Assessing the Dependency Between Claim Frequency and Claim Severity for a Dutch Home Content Insurance Portfolio.

A comparative case study of independent and dependent frequency-severity models

Leon Romeo Kingma
Univé Stad en Land
University of Twente

A thesis submitted for the degree of
Master of science

univé

UNIVERSITY OF TWENTE.

Enschede 2023

Master thesis Industrial Engineering and Management

Assessing the Dependency Between Claim Frequency and Claim Severity for a Dutch Home Content Insurance Portfolio.

Author:
L.R. Kingma (Leon)
L.r.kingma@student.utwente.nl

Univé Stad en Land
Christiaan Geurtsweg 8
7335 JV Apeldoorn
088 552 0300

University of Twente
Drienerlolaan 5
7522 NB Enschede
(053) 489 91 11

Supervisors Univé Stad en Land
H. Jansen (Harm)
harmen.jansen@unive.nl

S. Vrind (Sven)
sven.vrind@unive.nl

Supervisors University of Twente
L. Spierdijk (Laura)
l.spierdijk@utwente.nl

B. Roorda (Berend)
b.roorda@utwente.nl

Abstract

Faculty of Behavioural, Management and Social Sciences
Industrial Engineering & Management

This thesis explores the relationship between claim frequency and claim severity in home-content insurance. It investigates whether a model that doesn't assume independence can be preferred. The aim of this thesis is to compare the prediction ability of a frequency-severity model in both the dependent and independent setting. The model where the assumption of dependence between the claim frequency and severity is relaxed can be modelled by including the claim frequency as a covariate in the frequency-severity model. This study will create a framework for insurers in the same or a similar sector to research the relation between the claim frequency and severity. The framework is supported by a literature review that indicated what literature is present and how this study can complement the existing literature in this field. After the literature review, a preliminary analysis regarding the dependency is performed to create first insights into the correlation. In this analysis a negative correlation between the claim frequency and claim severity is discovered. After the literature review and the preliminary analysis, the marginal Generalized Linear Models (GLM) are modelled. With the marginal GLMs for the claim frequency, the claim severity in the independent setting (benchmark) and the claim severity in the dependent setting (adjusted), the total aggregate loss models can be made. These models predict the total aggregate loss on individual policyholder level. The model in the independent setting is a multiplication of the claim frequency GLM and the claim severity GLM. The total aggregate loss model in the dependent setting is similar, but the multiplication is with the adjusted claim severity GLM. Also, a dependency correction term is included. The results of the comparison between these two models regarding the prediction ability indicate a slight preference for the model in the dependent setting. This preference can be explained by the error measures and the distribution of the predicted values compared to the actual values. The model where dependence is allowed has more overlap with the actual values compared to the model where the independence is assumed. Both in-sample and out-of-sample modelling is performed. The validity of the results is therefore increased as the error measures of the in-sample modelling do not indicate a preference over out-of-sample modelling, indicating no overfitting is present. The adjustments in the total aggregate loss model could result in a small economic benefit for Univé Stad en Land as they have the possibility to better predict the total aggregate loss for all policyholders, and thus could be able to determine a more competitive premium.

Contents

List of Figures

List of Tables

ABBREVIATIONS

USL	Univé Stad en Land
GLM	Generalized Linear Model
CF	Claim Frequency
CS	Claim Severity
IBNR	Incurred But Not Reported
ML	Machine Learning
GAM	Generalized Additive Model
MLE	Maximum Likelihood Estimation

Chapter 1 Project Plan

The introduction chapter will provide the reader with the necessary information about the company, and how the research of this paper is structured. The Problem context and the core problem will be explained, the chapter ends with the research questions.

1.1. Univé Stad en Land

Univé Stad en Land (later on USL) is a Dutch insurance company that provides various insurance products to individuals, businesses, and organizations. The company has been in operation for over 200 years and has a strong presence in the Netherlands, with offices and agents from USL located in 'their' region of the Netherlands.

The Risk and Compliance and Business Control departments are crucial parts of USL's operations. The departments work closely with other departments within the company to develop risk management strategies and control strategies that are tailored to the specific needs of the company.

The main responsibility of Risk and Compliance is to monitor the financial performance and to ensure compliance with relevant regulations and standards. This includes managing the company's financial reporting, budgeting, and forecasting processes, as well as developing and implementing financial controls to mitigate the risks associated with financial transactions. The department of Business Control works closely with the risk and compliance department, as its main responsibility is to maintain and improve the strategic goals of USL. Tasks like financial planning, performance measurement and reporting are all included in this task description.

1.2. Problem context

The environment of the insurance companies is evolving quickly, from a 'small' industry trying to provide services to the peoples/businesses in need, to a vital sector supporting almost everyone in all sorts of services. To remain efficient, the insurance companies need to be flexible and eager to make changes over time. Important changes within an insurance company can be the type of products the company sells, the level of protection or the amount of premium that needs to be collected for each product. The latter is a vital aspect of an insurance company as pricing their product is one of the main drivers for maintaining a strong competitive position within the insurance market.

It is important to set a competitive premium for each individual policy, without being too cheap in the market. This relates to the basic principle in the economic market, where customers tend to shift direction when cheaper options are available that provide comparable services. In the insurance sector this principle is also present. For instance, Vanasse, Dionne & Gouriéroux (2001) show that in the auto insurance environment, if an insurance charges too little for young drivers and too much for old drivers, young drivers will be attracted while the old drivers will switch to competitors. This leaves the company with an unbalanced portfolio, which does not enhance the future perspectives and will ultimately result in economic losses.

This process of determining a premium for a policyholder is called ratemaking. A fair premium is a premium that objectively reflects the risks that the specific policyholder carries. Ratemaking is a critical process for insurance companies as it directly impacts their profitability and ability to remain financially stable and viable in the long term. The ratemaking process is based on frequency-severity modelling, which is often referred to as aggregate claim modelling. Frequency-severity modelling considers both the frequency and severity of the claims and is used to estimate the total (aggregate) losses that the insurer will need to pay out during a period. The Claim Frequency (CF) and the Claim Severity (CS) are modelled separately at USL and then combined to estimate the total losses the insurer is likely to experience.

Univé has a range of different coverages in the home insurance sector. In the case of home insurance, frequency-severity modelling involves developing models that can accurately assess and quantify the risks associated with insuring a property. These models are made with the variables that are present in the 'risk profile' of the policyholder. This risk profile includes personal factors such as their age and relationship status, but mainly property risk factors such as its location, the type of house or the type of roof have an impact on the rate-making process. By using data analysis and statistical modelling techniques, insurance companies can estimate the likelihood of a claim being made and the potential cost of that claim. This information can be used to set appropriate premiums.

The most common approach to model the CF and CS in home insurance is to use a multivariate analysis (Su & Bai, 2020), which allows insurers to consider multiple factors simultaneously and identify their individual contributions to risk. For example, an insurer might use a statistical model that considers the age of the home, the type of roof, and location of the property to estimate the likelihood of a claim occurring. This information, together with the prediction of the claim size, can then be used to set a premium that reflects the level of risk involved in insuring that property.

At Univé Stad en Land an important change is happening, the underlying model of the rate-making process is being re-evaluated and a new model is going to be implemented. This change is not something that happens yearly, so while in this transition period, it is of great importance that the new model is validated and well-implemented. Currently, one of the most traditional approaches for frequency-severity modelling is a two-step approach. Claim frequency & claim severity are modelled separately, instead of creating a joint model (Oeben, 2015). This happens with the use of generalized linear models (GLMs), these models are a generalisation of the ordinary linear regression models. Because GLMs are easy to interpret and flexile to use for the data of insurance companies, some insurance companies prefer this option over others, including USL.

At USL the validation of the frequency-severity models could benefit from further improvement, as one of the arguments used to defend the question about the model choice (GLM) is stated as "Almost all insurance companies use GLMs, so it should be effective". The frequency-severity modelling component at USL is fully align with the most traditional approach stated earlier: a two-step approach where the claim frequency and the claim severity are modelled separately using GLM techniques. This way of modelling makes assumptions that could be relaxed. An example of such an assumption

is the independence assumption between claim frequency and claim severity. To be able to state that USL implemented a new model which is validated, these assumptions should be researched and possibly relaxed.

1.3. Core problem

Refining the statistical risk model that functions as the basis of the ratemaking process can lead to financial benefits and a greater market share when implemented efficiently. A new model cannot be effectively implemented without validating the assumptions made. A lot of literature exists about the relation between the claim frequency and the claim severity not being independent, for example the studies of Frees (2016), Henckaerts (2019) and Yang (2022). However, the current model, which is making its way into the insurance market at USL, still makes the independence assumption between the CF and the CS. This assumption is made without studying the possible outcomes when the dependence is included in the frequency-severity model.

To be able to know whether the model that will be implemented at USL is as effective as it could be, the possible dependence deserves to be further researched. Also, failure to validate the model's assumptions could lead to biases and errors when interpreting the results. This will ultimately undermine the usefulness of the model.

The current model uses the two-step method and analyses the CF and CS separately using a Poisson distribution and a Gamma distribution respectively. Even though the independence is controversial to say the least, it is convenient in various statistical computations such as the maximum likelihood estimations (W. Lee et al., 2019). The dependency between the claim frequency and claim severity could be positive, negative or zero. A positive dependency implies that the condition that leads to a higher frequency would also lead to a higher severity, whereas a negative dependency implies that there are many relatively small claims, or a few big claims (Becker et al., 2022).

USL requires a frequency-severity model that predicts the total aggregate loss as accurate as possible. The total aggregate loss refers to the total amount of losses that have been incurred during a specified period because of the claims. The total aggregate loss is derived from the frequency-severity model. The independence assumption can limit the accuracy of the expected total aggregate loss. To address this limitation, USL wants to develop models that relax the independence assumption between CF and CS. This could allow them to better capture and understand the underlying dynamics of the claims process, which could lead to more accurate predictions of the aggregate claim loss.

Being able to obtain more accurate predictions will support USL with making more informed decisions about pricing their policies. If they can accurately predict the total aggregate loss, they can set premiums that are more in line with the risk they are taking on. This could help them attract more customers while also staying profitable. Ultimately, when the risks of the policyholders are predicted more accurately, the premium proposed can be determined more competitively, which can result in an economical benefit.

So, USL is currently in the transition period to switch to a newer frequency-severity model, this model is based on the two-step GLM method. This frequency-severity model is the underlying for determining the premium set for all policyholders. Without knowing whether there exists a dependency between the CF and the CS, USL has assumed them to be independent for the convenience of the model. For further readability of the report, we recall the current frequency-severity model where the independency is assumed the *benchmark model*. The proposed frequency-severity model, where the dependency is included, will be recalled as the *adjusted model*.

Thus, the importance of this study can be summarised that the adjusted model could predict the total aggregate loss better than the benchmark model. Then it could very well be the case that the premiums of the policyholders can differ from what they pay now following the frequency-severity model with the independency assumption.

1.4. The research problem

The adjusted model will remove uncertainty regarding the current model's performance without the assumption of independence. This research will study the relationship between the CF and the CS and propose a framework for modelling the dependence, and thus broaden the literature about the use of GLM models as a tool for frequency-severity modelling in the home insurance sector. The research question of this research can be formulated as

"How can a dependent frequency-severity model be used to predict the total aggregate loss and how does this adjusted model compare to the benchmark frequency-severity model?"

To be able to answer this question, this paper is divided into several chapters each containing their own specific sub-research questions. These questions form a framework for the research whereby each chapter contributes to answer the main research question. The sub-research questions are formulated as

Chapter 2 Literature review & hypotheses:

2.1 Which types of models does the literature describe, that are used as the underlying of the rate-making process?
2.2 How is the dependency between the claim frequency and the claim severity described in the literature?
2.3 How does the literature model the total aggregate loss with and without dependency between claim severity and claim frequency?

Chapter 3 Method:

3.1 How is the benchmark model used to model the total aggregate loss?
3.2 How does the benchmark model need to be adjusted to relax the independence assumption in the frequency-severity model?
3.3 Which goodness-of-fit measures can be used to compare the performance of the benchmark and the adjusted model?

Chapter 4 Data**:**

4.1 Is there a significant correlation between the CF and CS based on a correlation test?
4.2 Which risk factors are available, and which risk factors can be included in the benchmark and the adjusted frequency-severity model?

Chapter 5 Results**:**

5.1 How can the total aggregate loss with and without dependence be modelled?
5.2 How does the adjusted model perform in comparison to the benchmark model, with regards to the prediction quality and in terms of goodness of fit?

1.5. The problem Approach & Research Design

The main strategy for this research will be a quantitative approach, using several statistical methods to analyse the relationship between claim severity and claim frequency. The study will be based on a cross-sectional design, collecting data at a single point in time from a database of insurance claims over a period starting in 2015 and ending in 2023. This design is suitable for testing the research question and hypotheses, as it allows for the collection of data on both claim severity and claim frequency from the same set of claims.

Motivated by the problem context and the core problem, this work aims to fit the framework used for modelling the dependence in a frequency-severity model to data of the home-contents insurance of USL. The validation part of the model involves comparing the output of the model to historical data to confirm whether the model improves without the assumption. The adjusted model gets 'trained' with the data starting from 2015, and then this adjusted model will be compared to the most recent data available to determine the effectiveness of predicting the total aggregate loss and compare it to the benchmark frequency-severity model that did not take the dependency into account.

This study aims to answer the main research question, by answering each sub research question individually per chapter. In Chapter 2 *Literature review*, the existing methods as underlying of the rate-making process are described, this could include Machine learning or GLM. The research questions for chapter 2 will be answered with the use of a literature study, in this study the existing literature about the past, present and future research of both the dependency between CF and CS and the way of modelling the total aggregate loss will be explained.

The statistics and the theoretical background need to be evaluated in order to create a solid understanding of the framework this study will propose. So, in Chapter 3 the theory that is needed will be provided. The ending of the chapter will explain the models that will be both trained and tested. The research questions of chapter 3 will be answered supported by the literature review of chapter 2 and the framework proposed by Schulz (2013). The goodness-of-fit measures will be carefully chosen to best represent the accuracy of the models.

The next chapter of the paper will focus on the research data. An explanatory analysis of the data will be performed to provide the reader with a clear understanding of the data that forms the basis of this study. Research question 4.1 will be answered after conducting a correlation test. The data for the claim counts and their average claim

severity will be obtained to be able to perform the correlation test. This correlation test will create first insights in the relationship between CF and CS. The test will return a P-value, with this P-value it is possible to conclude at which significance level it is possible to state whether the correlation between the CF and CS is different from zero. Question 4.2 will be answered after the explanatory analysis, this analysis will highlight risk factors that could have a significant impact on the total aggregate loss.

The research question of chapter 5 will be answered after the GLM analysis is performed and the results can be quantified. The questions can be answered after the prediction of the total aggregate loss of both the benchmark and the adjusted models are compared, and the goodness-of-fit measures are done.

The discrepancy in knowledge in the literature is clear, how does the adjusted frequency-severity model compare to the benchmark frequency-severity model when predicting the total aggregate loss in the home insurance sector. Specifically, the *home contents insurance (All-risk)* is the insurance coverage type chosen to test this research on.

1.6. Data collection techniques and methods

Obtaining, analysing, and processing data about claim severity and claim frequency involves several steps. The first step is to collect all the necessary data. At USL all the claims that are made are tracked in a database, this is good starting point of the research. The data collected includes information on the number of claims, the severity of each claim, the types of claims, and many more relevant variables. A complementary database with the information of the policyholders is available, this is important to test for significant risk factors. Also data, in the form of knowledge, will be obtained from the literature. The literature will provide insights and will help to form a framework and create boundaries for this research.

The research will be performed with GLM analysis. GLM analysis will be used to determine whether claim frequency is dependent on claim severity or vice versa, and to control for potential significant variables such as the region or the age of the policyholder. The claim frequency will be added as a covariate in the regression of the claim severity to describe the possible dependency between the CF and the CS. Whether the adjusted model will improve the goodness-of-fit is determined with measures as the (scaled) deviance, Akaike Information Criterion and Bayesian Information Criterion.

The data analysis & preparation will be explained in chapter 4, this chapter is supported by the software of R. The correlation between the CF and CS will be analysed with a Pearson correlation test. Also, the Spearman correlation test will be used to capture non-linear effects. The results of these coefficients will enlighten the reader with a first impression of how the dependency between the CF and the CS holds.

The research will draw conclusions based on the results of the GLM analysis, considering the implications for insurance companies and their policyholders. The study may suggest that insurance companies need to adjust/finetune their models that function as an underlying in the process of calculating the risk premium, to account for the dependency between claim severity and claim frequency. As a result, premiums can differ from the current value. Policyholders may also benefit from the findings, as they

may be able to make more informed decisions about their insurance coverage based on their risk profile, but this is up to USL to decide whether to share the results of this research.

This study is based on the framework proposed by Schulz (2013). This framework uses the CF as a covariate in the model of the CS in order to account for the dependency. The study of Schulz is applied on data of a car insurance company, whereas this research will be focussing on a home content insurance. So, this research will complement the existing literature with new insights of using the framework proposed by Schulz (2013) and fitting it to the home content insurance sector. The framework will be validated by using it on new data.

Chapter 2 Literature review & hypotheses

The literature chapter will provide all the necessary knowledge and theory found in and supported by literature. This chapter is divided in several sections compromising this research. The ending of the chapter will state the hypotheses of this study.

2.1. Models used in the ratemaking process

Several methods have been either used or proposed by the literature to function as the underlying model for the ratemaking process. This section will provide the literature about the popular choice amongst insurance companies, GLMs. Also, the introduction and the use of Machine Learning models will be discussed.

Generalized Linear Models are a popular choice for ratemaking in non-life insurance. GLMs are statistical models that can be used to model a wide range of distributions, including those commonly used in insurance, namely Poisson and Gamma distributions. GLMs can incorporate multiple factors, such as age, gender, and location to predict future claims costs. They are particularly useful when modelling count or frequency data, such as the number of claims made by policyholders.

One of the first studies to fit GLMs to insurance data was by Green & Higgs, (1989), who used Poisson regression to model automobile/motor insurance claims. Since then, numerous studies have claimed the effectiveness of GLMs for ratemaking in non-life insurance. For example, Wüthrich & Merz (2008) compared the GLMs to traditional actuarial methods and found that GLMs were more accurate in predicting automobile insurance claim frequencies. This research also indicated that GLMs were capable in dealing with overdispersion, which is a common theme in insurance data.

Another study by England and Verrall (2002) used GLMs to model the frequency and severity of claims in commercial property insurance. They concluded that GLMs were able to capture the complex relationships between various risk factors and claims costs. In a similar study Karlis & Ntzoufras (2003) used GLMs to model the frequency and severity of claims in the liability insurance. They concluded that GLMs could handle/process the excess zeros often found in liability insurance data.

Whereas the user-friendly and interpretation are pros of GLMs, it also has cons. The predictability is better received when using ML models, stated in the presentation of Zhou & CPCU Debbie Deng (2019). Machine learning (ML) techniques such as neural networks and decision trees have become popular for ratemaking in non-life insurance. The literature is showing more and more studies about the comparison of GLMs with other methods such as ML. Also Generalized Additive models (GAM) or General Linear Mixed models are used in studies to compare the output of the different models. Studies such as (Eriksson, 2021; Oeben, 2015; Su & Bai, 2020) compare the GLM models to the GAM or ML models. The results of these studies all conclude that there is not much difference in results between the models, however, these studies have not been performed on house-content data. A study of Hu & Kuo (2019) used both GLMs as well as ML models to model the frequency and severity of claims in medical malpractice

insurance. They found that GLMs performed slightly better than other machine learning methods in terms of predictive accuracy and interpretability.

In summary, GLMs are perceived as the most traditional method of frequency-severity modelling, however, other methods such as GAM or ML are upcoming and there is no proof to state they are underperforming. Even though the more recent studies indicate the effectiveness of ML, Univé Stad en Land chooses to pursue GLMs as the foundation of the frequency-severity models.

2.2. Frequency-severity modelling

The process of using the 'best' models is based on the most efficient way to predict the expected aggregate claim loss, which is the result of the frequency-severity model. Frequency-severity modelling is a crucial aspect of non-life insurance as it helps insurers to estimate the expected losses and it helps to determine the premium rates for policies. This way of modelling is currently based on the assumption that the number of claims filed, and the amount paid per claim are independent of each other. This section of this chapter will examine the current research on frequency-severity modelling in non-life insurance and discuss its implications for the industry.

One of the most commonly used frequency severity models in non-life insurance is the Poisson model. This model assumes that the number of claims filed follows a Poisson distribution, and the amount paid per claim follows a separate distribution, often a Gamma distribution. The Poisson model has been widely studied and is well-understood by insurers, making it a popular choice for modelling claims in non-life insurance (Guillén et al., 2016).

However, more recent research has suggested that a Poisson model may not be the best fit for all types of claims. For example, claims incurred for catastrophic events such as hurricanes or earthquakes may follow a different distribution than the Poisson model, leading to bias, and thus inaccurate predictions of the number and severity of claims (Shiu et al., 2020). Other researchers have proposed alternative models such as the negative binomial or zero-inflated Poisson models, which may be more suitable for certain types of claims (Guillén et al., 2016).

The claim severity is often modelled using a Gamma distribution, this distribution will enable positive results only, which in the case of the severity of the claim is logical, as there is no such thing as a negative claim (Ghaddab et al., 2023). In addition to the choice of model, other factors such as data quality and selection bias also affect the accuracy of frequency severity modelling. For example, if the data used to fit the model is not representative of the claim's population, the resulting model may not accurately reflect the true distribution of claims (M. V. Wüthrich, 2015).

2.3. Dependency between claim frequency and claim severity

Understanding the relationship between claim frequency and claim severity is important for insurers to effectively price their products and manage their risk exposure. Several studies have explored the relationship between frequency and severity in non-life insurance. One of such a study is by Denuit & Boucher (2006), this study examined the

relationship between frequency and severity in the automobile insurance. They found that there was a negative correlation between frequency and severity, which means that as the frequency of claims increases, the severity of each claim tends to decrease. Similarly, another study by Wüthrich (2014) also found a negative correlation between frequency and severity in property insurance.

However, not all studies have found a negative correlation between frequency and severity. For instance, a study by Chen & Tzeng (2007) found a positive correlation between frequency and severity in the liability insurance. This means that as the frequency of claims increases, the severity of each claim also tends to increase. A similar positive correlation was also concluded in a study by Gagné & Dionne (2002) in the context of workers' compensation insurance.

Several other studies have also explored the relationship between frequency and severity in various types of non-life insurance, including fire insurance (Bühlmann, 1997), marine insurance (Merz & Wüthrich, 2008) and health insurance (M. V. Wüthrich, 2016). In general, these studies suggest that the relationship between frequency and severity depends on the specific type of insurance and the characteristics of the insured risks. This implies that a negative correlation in the automobile insurance does not correlate with the direction of the correlation in other insurance sectors.

In the case of home content insurance, several studies have found evidence/proof of a negative correlation between the frequency and severity of claims in the home content insurance. For example, a study by Jia & Eling (2016) analysed a dataset of Swiss household insurance claims and concluded that there exists a negative relationship between the frequency and severity of claims, with higher frequency of claims being associated with lower average claim amounts. Similarly, a study by Smit & Schmit (2012) using data from the Dutch insurance market found that there was a negative correlation between the number of claims and the average claim size.

2.4. Hypotheses

Based on the literature review in the previous sections, it can be stated that there is indeed a background found in the literature stating the dependence between the CF and CS. Therefore, the expectation of this research is that there exists a dependency between the CF and the CS in the data of USL. The prediction in which direction this dependency holds is complicated, but the most relatable studies by Jia & Eling (2016) and Smit & Schmit (2012) indicate that a negative dependency seems more likely. Thus, this implies that when a policyholder claims more frequently, the expectation is that the average amount incurred per claim is lower.

In the study of Schulz (2013), the model where the independence assumption is relaxed, performed slightly better than the model where the independence assumption holds. In that study the performance of the models are tested with the deviance of the models. And the difference between the two models is relatively small. The research of Schulz (2013) is conducted on data of automobile insurance policies in Canada. The research of this paper will be conducted on home content insurance policies from USL, thus the expectation is that the results of this research and that of Schulz (2013) are different.

Because the (negative) dependency is expected, the adjusted model should be able to score better in regard to the goodness-of-fit measures. Therefore, when using the adjusted model instead of the benchmark model could lead to differences in the process of pricing the policyholders. These conclusions form the following two hypotheses that will be either supported or rejected in this research

H1: There exists a negative dependency between the claim severity and the claim frequency in the home-content insurance at USL.

H2: The adjusted model will perform better in predicting the total aggregate loss, and it will propose a better fit with regards to the goodness-of-fit measures, than the benchmark model.

Chapter 3 Methods

The methods chapter will provide the reader with the derivation of the GLM models used to predict the total aggregate loss. The chapter will also provide all methods used in the latter stages of this research.

3.1. Preliminary correlation test

To state whether there exists a correlation between the claim frequency and the claim severity, two correlation tests will be performed. These tests will be included as a form of a preliminary analysis into the possible correlation between the CF and the CS. After this preliminary test, insights into the relationship are created which will help understanding the dependency between the CF and the CS better. The two correlation tests that will be used are the Pearson correlation test (r_p) and the Spearman correlation test (r_s). The results of the Spearman correlation test will be more valuable as it is able to capture the non-linear effects. However, for the completeness the Pearson correlation test will also be included.

3.1.1. Pearson correlation test

This method has less assumptions of the data in comparison with the Spearman method, the sample Pearson correlation coefficient r_p is defined so that the mean centering procedure on the *x* and *y* vectors is done first, where m_x and m_y represent the mean of the variables. Then the correlation is defined as

$$r_p = \frac{\sum(x-m_x)(y-m_y)}{\sqrt{\sum(x-m_x)^2(y-m_y)^2}} \quad (1)$$

For the test to conclude at with significance level the correlation is tested, the p-value needs to be determined. The p-value can be determined by first obtaining the degrees of freedom (*df = n-2*), where *n* is the number of observations in the dataset for variables *x* and *y*. Then, with the t value the corresponding p-value can be found in the t-distribution table (see appendix A1). The t-value can be calculated as follows:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

3.1.2. Spearman correlation test

The Spearman correlation test will capture non-linear effects which can result in better estimations of the real correlation than the Pearson correlation test. The Spearman correlation coefficient r_s is calculated in a comparable manner as r_p, except the r_s is calculated after both variables have been transformed to rank values (*x'* and *y'*). Again, a mean centering is performed first, where m_x and m_y represent means of the variables. The Spearman correlation coefficient can be calculated as

$$r_s = \frac{\sum(x\prime - m_{x\prime})(y\prime - m_{y\prime})}{\sqrt{\sum(x\prime - m_{x\prime})^2 (y\prime - m_{y\prime})^2}} \quad (2)$$

The test with the p-value (significance level) is similar to the approach with the Pearson correlation test.

3.2. GLM models

Generalized linear models are widely used in non-life insurance as a tool to determine the risk classification (W. Lee et al., 2019). GLM is an advanced statistical modelling technique formulated by John Nelder and Rober Wedderburn in 1972. It is an 'umbrella' term that encompasses many other models. Classic linear models attempt to fit a model to the mean response of some observed variable *Y* in the form of a linear predictor, whereas GLMs are an extension to this approach. This extension allows for greater flexibility in modelling observations in several ways (Schulz, 2013). First of all, GLMs allow for a non-linear function of the mean to be modelled in terms of a linear predictor.

Secondly, in classical linear regression, the error term is normally distributed with mean zero and a constant variance. GLMs allow the error distribution to be a member of the exponential dispersion family other than a normal distribution. Moreover, the GLMs allow for a mean-variance relation which is inherent in the exponential dispersion models density structure. Thus, when modelling the mean through a GLM, we indirectly model the variance as well.

The GLMs can all be described with the following three components:

1. **Random component** – defines the probability distribution of the response variable, such as the normal distribution used in the classical regression model, or the binomial distribution used in the binary logistic regression model. The model only includes this random component and does not have a separate error term.
2. **Systematic component –** defines the explanatory variables $(x_1, x_2, \dots, x_n)$ in the model and their linear combination, e.g. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$.
3. **Link function, η or $g(\mu)$ –** defines the link between the random and the systematic component. In simple terms, it maps a non-linear relation to a linear relationship. This needs to be done in order to fit a linear model. Classical linear regression is also a form of a GLM, where the link function is simply 'one': $\eta = g\big(E(Y_i)\big) = E(Y_i)$. The link function will be described in more detail in section 3.2.1.

3.2.1. Link function

In the classical linear regression, the link function is not visible if we solely look at the distribution. This is because the link function is there to define how the expected response will be mapped from the linear predictor scale to the mean scale through its inverse. The link function can be chosen to be able to map the mean so that it reflects the distribution. The 'mapped range' is chosen because of the (assumed) distribution of the response variable. So, in the case where the response variable is assumed to be Gamma distributed, *Y* has to be in the mapped range of (0, ∞).

Obviously, the link function chosen for a Gamma distributed response variable must ensure that the g^{-1}: (∞, ∞) → (0, ∞). This can be done with the use of the *Log link* function, this function would satisfy the mapped mean range of a Gamma distribution. The idea of a link is thus that it makes sure that $\mu > 0$. In Table 1 some link functions are shown, where *identity* is used with classical linear regression and the *log* is used in the Gamma example described earlier.

Table 1 Link functions

Name	*Link Function*	*Mean*	*Range of Mean*
Identity	$z = \mu$	$\mu = z$	*(-∞, ∞)*
Log	$z = Log(\mu)$	$\mu = e^z$	*(0, ∞)*
Inverse	$z = 1/\mu$	$\mu = 1/z$	*(-∞,∞)*
Inverse Squared	$z = 1/\mu^2$	$\mu = 1/\sqrt{z}$	*(0, ∞)*
Square root	$z = \sqrt{\mu}$	$\mu = z^2$	*(0, ∞)*

In this study, the only relevant link function is the *log* link. Therefore, Table 1 is included to provide possible link functions when other GLMs are used, but a deep understanding of the other link functions is not necessary.

3.3. Derivation of the total aggregate loss model under independence

The currently used frequency-severity model (benchmark model) to calculate the total aggregate loss at USL assumes the independency between the CF and CS. In this section the framework for using GLMs to predict the total aggregate loss will be described. This framework is based on relatively simple statistics which results in an accessible way of determining the total aggregate loss.

3.3.1. Benchmark frequency-severity model

As stated in section 3.2, the GLM technique is a frequently used method for modelling the process of prediction the total aggregate loss in an insurance company. This method is based on the multiplication of the expected CF and the expected CS (Garrido et al., 2016). Both the CF and the CS are separately modelled using the Poisson distribution and the Gamma distribution respectively at USL. Then, to predict the total aggregate loss, these marginal GLMs can be multiplied. Different distributions can be used to model the CF and CS, but the most generic distributions are the Poisson and the Gamma distribution.

The total aggerate loss per policyholder (individual level) can be determined by adding all of the claim severities of that specific policyholder, this can be formulated as

$$S_i = \sum_{j=1}^{N_i} Y_{ij},$$

Where

(1) S_i = the amount that is claimed throughout the chosen period for policyholder i.
(2) N is the number of claims, where N_i is the number of claims (CF) of policyholder i

(3) Y is the severity of the claim, where Y_{ij} is the severity of the claim for claims j $\in$ (1, 2, …, N_i) of policyholder i.
(4) $Y_{ij}, j = 1, \dots, N_i$ are conditionally i.i.d., given N_i

Specifically, the observed (individual) severity for policyholder i is

$$S_i = \begin{cases} (y_1 + y_2 + \cdots y_{N_i}) & N_i > 0 \\ \emptyset & N_i = 0 \end{cases}$$

With this formula it is easy to see that $S_i = 0$, when $N_i = 0$ (Schulz, 2013). Furthermore, it can then be determined that for the independent claim severities Y_{ij} we obtain

$$E(\bar{Y}_i) = E\left[\frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}\right] = \mu_i \qquad (3)$$

The total aggregate loss per policyholder can then be written as

$S_i = \sum_{j=1}^{N_i} Y_{ij} = N_i \bar{Y}_i$

As stated in the previous chapters, the insurers act as if the CF and the CS are mutually independent. The expected total aggregate loss can then be formulated as

$$\begin{aligned} E[S] &= E[E(S|N)] = E[E(\textstyle\sum_{j=1}^{N} Y_j|N)] \\ &= E[\textstyle\sum_{j=1}^{N} E(Y_j|N)] = [E\textstyle\sum_{j=1}^{N} E(Y_j)] \\ &= E[N\,E(Y)] = E(Y)E(N) \end{aligned}$$

This formula is not on the individual policyholder (*i*) level, but on the total level for the complete portfolio. The derivation of the expected aggregate losses holds, when E(Y|N) = E(Y) and the same thing can be stated for the expected frequency E(N|Y) = E(N). The research of Schulz (2013) indicates that with the moment generating function, the first two moments of S are determined by the first two moments of the frequency and the severity, the Var(S) can then be rewritten as

$$\begin{aligned} Var(S) &= E(S^2) - [E(S)]^2 \\ &= [E(Y)]^2 Var(N) - E(N)Var(Y) \end{aligned}$$

The CF and CS are modelled separately using a GLM, so when a vector of covariates is available at the individual level, this data will be used to determine the GLM functions. For both models the canonical log link function (see Section 3.2.1) is chosen (g_1 and g_2) so that the conditional means can be written as

(a) $E(N|x) = {g_1}^{-1}(x^T\alpha) = e^{x^T\alpha} = v$
(b) $E(Y|x) = {g_2}^{-1}(x^T\beta) = e^{x^T\beta} = \mu$

α and β represent the coefficient vectors. The vector of covariates (x) will not be the same for the CF and the CS, but it is formulated for simplicity. The covariate vector x^T is the transpose of the covariate vector x, this enables the multiplication of the coefficient and the covariate vector.

The assumption of independence between the CF and CS is present, the benchmark frequency-severity model for the aggregate loss, given the covariate vector x, can be written down as

$$E(S|x) = v\mu = {g_1}^{-1}(x^T\alpha) * {g_2}^{-1}(x^T\beta) \quad (4)$$

Because USL models the CF and CS with a Poisson and Gamma distribution respectively, the canonical link functions are Log-links (see Section 3.2.1). A canonical link function is a specific type of link function that is derived from the exponential family of distributions. The canonical link function is the link function that provides the best fit between the predictors and the response variable. Equation (4) reduces to

$$E(S|\mathrm{x}) = v\mu = e^{x^T\alpha + x^T\beta} \quad (5)$$

The advantages of this way of modelling are that it ensures that each covariate alters the baseline rate by a multiplicative factor (Garrido et al., 2016). Also, it simplifies the variable selection process (Shi et al., 2015). The use of the Log-link also supports the ease of modelling, as the log-link ensures a positive mean for the frequency and severity.

3.4. Derivation of the total aggregate loss model under dependence

Because the dependence between CF and CS is expected to be present (G. Y. Lee & Shi, 2019), the adjusted frequency-severity model where the dependency is included must be described. Existing literature about the topic of dependent frequency-severity modelling includes using the CF as a covariate (Renshaw, 1994), Copula-based models (Frees et al., 2016), and bivariate random effect-based models (Lu, 2019).

The covariate method includes the dependence because the variable *Claim Count*, which reflects the CF, is included in the marginal GLM of the CS. This method states that if the variable *Claim Count* is significant, then the CF does have an effect on the CS, meaning the dependence should be included. With the bivariate random effect-model, the dependency between the CF and CS can explicitly be modelled. This is achieved by including random effects for both variables and allowing them to be correlated at the individual or group level.

In this research the covariate method will be tested on data in the home insurance industry, this method is chosen because it is relatively simple to implement in comparison with the Copula or the bivariate random effect-based model. Also, this method is able to capture the full effects of the dependency as the CF will be included in the GLM of the CS.

3.4.1. Adjusted frequency-severity model

The decomposition of using the CF as a covariate is relatively straightforward, with the use of conditional probability. In the derivation of the adjusted total aggregate loss model, a covariate vector X is used. This covariate vector does not necessarily have to be the same as the covariate vector X used in the benchmark model, but for simplicity and readability, the X is reused. Based on the research conducted by (Frees et al., 2016b), the decomposition can be formulated as follows

For N > 0, let $\bar{Y} = \frac{(Y_1+Y_2+\ldots+Y_N)}{N}$ be the average claim severity for the claim frequency N. then the total aggregate loss can simply be written as S=$N\bar{Y}$.

$$P(N = n, S = s|x) = P(N = n|x) * P(S = s|N = n, x)$$

This implies that the expected value of the total aggregate loss is no longer simply the product of the marginal means of the frequency and severity components, but it can be described as

$$\begin{aligned} E(S|x) & \qquad (6) \\ &= E(N\bar{Y}|\, x) \\ &= E(E[N\bar{Y}|x, N]|x) \\ &= E(NE[\bar{Y}|x, N]|x) \end{aligned}$$

Equation (6) is not the same as writing it down how the model is composed in the independent frequency-severity model

$$E(S|x) = E(NE(\bar{Y}|N, x)\,|x) \neq E(N|x)E(Y|x)$$

This leads to different estimators for v & μ

(a) $E(N|x) = {g_1}^{-1}(x\alpha) = e^{x\alpha} = v$
(b) $E(Y|x, N) = {g_2}^{-1}\left(x\tilde{\beta} + \theta N\right) = e^{x\tilde{\beta}+\theta N} = \mu^A$

Where the link functions are again described as g_1 and g_2 and chosen to be the canonical log link function. $\alpha\ and\ \tilde{\beta}$ are vectors of the regression coefficients. μ^A represents the modified marginal GLM for the CS so that the dependence is included. Note that the regression parameters $\tilde{\beta}$ are different from the regression parameters in the independent setting β. This is because the inclusion of the CF as a covariate will affect the regression parameters and their estimates. The interesting part of formula (b) is the θ which indicates the degree of dependence between the CF and CS. The interpretation of this dependence variable is such that when $\theta = 0$, then $\mu^A = \mu$ which thus reflects the same estimator as in the benchmark model. This is true because the regression parameters of the marginal GLM for CS in the adjusted setting will be identical to the benchmark GLM($\beta = \tilde{\beta}$). The regression parameters are equal because both means will be modelled using the same covariates and the CF is not included in the adjusted setting (when the degree of dependence is 0, then both GLMs are identical). When the estimators for the CF and CS are equal, the total aggregate loss model is also equal.

When the dependence parameter θ is greater than zero, the total aggregate loss model is defined differently. With the information known that the log-link is chosen, the conditional mean severity μ^A follows

$$e^{x\tilde{\beta}+\theta N} = \tilde{\mu}e^{\theta N}$$

$\tilde{\mu}$ is the marginal GLM model of the CS where the CF is included as a covariate. The total aggregate loss model then becomes

$$E[S|\mathrm{x}] = [\mathrm{NE}(\bar{Y}|\mathrm{X},\mathrm{N})|\mathrm{x}] = E\left[N\tilde{\mu}e^{\theta N}|\mathrm{x}\right] = \tilde{\mu}E[Ne^{\theta N}|\mathrm{x}] \quad (7)$$

Further derivation of equation *(7)*, with the use of the moment generating function, is described in the paper of Schulz (2013, p50-51). Based on the research of Garrido (2016) and Schulz (2013, p50-51) the expected total aggregate loss is formulated as

$$E(S|x) = v\tilde{\mu}\exp\{v(e^{\theta}-1)+\theta\} \quad (8)$$

This model differs from equation (5) with the addition of the dependence correction formulated as $\exp\{v(e^{\theta}-1)+\theta\}$. The model assumes that the CF follows a Poisson distribution, this assumption is valid for this research as the marginal GLM for the CF is indifferent between the benchmark and the adjusted model. The final formulation of the adjusted model makes no (distributional) assumptions for the severity if it belongs to the exponential dispersion family. The mean can then be modelled via a GLM, this is the case as the severity is modelled using the Gamma distribution as 'family'.

3.5. MLE for determining the coefficients

To use the benchmark and adjusted models to predict the total aggregate loss, the values of the coefficients need to be determined. The method used to estimate these values is the Maximum Likelihood Estimation (MLE). The MLE is used to estimate the parameters of a probability distribution.

The process of estimating the coefficients involves iteratively optimising the likelihood function that states the probability of observing the data given the values of the coefficients. The specific method used for optimisation depends on the distributional assumption of the response variable and the link function chosen in the GLM model. This study provides a method to model the total aggregate loss function with the use of the likelihood functions for the claim frequency and claim severity. Therefore, in this section the Poisson likelihood function will be derived, as the Poisson distribution is assumed for the claim frequency. The derivation for the likelihood function for a Gamma distribution can be found in the supported literature of Dobsen Anette J. (2002, p68-p73).

In a Poisson GLM with a log link function, the likelihood function is the Poisson likelihood, and the optimisation can be done using an iterative algorithm such as Fisher scoring. The process of determining the Poisson likelihood function can be seen as

$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$ *Write the PDF*

$L(\lambda; x_1; \dots; x_n) = \prod_{j=1}^{n} \frac{\lambda^{x_j} e^{-\lambda}}{x_j!}$ *Likelihood function*

$L(\lambda; x_1; \dots; x_n) = -n\lambda + \ln(\lambda)\sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \ln(x_j!)$ *Natural log likelihood function*

$\frac{d}{d\lambda} l(\lambda; x_1; \dots; x_n) = -n + \frac{1}{\lambda}\sum_{j=1}^{n} x_j$ *Derivative natural log likelihood function*

$\lambda = \frac{1}{n}\sum_{j=1}^{n} x_j$ *Derivative equal to zero and solve for λ*

The steps to determine the MLE of a Poisson likelihood function are written down next to the formulas. The coefficients that maximize the log-likelihood function are the MLEs. However, the log-likelihood function does not have a closed-form solution, so we need to use an iterative algorithm such as Fisher scoring or Newton-Raphson to maximize it. For understanding this research, it is not of importance that the concepts of algorithms such as Fisher scoring is fully clear. R has built-in functions for fitting GLM models, which uses MLE to estimate the regression coefficients. The function takes the form glm(formula, data, family, ...), where the formula specifies the model equation, data is the input data, and family specifies the distributional and link function for the response variable. Once the model is fit, the estimated coefficients and their standard errors can be obtained.

For a more in-depth explanation of how the Log-likelihood functions looks like and how the MLEs can be computed, see the GLM book of Dobsen Anette J. (2002, p68-p73).

3.6. Goodness of Fit

To be able to make a comparison between the benchmark model and the adjusted model, several goodness-of-fit measures are used. The different goodness-of-fit measures are explained in this section.

3.6.1. Deviance function

The goal of modelling data is to obtain fitted values $\hat{\mu}$, for the mean of the response values Y (Schulz, 2013). The fitted values will not exactly coincide with the real data values. The significance of the discrepancy between the real values Y and the estimated values $\hat{\mu}$ can be measured and analysed with the use of the deviance.

The deviance function is a statistical measure, used in this study to assess the goodness-of-fit of a GLM. The deviance function is a generalisation of the idea of using the sum of squares of residuals, it functions as a measure of goodness-of-fit (Al-Mosawi, 2017). In the context of GLMs, residuals represent the deviation between the observed data and the expected values based on the fitted GLM. The deviance function can be observed as the distance between two probability distributions, and it can be used to perform model comparisons.

In the case of determining the performance of a GLM model compared to the reality, the deviance function can define the difference between the maximum log-likelihood of the fitted model and the maximum log-likelihood of a ‘saturated’ model. A so-called saturated model is a model that perfectly reflects the real data, this happens when the number of parameters (r) is equal to the number of observations (n). Therefore, the saturated model is only used as a benchmark, and the deviance represents the degree of lack of fit of the GLM compared to the real data.

The deviance will also be used for model selection, where several GLMs with different sets of explanatory variables are fitted to the data. The deviance is useful when determining whether a simplification of the model leads to more biased estimates. The deviance of these models can be calculated and compared, the model with the smallest deviance is used. This method can be used when the simplification is still a nested model of the more extensive model.

The deviance of the fitted model can be calculated as twice the difference of the log-likelihood with the saturated and the fitted model. If we denote the log-likelihood of the fitted model as $L(\gamma)$ where γ represents the model parameters. The log-likelihood of the saturated model is denoted as $L(\hat{\gamma})$. The deviance function can be written as (*Glmbook*, n.d.)

$$D^* = -2(L(\gamma) - L(\hat{\gamma})) \quad (9)$$

The scaled deviance function can be determined by dividing the deviance with the dispersion parameter. The dispersion parameter in simple terms represents how much the observed data points spread or vary around the predicted values in a statistical model. It quantifies the relationship between the mean and the variability of the response variable. In Table 2 the deviance functions for the Poisson and Gamma distributions are shown.

Table 2 Deviance functions for the Poisson and the Gamma distribution

Distribution	$\boldsymbol{D(y,\hat{\mu})}$
Poisson	$2\sum\{ylog\left(\frac{y}{\hat{\mu}}\right) - (y - \hat{\mu})\}$
Gamma	$2\sum\{-log\left(\frac{y}{\hat{\mu}}\right) + \frac{y-\hat{\mu}}{\hat{\mu}}\}$

3.6.2. AIC

The Akaike Information Criterion (AIC) is a measure of the relative quality of the GLM model for a given dataset. AIC is a goodness-of-fit measure which also takes the complexity of the model into account. The complexity of the model is based on the number of variables included. The less variables used, the simpler the model. A simpler model is preferred over a more complex model, provided that the models fit the data similarly well. This explains the concept of a parsimonious model. The model with the smallest AIC is preferred (Mcleod & Xu, n.d.).

Again, the same as for the deviance holds, that the AIC is working with the log-likelihood of the fitted model. H. Akaike (1974) formulated the AIC as

$$AIC = -2 * l(\gamma) + 2k \quad (10)$$

Where the *k* is the number of parameters used in the model. The AIC will be used to compare the GLMs. This is possible because all models are fitted to the same data. The AIC is not an absolute measure of the quality of the model, therefore, it should be used in combination with other (statistical) measures, in this research it will be combined with the (scaled) deviance and the BIC.

3.6.3. BIC

The Bayesian Information Criterion (BIC) is comparable to AIC and thus also used as a measure to assess the goodness-of-fit. The BIC can be derived using Bayesian methods (Schwarz G, 1978). Again, the smallest BIC is referred as the most compatible model. The BIC is defined as

$$BIC = -2 * l(\gamma) + k\ log(n) \quad (11)$$

Where k is the number of parameters in the model and n is the sample size. The difference with the AIC formula can be seen as that the logarithm of the sample size is used, instead of the coefficient *two* used in the AIC formula. The BIC 'penalises' models with a larger number of parameters as the 'penalty term' (k log(n)) increases when the number of parameters increases. Also, when the sample size increases significantly, the penalty term will increase.

So in summary, all three of the goodness of fit measures used in this research are based on the likelihood function of the model and the total number of parameters used. Especially AIC and BIC propose a measure that indicates the goodness-of-fit with taking the trade-off between model fit and model complexity into account.

Chapter 4 Data

This chapter will provide the information and context needed to understand the data that this study is performed on. The chapter will describe the process of obtaining and preparing the data for this research. The chapter is divided into several sections, the first will enlighten the reader with information about the data. The second section focusses on the relation/correlation between the CF and CS. Additionally, the risk factors used in the GLMs are described with a correlation matrix between the numerical risk factors.

4.1.1. Policyholder datasets

All the data used in this research is provided by USL, with observations ranging from 2015 to 2023. The information about all the policies and their policyholders are given with sheets of the portfolio of USL. With the most recent portfolio (2023) obtaining over 90.000 observations for the private home insurance. Every observation is described in detail with 297 variables. In Table 3 a visual (example) representation of the data set is shown, with only 6 out of the possible 297 variables shown.

Table 3 Example of the Policyholder information

Age	**Ownership status**	**Gender**	**Urbanisation**	**Home type**	**WOZ value**	**…**
34	No	M	20.000 – 50.000	Apartment	125.000	…
45	Yes	M	5.000 – 10.000	Duplex homes	287.000	…
68	Yes	F	< 5.000	Apartment	312.000	…
…	…	…	…	…	…	…

Besides the dataset with the information on all policies, there is a dataset with all claims that have been incurred. This dataset also ranges from 2015 to 2023. For the analysis later on in the study, these datasets need to be combined in order to find trends and patterns between the policyholder's profile/characteristics and their claim behaviour.

The dataset containing policyholder information incorporates various noteworthy variables that would enable a quick overview of the dataset. Later on, in this chapter the dataset of the claims will be touched upon, however, to create the best understanding of the data this researched is performed on, some risk factors are highlighted to create insights in the data.

First of all, the age of the policyholders in the portfolio is an interesting variable to look into. In Figure 1, two histograms are shown, the left histogram shows the count for the age of the policyholder in the portfolio dataset. The histogram on the right side indicates the count for the age of the dataset of the claims. The histogram of the Policyholder age is more skewed on the right side, which indicates a higher average age of policyholders in the portfolio dataset compared to the Claim dataset.

In the histograms of the claims, the histogram is more left skewed. This can be explained due to the fact that younger people tend to file a claim more often than older people.

This is a common phenomenon in the insurance industry supported by the study of Joksch (1980).

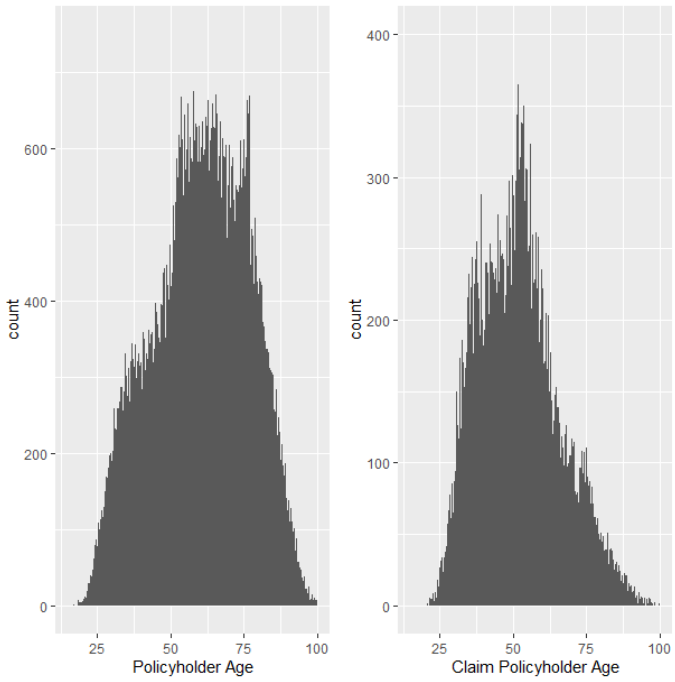


Figure 1 Histograms of the (claim) policyholder age

In the beginnining of the report USL is described as a regional insurer, the choropleth map in Figure 2 shows the density of the living and claim location of the policyholders. Noticably, there are polichholders trhoughout the whole country, this is due to the fact that policyholders move to a house outside of the region of USL. Whereas these policyholders are currently living outside of the region, they could remain insured by USL. The choropleth map on the right side indicates where the most claims are incurred. The municipalities with the most policyholders, seem to be filing in the most claims as well. These darker organge/red municipalities are the places where the urbanisation variable is also highest (>100.000 residents).

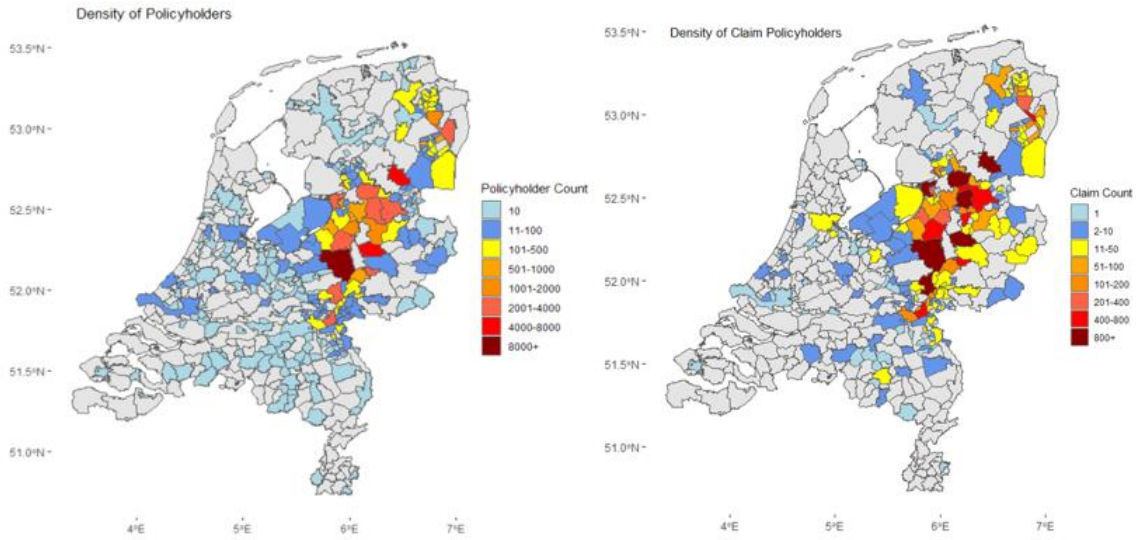


Figure 2 Policyholder density in the Netherlands

Over the past few years, the Netherlands has seen a steady increase in the WOZ (Wet Waardering Onroerende Zaken) values of houses. The WOZ value is the value that is assigned to a property by the municipality for tax purposes. This value is determined based on several indicators, so that it reflects the market value of the property. The WOZ value can be calculated by analysing the sales prices of similar properties in the area.

From 2019 to 2023, the WOZ values of houses in the Netherlands have increased steadily. This increase can be attributed to a number of factors, including a strong housing market and low interest rates. The demand for housing in the Netherlands has been steadily increasing, leading to an increase in housing prices. This, in turn, has led to an increase in the WOZ values of houses. Because the WOZ values are determined over the past year, the WOZ of January 2023 reflects the previous years which could lead to biases when interpreting the results. For example, recent studies show that the average house prices have dropped in 2023, but the WOZ values have failed to represent this behaviour as the measurement needs to be finished (Waarderingskamer, 2023).

This increase in WOZ values of houses can also be seen in the dataset of the policyholders, this trend can be found in Figure 3.

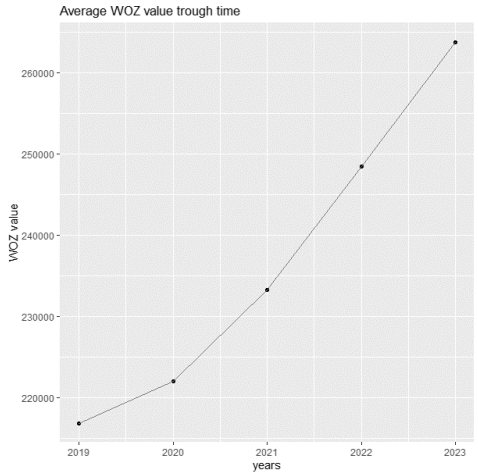


Figure 3 Average WOZ values for the houses of the policyholders

Overall, the increasing WOZ values of houses in the Netherlands from 2019 to 2023 can be attributed to a combination of factors, including a strong housing market, low interest rates, and the impact of the COVID-19 pandemic. It remains to be seen whether this trend will continue in the years to come.

For this research, it is important to take this trend into account, as working with the WOZ value can be misleading because of the increasing prices. New policyholders with similar houses as others in the portfolio, can still experience a higher WOZ value as for policyholders that have been insured for a few years, due to the fact that the WOZ value has not been re-evaluated recently. The WOZ value is updated every few years at USL, which could thus result in bias as the average WOZ value of houses has increased over the years.

Figure 4 consists of three graphs that provide insights into the characteristics of policyholders. The first graph shows the count of the type of houses that the policyholders have. The graph indicates that the majority of policyholders own a single-family home, followed by a townhouse and a condo.

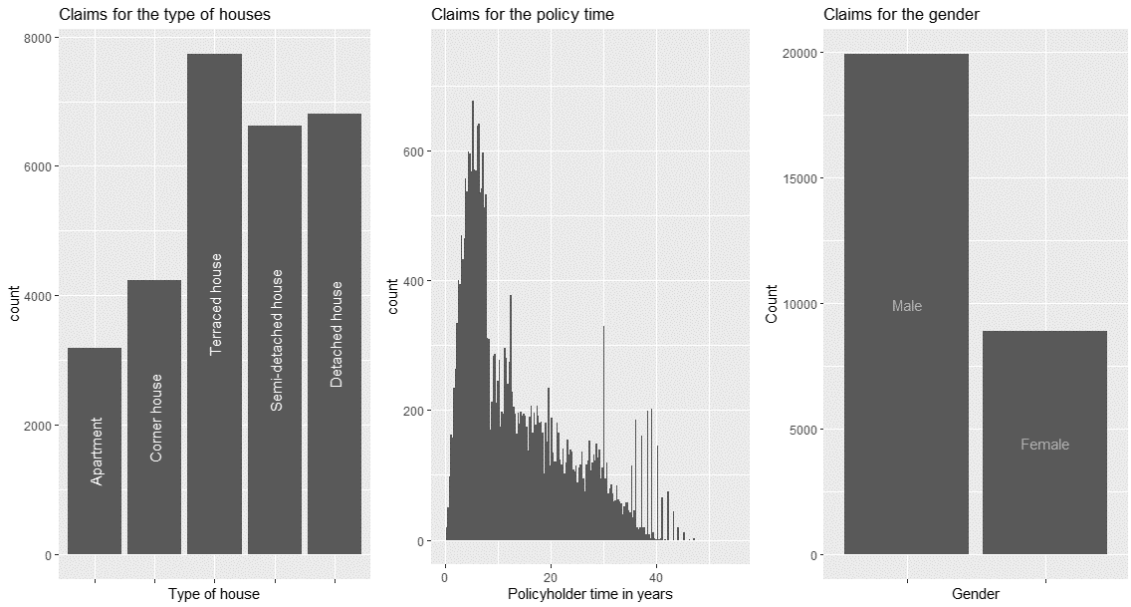


Figure 4 Grid consisting of three figures about the type of house, the policy duration and the gender distribution.

The second graph in the Figure 4 shows the count for how long policyholders have their policy, measured in years. The graph reveals that a large group of policyholders have been with USL for less than five/ten years, with a smaller number of policyholders being with the company for longer periods. This information can be used by USL to evaluate customer retention rates and to identify potential target groups for improvement in customer service. The last graph in the figure is a bar chart that shows the number of female and male policyholders in the dataset. The chart indicates that the dataset has more male policyholders than female policyholders.

4.1.2. Claim datasets

Now that the datasets have been described regarding the general information of policyholders, this section will cover the frequency and severity of the claims incurred. A total of 28.889 claims are incurred over a period starting in 2015 and ending in 2023. In the dataset of the home-content (all risk) insurance since 2015. Just over 8.000 of those claims are incurred with a severity of 0. The count histogram of the severity of all claims with a positive value can be seen in Figure 5.

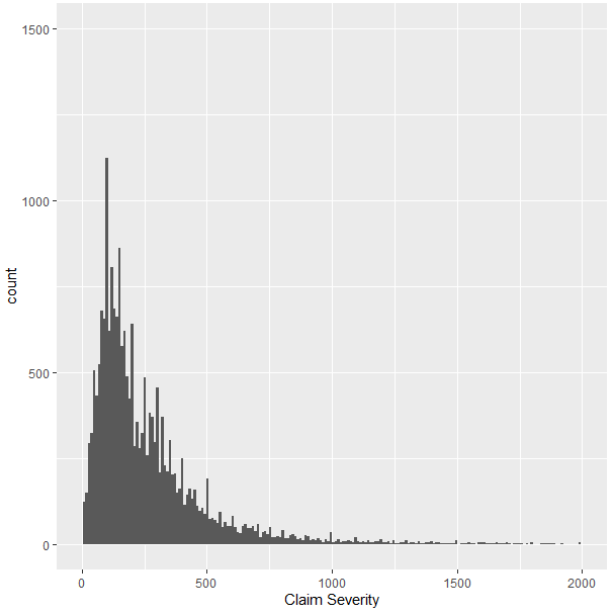


Figure 5 Count histogram for the Claim severity.

As described previously in this report, USL models the CS assuming a Gamma distribution for the claims, in Figure 5 the real distribution can be seen. The distribution of the claim frequency can be seen in Figure 6, the CF is assumed to be a Poisson distributed, which does not allow for negative observations.

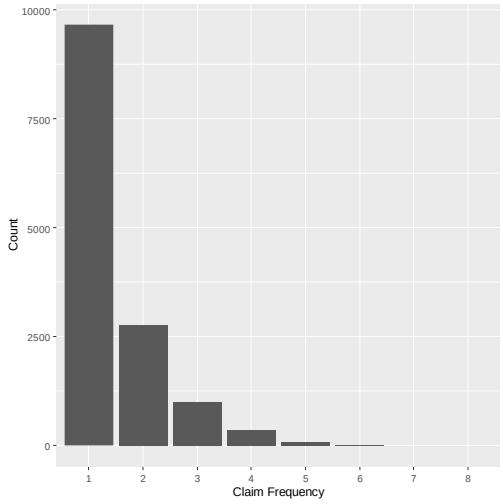


Figure 6 Count histogram for the Claim frequency

When combined, the Gamma distributed claim severity and the Poisson distributed claim frequency can provide a model for the total aggregate loss model in the home content insurance. It can also help to make predictions about the frequency and severity of future claims, which in turn can help set appropriate premiums and possible reserves.

In Figure 7, more explanatory visualisations of several risk indicators that have not been touched upon much in this chapter can be found. In this grid the relationship between the risk indicator and the claim frequency is shown with the use of a scatter plot (Left column of the grid). Also, the relationship with the claim severity is shown with a scatter plot (right column of the grid).

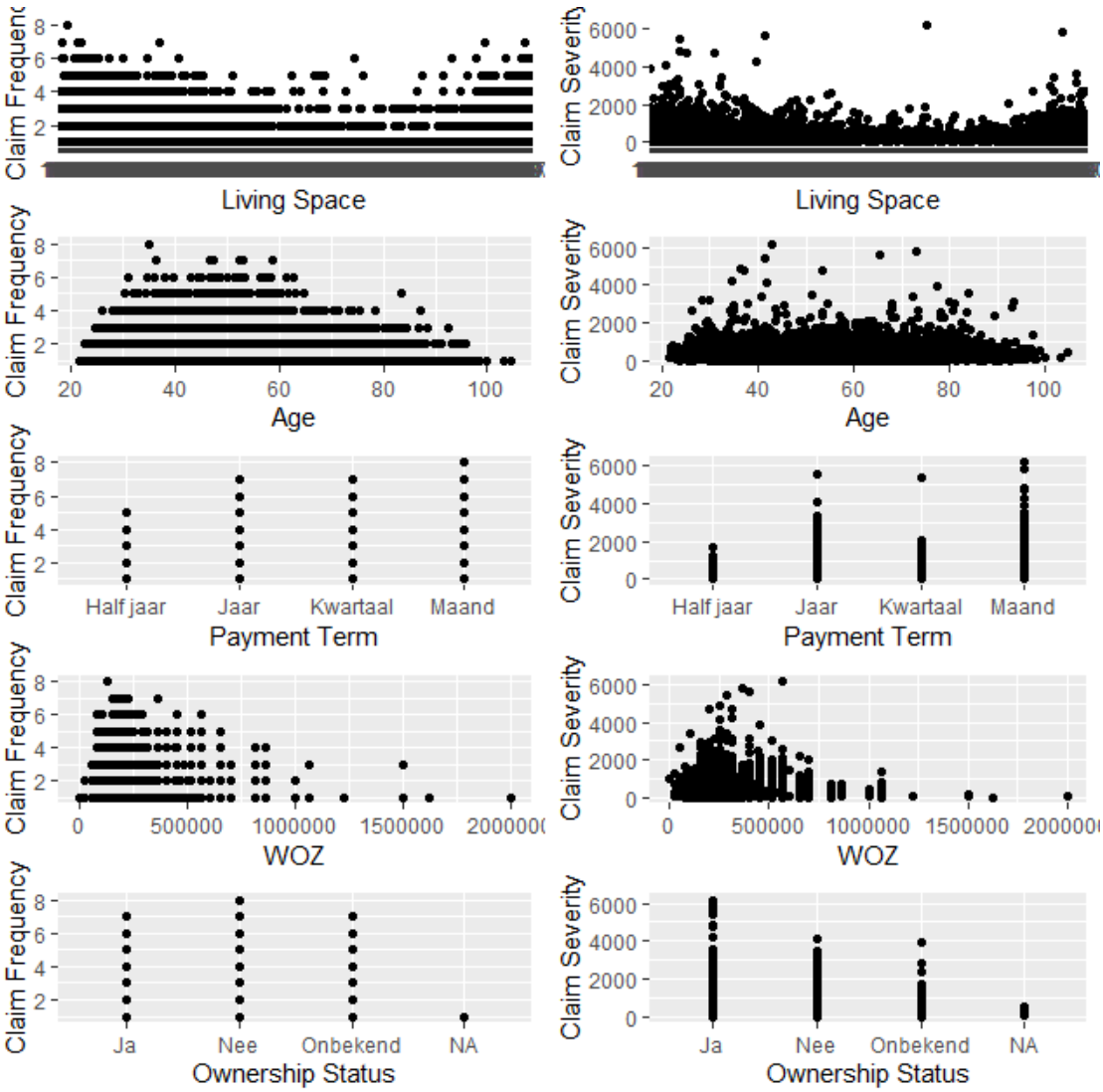


Figure 7 Grid of scatter plots and density plots of some risk indicators, the scatter plots are visualised against the claim frequency and claim severity.

4.2. Correlation claim frequency and claim severity

This section will provide insights in the observations of the average claim severity conditioning on the claim frequency. The dataset contains over 28.000 observations and the total severity average is 189,53 euros. Figure 8 provides a first impression on the relationship between the CF and the CS, the claim counts are divided in groups to improve the visibility of the figure. Because of the outliers, the figure would have been difficult to interpret. Therefore, observations with a severity higher than 500 are excluded. It is hard to draw conclusions from this graph, as some observations are not included (CS > 500) and the lines overlap on many occasions. However, the claim counts 7 and 8 (Group 7-8) do have their 'peak' more to the right than the other groups, indicating a possible higher claim severity.

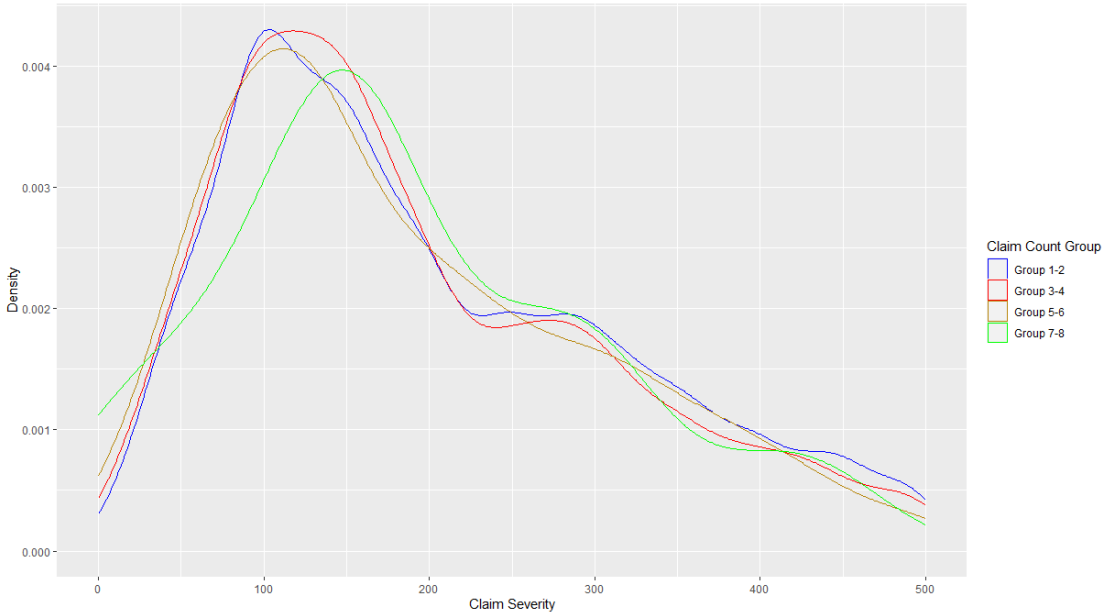


Figure 8 Density for the claim severity per claim count (in groups)

To get a better understanding of the data, Table 4 is included. No noticeable evidence for a correlation between the CF and CS can be seen in Table 4. The average CS is lowest for a claim frequency of 8/9, however, for the greater claim counts, the amount of observations is substantially reduced. The sample group is not big enough to draw conclusions from this table. For the remaining values of the claim counts, the average CS seems to be randomly distributed.

Table 4 Average CS conditioning on CF

Claim Frequency	Count	Average CS
1	12939	178.14
2	3718	195.60
3	1438	196.47
4	571	192.55
5	216	179.95
6	78	181.54
7	28	148.86
8	18	164.29

9	2	162.52
10	1	295.36

As a form of a preliminary analysis and to be able to tell whether a correlation between the CF and the CS is present, this study will perform a Pearson correlation test and a Spearman correlation test. The Pearson correlation coefficient explains the linear relationship between the variables, whereas the Spearman correlation coefficient is able to capture non-linear relationships. The Pearson test is not fully suited for this type of data, but for the sake of completeness the Pearson test is still included. The Spearman correlation test is preferred, as it is able to describe the possible non-linear relationship between the CF and CS. When using a correlation test between the variables CF and CS as seen in Table 4, the following results are obtained

Pearson's $p_{N,\bar{S}}$ = 0.243
Spearman's $p_{N,\bar{S}}$ = -0.212

These values both tell us a different story as to which the direction of the correlation holds. A t-test is performed with the H0 (null hypothesis) stated that the correlation between the CF and the CS is not different from 0, and the H1 (alternative hypothesis) stated that the correlation is not equal to zero. For both t-tests, Pearson and Spearman, the P-values are 0.499 and 0.560 respectively, meaning that the H0 cannot be rejected. This implies that because of the tests, the correlation between the CF and CS is not different from zero.

In Table 4 all observations are included, also the claims with an incurred amount of zero. Earlier in this report the idea of IBNR is already highlighted, in short, this happens at insurance companies when a claim is incurred, but not reported, so no value of the CS is saved. Also, a lot of claims have been reported with a CS of zero. This is due to the fact that a policyholder files a claim but fails to fill in all additional data that is required handle the claim. These claims will not be pursued and thus end up with a claim severity of zero. These observations do not enhance the study proposed in this report. So, in Table 5, only observations with a CS greater than zero are included.

Table 5 Average CS (>0) conditioning on CF

Claim Frequency	Count	Average CS
1	9654	278.27
2	2776	260.52
3	1013	246.02
4	363	241.82
5	92	229.24
6	25	240.73
7	7	236.07
8	1	126.26

There is a big difference compared to the table where claims that equal zero are also included. Logically, the count per CF reduced, but also the maximum of the claim frequency decreased. The maximum claim frequency of a policyholders with claims

greater than zero is eight, this means that the policyholders that claimed 9,10 or 11 times must have filed claims and not pursued with them. Also, the average severity seems to be decreasing when the claim frequency increases. Again, the Pearson and Spearman correlation t-tests are performed with the correlation coefficients stated as

Pearson's $p_{N,\bar{S}}$ = -0.778
Spearman's $p_{N,\bar{S}}$ = -0.929

The correlation coefficients seem to be indicating a negative relationship, the results of the t-test support this observation. The p-values for the Pearson correlation test and the Spearman correlation test are 0.023 and 0.002 respectively. The Pearson correlation test indicates that the H0 can be rejected at a 5% significance level, whereas the Spearman correlation test rejects the H0 at a 1% significance level. The conclusion is thus that correlation between the claim frequency and the claim severity is not equal to zero.

The dependency between the CF and CS cannot be described solely with these values for the correlation, however, they provide an insight in the data. The dependency has many reasons and indicators as to why it behaves a certain way, whereas these correlation tests only determine the correlation of one variable with another, regardless of all covariates. That is why this is not a conclusion about whether there exists a dependency between the claim frequency and the claim severity.

Another notable 'bias' in these calculations is the fact that the exposure of all policyholders during the period of 2015 till 2023 can be different. A policyholder that claims four time during a year is filing claims more frequently than someone filing four claims over a period of five years, which is why in the modelling part of this study an exposure variable needs to be included. This variable will correct for errors regarding policyholders having different policy duration in the dataset. This variable should be included in the model for the claim frequency.

4.3. Descriptive statistics and risk factors

The variables/ risk factors that will be used as covariates in the regression analysis can be numerical (e.g. age or the WOZ-value) or categorical (e.g. gender house type). This section will highlight the key risk factors that could be significant when modelling the frequency-severity component. The categorical variables must be clustered so that the model can be both parsimonious and efficient.

In order to accurately predict the risk factors, it is important to consider a range of variables that may impact insurance claims. Through careful consideration and consultation with USL, and the support of papers from the literature such as Becker et al., 2022; Lee et al., 2019a, the following risk factors have been predicted to be included in the regression part of Chapter 5: Own Risk, gender, age, policy duration, payment term, house type, urbanisation, living space, property value assessment (WOZ), location in a major city, year of construction, ownership status, and postcode risk class. Additionally, the presence of 'Thatched roof' has been identified as an important factor in assessing the risk of property damage and will also be included in the analysis.

The categorical risk factors are clustered individually, each having different number of levels. Table 6 indicates the variables and the type of variable, whether the variable is numerical or categorical/binary. Also, the abbreviation (e.g. x_1) is included in Table 6 to enhance the readability for later use. The last two columns show the minimum and maximum observation of the variable when possible.

Table 6 Risk factors for the GLM regression

Variable	Risk factor	Type of variable	Min.	Max.
X_1	Own Risk	Numerical	0	300
X_2	Gender	Binary	-	-
X_3	City	Categorical (516)	-	-
X_4	Payment term	Categorical (4)	-	-
X_5	House type	Categorical (10)	-	-
X_6	Urbanisation	Categorical (10)	<5.000	>500.000
X_7	Living space	Numerical	15	2073
X_8	Property Value Assessment (WOZ)	Numerical	25000	2000000
X_9	Insured amount precious	Categorical (14)	0	>50000
X_{10}	Insured amount jewellery	Categorical (11)	0	>50000
X_{11}	Ownership status	Binary	-	-
X_{12}	Postcode risk class	Categorical (9)	-	-
X_{13}	Thatched roof	Binary	-	-
X_{14}	Extra living space	Categorical (14)	0	<500
X_{15}	Claim count	Numerical	0	8
X_{16}	Age	Numerical	21	104
X_{17}	Policy duration	Numerical	0.5	47
X_{18}	Construction year	Categorical (16)	Before 1900	2020-2024
X_{19}	Firewood heater	Binary	-	-

The dashes in Table 6 represent that there is no minimum/maximum for the variable. The values between the brackets in the column of 'type of variable' indicate how many categories are used to describe the categorical variables.

Several risk indicators seem to be numerical, but are reported as categorical. Examples of these risk indicators are Urbanisation and Own Risk. These indicators are measured numerical but for the ease of use they are reported as a categorical variable, for example, the urbanisation is divided in groups as *<5.000 residents, 5.000 up to 10.000 residents,* etc.

Several drawbacks arise when using the above mentioned risk factors. The Property Value Assessment (WOZ) is not up to date for every policyholder, this means that a policyholder can be held back in the wrong group for their WOZ-value. As discussed in section 4.1.1, the WOZ-values is averaging a yearly increase, so if the data is not updated regularly these values are not representative for their current WOZ-value anymore, meaning that the premium that they should pay on their correct risk profile could differ from the premium that they pay currently.

Another drawback arises with the risk factor Gender, as many households have a combined insurance, meaning that not only the male or female is responsible for a claim, but the claim could be incurred because of anyone that lives in the household. So, the kids, the partner or the policyholders could be responsible for the damage, which leaves the risk factor gender with a lot of uncertainties. It is not fair to charge policyholders more premium when the policyholder is male/female, but he/she lives in a household with both males and females. The only situation which is interesting to study is when policyholders live by themselves, how the claim behaviour of males and females differ. Although this is interesting information for USL to study, for this research we stick to the sample group of all (claim) policyholders of the home-content insurance.

4.4. Correlation matrix

In this study, a correlation matrix is employed as an addition to the preliminary analysis to explore the interrelationships among variables in the dataset. The correlation matrix allowed to identify and quantify the strength and direction of linear associations between different pairs of variables. By examining the matrix, it is possible to gain insights into the data structure, potential multicollinearity, and possible patterns within the dataset.

The numerical risk factors and the categorical risk factors that could be transformed to numerical are used in this correlation matrix. This correlation matrix indicates the Spearman correlation between the risk indicators x_1 = Own risk, x_7 = Living space, x_8 = Property value assessment (WOZ), x_9 = Insured amount precious, x_{10} = Insured amount jewellery, x_{15} = Claim count, $x_{16=}$ Age and x_{17} = Policy duration. When risk factors are highly correlated, this could lead to biases such as unstable coefficients or inflated standard errors. What should be noted is that a measure of linear correlation as shown in this correlation matrix does not capture complex interactions or causal relationships.

Table 7 Correlation matrix

VARIABLES	X_1	X_7	X_8	X_9	X_{10}	X_{15}	X_{16}	X_{17}
X_1	1							
X_7	-0.005	1						
X_8	0.007	0.734	1					
X_9	0.080	0.088	0.093	1				
X_{10}	0.035	0.038	0.049	0.151	1			
X_{15}	-0.046	0.057	0.039	-0.0172	0.006	1		
X_{16}	0.051	0.143	0.178	0.088	0.042	-0.148	1	
X_{17}	0.137	0.183	0.175	0.104	0.014	0.060	0.451	1

Table 7 presents a correlation matrix of several indicators in the dataset. The correlations are indicated by numerical values ranging from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation.

Firstly, examining the diagonal elements, we observe that each risk indicators correlation with itself is denoted as 1, this is because this observation indicates the correlation with itself.

Analysing the off-diagonal elements, we find various interesting patterns. For instance, the variable "Own risk" (x1) indicate relatively weak correlations with the other risk indicators. Conversely, "Living space" (x7) and "Property value assessment" (x8) display a relatively strong positive correlation (0.734), indicating that these two variables tend to increase together. This is easy explainable as houses with larger living spaces and all remaining indicators constant will have a higher property value assessment.

Moreover, the risk indicators "Claim count" (x15), "Age" (x16), and "Policy duration" (x17), seem to have relatively weak correlations with other indicators, except for a moderate positive correlation (0.451) between "Policy duration" (x17) and "Age" (x16). This correlation is also explainable as people that are older will have a larger span of a possible home-content insurance meaning a larger probability of the policy duration being longer.

Overall, this correlation matrix provides insights into the relationships between various indicators, helping to understand potential interdependencies and supporting further analyses and modelling decisions in the next part of this study.

Chapter 5 Results

This section will comprise the results from the modelling part of this study. Both the benchmark model and the adjusted model are fitted to predict the total aggregate loss. Both models will be derived by joining the marginal GLMs for the claim frequency and claim severity. The dependency analysis is performed so that the prediction accuracy and the goodness-of-fit of both models are compared.

5.1. GLM modelling in R

The process of selecting the risk factors to include in the GLM models already started in the previous chapter. In the previous chapter some, on first glance, interesting risk factors were shown to be an ineffective predictor of the total aggregate loss, for example gender.

The data that is used for both the claim frequency and claim severity analysis is a subset of all policyholders' information with a claim frequency greater than zero. This subset is chosen for several reasons. Firstly, it allows to focus specifically on policyholders who have experienced at least one claim, which can provide valuable insights into the factors influencing claim occurrence and severity. By excluding policyholders with no claims, the noise and potential bias introduced by those individuals who have not experienced any claims will be eliminated, which aligns with the focus of this study. This targeted approach helps to ensure that the analysis captures the patterns and relationships that are relevant to the occurrence and severity of claims.

Different GLMs will be used to seek significant risk factors. With the use of deviance, AIC, BIC, the standard errors and the significance level of the risk factors, nestled models can be compared to be able to state the most parsimonious model. Even though significance is a useful criterion to choose the best predicting risk factors, also theoretical and practical considerations should be taken into account. Sometimes the data fails to reflect a common phenomenon meaning that a possible predictor is excluded from the model which could have been good to include. The considerations have been made with USL, as they are the most familiar with the data and the claim behaviour of their policyholders.

The modelling is performed with the software of R, where an existing function for *glm* models is incorporated. When using this function, the risk factors, the dataset and the 'family' need to be chosen. The family is, as stated earlier, assumed to be Poisson distributed for the claim frequency and Gamma distributed for the claim severity. The GLM models are trained on a dataset that contains 70% of the total observations, the remaining 30% will be used as a 'hold-out' dataset where both the benchmark and the adjusted model can be compared on their prediction accuracy.

To be able to capture the behaviour of policyholders which reflects the exposure to risk per policyholder, an exposure offset needs to be included. This will reflect the amount of time that the policy was in force during the year (W. Lee et al., 2019).This is of importance because it could be the case that policyholders have not been insured for the

full year. Therefore, in the claim frequency analysis, an explanation for this exposure term is elaborated.

The last modification to the dataset is done by removing duplicates in the data. The data of the claims include an observation for every claim, while all variables (except for the severity) are the same for a policyholder that filed in multiple claims. For the Claim frequency analysis, these duplicates do not add value and will result in bias in the results. The severity variable of the claims is not of interest for the claim frequency GLM. Therefore, for the claim frequency GLM analysis, these duplicates have been dropped.

To improve the readability of the report, there are no formulas of all GLMs included in the sections that explain the GLM models. In these sections, the covariates that are included are described, and the formulas can be found in the Appendix.

5.2. Claim frequency analysis

5.2.1. Exposure offset term

An offset term is simply an additional model variable, whose coefficient is constrained to be one (Yan et al., 2009). The offset term corrects the claim frequency GLM so that it reflects what portion of the year the individual was insured for. When the exposure term equals one, then the policyholder has been insured during the full year. Because of the already assumed Poisson distribution for the claim frequency, as the exposure increases, the expected claim count will increase proportionally (Schulz, 2013). By including an offset term for insurance duration, it can explicitly be accounted for the varying exposure times of the policyholders. This adjustment allows to estimate the claim frequency rate, which is the number of claims per unit of exposure.

Including the offset term with a coefficient of one effectively scales the claim frequency rate by the insurance duration (Yan et al., 2009). The estimated coefficients for the other predictors in the claim frequency model can be interpreted as the change in the claim frequency rate for a unit change in the predictor, holding the exposure time constant.

The first step to creating the offset term is to create a new variable in the R data frame which reflects the amount of years that the policyholder has been insured during the claim data period (Claim exposure). This is different from the policy duration variable that is explained in chapter 4, because the claim exposure variable starts in 2015. The data for all the claims begins in 2015, so the claim exposure variable returns approximately 8.5 for all policyholders that are insured on or before January the first of 2015. For all policyholders that have been insured since after January 2015, the exposure variable is the same as the policy duration variable.

Finally, the offset term for the claim exposure variable is chosen to be the logarithm of the exposure variable. This is chosen because of several reasons. Firstly, it captures the assumed proportional relationship between the expected claim frequency and the exposure, as the logarithm of the exposure helps to represent this proportionality. Secondly, for count data with a Poisson distribution, the logarithm of the exposure aligns with the canonical link function described in the previous chapter (log link). This will

ensure a linear relationship between the linear predictor and the logarithm of the (expected) claim frequency.

5.2.2. Claim frequency GLMs

In this section several (nestled) GLMs are formulated, in the beginning of the section the formulation and the results of the models are stated. The last part of this section will provide a comparison between the models and will ultimately choose one GLM to be the final marginal GLM for the claim frequency. Table 8 shows all the covariates that are used in the GLMs for the claim frequency. The covariates used in the first model can be seen in Table 8. The other GLMs for the claim frequency are nestled within the first (full) model, therefore only the excluded covariates in comparison with the first model are mentioned.

Table 8 List of covariates for the CF GLM

Covariate	**Included (X = yes)**
Claim Exposure	X
Own Risk	X
Gender	
City	X
Payment term	X
House type	X
Urbanisation	X
Living space	X
Property Value Assessment (WOZ)	X
Insured amount precious	X
Insured amount jewellery	
Ownership status	X
Postcode risk class	X
Thatched roof	X
Extra living space	X
Claim count	
Age	X
Policy duration	X
Construction year	X
Firewood heater	

In this model almost all risk factors are included to capture the effects of the model under circumstances where most risk factors are included. The risk factors not used as a covariate are neglected because of the relevance for this type of modelling. Gender for example, can be an interesting covariate. However, as stated earlier, there is a lot of bias included with the covariate gender. Policies often represent the complete household, therefore it cannot be used to reflect one gender. The residual deviance with the covariates as stated in Table 8 is found to be 3192.6, with an AIC of 25614 and a BIC of 29332. As some risk factors are categorical, the overall significance of the risk factors is determined with the use of the ANOVA function in R, which allows to perform a hypothesis test using the chi-square test. This test compares the model with the categorical variable to a reduced model without the variable. ANOVA uses scaled

deviance to compare models. The p-value associated with the test will indicate the overall significance of the variable. The significance of the other variables is computed by evaluating the summary of the models. Therefore, when stating about a 'significance test' this is the combination of interpreting the summary results and performing an ANOVA test.

The significance test indicated that some variables are not significant, together with USL is chosen to exclude the variable *Urbanisation* from the next models to compare whether these models are in fact an effective simplification and make the model parsimonious. The second GLM excluded the variable *Urbanisation.* The residual deviance is found to be 3196.2 with an AIC of 25606 and a BIC of 29280. After performing a significance test again, the risk factors *Living space* and *Extra living space* is found to be insignificant.

The third GLM excluded the variables *Living Space and Extra living space.* The residual deviance is found to be 3200.5 with an AIC of 25586 and a BIC of 29175. The variable *City* is said to be insignificant, also the possibility of correlation with the variable *Postalcode Risk Class* supported the reasoning to exclude the variable *City.* As this variable has 453 degrees of freedom, the residual deviance is expected to increase, therefore the focus will be on the AIC and the BIC which also take the simplification into account.

The following GLM is performed without the variable City. The residual deviance is found to be 3392.1 with an AIC of 24878 and a BIC of 25243. The last variables to be excluded from the model are *Insurance amount home-content and Thatched roof.* The residual deviance is found to be 3410.6 with an AIC of 25015 and a BIC of 25294.

5.2.3. Final claim frequency model

When comparing the models, it's important to consider the goodness-of-fit measures and the principle of parsimony. In Table 9- it can be seen that model 1 has the lowest deviance value, indicating a better fit to the data. However, it also has the highest AIC and BIC, suggesting a potential penalty for including more variables.

Model 3, on the other hand, exhibits similar deviance and AIC values to model 2 but achieves a lower BIC value. This suggests that model 3 is a more parsimonious choice compared to model 2, as it achieves comparable goodness of fit while using fewer variables. Model 4 excludes the variable *City* with reduces the degrees of freedom a lot. The result of this exclusion can be seen in the increase in the deviance. However, the AIC and the BIC suggest that model 4 is indeed a simplification and it makes the model more parsimonious.

Considering both the goodness-of-fit measures and the principle of parsimony, model 4 appears to be a favourable choice. It retains the significant variables while excluding non-significant ones, resulting in a simpler and more interpretable model.

Table 9 Comparison claim frequency models.

Model	Variables Included	Residual Deviance	AIC	BIC

1	*Full Model*	3193	25614	29332
2	*Excluding Urbanisation*	3196	25606	29280
3	*Excluding Living Space*	3201	25586	29175
4*	*Excluding City*	3392	24878	25243
5	*Excluding Insurance amount & Thatched roof*	3410	25015	25294

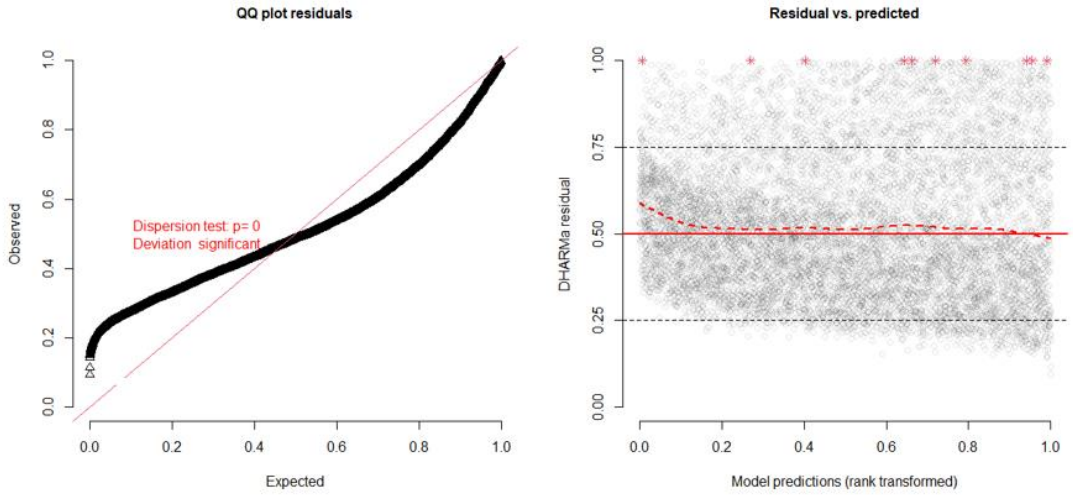


Figure 9 QQ plot & residual vs predicted plot for the CF GLM

In a QQ plot, the expected quantiles are plotted on the x-axis, while the observed quantiles are plotted on the y-axis. In the case of this GLM (a GLM with a Poisson distribution and a log link), the expected quantiles are generated based on the assumption that the residuals should follow a Poisson distribution. The context regarding the dispersion test in the QQ plot can be neglected as this report will not use this method for describing the models.

For a well-fitted model, the points on the QQ plot should generally fall along a straight line, indicating that the observed quantiles align with the expected quantiles. In it is clear to see that the straight line is not followed strictly, indicating that the Poisson distribution might not be the best fit. This is evidence of possible under dispersion in the model (Sáez-Castillo et al., 2022). In Figure 9, the QQ plot shows curving behaviour around the straight line, this could also indicate a misspecification of the model. A possible misspecification is that the relationship between the predictors and the response is not sufficiently captured by the log link function, or that additional covariates should be included in the model. Since there is a relatively large amount of policies with zero claim count, a zero-inflated Poisson distribution assumption could have potentially provided a better fit.

The second plot in Figure 9 (fitted versus residual plot) is used to examine the relationship between the predicted values and the corresponding residuals from the

GLM. It helps to identify patterns in the residuals, this can help to create an assessment of the models' assumptions and performance.

The plot should be interpreted that when the residuals are randomly scattered around zero, then it suggests that the model meets the assumptions of the constant variance. In this case, the scatter is not around zero, but around the line y = 0.50. There could be several reasons for this.

The first reason could be that the bias in the residuals may indicate that the model is not capturing all the underlying patterns or relationships in the data. This could again be due to missing predictors. The second potential reason could be heteroscedasticity, which means that the variance of the residuals is not constant across the range of the predicted values. Also, outliers in the data can cause the scatter to be offset from y = 0. If there are extreme values in the response variable that are not accounted for by the model, they can contribute to the bias observed in the residuals.

5.3. Benchmark claim severity analysis

The next step is to formulate the GLM for the claim severity. This section is divided into two subsections, one for the benchmark model and one for the adjusted model. The model for the claim severity is different as the claim frequency will be included as a covariate in the adjusted model. Both GLMs are assumed to be Gamma distributed with the log link as the canonical link function.

5.3.1. Claim severity GLMs

Same as for the claim frequency GLM, we start with almost all risk factors included to see the effect on the claim severity. The *Claim exposure* variable does not have to be included anymore, as the focus (response) is only on the severity when a claim is incurred, not the frequency of it happening. Again, Table 10 is included to indicate all the covariates used in the first model. The other GLMs derived are nestled models from the first model, therefore only the excluded covariates will be mentioned.

Table 10 List of covariates for the independent claim severity GLM

Covariate	**Included (X = yes)**
Claim Exposure	
Own Risk	X
Gender	
City	X
Payment term	X
House type	X
Urbanisation	X
Living space	X
Property Value Assessment (WOZ)	X
Insured amount precious	X
Insured amount jewellery	
Ownership status	X
Postcode risk class	X
Thatched roof	X
Extra living space	X

Claim count	
Age	X
Policy duration	X
Construction year	X
Firewood heater	

The first GLM results in a residual deviance of 9078 and an AIX of 183329, which is relatively high compared to the claim frequency GLM. Also, the BIC reflects these relatively high values with a result of 187337. Possible reasoning could be over/underfitting, also a correlation between the covariates could be a reason for this. Possible correlation could arise with the variables *City* and *Postalcode Risk Class.* Therefore, to reduce the amount of degrees of freedom (465), the variable *City* is excluded from the GLM similar to the GLM of the claim frequency. This simplification increases the residual deviance to 9469, but R states that this model provides a better fit as the *glm* algorithm had difficulties with converging for the previous model and with the model without *City* it did converge. The AIC and BIC resulted in 183057 and 183570 respectively. In the third model for the claim frequency, the variables *Payment term and Own risk* are excluded. The residual deviance of the claim severity model 3 resulted in 9472. The AIC and the BIC respectively are 183054 and 183537. When running a significance test again, most variables are significant, the variable *Type of house* is not significant and may be excluded from the GLM as USL stated. This GLM has a residual deviance of 9499 and an AIC and BIC of 183790 and 184228 respectively.

5.3.2. Final claim severity model (benchmark)

The full GLM had relatively high AIC, and BIC values, suggesting potential issues with overfitting. To address these concerns, we excluded the variable *City* from the model, this gave the second model.

The second model demonstrated improved model fit compared to the first model, with regards to the AIC and BIC values. Although the residual deviance slightly increased, this trade-off was considered acceptable because of the concept of parsimonious models. However, further analysis revealed that several risk factors were not statistically significant. Ultimately, the third GLM emerged as the preferred model after excluding *Payment term* and *Own risk*, leading to a decrease in residual deviance compared to models 2 and 3, and also lower AIC and BIC values compared to the other models. This simplification made the model more parsimonious which is desired. The third model indicated a favourable balance between model fit and simplicity.

Table 11 Comparison claim severity models in the independent setting.

Model	Variables Included	Residual Deviance	AIC	BIC
1	*Full model*	9078	183329	187337
2	*Excluding City*	9469	183057	183570
3*	*Excluding Payment term and Own risk*	9472	183054	183537
4	*Excluding Type of house*	9499	183790	184228

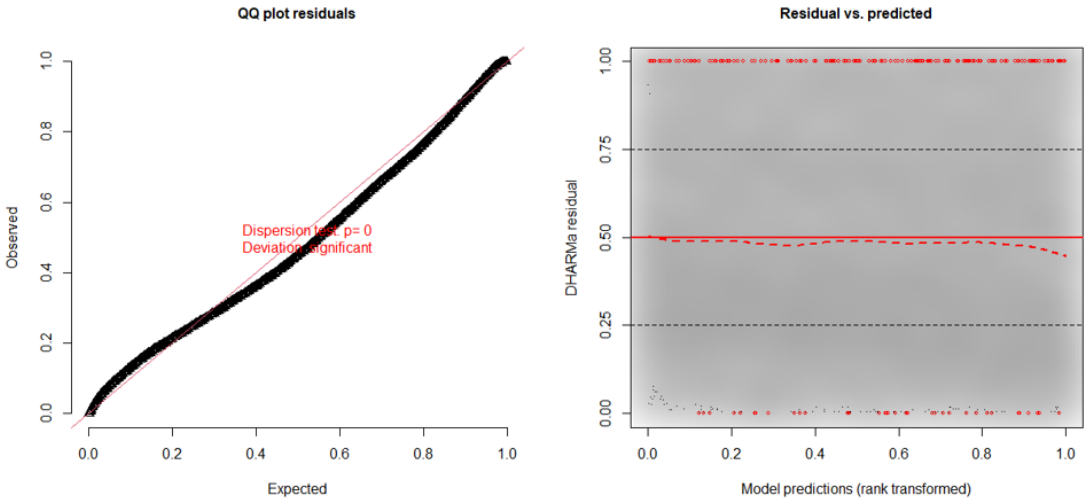


Figure 10 QQ plot & residual vs predicted plot for the CS GLM

The QQ plot in Figure 10 aligns more closely to the straight line in comparison with the QQ plot for the CF GLM. The straight line suggests that the residuals of the model indicate a distribution similar to the Gamma distribution assumed in this study. This alignment indicates a decent fit between the model's assumptions and the observed data. The closer the points to the line, the better that the residuals follow the expected quantiles, supporting the validity of the underlying assumptions. This alignment provides confidence that the model captures the patterns and variability present in the data. The Residuals against the predicted plot show similar behaviour as for the CF GLM, where the scatter is around the line y = 0.50.

5.4. Adjusted claim severity analysis

The method for determining the claim severity GLM for the adjusted model is comparable to the previous section. In the case of the dependent setting, the *Claim count* variable is included as a covariate in the GLM. This will create a GLM for the severity depending on the frequency of claims. The aim of this research is to compare the total aggregate models, in the dependent and independent setting. The total aggregate loss models are described later on in this chapter. Again, the variable *City* is excluded to account for the possible intercorrelation. The risk factors used as covariates in the first GLM can be seen in Table 12. The simplifications of these models are described by mentioning the excluded covariates.

Table 12 List of covariates for the dependent claim severity GLM

Covariate	**Included (X = yes)**
Claim Exposure	
Own Risk	X
Gender	
City	X
Payment term	X
House type	X

Urbanisation	X
Living space	X
Property Value Assessment (WOZ)	X
Insured amount precious	X
Insured amount jewellery	
Ownership status	X
Postcode risk class	X
Thatched roof	X
Extra living space	X
Claim count	
Age	X
Policy duration	X
Construction year	X
Firewood heater	

This GLM ended with a residual deviance of 9431, which is lower than the previous claim severity GLM with the exclusion of *City*. The AIC and the BIC resulted in 183680 and 184217. After the significance test the variables *Payment term, Own risk* and *Type of house* are not significant and thus are not expected to have significant effect on the claim severity.

In the second model the same simplification as in the previous section is made, by excluding the variables *Payment term and Own risk*. This residual deviance of this GLM is 9436, and the AIC and the BIC resulted in 183682 and 184188. After the significance test the variable *Type of house* is still not significant and thus not expected to have significant effect on the claim severity. The third model has a residual deviance of 9456 and AIC and BIC of 183721and 184167 respectively.

5.4.1. Final claim severity model (adjusted)

While model 1 achieves the lowest residual deviance of 9306, indicating a better fit to the data, the differences in residual deviances between the models are relatively small. When considering the information criteria, model 2 stands out as the preferred option. It has a slightly higher residual deviance of 9313 compared to model 1 and a comparable AIC, but its BIC value is lower at 182788. model 3 on the other hand, has the highest deviance and AIC values among the three models.

Considering the comparable deviance and AIC values and the lower BIC value of model 2, it emerges as the preferred option. This model strikes the balance between simplicity and performance, as it excludes non-significant variables while maintaining a reasonable fit to the data.

Table 13 Comparison claim severity models in the dependent setting.

Model	Variables Included	Residual Deviance	AIC	BIC
1	*Full model (excluding City)*	9306	182270	182813
2*	*Excluding Payment term & Own risk*	9313	182275	182788

3	*Excluding Type of house*	9330	182325	182777

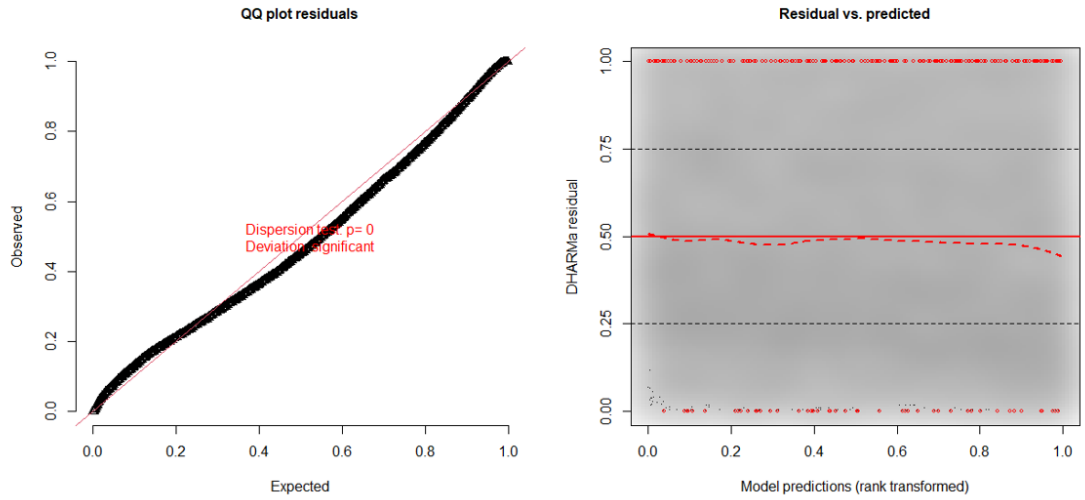


Figure 11 QQ plot & residual vs predicted plot for the modified CS GLM

Figure 11 shows that for both the QQ plot and the residuals against the predicted plot, the behaviour is similar in the dependent and independent setting for the claim severity GLM. So, no conclusions can be drawn based on Figure 11.

5.5. Total aggregate loss models

This section will show the derivations of the total aggregate loss models. The total aggregate loss models are derived with the multiplication of the marginal GLMs. The first section will describe the benchmark total aggregate loss model, this is the model that is currently used by USL. The second section will describe the derivation of the adjusted total aggregate loss model, this is the model that is able to capture the effects of the possible dependency between the CF and the CS.

5.5.1. Benchmark model

The benchmark model is the 'simple' multiplication of the claim frequency GLM and the claim severity GLM. The marginal frequency and severity GLMs are described in the previous sections and can be written as

$$\boldsymbol{E(N|x) = g_1^{-1}(x^T\alpha) = v} = \exp(\alpha_0 + \alpha_1 * Own\ risk + \alpha_2 Payment\ term + \alpha_3 Type\ of\ house + \alpha_4 WOZ + \alpha_5 Construction\ Year + \alpha_6 Ownership\ status + \alpha_7 Postalcode\ risk\ class + \alpha_8 Thatched\ roof + \alpha_9 Extra\ living\ space + \alpha_{10} Insurance\ amount\ homecontent + \alpha_{11} Age + \alpha_{12} Policy\ duration)$$

(12)

$$\boldsymbol{E(Y|x) = g_2^{-1}(x^T\beta) = \mu} = \exp(\beta_0 + \beta_1 Type\ of\ house + \beta_2 Urbanisation + \beta_3 Living\ space + \beta_4 WOZ + \beta_5 Construction\ Year + \beta_6 Ownership\ status + \beta_7 Postalcode\ risk\ class + \beta_8 Thatched\ roof + \beta_9 Extra\ living\ space + \beta_{10} Insurance\ amount\ homecontent + \beta_{11} Age + \beta_{12} Policy\ duration)$$

(13)

The total aggregate model in the independent setting can then be formulated as

$$E[S|x] = E(Y|x) * E(N|x) = v\mu$$

5.5.2. Adjusted model

The adjusted model is comparable to the benchmark model, however, in the dependent setting, the marginal GLM for the CS is be adjusted. Also, the dependency correction term as described in 3.4.1 will be included for the total aggregate loss model. The same marginal GLM for the CF is used in both the dependent as the independent setting as there is no difference between these settings when predicting the claim frequency. The marginal CS GLM is different and can be written as

$$\boldsymbol{E}(\bar{\boldsymbol{Y}}|\boldsymbol{x},\boldsymbol{N}) = \boldsymbol{g_2}^{-1}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta} + \boldsymbol{\theta N}) = \mu^A = \exp(\beta_0 + \beta_1 Type\ of\ house + \beta_2 Urbanisation + \beta_3 Living\ space + \beta_4 WOZ + \beta_5 Construction\ Year + \beta_6 Ownership\ status + \beta_7 Postalcode\ risk\ class + \beta_8 Thatched\ roof + \beta_9 Extra\ living\ space + \beta_{10} Insurance\ amount\ homecontent + \beta_{11} Age + \beta_{12} Policy\ duration)$$

(14)

The total aggregate model in the dependent setting is formulated with the addition of the correction term as

$$E(S|x) = v\mu^A \exp\{v(e^{\theta} - 1) + \theta\}$$

5.6. Comparison and dependence analysis

For the train dataset, which is a proportion of all the data (70%) the benchmark and the adjusted models have been trained. Now that the models are trained, the test dataset comes into play. The test dataset is the remaining 30% of the data that has not been included into determining the models. This will enable to test the prediction ability of both models and compare it to real life observations. In R this is done by obtaining the predictions for the CF GLM and the predictions of the (adjusted) CS GLM. Then, with the total aggregate loss models shown in the previous section, predictions for the total aggregate loss can be made. Both result in a vector of predictions, the vectors of the benchmark model and the adjusted model can be compared to the real total aggregate loss. This section will first highlight the comparison between the marginal GLM of the claim severity in the independent and the dependent setting. After the comparison of the marginal GLMs, the total aggregate loss models will be compared.

5.6.1. Comparison of the marginal GLMs

To compare the marginal GLMs of the claim severity Table 14 is included. This table provides the values of different error measures of the GLMs indicating a better fit when the GLM is in the dependent setting.

Table 14 Error measures of the claim severity GLMs

Measure	Independent	Dependent
Mean Squared Error	7.778880e+04	7.758447e+04
Root Mean Squared Error	2.789064e+02	2.785399e+02
Mean Absolute Error	1.687073e+02	1.682158e+02
Mean Absolute Percentage Error	6.386475e-01	6.366326e-01
Mean Absolute Scaled Error	5.148090e+00	4.858113e+00
Mean Percentage Error	1.144274e+02	1.133235e+02

Symmetrically V-shaped Root Mean Squared Error	1.045152e+02	1.043779e+02
Mean Squared Logarithmic Error	8.104678e-01	8.059220e-01

Table 14 suggests that the inclusion of *Claim Count* as a covariate improves the GLM, as all the error measures are slightly better. Overall, these findings indicate that the adjusted model may have a slight edge in terms of accuracy compared to the benchmark model.

5.6.2. Comparison between the benchmark and the adjusted model

To get insights in the prediction ability of the total aggregate loss models, the same error measures have been determined. Table 15 shows the values of the error terms for both models.

Table 15 Error measures of the Benchmark and the Adjusted model

Measure	Benchmark	Adjusted
Mean Squared Error	1.479069e+05	1.540987e+05
Root Mean Squared Error	3.845866e+02	3.925540e+02
Mean Absolute Error	2.539984e+02	2.519499e+02
Mean Absolute Percentage Error	6.601909e-01	7.247229e-01
Mean Absolute Scaled Error	4.598486e+00	4.874384e+00
Mean Percentage Error	1.260373e+02	1.097466e+02
Symmetrically V-shaped Root Mean Squared Error	9.904840e+01	1.011004e+02
Mean Squared Logarithmic Error	9.101946e-01	8.881002e-01

This table might indicate a slight preference for the benchmark model, but this does not tell the complete story. The dataset contains multiple outliers, which can result in greater error terms. Especially a measure like Mean Squared Error punishes a model for having outliers as it squares the errors. The Mean Absolute Error is better for determining the error term in comparison to the Mean Squared Error when the dataset contains outliers. So, this table is a first step in comparing the models, but it cannot be used to pick a decisive model. To be able to see how the models perform and behave, several visualisations are made. The remainder of this section will highlight these visualisations.

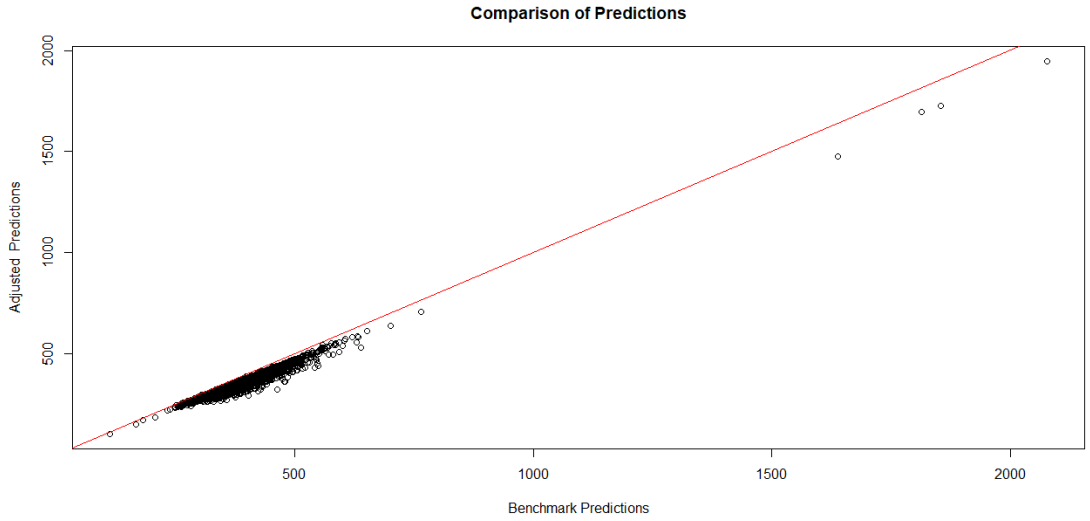


Figure 12 Comparison of predictors between the Adjusted and the Benchmark model

In Figure 12 the predictions of the adjusted model and the benchmark model are compared. The red line represents the diagonal of y = x, this helps to see the differences more clearly. Noticeable is that most observations are below the red line, indicating that the predictions of the benchmark model are higher than the predictions of the adjusted model. This graph does not help us determine which model is preferred, but is does create awareness that if the adjusted model is preferred, the predictions could return be lower.

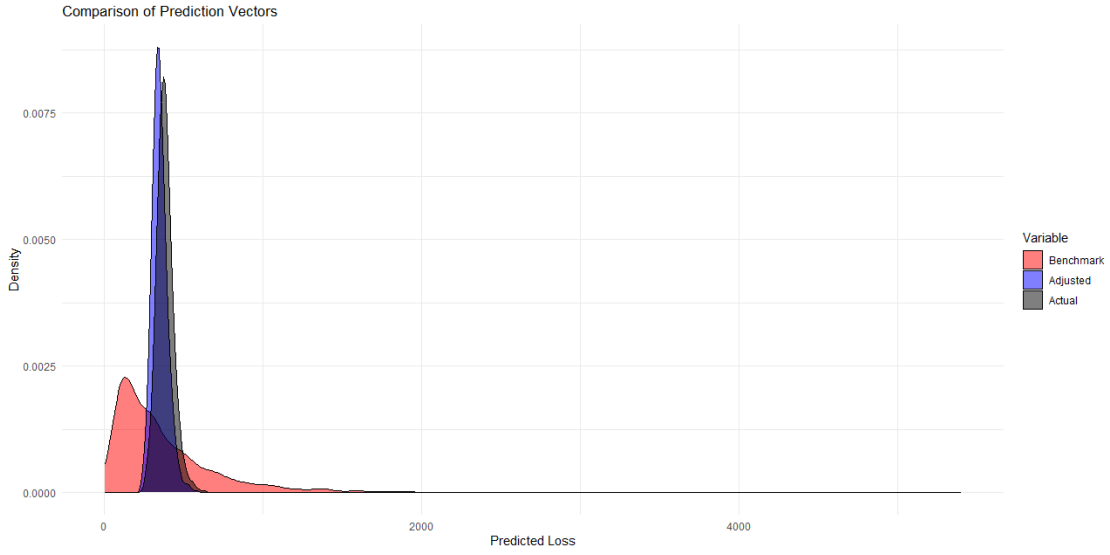


Figure 13 Density plots of the actual total aggregate loss values and the predicted values

As stated earlier, the outliers make it hard to draw conclusions based on measures only. One way of comparing the predictions to the actual values is with a density plot. A density plot simplifies the distribution of data into a smooth curve, allowing to understand how values are spread out. It helps to visualize the shape of the distribution and compare multiple distributions in a clear and concise manner. In Figure 13 such a density plot can be seen, this density plot displays the density distributions of the predictions of the two models and the actual values. It is noticeable that the adjusted

model is able to capture the distribution of the actual values better, resulting in the most overlap. This density plot shows how outliers can mislead the error values as seen in Table 15. Even though some error measures seem more leaning towards the benchmark model being the better predictor, looking at Figure 13, the adjusted model manages to capture the 'density' better.

The density plot makes some assumptions that could make the interpretation of the figure more complicated. The first assumption being continuity, this assumption is satisfied as the total aggregate loss predictions are continuous. The second assumption is that the underlying distributed data is smooth, without any sudden spikes or irregularities.

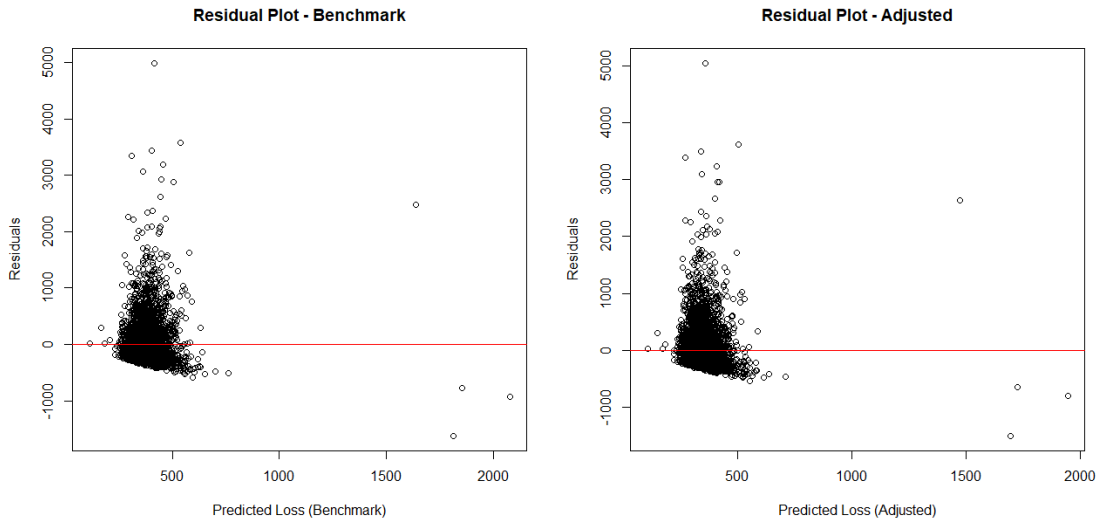


Figure 14 Residual plot of the Benchmark and the Adjusted model

A residual plot, as shown in Figure 14, provides insights into the performance and accuracy of the two models attempting to predict the total aggregate loss. When comparing two models using a residual plot, we can examine the distribution and patterns of the residuals. The residuals are the differences between the predicted values and the actual observed values.

In general, a well-fitted model should have residuals that scatter randomly and do not indicate any patterns or trends in the plot. Even though both residual plots are different, they do show similar characteristics, this suggests that they are capturing the underlying patterns and relationships of the data to a similar level.

When examining the two residual plots , they are found to be similar. It becomes challenging to conclude a preference between them. It is difficult due to the absence of distinguishing features and characteristics within the plots.

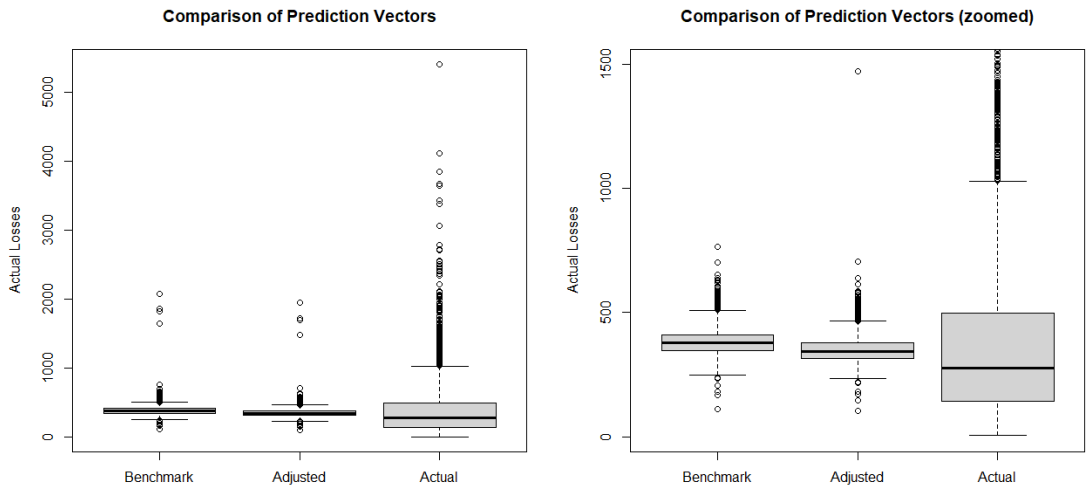


Figure 15 Box plots of the actual values and the predicted values

Figure 15 shows the comparison of three box plots, where one represents the actual values and the other two show the predicted values of the benchmark and adjusted models. While the two predicted box plots indicate similarities, it is evident that the median of the adjusted model is closer to the median of the actual values, which might indicate a higher level of accuracy in the adjusted model.

However, it is noticeable that both predicted models fail to capture the outliers present in the real data. These outliers, representing extreme values or unusual observations, are not accounted for in the predictions. This suggests that the models may struggle to capture the full range of variability of the data, especially in the tails of the distribution.

Despite this possible limitation, the two models share a similar pattern and distribution. While the median of the adjusted model is slightly closer to the actual values, the difference is relatively small. Therefore, the choice between these two models based solely on the box plots is not possible.

5.7. In sample comparison

In-sample testing is used in the process of measuring the models performance and comparing the benchmark and the adjusted model. In-sample testing involves the assessment of the benchmark and the adjusted model using the same dataset that was employed for the training of these models. While it may sound questionable, the necessity of in-sample testing in the model evaluation process remains important.

The primary motivation behind in-sample testing lies in its ability to provide a foundational understanding of a model's performance within the context of the training data. It serves as an ‘initial model’, offering insights into how well the model aligns with the data it was exposed to during the training phase.

However, it is essential to acknowledge the limitations of relying solely on in-sample testing for model selection. Models that excel in minimizing error measures, such as Mean Absolute Percentage Error (MAPE) or Root Mean Squared Error (RMSE), within the training data may not necessarily demonstrate similar excellence when applied to

new, unseen data (out of sample). This phenomenon is known as overfitting, where a model may become overly tailored to the training dataset, thus failing to generalize well to real-world scenarios.

Table 16 Error measures of the Benchmark and the Adjusted model (in-sample)

Measure	Benchmark	Adjusted
Mean Squared Error	1.589477e+05	1.650682e+05
Root Mean Squared Error	3.986825e+02	4.062859e+02
Mean Absolute Error	2.575973e+02	2.541201e+02
Mean Absolute Percentage Error	6.636587e-01	7.264306e-01
Mean Absolute Scaled Error	4.581965e+00	4.857243e+00
Mean Percentage Error	3.540612e+13	3.301849e+13
Symmetrically V-shaped Root Mean Squared Error	1.034375e+02	1.054102e+02
Mean Squared Logarithmic Error	9.660888e-01	9.376537e-01

Table 16 shows the (in-sample) error measures of the benchmark and the adjusted model. To check for possible overfitting, it is necessary to compare these measures to Table 15, where the error measures for the models out-of-sample are determined. Both the RMSE and the MAPE measures are lower in the out-of-sample models indicating that there is no case of overfitting with the benchmark and the adjusted models.

Chapter 6 Conclusion

This chapter is divided into three sections, first in the discussion will draw conclusions based on the modelling results in Chapter 5. Then in the section named contributions, it will be stated how this study contributed to the literature. The last part of this chapter will discuss the limitations of this study and how this study can be expanded with future research.

6.1. Discussion

The main focus of this study is to compare the currently used model that predicts the total aggregate loss on individual level in the independent setting, to the one in the dependent setting. The goal was not necessarily to find the best possible model to describe the data but rather to compare the effects of extending the independent model to the dependent setting. The process of choosing the risk factors and formulating the models is done in the same way, which enables us to compare the models effectively.

This study proposes a model and method for allowing the dependence between claim frequency and claim severity. In this method, the benchmark model is nested in the adjusted model, which makes the independence method a 'special case' or a simplification of the adjusted model. The adjusted model is thus an expansion of the benchmark model by integrating the claim frequency as a covariate.

In Chapter 5 the results of the analysis are described and explained. Firstly, the comparison between the marginal claim severity models is of interest as the adjusted marginal GLM for the claim severity is formulated in the dependent setting. The results of this model indicate that the *Claim count* is a significant variable at the 0,1% level, indicating that this variable is indeed helping to explain the behaviour of the claim severity. Together with the preliminary analysis of the correlation tests, it is concluded that the first hypothesis holds. This conclusion is drawn due to the fact that the *Claim count* is a significant covariate and the Spearman test indicate a strong (negative) correlation between the CF and the CS. The coefficient that R provided for the covariate *Claim count* is -0.0478, all other coefficients and standard error values of the models can be found in Appendix 6, 7 and 8. The coefficient of the *Claim count* supports the results of the preliminary analysis of the Spearman correlation test by it being negative. This implies that when the claim frequency increases, the severity of these claims tends to decrease.

The CS GLM in the dependent setting manages to outperform the GLM in the independent setting in all three measures (deviance, AIC and BIC) used to choose the most parsimonious model. This could suggest that including the claim frequency in the form of a covariate will indeed improve the goodness of fit of the models.

So, the goodness of fit of the marginal CS GLM is improved, when fitting the model in the dependent setting. The error measures shown in Table 14 also suggests that the GLM with *Claim Count* as covariate has better prediction accuracy than the model in the independent setting. All measures used in this thesis indicate the preference for the GLM that allowed for dependence.

The first conclusion that can be drawn from comparing the adjusted and the benchmark model is that following the error measures (Table 15), there is no immediate preference between the models. The benchmark model seems to score a bit better, however, the outliers could explain these results. Therefore, visualisations should better explain the prediction ability of the models. The residuals (Figure 14) do not indicate any preference between the models, same can be stated for the box plot (Figure 15). However, the density plot (Figure 13) of the models is indicating a preference between the models. The Adjusted model seems to capture the variation of the severities per policyholder more accurately than the benchmark model.

To reflect on the second hypotheses, the adjusted model did not decisively outperform the benchmark model regarding the goodness-of-fit measures. However, the mean absolute error is lower than in the benchmark model, and the visualisations suggest that the adjusted model has a better fit. Even though the improvements are small, these improvements can have a significant impact on the pure premium asked to the policyholders. To illustrate this, Figure 12 indicates that the adjusted model is averaging a lower predicted severity. When these results form as the underlying of the ratemaking process, this can have an economical effect. This holds because the premiums asked to policyholders can be lowered to cover for the risk profiles of the policyholders. If it is possible to lower the premiums, then a more competitive position in the insurance market could be obtained.

6.2. Contributions

This study provided a way of modelling the total aggregate loss of policyholders while allowing for dependence between the claim frequency and claim severity in the home-content insurance sector. This case study can be generalized for other insurers with the same coverage types and similar characteristics. The contribution to the literature can be summarised as a ‘deep-dive’ in the relationship between the CF and the CS and how the total aggregate loss can be predicted in a dependent setting. The framework used in this study is a relatively easy to follow method, which allows other insurers to try and use this framework for their own. Furthermore, this paper extended the framework of Schulz (2013) so that it is applicable to the home-content insurance sector, which makes the framework more valid and credible.

The preliminary correlation analysis provides insights to the relation between the CF and the CS, already suggesting a (negative) correlation between them. In the modelling part of this study, this correlation/relation is again noticeable. It is therefore of importance for insurers to understand this relation before setting premiums based on assumptions of independence.

6.3. Limitations & Future research

The study examining the relationship between claim frequency and claim severity and its impact on total loss modelling has provided valuable insights. However, certain limitations must be stated to ensure a comprehensive understanding of the research’s outcomes.

Firstly, the datasets used in this study were relatively limited in size and scope, because this study examined only one coverage of one insurer. Expanding the dataset to include a larger and more diverse sample group from different sources could enhance the robustness of the findings.

Secondly, this study used the conventional frequency-severity models to create a situation where the results could be implemented immediately. Exploring alternative methodologies, such as copula models, will broaden the analysis and could have been a good comparison of the prediction capabilities. Copulas can capture complex dependence structures between claim frequency and claim severity, potentially leading to more accurate total aggregate loss predictions.

Furthermore, the current research considered various risk factors, but further investigation into additional variables could provide deeper insights into their influence. For example, the risk factor *Gender* is excluded from the research for reasons mentioned earlier in Chapter 4. Creating possible interactions between risk factors could solve problems of not being able to include risk factors.

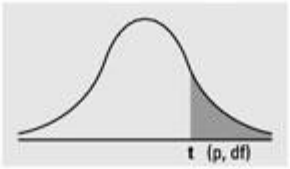
Lastly, the study used the standard distributions that were assumed by USL (Poisson and Gamma). Exploring alternative distribution models, such as negative Binomial or non-zero Poisson inflated models, could potentially improve the goodness-of-fit of the models, leading to more precise total aggregate loss predictions.

In conclusion, while this research has shed light on the relationship between claim frequency and claim severity and its implications for total aggregrate loss modelling, addressing these limitations in future studies will refine the accuracy and applicability of the models. Also, the sensitivity and the robustness of the study can then be improved.

Chapter 7 Appendix

A1. t-distribution table

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

A2. Correlation term for Poisson distribution

In this section the derivation of the correction term for the Poisson distribution is described. The theory presented is based on the study of (Garrido et al., 2016).

If N is a Poisson(λ), then for all $\theta \in R$,

$$M_N(\theta) = \exp\left(\lambda\left(e^{\theta} - 1\right)\right)$$

In this case, the expected severity can be formulated as

$$E(S|X) = u\lambda \exp\left(\lambda\left(e^{\theta} - 1\right) + \theta\right)$$

This represents the correction term used in this study, as long as the Poisson distribution is assumed. Further insights into the derivation of the correction term, can be found in the study of (Garrido et al., p215, 2016)

A3 Claim frequency GLMs

```
mod1 <- glm(ClaimCount ~ log(ClaimExposure) + `Own Risk` +
              City + `Payment Term` + `Type of house` + Urbanisation +
              as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` +
              `Ownership Status` + `Postalcode Risk Class` + `Thatched Roof` +
              `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime,
            data = TrainDataFREQ,
            family = poisson(link = "log"))

mod2 <- glm(ClaimCount ~ log(ClaimExposure) + `Own Risk` +
              City + `Payment Term` + `Type of house` +
              as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` +
              `Ownership Status` + `Thatched Roof` +
              `Extra Living Space` + `Postalcode Risk Class`+ `Insurance Amount Home-
            content`+ Age + PolicyTime,
            data = TrainDataFREQ,
            family = poisson(link = "log"))

mod3 <- glm(ClaimCount ~ log(ClaimExposure) + `Own Risk` +
              `Payment Term` + `Type of house` + City +
              + as.numeric(WOZ) + `Construction Year` +
              `Ownership Status` + `Thatched Roof` + `Postalcode Risk Class`+
              `Insurance Amount Home-content`+ Age + PolicyTime,
            data = TrainDataFREQ,
            family = poisson(link = "log"))

mod4 <- glm(ClaimCount ~ log(ClaimExposure) + `Own Risk` +
              `Payment Term` + `Type of house` +
              + as.numeric(WOZ) + `Construction Year` +
              `Ownership Status` + `Thatched Roof` + `Postalcode Risk Class`+
```

```
  `Insurance Amount Home-content`+ Age + PolicyTime,
 data = TrainDataFREQ,
 family = poisson(link = "log"))

mod5 <- glm(ClaimCount ~ log(ClaimExposure) + `Own Risk` +
  `Payment Term` + `Type of house` +
  + as.numeric(WOZ) + `Construction Year` +
  `Ownership Status` + `Postalcode Risk Class`+
  Age + PolicyTime,
 data = TrainDataFREQ,
 family = poisson(link = "log"))
```

A4 Claim severity GLMs (independent)

```
modS1 <- glm(SchadelastMutatie ~  `Own Risk` + City +
  `Payment Term` + `Type of house` + Urbanisation +
  as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` +
  `Ownership Status` + `Postalcode Risk Class` + `Thatched Roof` +
  `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime,
 data = TrainDataSEV,
 family = Gamma(link = "log"))

modS2 <- glm(SchadelastMutatie ~  `Own Risk` +
  `Payment Term` + `Type of house` + Urbanisation +
  as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` +
  `Ownership Status` + `Postalcode Risk Class` + `Thatched Roof` +
  `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime,
 data = TrainDataSEV,
 family = Gamma(link = "log"))

modS3 <- glm(SchadelastMutatie ~
  `Type of house` + Urbanisation +
  as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` +
  `Ownership Status`+ `Postalcode Risk Class` + `Thatched Roof` +
  `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime,
 data = TrainDataSEV,
 family = Gamma(link = "log"))

modS4 <- glm(SchadelastMutatie ~
  Urbanisation +
  as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` +
  `Ownership Status`+ `Postalcode Risk Class` + `Thatched Roof` +
  `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime,
 data = TrainDataSEV,
 family = Gamma(link = "log"))
```

A5 Claim severity GLMs (dependent)

modSD1 <- glm(*SchadelastMutatie ~ ClaimCount+ `Own Risk` + `Payment Term` + `Type of house` + Urbanisation + as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` + `Ownership Status` + `Postalcode Risk Class` + `Thatched Roof` + `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime, data = TrainDataSEV,*
family = Gamma(link = "log"))

modSD2 <- glm(*SchadelastMutatie ~ ClaimCount+ `Type of house` + Urbanisation + as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` + `Ownership Status` + `Postalcode Risk Class` + `Thatched Roof` + `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime, data = TrainDataSEV,*
family = Gamma(link = "log"))

modSD3 <- glm(*SchadelastMutatie ~ ClaimCount+ Urbanisation + as.numeric(`Living Space`) + as.numeric(WOZ) + `Construction Year` + `Ownership Status` + `Postalcode Risk Class` + `Thatched Roof` + `Extra Living Space` + `Insurance Amount Home-content`+ Age + PolicyTime, data = TrainDataSEV,*
family = Gamma(link = "log"))

A6. Results claim frequency GLM

Variable	**Coefficient**	**Standard_Error**
(Intercept)	-0.1727	1.1247
log(ClaimExposure)	0.2133	0.0256
`Own Risk`	-0.0008	0.0003
`Payment Term`Jaar	0.0278	0.0622
`Payment Term`Kwartaal	0.0854	0.0668
`Payment Term`Maand	0.0533	0.0601
`Type of house`Appartement/etagewoning	-0.0107	0.5020
`Type of house`Geschakelde woning	0.1033	0.5093
`Type of house`Hoekwoning	0.0617	0.5015
`Type of house`Tussenwoning	0.0715	0.5013
`Type of house`Twee-onder-een-kap	0.0751	0.5013
`Type of house`Vrijstaande woning	0.0595	0.5011
`Type of house`Woning zakelijk	-0.1567	0.6716
`Type of house`Zakelijk 1, 2, 3 of 4	-0.3688	1.1191
as.numeric(WOZ)	0.0000	0.0000
`Construction Year`1920-1939	-0.0131	0.0567
`Construction Year`1940-1959	0.0379	0.0571
`Construction Year`1960-1969	0.0131	0.0546

`Construction Year`1970-1979	0.0468	0.0533
`Construction Year`1980-1989	-0.0079	0.0548
`Construction Year`1990-1994	0.0247	0.0603
`Construction Year`1995-1999	0.0677	0.0590
`Construction Year`2000-2004	0.0589	0.0611
`Construction Year`2005-2009	0.0117	0.0612
`Construction Year`2010-2014	-0.0166	0.0648
`Construction Year`2015-2019	0.0086	0.0626
`Construction Year`2020-2024	-0.0622	0.1091
`Construction Year`voor 1900	0.0136	0.0821
`Ownership Status`Nee	0.0601	0.0260
`Ownership Status`Onbekend	-0.0125	0.0322
`Thatched Roof`true	0.0322	0.0599
`Postalcode Risk Class`Risicoklasse 2	-0.0090	0.0363
`Postalcode Risk Class`Risicoklasse 3	-0.0006	0.1832
`Postalcode Risk Class`Risicoklasse 4	-0.0827	0.1972
`Postalcode Risk Class`Risicoklasse 5	-0.3694	0.5005
`Postalcode Risk Class`Risicoklasse 7	-0.2771	1.0010
`Postalcode Risk Class`Risicoklasse 8	-0.4525	0.7083
`Insurance Amount Home-content`€ 10.000	0.3348	1.0011
`Insurance Amount Home-content`€ 15.000	0.3375	1.0021
`Insurance Amount Home-content`€ 20.000	0.3043	1.0031
`Insurance Amount Home-content`€ 25.000	0.2106	1.0153
`Insurance Amount Home-content`€ 30.000	0.1832	1.0122
`Insurance Amount Home-content`€ 35.000	0.2148	1.0620
`Insurance Amount Home-content`€ 40.000	0.4052	1.0420
`Insurance Amount Home-content`€ 45.000	0.2014	1.0814
`Insurance Amount Home-content`€ 5.000	-0.0045	1.4292
`Insurance Amount Home-content`€ 50.000	0.7267	1.0390
`Insurance Amount Home-content`> € 50.000	0.2383	1.0619
Age	-0.0068	0.0007
PolicyTime	-0.0005	0.0012

A7. Results independent claim severity GLM

Variable	**Coefficient**	**Standard_Error**
(Intercept)	5.1552	0.7160
`Type of house`Appartement/etagewoning	0.1101	0.7071
`Type of house`Geschakelde woning	0.1846	0.7132
`Type of house`Hoekwoning	-0.0033	0.7067
`Type of house`Tussenwoning	0.0539	0.7066
`Type of house`Twee-onder-een-kap	0.0435	0.7063
`Type of house`Vrijstaande woning	0.0669	0.7062
`Type of house`Woning zakelijk	0.2054	0.8650
`Type of house`Zakelijk 1, 2, 3 of 4	0.4647	1.2354

Urbanisation100.000 tot 250.000 inwoners	0.0750	0.0315
Urbanisation20.000 tot 50.000 inwoners	0.0286	0.0257
Urbanisation250.000 tot 500.000 inwoners	2.1220	1.4115
Urbanisation5.000 tot 10.000 inwoners	0.0120	0.0291
Urbanisation50.000 tot 100.000 inwoners	0.0533	0.0458
UrbanisationMeer dan 500.000 inwoners	-0.4721	0.3619
UrbanisationMinder dan 5.000 inwoners	-0.0002	0.0262
as.numeric(`Living Space`)	0.0002	0.0002
as.numeric(WOZ)	0.0000	0.0000
`Construction Year`1920-1939	-0.0177	0.0572
`Construction Year`1940-1959	-0.0624	0.0577
`Construction Year`1960-1969	-0.0561	0.0553
`Construction Year`1970-1979	-0.0946	0.0540
`Construction Year`1980-1989	-0.0827	0.0553
`Construction Year`1990-1994	-0.0425	0.0610
`Construction Year`1995-1999	-0.0969	0.0598
`Construction Year`2000-2004	-0.1260	0.0625
`Construction Year`2005-2009	-0.0868	0.0612
`Construction Year`2010-2014	-0.0837	0.0660
`Construction Year`2015-2019	0.0181	0.0626
`Construction Year`2020-2024	-0.0613	0.1036
`Construction Year`voor 1900	-0.1639	0.0837
`Ownership Status`Nee	0.0977	0.0258
`Ownership Status`Onbekend	0.0133	0.0321
`Postalcode Risk Class`Risicoklasse 2	0.0925	0.0426
`Postalcode Risk Class`Risicoklasse 3	-0.0218	0.1766
`Postalcode Risk Class`Risicoklasse 4	0.0025	0.1978
`Postalcode Risk Class`Risicoklasse 5	-0.1249	0.5761
`Postalcode Risk Class`Risicoklasse 6	-0.2044	0.9983
`Postalcode Risk Class`Risicoklasse 7	0.7397	0.4997
`Postalcode Risk Class`Risicoklasse 8	-0.8925	0.9989
`Thatched Roof`true	0.0155	0.0614
`Extra Living Space`< 120m2	0.3022	0.1533
`Extra Living Space`< 140m2	0.3465	0.1867
`Extra Living Space`< 160m2	0.3053	0.1936
`Extra Living Space`< 180m2	0.2014	0.1980
`Extra Living Space`< 200m2	0.2491	0.1870
`Extra Living Space`< 250m2	-0.0418	0.2341
`Extra Living Space`< 300m2	0.7436	0.3882
`Extra Living Space`< 350m2	0.7007	0.5069
`Extra Living Space`< 450m2	-0.2551	1.0036
`Extra Living Space`< 60m2	0.2228	0.0906
`Extra Living Space`< 80m2	0.0884	0.1192
`Extra Living Space`0m2	0.2848	0.1522
`Insurance Amount Home-content`€ 15.000	0.1893	0.0454
`Insurance Amount Home-content`€ 20.000	0.3625	0.0632

`Insurance Amount Home-content`€ 25.000	0.4651	0.1565
`Insurance Amount Home-content`€ 30.000	0.0254	0.1403
`Insurance Amount Home-content`€ 35.000	0.2813	0.3330
`Insurance Amount Home-content`€ 40.000	0.2194	0.2772
`Insurance Amount Home-content`€ 45.000	0.3385	0.4086
`Insurance Amount Home-content`€ 5.000	-0.1910	1.0540
`Insurance Amount Home-content`€ 50.000	0.0093	0.3020
`Insurance Amount Home-content`> € 50.000	1.9164	0.4139
Age	0.0028	0.0007
PolicyTime	-0.0044	0.0010

A8. Results dependent claim severity GLM

Variable	**Coefficient**	**Standard_Error**
(Intercept)	3.2265	1.0061
ClaimCount	-0.0478	0.0074
`Type of house`Appartement/etagewoning	1.0411	0.7082
`Type of house`Geschakelde woning	1.1070	0.7139
`Type of house`Hoekwoning	0.9925	0.7078
`Type of house`Tussenwoning	1.0277	0.7077
`Type of house`Twee-onder-een-kap	0.9756	0.7076
`Type of house`Vrijstaande woning	1.0242	0.7074
`Type of house`Woning zakelijk	1.0973	0.8664
`Type of house`Zakelijk 1, 2, 3 of 4	1.4601	1.0134
Urbanisation100.000 tot 250.000 inwoners	0.0819	0.0310
Urbanisation20.000 tot 50.000 inwoners	0.0331	0.0254
Urbanisation250.000 tot 500.000 inwoners	2.1214	1.4106
Urbanisation5.000 tot 10.000 inwoners	0.0179	0.0289
Urbanisation50.000 tot 100.000 inwoners	0.0751	0.0465
UrbanisationMeer dan 500.000 inwoners	-0.4493	0.3582
UrbanisationMinder dan 5.000 inwoners	0.0116	0.0259
as.numeric(`Living Space`)	0.0001	0.0002
as.numeric(WOZ)	0.0000	0.0000
`Construction Year`1920-1939	-0.0279	0.0573
`Construction Year`1940-1959	-0.0725	0.0579
`Construction Year`1960-1969	-0.0449	0.0556
`Construction Year`1970-1979	-0.0722	0.0543
`Construction Year`1980-1989	-0.0473	0.0554
`Construction Year`1990-1994	-0.0237	0.0611
`Construction Year`1995-1999	-0.0490	0.0598
`Construction Year`2000-2004	-0.1304	0.0621
`Construction Year`2005-2009	-0.0620	0.0617
`Construction Year`2010-2014	-0.0802	0.0656
`Construction Year`2015-2019	0.0294	0.0630
`Construction Year`2020-2024	-0.0603	0.1063

`Construction Year`voor 1900	-0.1349	0.0825
`Ownership Status`Nee	0.0879	0.0255
`Ownership Status`Onbekend	0.0073	0.0317
`Postalcode Risk Class`Risicoklasse 2	0.1058	0.0414
`Postalcode Risk Class`Risicoklasse 3	-0.0585	0.1713
`Postalcode Risk Class`Risicoklasse 4	0.1163	0.2046
`Postalcode Risk Class`Risicoklasse 5	-0.4420	0.4462
`Postalcode Risk Class`Risicoklasse 6	0.2243	0.9977
`Postalcode Risk Class`Risicoklasse 7	0.9592	0.4998
`Postalcode Risk Class`Risicoklasse 8	-0.9490	0.9983
`Thatched Roof`true	0.1358	0.0588
`Extra Living Space`< 120m2	0.2292	0.1432
`Extra Living Space`< 140m2	0.5233	0.1742
`Extra Living Space`< 160m2	0.2067	0.1812
`Extra Living Space`< 180m2	0.1516	0.2055
`Extra Living Space`< 200m2	0.2352	0.1972
`Extra Living Space`< 250m2	0.0107	0.2132
`Extra Living Space`< 300m2	0.8542	0.3072
`Extra Living Space`< 350m2	0.3067	0.4546
`Extra Living Space`< 450m2	-0.1822	1.0028
`Extra Living Space`< 60m2	0.2771	0.0892
`Extra Living Space`< 80m2	0.1609	0.1180
`Extra Living Space`0m2	0.4934	0.1466
`Insurance Amount Home-content`€ 10.000	1.0523	0.7054
`Insurance Amount Home-content`€ 15.000	1.2683	0.7068
`Insurance Amount Home-content`€ 20.000	1.3942	0.7080
`Insurance Amount Home-content`€ 25.000	1.5201	0.7227
`Insurance Amount Home-content`€ 30.000	1.2010	0.7190
`Insurance Amount Home-content`€ 35.000	0.9027	0.7796
`Insurance Amount Home-content`€ 40.000	1.2309	0.7667
`Insurance Amount Home-content`€ 45.000	1.5678	0.8650
`Insurance Amount Home-content`€ 5.000	1.6439	1.2706
`Insurance Amount Home-content`€ 50.000	1.3803	0.7803
`Insurance Amount Home-content`> € 50.000	2.7653	0.8018
Age	0.0014	0.0007
PolicyTime	-0.0037	0.0010

Chapter 8 References

Al-Mosawi, M. (2017). *An Extension of Generalized Linear Models for dependent frequency and severity*. www.math.su.se

Becker, D. G., Woolford, D. G., & Dean, C. B. (2022). Assessing dependence between frequency and severity through shared random effects. *PLoS ONE*, *17*(8 August). https://doi.org/10.1371/journal.pone.0271904

Bühlmann, H. (1997). An overview of frequency-severity models. *Casualty Actuarial Society*, 1–23.

Chatfield, S. E. C., & Zidek, J. (2002). *CHAPMAN & HALL/CRC Texts in Statistical Science Series*.

Chen, Y. T., & Tzeng, L. Y. (2007). The frequency and severity relationship of liability insurance. . *Journal of Risk and Insurance*, 167–185.

Denuit, M., & Boucher, J.-Ph. (2006). The relationship between frequency and severity in automobile insurance. *ASTIN Bulletin*, 179–196.

Dobsen Anette J. (2002). *An Introduction to Generalized Linear Models*.

England, P., & Verrall, R. J. (2002). Modelling commercial property insurance claims and the effects of policy excesses using generalized linear models. *Mathematics and Economics*, 69–87.

Eriksson, A. (2021). *A Comparison of Gradient Boosting Machines and Generalized Linear Models for Non-Life Insurance Pricing*. www.math.su.se

Frees, E. W., Lee, G., & Yang, L. (2016a). Multivariate frequency-severity regression models in insurance. *Risks*, *4*(1). https://doi.org/10.3390/risks4010004

Frees, E. W., Lee, G., & Yang, L. (2016b). Multivariate frequency-severity regression models in insurance. *Risks*, *4*(1). https://doi.org/10.3390/risks4010004

Gagné, R., & Dionne, G. (2002). Controlling for the frequency/severity interaction in workers' compensation. *Journal of Risk and Insurance*, 37–55.

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, *70*, 205–215. https://doi.org/10.1016/j.insmatheco.2016.06.006

Georges Dionne, Christian Gouriéroux, & Charles Vanasse. (2001). Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment. *Journal of Political Economy 109*, 444–453.

Ghaddab, S., Kacem, M., de Peretti, C., & Belkacem, L. (2023). Extreme severity modeling using a GLM-GPD combination: application to an excess of loss reinsurance treaty. *Empirical Economics*. https://doi.org/10.1007/s00181-023-02371-4

glmbook. (n.d.).

Green, S. G., & Higgs, P. J. (1989). Motor insurance rating using generalised linear models. *The Statistician*, 317–332.

Guillén, M., Denuit, M., & Boucher, J. (2016). Risk modeling in insurance. *Wiley*.

H. Akaike. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2019). *Boosting insights in insurance tariff plans with tree-based machine learning methods*. http://arxiv.org/abs/1904.10890

Hu, J., & Kuo, T. C. (2019). Predictive modeling for medical malpractice insurance: A comparison of GLMs and machine learning algorithms. *Journal of Risk and Insurance*, 913–938.

Jia, R., & Eling, M. (2016). The relationship between frequency and severity of insured losses: A study of Swiss household insurance. *Journal of Risk and Insurance*, 397–421.

Joksch, H. C. (1980). *A RELATION BETWEEN COLLISION CLAIM FREQUENCY AND DRIVER AGE DISTRIBUTION IN HLDI DATA*.

Karlis, D., & Ntzoufras, I. (2003). Bayesian modelling of insurance claims counts using scale mixtures of Poisson distributions. Insurance. *Insurance: Mathematics and Economics*, 249–265.

Lee, G. Y., & Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics*, *87*, 115–129. https://doi.org/10.1016/j.insmatheco.2019.04.004

Lee, W., Park, S. C., & Ahn, J. Y. (2019a). Investigating dependence between frequency and severity via simple generalized linear models. *Journal of the Korean Statistical Society*, *48*(1), 13–28. https://doi.org/10.1016/j.jkss.2018.07.003

Lee, W., Park, S. C., & Ahn, J. Y. (2019b). Investigating dependence between frequency and severity via simple generalized linear models. *Journal of the Korean Statistical Society*, *48*(1), 13–28. https://doi.org/10.1016/j.jkss.2018.07.003

Lu, Y. (2019). Flexible (panel) regression models for bivariate count-continuous data with an insurance application. In *J. R. Statist. Soc. A* (Vol. 182). https://academic.oup.com/jrsssa/article/182/4/1503/7068327

McCullagh P., & Nelder J.A. (n.d.). glmbook (1). *Second Edition*.

Mcleod, A. I., & Xu, C. (n.d.). *bestglm: Best Subset GLM*.

Merz, M., & Wüthrich, M. V. (2008). Frequency-severity modeling of insurance claim sizes using generalized linear models. *Insurance: Mathematics and Economics*, 332.

Oeben, M. (2015). *Generalized Linear Mixed Models in the competitive non-life insurance market*.

Overzicht marktontwikkelingen 2021 - 2022. (2023, March). Waarderingskamer.

Renshaw, A. E. (1994). Modelling the Claims Process in the Presence of Covariates. *ASTIN Bulletin*, *24*(2), 265–285. https://doi.org/10.2143/ast.24.2.2005070

Sáez-Castillo, A. J., Conde-Sánchez, A., & Martínez, F. (2022). *DGLMExtPois: Advances in Dealing with Over and Under-dispersion in a Double GLM Framework*.

Schulz, J. (2013). *Generalized Linear Models for a Dependent Aggregate Claims Model*.

Schwarz G. (1978). Estimation the Dimension of a Model. *Annals of Statistics*, 461–464.

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance: Mathematics and Economics*, *64*, 417–428. https://doi.org/10.1016/j.insmatheco.2015.07.006

Shiu, Y., Lai, C., & Chen, J. (2020). Frequency-severity modeling of large insurance claims: A spatial model with time-varying parameters. *Journal of Risk and Insurance*, 7–30.

Smit, H. T., & Schmit, J. T. (2012). Claims frequency and claims severity in the Dutch insurance market. *Journal of Risk and Insurance*, 1073–1095.

Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PLoS ONE*, *15*(8 August 2020). https://doi.org/10.1371/journal.pone.0238000

Wüthrich, M. V. (2016). Statistical methods for the analysis of the frequency-severity relationship in health insurance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 951–974.

Wüthrich, M. V. (2015). Multivariate frequency-severity modeling. *Springer*.

Wüthrich, M. V., & Merz, M. (2008). Modeling frequency and severity for non-life insurance claims using Poisson and gamma distributions. *Astin Bulletin*, 105–133.

Yan, J., Guszcza, J., Flynn, M., & Peter Wu, C.-S. (2009). *Applications of the Offset in Property-Casualty Predictive Modeling*.

Yang, L. (2022). Nonparametric Copula Estimation for Mixed Insurance Claim Data. *Journal of Business and Economic Statistics*, *40*(2), 537–546. https://doi.org/10.1080/07350015.2020.1835668

Zhou, J., & CPCU Debbie Deng, F. (2019). *GLM vs. Machine Leaning-with Case Studies in Pricing*.