
October	-0.17788	0.09823	-1.811	0.07017	
November	-0.10865	0.10011	-1.085	0.27777	
December	0.14751	0.10260	1.438	0.15046	

Null deviance:	40865 on 29477 degrees of freedom
Residual deviance:	21105 on 29453 degrees of freedom
AIC:	21155