THE IMPACT OF DEPENDENCE BETWEEN CLAIM FREQUENCY AND SEVERITY ON EXPECTED LOSS USING GLM: MTPL APPLICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

İLKYAZ ASLANÖZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ACTUARIAL SCIENCES

JULY 2024

Approval of the thesis:

THE IMPACT OF DEPENDENCE BETWEEN CLAIM FREQUENCY AND SEVERITY ON EXPECTED LOSS USING GLM: MTPL APPLICATION

submitted by **İLKYAZ ASLANÖZ** in partial fulfillment of the requirements for the degree of **Master of Science in Actuarial Sciences Department, Middle East Technical University** by,

Prof. Dr. A. Sevtap KESTEL _____
Dean, Graduate School of **Applied Mathematics**

Assist. Prof. Dr. Büşra Z. TEMOÇİN _____
Head of Department, **Actuarial Sciences**

Prof. Dr. A. Sevtap KESTEL _____
Supervisor, **Actuarial Sciences, METU**

Dr. Bükre YILDIRIM KÜLEKCİ _____
Co-supervisor, **Actuarial Sciences, METU**

Examining Committee Members:

Assist. Prof. Dr. Başak BULUT KARAGEYİK
Actuarial Sciences, Hacettepe University _____

Prof. Dr. A. Sevtap KESTEL
Actuarial Sciences, METU _____

Assist. Prof. Dr. Büşra Z. TEMOÇİN
Actuarial Sciences, METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: İLKYAZ ASLANÖZ

Signature :

ABSTRACT

THE IMPACT OF DEPENDENCE BETWEEN CLAIM FREQUENCY AND SEVERITY ON EXPECTED LOSS USING GLM: MTPL APPLICATION

ASLANÖZ, İLKYAZ
M.S., Department of Actuarial Sciences
Supervisor : Prof. Dr. A. Sevtap KESTEL
Co-Supervisor : Dr. Bükre YILDIRIM KÜLEKCİ

July 2024, 52 pages

In non-life insurance, the accurate estimation of total loss is extremely important for companies' asset-liability management. To estimate the total loss, insurance companies use generalized linear model (GLM) as it is compatible with insurance data and hence makes considerably consistent predictions. The common practice is constructing a GLM for frequency, usually using the Poisson distribution, and another GLM for severity, usually using the Gamma distribution, and then multiplying the results of these two models. However, this multiplication is only possible under the assumption of independence of the claim frequency and severity. Although the independence assumption simplifies the modeling, it also causes deviations from the real loss value. In this thesis, constructing a GLM, which can incorporate the dependence between claim frequency and severity, to predict the total loss is aimed. Two GLMs are built and tested; dependent-GLM and copula-GLM, and they are compared with the regular independent-GLM. To examine these models, non-life motor third party liability (MTPL) insurance data is used. In the first model, the dependency is provided by taking the claim number as a covariate of marginal severity GLM. The second model provides the dependence between the marginal frequency and severity GLMs by using a copula function. To compare these two dependent models and the independent model, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used, and also the means of the estimations of the models are compared

to the means of the real observations. The findings show that the independent-GLM deviates more from the real value compared to the other two. On the other hand, the dependent-GLM model is quite close to the copula-GLM model but gives slightly better results.

Keywords: Generalized Linear Model, Copula, Dependence, Claim Frequency, Claim Severity.

ÖZ

HASAR SIKLIĞI VE ŞİDDETİ ARASINDAKİ BAĞIMLILIĞIN GLM KULLANILARAK TOPLAM HASAR ÜZERİNDEKİ ETKİSİ: MTPL UYGULAMASI

ASLANÖZ, İLKYAZ
Yüksek Lisans, Aktüerya Bilimleri Bölümü
Tez Yöneticisi : Prof. Dr. A. Sevtap KESTEL
Ortak Tez Yöneticisi : Dr. Bükre YILDIRIM KÜLEKCİ

Temmuz 2024, 52 sayfa

Hayat dışı sigortalarda toplam hasarın doğru tahmin edilmesi şirketlerin aktif-pasif yönetimi açısından son derece önemlidir. Sigorta şirketleri, toplam hasarı tahmin etmek için, sigorta verileriyle uyumlu olması ve dolayısıyla da oldukça tutarlı tahminler yapması sebebiyle genelleştirilmiş doğrusal model (GLM) kullanır. Yaygın uygulama, Poisson dağılımını kullanarak hasar sıklığı için bir GLM ve Gamma dağılımını kullanarak hasar şiddeti için bir başka GLM oluşturmak ve ardından bu iki modelin sonuçlarını birbiriyle çarpmak şeklindedir. Fakat bu çarpma işlemi sadece hasar sıklığı ve hasar şiddetinin bağımsızlığı varsayımı altında mümkündür. Bağımsızlık varsayımı modellemeyi basitleştirse de tahminlerin gerçek hasar değerinden uzaklaşmasına neden olur. Bu tezde, toplam hasarı tahmin etmek için hasar sıklığı ve şiddeti arasındaki bağımlılığı da içeren bir GLM kurmak hedeflenmiştir. Bu amaçla bağımlı-GLM ve kopula-GLM olmak üzere iki model kuruldu ve hem kendi aralarında karşılaştırıldılar, hem de yaygın uygulama olan bağımsız-GLM ile kıyaslanmıştır. Kurulan modellerin incelenmesi için üçüncü şahıs mali mesuliyet sigortası (MTPL) verileri kullanılmıştır. Bağımlı-GLM'de bağımlılık, hasar sayısının marjinal hasar şiddeti GLM'inde ortak değişken olarak alınması ile sağlanmaktadır. Kopula-GLM ise, marjinal frekans ve şiddet GLM'leri arasındaki bağımlılığı bir kopula fonksiyonu kullanarak sağlamaktadır. Bu üç modelin kıyaslanmasında Akaike Bilgi Ölçütü (AIC)

ve Bayesçi Bilgi Ölçütü (BIC) kullanılmış olup, ayrıca model tahminlerinin ortalamaları gerçek verinin ortalamalarıyla karşılaştırılmıştır. Üç model karşılaştırıldığında bağımsız-GLM'in gerçek veriden diğer modellere göre daha çok saptığı belirlenmiştir. Bağımlı-GLM sonuçları ise kopula-GLM ile oldukça yakın olmakla birlikte biraz daha gerçeğe yakın değerler vermiştir.

Anahtar Kelimeler: Genelleştirilmiş Doğrusal Model, Kopula, Bağımlılık, Hasar Sıklığı, Hasar Şiddeti

To my grandparents; Nermin & Abdulbaki

ACKNOWLEDGMENTS

I would like to express my great appreciation to my supervisor Prof. Dr. A. Sevtap Kestel for her valuable advice and encouragement during the development and preparation of this thesis. Without her patient guidance, this work would not be possible.

I also would like to thank my co-supervisor Dr. Bükre Yıldırım Külekci for her valuable advice, guidance, willingness to give her time and share precious experiences which have brightened my path.

I also would like to thank my jury members, Assist. Prof. Dr. Başak Bulut Karageyik, and Assist. Prof. Dr. Büşra Z. Temoçin for their valuable comments and discussions about this work.

I am sincerely grateful for the endless academic and emotional support provided by Res. Assist. Cem Yavrum throughout my master's degree and thesis process. His advice, guidance, and friendship have been invaluable, and I have learned so much from him.

I want to express my gratitude to Dr. Meral Şimşek for her support, advice, and motivation for my thesis whenever I needed it.

I would like to thank my dear friend Şeref Kutay Yakut for his everlasting help, encouragement and friendship throughout my academic journey.

I wish to thank my dear sister İdil Aslanöz for being my lifetime support and my greatest source of happiness.

I also thank my parents Şerife Aslanöz and Sıtkı Mutlu Aslanöz for their great support.

I wish to thank my dear friend and music teacher Barış Yaman for his great emotional support through his friendship, encouragement, concerts, and music lessons during my thesis process.

Finally, I would like to thank my bandmates from our band Flight; Ege, Kadir, Melihcan, Oğuz Kerem, and Umut Güneş. Making music with them was one of my biggest emotional supports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
EDF	Exponential Dispersion Family
GLM	Generalized Linear Model
LM	Linear Model
MLE	Maximum Likelihood Estimation
MTPL	Motor Third Party Liability

CHAPTER 1

INTRODUCTION

An insurance policy agreement, between an insurance company and a policyholder, transfers the economic risk of unpredictable losses from the policyholder to the insurance company against a predefined premium fee. Non-life insurance comprises damages to tangible assets such as car, house, or property damages due to fire, crash, natural disaster, etc, or losses due to bodily injury of the insured or another person (third party liability), or losses caused by interruption of a company or business, health problems of the insured or employees, etc.

Determining the premium value of such a contract is important for an insurance company because if the company determines an excessive premium, the customer would prefer another company with a fair price. On the other hand, if prices are undervalued, this can cause a loss in profitable policies and gaining underpriced ones. The premium of non-life insurance is determined based on the expected loss of claims and loadings like administration costs, cost of capital, and provision for future costs of the company, etc. Therefore, to determine the most appropriate price, the expected loss should be calculated fairly.

Expected loss varies between policies as each policy has unique properties. For instance, each insured has different properties like age, health situation, educational status, marital status, etc. On the other side, properties have different features, like the value of a car or a house, the presence or absence of security measures like an alarm, whether the property is old or new, whether there is any damage on it, etc. These differences can be dealt with with statistical models that bring concrete ev-

idence in calculating the expected loss, in other words, the pure premium. These policy-specific features that affect the price of the policy are called rating factors, and some of them are used as determining variables, in other words, covariates of the premium model [4].

As in [12] Kramer et. al. emphasize, modeling claim severity and frequency separately, and determining the total loss by multiplying the expected values of these two is a prevalent approach to designating the pure premium of a policy. However, this approach requires acceptance of independence between the claim frequency and severity, and this is not the case most of the time. For instance, in [9] Garrido states that, when the motor insurance claim data of the companies is examined, it is seen that the drivers usually tend to have either a few minor accidents, or a single major accident, which shows that the claim number and claim size generally have negative association in car insurance. On the other side, in the case of home insurance, sewer backup or flooding causes both large and frequent claims in problematic areas, which means severity and frequency have a positive association. Therefore, the independence assumption on claim frequency and severity may cause over or underestimation of the total loss of individual policies and eventually low accuracy in risk calculation of the whole portfolio. Hence, while fitting a model to the frequency and severity of individual claims to determine the premium of a policy, this dependency should be first tested to see if it exists and then according to the results, an appropriate model should be constructed.

GLM is the most preferred modeling method in the insurance industry because it adapts well to insurance data that is non-normal, such as Poisson distributed positive integer values of claim counts, and Gamma distributed, right-skewed, continuous, and positive values of claim amounts. GLM is the generalized version of the linear model (LM), and makes better predictions than LM due to its good adaptability [15].

There are two GLM approaches; modeling burning cost with a Tweedie distribution, and modeling frequency with a Poisson distribution and the severity with a Gamma distribution, separately. Although the Tweedie model is more handy because it has one model and hence it is simpler than the Poisson-Gamma model, which has two

different models for the frequency and the severity, Tweedie cannot catch the trends as accurately as the Poisson-Gamma model because of this simplicity.

What both Tweedie and Possion-Gamma models have in common is that they model frequency and severity assuming they are independent. To understand how dependence should be handled, to see what the consequences might be if it is not taken into account, and to try to create a more accurate model by taking dependency into account, dependent models should be established. There are two common approaches to account for the dependence between frequency and severity. The first one is a revised version of the independent Poisson-Gamma model and provides the dependency by taking the frequency as a covariate of the severity model while modeling the average claim amount distribution. The second approach uses a copula function to link the marginal GLMs of frequency and severity. Copulas are functions that are used to model the dependence between random variables.

When building a Poisson-Gamma GLM to estimate the total loss, severity and frequency are separately expressed in terms of a linear combination of rating variables, in other words, covariates, such as age, sex, brand, etc., via a link function. Then, the expected average severity and expected frequency are multiplied to get loss cost, namely the pure premium (See [17]).

Jorgensen & De Souzaa in [11] use the Tweedie distribution which parametrizes the compound Poisson-Gamma distribution as a member of the exponential dispersion family, and then fit a GLM to determine the claim rate. This model consists of three parameters, namely, the mean claim rate, a dispersion parameter, and a shape parameter. The results of the study indicate that the Tweedie model is efficient due to its simplicity.

In addition to these studies, Dimakos & Frigessi in [5] use a Bayesian approach, again with independence assumption, to be able to consider all randomness in premium estimation. They calculate the premium by multiplying the expected claim number and expected claim size. The expected claim number is based on a spatial Poisson regression model, and the expected claim size is based on a spatial Gamma

regression model, and they use an improper Markov Random Field to model the spatial structure. However, when they apply their model to a real car insurance data, they do not detect any noticeable improvement.

On the other hand, there is a lot of work trying to take into account the dependency between frequency and severity. For instance, in [10] Gschlößl & Czado furthered the Bayesian approach study of Dimakos & Frigessi in [5], and modeled claim size conditionally on the claim number using a car insurance dataset. In this manner, they take account of dependency between frequency and severity; and as a result, they detect that dependency has a remarkable effect.

Frees et al. study health care expenditures by considering the claim number as a covariate of average claim size distribution, and find that their model fit better to data and gives better results than the independent Poisson-Gamma model [6]. Garrido et al. apply this approach to car insurance data and construct a GLM for conditional severity, which takes marginal frequency as a covariate to consider the dependency between severity and frequency, and find that their dependent model is better than the independent Poisson-Gamma model [9].

As mentioned above, another method to determine dependency is using copulas. Frees examines the copula concept theoretically to adapt it to actuarial issues and uses copula to determine the loss and expense of an insurance company. In addition, they study additional applications of copulas like stochastic ordering, fuzzy sets, insurance pricing, etc [7].

Song uses Gaussian copula to link univariate dispersion models (Gamma, Poisson, etc.) and creates multivariate dispersion models [23]. With this idea, Czado et al. construct a marginal Gamma GLM for claim size and a marginal Poisson GLM for claims number and link them through a Gaussian copula using car insurance policies [3].

Shi et al.compare Tweedie, two-part, and mixed copula models and find that the copula model performs better than the other two models in [19].

When pricing policies in the insurance industry, it is crucial to estimate the total loss as close to reality as possible. Therefore, companies want to build accurate models to make total loss predictions. In this context, it is essential to know whether an independent model or a dependent model gives better results, and if it turns out that the dependent model gives a better result, which dependence structure brings the prediction closer to reality in terms of accurate total loss prediction. In this manner, this thesis aims to compare the independent model, and the first and second dependent approaches mentioned above to determine which one gives the best results, and must be preferred by the insurance companies. To this end, by using an MTPL dataset; firstly an independent Poisson-Gamma model, secondly, by adding claim number to severity GLM as a covariate, a dependent Poisson-Gamma model, and finally, using a copula function to link the Poisson and Gamma models, a copula model is constructed. By comparing their results, their accuracy is examined and tried to understand which one is better to use for the insurance industry.

The rest of the thesis is organized as follows: In Chapter 2; LM, GLM, independent and dependent cases of GLM, and the copula approach are overviewed. In Chapter 3; these models are applied to MTPL insurance data for comparison, and the results are compared based on the reliability of the expected premium estimations. Finally, Chapter 4 concludes the thesis study and its results.

CHAPTER 2

PRELIMINARIES

In this chapter, the models and methodology that are used in this thesis will be introduced. Firstly in Section 2.1, the exponential dispersion family (EDF), which is one of the most important features of GLM, is introduced. In Section 2.2 LM, then in Section 2.3 GLM and the main features that distinguish it from LM are explained. Finally, the copula approach to consider the dependence between the claim frequency and severity is discussed in Section 2.4.

2.1 Exponential Dispersion Family

The EDF consists of probability distributions whose probability density function has a special form of:

$$f_Y(y;\theta,\phi) = \exp\left[\frac{y\theta - k(\theta)}{a(\phi)} + C(y,\phi)\right] \tag{2.1}$$

Here $Y \sim$ EDF(θ,ϕ) is the response variable, θ is the canonical parameter, $\phi > 0$ is the dispersion parameter, k(θ) is the cumulant function which is twice differentiable with an invertible second derivative. In addition, every different choice of cumulant function gives a different family of probability distribution.

Members of the exponential dispersion family can be either continuous distributions like Gamma, or discrete distributions like Poisson.

2.1.1 Mean and Variance of EDF

Let Y $\sim$ EDF, and $\ln f_Y(y;\theta,\phi) = l(y;\theta,\phi) = \big[\big(y\theta - k(\theta)\big)/a(\phi)\big] + C(y,\phi)$ be the log-likelihood function of Y.

Then, the mean of Y is found from the well-known relation:

$$\mathbb{E}\left[\frac{\partial}{\partial\theta} l(y;\theta,\phi)\right] = 0. \tag{2.2}$$

Since

$$\begin{aligned}\frac{\partial}{\partial\theta} l(y;\theta,\phi) &= \frac{\partial}{\partial\theta}\Big(\frac{y\theta - k(\theta)}{a(\phi)} + C(y,\phi)\Big) \\ &= \frac{y - k'(\theta)}{a(\phi)},\end{aligned} \tag{2.3}$$

then

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\theta} l(y;\theta,\phi)\right] &= \mathbb{E}\left[\frac{Y - k'(\theta)}{a(\phi)}\right] \\ &= \frac{\mathbb{E}(Y) - k'(\theta)}{a(\phi)} \\ &= 0.\end{aligned} \tag{2.4}$$

This statement implies that the mean of Y is equal to the first derivative of the cumulant function:

$$\mathbb{E}(Y) = k'(\theta). \tag{2.5}$$

Another well-known relation is:

$$\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(y;\theta,\phi)\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta} l(y;\theta,\phi)^2\right] = 0. \tag{2.6}$$

The second derivative of the likelihood function is

$$\frac{\partial^2}{\partial\theta^2} l(y;\theta,\phi) = \frac{-k''(\theta)}{a(\phi)}, \tag{2.7}$$

as

$$\frac{\partial}{\partial\theta} l(y;\theta,\phi) = \frac{y - k'(\theta)}{a(\phi)}. \tag{2.8}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{-k''(\theta)}{a(\phi)}\right]+\mathbb{E}\left[\Big(\frac{Y-k'(\theta)}{a(\phi)}\Big)^2\right] &= \frac{-k''(\theta)}{a(\phi)}+\mathbb{E}\left[\Big(\frac{\big(Y-\mathbb{E}(y)\big)^2}{a(\phi)^2}\Big)\right] \\ &= \frac{-k''(\theta)}{a(\phi)}+\frac{\mathbb{V}ar(Y)}{a(\phi)^2} \\ &= 0\end{aligned} \tag{2.9}$$

is concluded as

$$\mathbb{V}ar(Y)=a(\phi)k''(\theta). \tag{2.10}$$

2.1.2 Normal Distribution

Normal distribution is a continuous distribution which belongs to EDF. Although it is the most common distribution in nature, it is barely seen in insurance since the insurance data such as claim size, income, claim number, etc. naturally exhibit non-negativity and high skewness. However, normal distribution is used while analyzing the insurance data by applying a transformation to data to achieve normality. Normal distribution is denoted by $\mathrm{Y} \sim \mathrm{N}(\mu, \sigma^2)$. Here μ indicates mean, and σ^2 indicates variance.

Probability density function of the normal distribution is;

$$f_Y(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad -\infty<y<\infty. \tag{2.11}$$

Then,

$$\begin{aligned}f_Y(y;\mu,\sigma^2) &= \exp\left[\ln\left(\left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}\right)\right], \\ &= \exp\left[-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{(y-\mu)^2}{2\sigma^2}\right], \\ &= \exp\left[-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{y^2-2y\mu+\mu^2}{2\sigma^2}\right], \\ &= \exp\left[-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}+\frac{y\mu-\frac{1}{2}\mu^2}{\sigma^2}\right].\end{aligned} \tag{2.12}$$

The canonical parameter is μ, the dispersion parameter $a(\phi)$ is σ^2, and the cumulant function is $k(\theta) = \frac{1}{2}\theta^2$. Finally, mean and variance are found as;

$$\mathbb{E}(Y) = k'(\theta) = \theta = \mu, \tag{2.13}$$

$$\mathbb{V}ar(Y) = a(\phi)k''(\theta) = \sigma^2 1 = \sigma^2. \tag{2.14}$$

2.1.3 Poisson Distribution

Poisson is a distribution that belongs to EDF and is denoted by Y$\sim$ Poisson(λ). Poisson as a discrete distribution is used to represent the probability of the number of events. In this thesis, we use it to model the number of claims, in other words, frequency. Both mean and variance are equal to λ and the probability density function is

$$f_Y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}. \tag{2.15}$$

It can be written as

$$\begin{aligned} f_Y(y;\lambda) &= \frac{1}{y!}\exp\{y\ln(\lambda) - \lambda\} \\ &= \frac{1}{y!}\exp\{y\theta - e^{\theta}\}. \end{aligned} \tag{2.16}$$

by taking $\ln(\lambda) = \theta$ which shows that $f_Y(y;\lambda)$ is a member of exponential family. Here; the canonical parameter is $\ln(\lambda)$, the dispersion parameter ϕ is 1, and the cumulant function is $k(\theta) = \lambda = e^{\theta}$. Finally, mean and variance are found as;

$$\mathbb{E}(Y) = k'(\theta) = e^{\theta} = \lambda, \tag{2.17}$$

$$\mathbb{V}ar(Y) = a(\phi)k''(\theta) = 1e^{\theta} = \lambda. \tag{2.18}$$

2.1.4 Gamma Distribution

Gamma is a continuous, right-skewed distribution that belongs to the exponential family, defined only on positive numbers, and denoted by Y$\sim$ Gamma (α, β). In this thesis, we use it to model the claim amount, in other words, the severity. The

probability density function is:

$$\begin{aligned} f_Y(y;\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \\ &= \exp\big\{(\alpha-1)\ln y - \beta y + \alpha\ln(\beta) - \ln\Gamma(\alpha)\big\}, \quad y>0 \end{aligned} \tag{2.19}$$

After rearranging the equation as $\sigma^2 = \frac{1}{\alpha}$, and $\mu = \frac{\alpha}{\beta}$, the density function is:

$$f_Y\left(y;\mu,\sigma^2\right) = a(y;\sigma^2)\exp\Big\{-\frac{1}{\sigma^2}\big(\frac{y}{\mu} - \ln\frac{y}{\mu} - 1\big)\Big\}, \quad y>0 \tag{2.20}$$

where

$$a\left(y;\sigma^2\right) = a(y;1/\lambda) = \frac{\lambda^\lambda e^{-\lambda}}{y\Gamma(\lambda)}, \quad y>0. \tag{2.21}$$

This shows that $f_Y(y;\mu,\sigma^2)$ is a member of the exponential dispersion family. The canonical parameter is $-\frac{1}{\mu}$, the dispersion parameter ϕ is $\frac{1}{\alpha}$, and the cumulant function is $k(\theta) = -\ln\Big(\frac{1}{\mu}\Big) = -\ln(-\theta)$. Finally, mean and variance are found as;

$$\mathbb{E}(Y) = k'(\theta) = -\frac{1}{\theta} = \mu = \frac{\alpha}{\beta}, \tag{2.22}$$

$$\mathbb{V}ar(Y) = a(\phi)k''(\theta) = \frac{1}{\alpha}\frac{1}{\theta^2} = \frac{1}{\alpha}\mu^2 = \frac{\alpha}{\beta^2}. \tag{2.23}$$

2.2 Linear Models

Both LM and GLM express the relationship between the mean response of an observed variable Y, also called the response variable, and covariate matrix X via regression parameters β_is:

$$\mathbb{E}[Y|X] = X^T\beta. \tag{2.24}$$

Y is denoted by the matrix $Y = \{y_1, y_2, \ldots, y_n\}^T$, and X is denoted by an $n \times p$ matrix whose rows refer to different observations, and columns refer to different covariates. Lastly, $\beta = \{\beta_1, \beta_2, ..., \beta_p\}^T$ is the matrix of regression parameters to be estimated by the model, and $\mu = \mathbb{E}[Y|X]$.

A LM is denoted by

$$Y = X\beta + \epsilon. \tag{2.25}$$

Here, ϵ is the error term, normally distributed with zero mean and constant variance. In addition to error term, Y also has normal distribution; $Y \sim \text{N}(\mu, \sigma^2)$.

The expected value of y_i is expressed as;

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \eta_i. \tag{2.26}$$

Here, η is the linear predictor and briefly expressed as $\eta = X^T\beta$. If we combine the above-mentioned together;

$$\mu = \mathbb{E}[Y|X] = X^T\beta = \eta. \tag{2.27}$$

2.3 Generalized Linear Model

GLM is a generalization of LM. Firstly, while LM describes the mean as a linear function of covariates and regression parameters, GLM uses non-linear functions, too. Secondly, unlike LM, the error term has not to be normally distributed in GLM. Their distribution can be a member of the EDF. Also, LM can only model normally distributed response variable Y, while GLM can model response variable which is a member of the exponential dispersion family. This situation gives a more flexible framework to insurance studies as most of the data in the insurance industry is not normally distributed such as claim amount, claim number, etc.

While linear models are in the form of $\mathbb{E}[Y|X] = X^T\beta$, GLMs are in the form of $g\{\mathbb{E}[Y|X]\} = X^T\beta$; g is a link function, which is monotonic and differentiable. It determines the relation between η, the linear predictor, and μ, the expected value of the response variable y.

$$g(\mathbb{E}[Y|X]) = g(\mu) = \eta = X^T\beta, \tag{2.28}$$

which implies;

$$\mu = g^{-1}(\eta). \tag{2.29}$$

Putting the above-mentioned together, we can show the structure of GLM as follows [15];

$$\mu = f(y) = \exp\Big\{\frac{y\theta - k(\theta)}{a(\phi)} + c(y,\phi)\Big\}, \quad g(\mu) = X^T\beta = \eta. \tag{2.30}$$

Table 2.1: Canonical Link Functions of Distributions

Distribution	Link Function
Normal	identity: $\eta = \mu$
Poisson	log: $\eta = log\mu$
Gamma	reciprocal: $\eta = \mu^{-1}$

2.3.1 Link Function

In LMs, the link function is always the identity function. However, in the case of GLMs, it can be log, logit, inverse, reciprocal, or another function different from identity. This relaxation in conditions of GLM has great benefits. For instance, Poisson distribution deals with counts, a positive data, implying that the mean μ is greater than zero all the time. However, in the case of the identity function, as $-\infty < \eta < \infty$, μ has to lie on the whole real line, as the identity link function implies that $\mu = \eta$. Therefore, the identity link function is not a convenient choice for Poisson distribution. On the other hand, the log-link function allows η to be less than zero while μ is greater than zero since it is defined as $\eta = log\mu$ and implies that $\mu = e^{\eta}$. Therefore, log-link is a more appropriate link function for Poisson distribution than the identity function.

2.3.2 Canonical Link Function

If the link function is chosen as linear predictor η is equal to canonical parameter θ, then it is called the canonical link. Canonical links for Normal, Poisson, and Gamma distributions are given in Table 2.1. The canonical link function of the Normal distribution is the identity function, that of the Poisson distribution is the log function, and that of the Gamma distribution is the reciprocal function.

Although using canonical links for the model is generally intensely plausible in terms of leading the desirable statistical properties, there may be cases in which non-canonical links are more suitable for the model, therefore the most suitable link should be searched without using the canonical link directly.

2.3.3 Maximum Likelihood Estimation

To determine the regression parameters β_i of GLM, one of the most common methods is maximum likelihood estimation (MLE). In this section how it works is explained. As mentioned above, the density function of an EDF member is in the form;

$$f_Y(y_i;\theta,\phi) = \exp\left[\frac{y_i\theta_i - k(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right] \tag{2.31}$$

The likelihood is denoted by;

$$\begin{aligned} L(\theta,\phi;y_i) &= \prod_{i=1}^{n} f_Y(y_i;\theta,\phi) \\ &= \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - k(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right]. \end{aligned} \tag{2.32}$$

By taking the logarithm of likelihood, log-likelihood is obtained;

$$\begin{aligned} l(\theta,\phi;y_i) &= \ln L(\theta,\phi;y_i) \\ &= \ln \prod_{i=1}^{n} f_Y(y_i;\theta,\phi) \\ &= \sum_{i=1}^{n} \ln f_Y(y_i;\theta,\phi) \\ &= \sum_{i=1}^{n} \ln\ exp\left[\frac{y_i\theta_i - k(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right] \\ &= \sum_{i=1}^{n} \left[\frac{y_i\theta_i - k(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right]. \end{aligned} \tag{2.33}$$

Definition 1 (Invariance Property)**.** *If $n \times 1$ vector $\hat{v} = (\hat{v_1}, ..., \hat{v_n})$ is the MLE of the vector $v = (v_1, ..., v_n)$, then $f(\hat{v})$ is the MLE of $f(v)$ for a function f.*

We show that $k'(\theta) = \mu$. Let define a function t as $t(\theta) = k'(\theta) = \mu$. By invariance property, if we know MLE of θ, then we know MLE of $t(\theta) = k'(\theta) = \mu$.

The relationship between μ, η, θ, and β is as follows;

$$\mu = k'(\theta) = t(\theta) = g^{-1}(\eta) = g^{-1}\{X^T\beta\}. \tag{2.34}$$

Then,

$$\theta = t^{-1}(\mu) = t^{-1}[g^{-1}(\eta)] = t^{-1}[g^{-1}(X^T\beta)]. \tag{2.35}$$

From these equations, we conclude that;

$$k(\theta) = k\big(t^{-1}[g^{-1}(X^T\beta)]\big). \tag{2.36}$$

According to the above equations, log-likelihood can be written as;

$$\begin{aligned} l(\beta;\theta,y) &= \sum_{i=1}^{n}\left[\frac{y_i\theta_i - k(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right] \\ &= \sum_{i=1}^{n}\left[\frac{y_i t^{-1}[g^{-1}(X^T\beta)] - k\big(t^{-1}[g^{-1}(X^T\beta)]\big)}{a_i(\phi)} + C(y_i,\phi)\right]. \end{aligned} \tag{2.37}$$

Since the points where the derivative of the function is equal to zero are the maximum and minimum points, regression parameters β_is are found by solving

$$\frac{\partial^2}{\partial\beta_i^2} l(\beta;\theta,y) = 0 \tag{2.38}$$

equation system for all β_is.

A function has its maximum at a point if the second derivative of the function is less than zero at that point. Therefore, equation $\dfrac{\partial^2}{\partial\theta^2} l(\beta;\theta,y) < 0$ has to be satisfied for all β_is as we seek maximums.

2.3.4 Independent-GLM

Let N_i be the claim count, for a policy $i = 1, 2, \ldots, n$, and Y_{ij} be the claim amounts for $j = 1, \ldots, N_i$. Then the aggregate loss is denoted by;

$$S_i = \sum_{j=1}^{N_i} Y_{ij}, \tag{2.39}$$

and mean claim size is denoted by;

$$\bar{Y_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}. \tag{2.40}$$

Equation 2.40 shows that average severity depends on N_i. Then the total loss is;

$$S_i = \sum_{j=1}^{N_i} Y_{ij} = N_i\frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij} = N_i\bar{Y_i}. \tag{2.41}$$

shows that loss cost is the multiplication of the claim severity and claim frequency.

The frequency and severity GLMs are constructed to estimate the expected total loss, $\mathbb{E}[S_i]$. However, to simplify the GLM structure, common practice is assuming the claim frequency, N_i, and mean claim severity, $\bar{Y}_i$, are independent. Under the assumption of independence, the total severity is computed as [15];

$$\mathbb{E}[S_i] = \mathbb{E}[N_i]\mathbb{E}[\bar{Y}_i]. \tag{2.42}$$

This assumption allows us to easily construct and combine frequency and severity GLMs.

Let $X = (x_1, x_2, \ldots, x_n)$ be the covariates, μ_1 and μ_2 be means, g_1 and g_2 be the link functions, β_1 and β_2 be unknown regression parameter vectors for the frequency and severity, respectively. Then, by the Equation 2.28, the marginal GLM for frequency is:

$$g_N(\mathbb{E}[N_i|X_i]) = g_1(\mu_{iN}) = \eta_{iN} = X_{iN}^T\beta_N \quad \Rightarrow \quad \mu_{iN} = g_N^{-1}(X_{iN}^T\beta_N), \tag{2.43}$$

and the marginal GLM for severity is:

$$g_Y(\mathbb{E}[\bar{Y}_i|X_i]) = g_Y(\mu_{iY}) = \eta_{iY} = X_{iY}^T\beta_Y \quad \Rightarrow \quad \mu_{iY} = g_Y^{-1}(X_{iY}^T\beta_Y). \tag{2.44}$$

Then, the total claim cost is:

$$\begin{aligned} \mathbb{E}[S_i|X_i] &= \mathbb{E}[N_i]\mathbb{E}[\bar{Y}_i] \\ &= \mu_{iN}\mu_{iY} \\ &= g_N^{-1}(X_{iN}^T\beta_N)g_Y^{-1}(X_{iY}^T\beta_Y). \end{aligned} \tag{2.45}$$

Taking g as a log-link, we have:

$$\ln(\mathbb{E}[N_i|X_i]) = \ln(\mu_{iN}) = \eta_{iN} = X_{iN}^T\beta_N \quad \Rightarrow \quad \mu_{iN} = \exp(X_{iN}^T\beta_N), \tag{2.46}$$

and

$$\ln(\mathbb{E}[\bar{Y}_i|X_i]) = \ln(\mu_{iY}) = \eta_{iY} = X_{iY}^T\beta_Y \quad \Rightarrow \quad \mu_{iY} = \exp(X_{iY}^T\beta_Y). \tag{2.47}$$

Then the total claim cost is:

$$\begin{aligned} \mathbb{E}[S_i|X_i] &= \mu_{iN}\mu_{iY} \\ &= \exp(X_{iN}^T\beta_N)\exp(X_{iY}^T\beta_Y) \\ &= \exp(X_{iN}^T\beta_N + X_{iY}^T\beta_Y). \end{aligned} \tag{2.48}$$

2.3.5 Dependent-GLM

As introduced earlier, there are different approaches to consider the dependency between claim frequency and claim severity. In this chapter, the approach of Garrido in, [9] a conditional GLM is used. By taking claim count as a covariate in the severity GLM, mean severity is assumed as a function of frequency. Conditional severity is denoted by $\mathbb{E}[\bar{Y}_i|X_i, N_i]$, and the GLM is

$$\begin{aligned} g\{\mathbb{E}[\bar{Y}_i|X_i, N_i]\} &= g\{\tilde{\mu}_{iY}\} \\ &= \hat{X}_{iY}^T\hat{\beta}_Y + N_i\beta_D. \end{aligned} \tag{2.49}$$

Since covariate matrix X_{iY}, mean μ and the regression parameters β_is are different for the dependent-GLM than the independent-GLM, we denote μ with a tilde and the others with hats. Again, taking g as a log-link;

$$\begin{aligned} \ln\{\mathbb{E}[\bar{Y}_i|X_i, N_i]\} &= \ln\{\tilde{\mu}_{iY}\} \\ &= \hat{X}_{iY}^T\hat{\beta}_Y + N_i\beta_D. \end{aligned} \tag{2.50}$$

and the conditional mean severity is;

$$\begin{aligned} \mathbb{E}[\bar{Y}_i|X_i, N_i] &= \exp\{\hat{X}_{iY}^T\hat{\beta}_Y + N_i\beta_D\} \\ &= \exp\{\hat{X}_{iY}^T\hat{\beta}_Y\}\exp\{N_i\beta_D\} \\ &= \hat{\mu}_{iY}\exp\{N_i\beta_D\}. \end{aligned} \tag{2.51}$$

Then, the expected total claim is;

$$\begin{aligned} \mathbb{E}[S_i] &= \mathbb{E}[N_i\bar{Y}_i] \\ &= \mathbb{E}\big[N_i\mathbb{E}[\bar{Y}_i|N_i]\big] \\ &= \mathbb{E}\big[N_i\hat{\mu}_{iY}\exp\{N_i\beta_D\}\big] \\ &= \hat{\mu}_{iY}M'_{N_i}(\beta_D). \end{aligned} \tag{2.52}$$

Here, $M'_{N_i}(\beta_N)$ is the derivative of the moment generating function of N_i at the point β_N.

2.4 Copula

Copulas, firstly introduced by Sklar in [20], are multivariate distribution functions that are used to couple or connect two marginal uniform distribution functions and

are widely used in finance, statistics, economics, etc. to model the dependency between variables. Its first use in actuarial sciences was when Song used Gaussian copula and Kramer et al. used it to link the marginal distributions of claim frequency and claim severity, and construct a joint model [3, 12].

There are two main copula families; the Archimedean family such as Clayton, Gumbel, and Frank copulas, and the elliptical family such as Gaussian and t copulas. Archimedean copulas are generally used for similar and large numbers of variables and have only a single parameter. This parameter expresses the strength of the dependence or the degree of the spread. Archimedean copulas can model only the data that has a positive correlation, and the rotated versions of them can model the negatively correlated data, too. Elliptical copulas are defined for elliptical distributions, and they are radially symmetric [13].

Definition 2 (Copula)**.** *Let* $u = (u_1, u_2)$*, and* $v = (v_1, v_2)$*. Then C is bivariate copula defined from* $[0, 1] \times [0, 1]$ *to* $[0, 1]$*. which satisfies:*

(i) For every u, v in [0, 1]:
C(u, 0) = 0 = C(0, v) and
C(u, 1) = u and C(1, v) = v;

(ii) For every u_1, u_2, v_1, v_2 *in* $[0, 1]$ *such that* $u_1 \leq u_2$ *and* $v_1 \leq v_2$*:*
$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$ *is satisfied.*

Copula can be interpreted as a bivariate cumulative distribution function that couples two marginal functions in one distribution, and with the help of Sklar's theorem, can be defined by the distribution functions F and G instead of u and v.

Theorem 1 (Sklar's Theorem)**.** *Let F and G be the marginal functions, and H be the joint distribution function of them. Then there exists a copula C which satisfies*

$$H(x, y) = C(F(x), G(y)),$$

with $x \in F^{-1}([0, 1])$ *and* $y \in G^{-1}([0, 1])$*. Conversely, If F and G are distribution functions and C is a copula function, then H is a joint distribution function of F and G.*

Table 2.2: Functions of Copula Families

Copula Family	Copula Function $C(u, v\|\theta)$	Range of θ
Gaussian	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v)\|\theta)$	$\theta \in (-1, 1)$
Frank	$-\frac{1}{\theta}\log\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right)$	$\theta \in \mathbb{R}\backslash\{0\}$
Gumbel	$exp\Big\{-\big((-(logu)^{\theta}+(-logv)^{\theta}\big)^{\frac{1}{\theta}}\Big\}$	$\theta \in [1, \infty)$
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-\frac{1}{\theta}}$	$\theta \in (0, \infty)$

In addition, the copula function C is unique if both F and G are continuous. If one or two of them are not continuous, then it is unique in the range of F and G only.

According to the properties of the marginal distributions, different copula families are used to combine them, like Gaussian, Gumbel, Clayton, and Frank. A brief information about these copulas can be found in Table 2.2. Here, θ changes for each copula family, and is related to Kendall's τ of that family, which is an association measure. To choose the most convenient copula family, the properties of data should be taken into account. For instance the dependence structure, the tail behavior, etc.

2.4.1 The Joint Density Function

A copula function is used to link the marginal distributions of the claim number and the mean claim amount, namely the results of the frequency GLM and the severity GLM to get aggregate loss. Although the claim amount is a continuous variable, the claim number is discrete. Therefore, the copula function is arranged to contain a discrete random variable and a continuous random variable.

Let X_1 be a continuous and X_2 be a discrete random variable. The joint distribution $F_{X_1,X_2|\theta}$ is expressed as;

$$F_{X_1,X_2|\theta} = C(F_{X_1}(x_1), F_{X_2}(x_2)|\theta). \tag{2.53}$$

Here, C is the copula which depends on the parameter θ. The joint density function of X_1 and X_2, denoted by $f(x_1, x_2)$, is defined as $\frac{\partial}{\partial x_1} P(X_1 \leq x_1, X_2 = x_2)$ and satisfies;

$$\frac{\partial}{\partial x_1} P(X_1 \leq x_1, X_2 = x_2) \tag{2.54}$$

$$= \frac{\partial}{\partial x_1} P(X_1 \leq x_1, X_2 \leq x_2) - \frac{\partial}{\partial x_1} P(X_1 \leq x_1, X_2 \leq x_2 - 1) \tag{2.55}$$

$$= \frac{\partial}{\partial x_1} C(F_{X_1}(x_1), F_{X_2}(x_2)|\theta) - \frac{\partial}{\partial x_1} C(F_{X_1}(x_1), F_{X_2}(x_2 - 1)|\theta) \tag{2.56}$$

$$= f_{X_1}(x_1)\big[(D_1(F_{X_1}(x_1), F_{X_2}(x_2), \theta)) - D_1(F_{X_1}(x_1), (F_{X_2}(x_2 - 1)|\theta)\big], \tag{2.57}$$

where $D_1(u_1, u_2) = \frac{\partial}{\partial u_1} C(u_1, u_2)$.

2.4.2 Copula-GLM

To connect the claim frequency and severity with a copula function, we propose that their marginal distributions are first modeled via GLM, and then their joint distribution is obtained using a copula function. In Section 2.4.1, the construction of the joint density function for two random variables is explained. To get the dependent copula model for the claim severity and the frequency, continuous random variable X_1 is considered as the claim severity distribution obtained from severity GLM, and discrete random variable X_2 is regarded as the claim frequency distribution obtained from the frequency GLM in Equation 2.53.

$$F_{Y,N|\theta} = C(Gamma\ Severity\ GLM, ZTP\ Frequency\ GLM|\theta) \tag{2.58}$$

After finding the joint density function, the regression parameters are determined using maximum likelihood estimation that is explained in Section 2.3.3. (See [8, 21]).

CHAPTER 3

APPLICATION: MTPL INSURANCE

This chapter presents the application of the models described in the previous chapter to MTPL the claim number and the claim severity insurance data. Firstly in Chapter 3.1, the data and its features are introduced. In 3.2 the independent-GLM is built, and in 3.3 the dependent-GLM is constructed by adding the number of claims as a covariate to independent GLM. Then, a copula function is applied to independent-GLM and get the copula-GLM in 3.4. Finally, the results of these three models are compared in 3.5 and the utilization of the models in premium calculation is discussed in 3.6.

3.1 Data Description

An MTPL insurance data, *dataCar*, is used from the R package *insuranceData* [22], which includes one-year vehicle insurance policies between 2004 and 2005, and due to our best knowledge, there is no study in the literature that employs this dataset for such comparative models.

Data originally consists of 67,856 policies, and all claims are independent of each other. Duplicate rows and the rows whose vehicle value is 0 are removed as they are thought to be input errors, and 67,358 rows are left. The variables that are evaluated to build the independent and dependent GLMs are; *vehicle value, exposure, number of claims, total claim amount, vehicle body type, vehicle age, gender, area,* and *driver age.* In Table 3.1, these variables and their subcategories and intervals can be seen. A

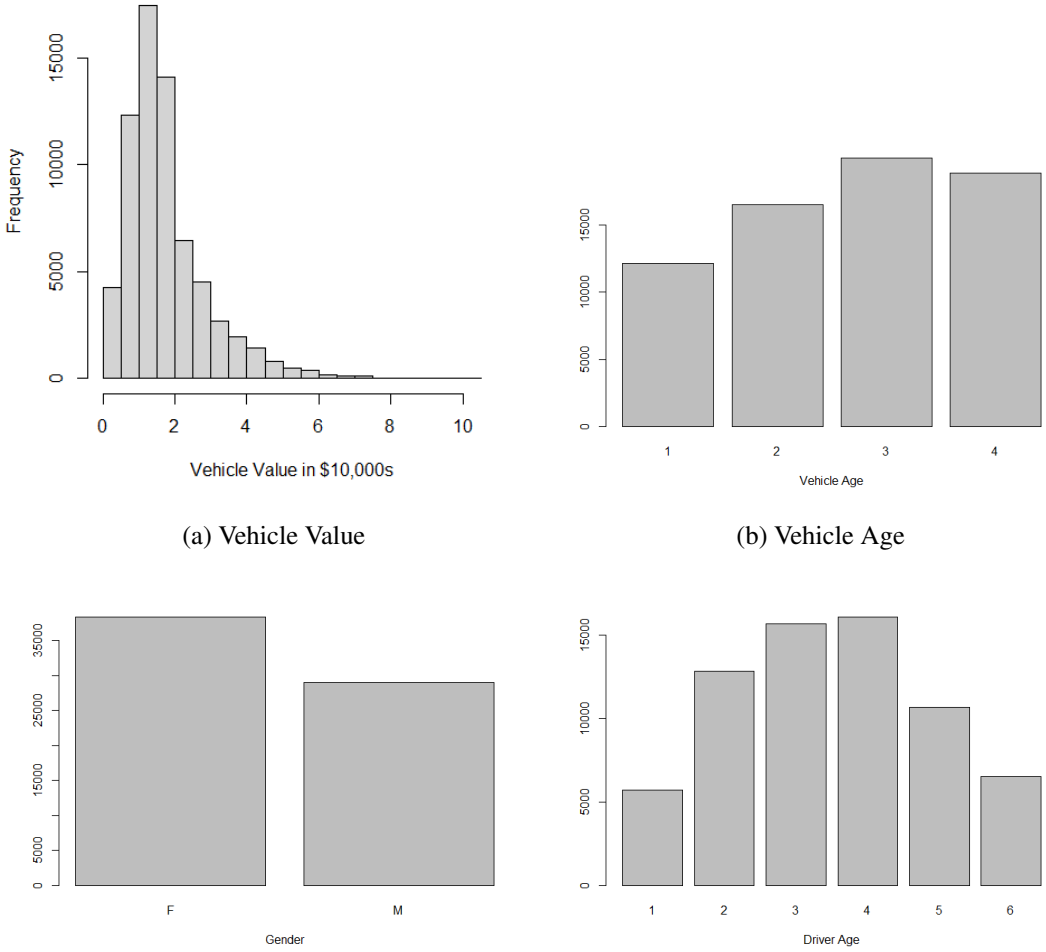
Table 3.1: Variables and Their Categories / Intervals

Variable	**Categories**
Total Claim Amount	0 - 55,922.13
Claim Number	1, 2, 3, 4
Exposure	(0, 1]
Driver Age	1, 2, 3, 4, 5, 6
Vehicle Age	1, 2, 3, 4
Gender	Female, Male
Vehicle Body	BUS, CONVT, COUPE, HBACK, HDTOP, MCARA, MIBUS, PANVN, RDSTR, SEDAN, STNWG, TRUCK, UTE
Area	A, B, C, D, E, F
Vehicle Value	0.18 - 10.360 in unit of 10,000

snapshot of the MTPL dataset is presented in Appendix A.

Exposure and vehicle value are continuous, vehicle body type, vehicle age, gender, area, and driver age are factor covariates. By dividing the total claim amount by the claim number, we calculate the mean claim amount of each policy. The total claim amount varies between $0 and $55,922.13, while the claim number varies between 0 and 4. The average claim amount per policy varies between $0 and $55,922.13, and exposure varies between 0 and 1. Driver age has six, vehicle age has four, gender has two, and vehicle body has thirteen categories, as shown in Table 3.1. Vehicle body types BUS, CONVT, HDTOP, MCARA, MIBUS, PANVN, and RDSTR are banded as *OTHER*, and seven categories are left. In addition, A, B, E categories of area are banded and named *ABE;* and C, D, F categories are banded and named *CDF*. Thus, there are two categories left. These are made by grouping the variables that are too few in number to give significant results in the model. As a result, the subcategories of the data become large enough to give meaningful results in the model and become significant in terms of modeling. When making these bandings, whether the variables have similar characteristics should be taken into consideration. However, since this data set does not disclose this information, we can only consider their numbers. Due to these limitations of the data, our main goal in this study is not constructing the best possible model, but determining the effect of dependency on independent GLM, and the differences between constructing this dependency by only using GLMs, and by

taking advantage of copulas.



(a) Vehicle Value

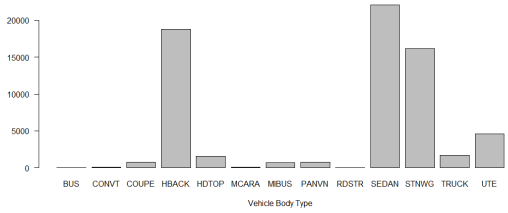
(b) Vehicle Age

(c) Gender

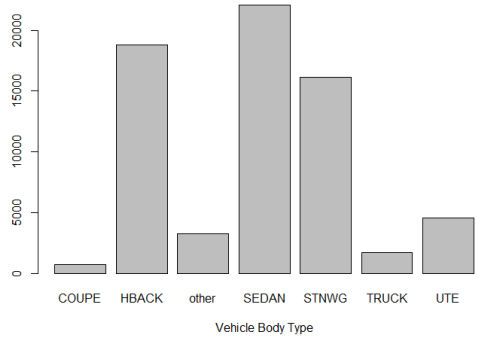
(d) Driver Age

Figure 3.1: Variable Distributions

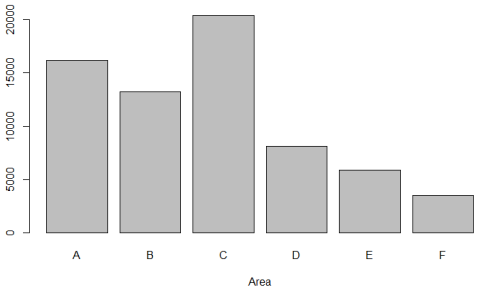
Distributions and descriptive statistics of the variables are presented in the following figures and tables. In Figure 3.1a, we see that the vehicle value distribution is right-skewed, which indicates that most of the portfolio consists of mid-priced vehicles. In Figure 3.1b, all four vehicle age categories contain similar amounts of vehicles. Similarly, there is no striking difference between gender groups regarding number of policies (Figure 3.1c). Finally in Figure 3.1d, we can see that there is a concentration of middle-aged drivers.



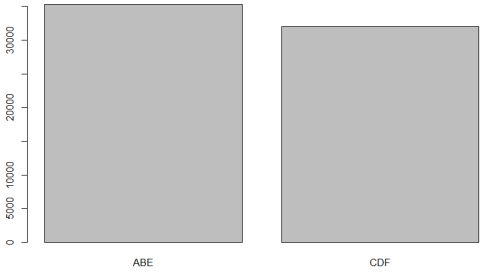
(a) Vehicle Body Type Before Banding



(b) Vehicle Body Type After Banding



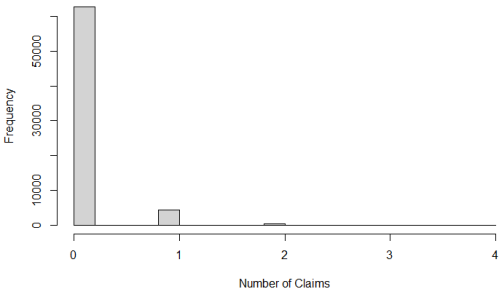
(c) Area Before Banding



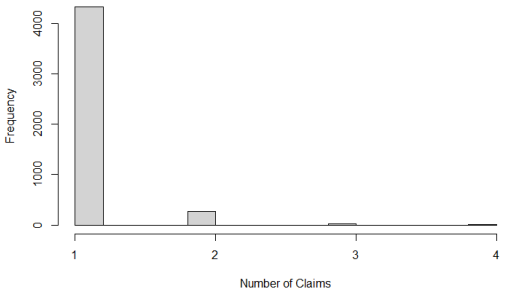
(d) Area After Banding

Figure 3.2: Variable Distributions Before and After Banding

In Figure 3.2a, we can see the distribution of vehicle body types of the original data, and in Figure 3.2b we can see the distribution of vehicle body types after we band the data. In Figure 3.2c, there is the distribution of areas of the original data, while Figure 3.2d has the distribution of areas after the banding procedure.



(a) Claim Number Including Zero Claims



(b) Claim Number Zero Claims Truncated

Figure 3.3: Claim Number Distribution

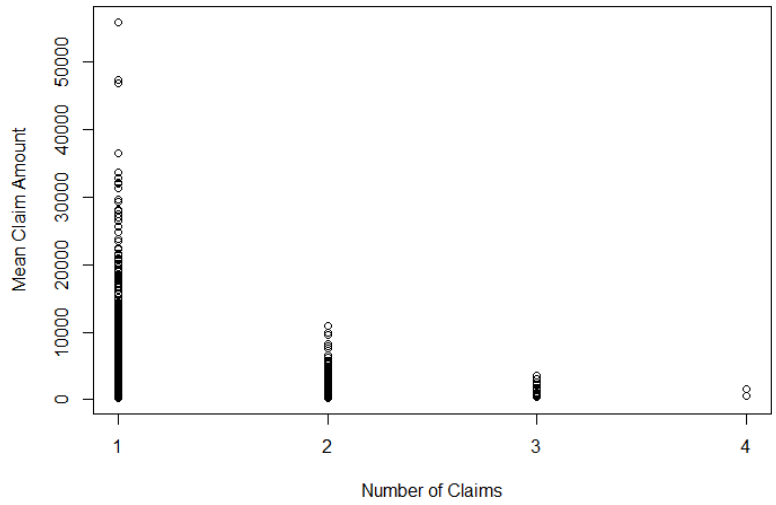


Figure 3.4: Mean Claim Amount per Claim Number

The distribution of claim numbers including the zero-claims can be seen in Figure 3.3a. After excluding the zero-claims, we can see there is an accumulation in claim number 1 and a dramatic decline in the claim number from 1 to 4 (Figure 3.3b).

In Figure 3.4, mean claim amounts per claim number can be seen. We can say that, there is a noticeable decrease in mean claim amount as the number of claims increases. This situation can also be seen with numerical data in Table 3.6. The distribution of the mean claim amount can be seen in Figure 3.5. The data is right-skewed, which means low claim amounts outnumber high claim amounts.

Some descriptive statistics of the data including and excluding zero claims are given in Table 3.2, and in Table 3.3 respectively. Since most of the data is stacked at zero, the first quartile, median, and third quartile of both the claim number and mean claim amount are zero, and means are very low; 0.0731 for the claim number, and $131.86 for the mean claim amount (Table 3.2). When zeros are excluded, the data is stacked at one. In this case, the first quartile, median, and third quartile of the claim number are one, and the mean is 1.0673. For the mean claim amount, the values are again higher compared to data including zero; the first quartile is $353.77, median is $711.51, mean is $1913.259, and the third quartile is $1948.48 (Table 3.3).

The number of policies according to claim numbers and percentages to all data are

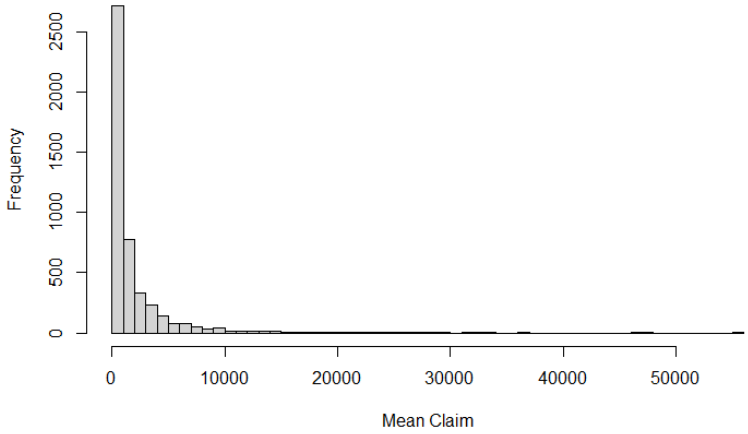


Figure 3.5: Mean Claim Distribution

Table 3.2: Descriptive Statistics of Data Including Zero Claims

Variable	**Claim Number**	**Mean Claim Amount ($)**
Minimum	0	0
1st Quartile	0	0
Median	0	0
Mean	0.0731	131.86
3rd Quartile	0	0
Maximum	4	55,922.13
Variance	0.0778	1,053,308
Skewness	4.0534	18.39

Table 3.3: Descriptive Statistics of Data Excluding Zero Claims

Variable	**Claim Number**	**Mean Claim Amount ($)**
Minimum	1	200
1st Quartile	1	353.77
Median	1	711.51
Mean	1.0673	1913.26
3rd Quartile	1	1948.48
Maximum	4	55,922.13
Variance	0.0733	11,966,112
Skewness	4.342	5.33

Table 3.4: Amounts of Policies per Claim Number

Claim Number	**Number of Policies**	**Percentage**
0	62743	93.15
1	4326	6.42
2	269	0.40
3	18	0.03
4	2	0.00

Table 3.5: Contingency Table

Mean Claim	**Number of Claims**				
Amount ($)	**0**	**1**	**2**	**3**	**4**
0 - 5000	62743	3919	254	18	2
5000 - 10000	0	274	14	0	0
10000 - 15000	0	74	1	0	0
15000 - 20000	0	27	0	0	0
20000 - 25000	0	15	0	0	0
25000 - 30000	0	8	0	0	0
30000 - 35000	0	5	0	0	0
35000 - 40000	0	1	0	0	0
40000 - 45000	0	0	0	0	0
45000 - 50000	0	2	0	0	0
50000 - 55000	0	0	0	0	0
55000 - 60000	0	1	0	0	0

given in the Table 3.4. 93.15% of policies resulted in no claim, and the vast majority of the policies with claim has one claim only. The distribution of the number of policies according to the number of claims and the amount of claim is shown in the Table 3.5. It is clear from the table that majority of the policies have zero claim, and most of the policies whose claim number is different from zero resulted in one claim. Non-zero claim amounts are collected in the range of 0-25000 and concentrated in the range of 0-5000.

Distribution of the mean of the policies claim amounts according to claim numbers are given in Table 3.6. There are 4615 policies with claims, which is 6,85% of the whole data, and as the number of claims increases, the mean amount of claims per claim decreases.

Pearson's and Spearman's correlation coefficients between the claim frequency and severity of the data containing zero claims and the data with zero claims removed are

Table 3.6: Mean Claim Amounts per Claim Number

Claim Number	**Mean of Claim Amounts in $**
0	0
1	1943.214
2	1475.758
3	1341.34
4	1109.745

Table 3.7: Correlation Coefficients of the Data

	Pearson's	**Spearman's**
Data with Zero Claims	0.4815	0.999
Data without Zero Claims	-0.0331	0.045

given in Table 3.7. According to these correlation coefficients, frequency and severity are highly positively correlated in the data with zero claims, and one of the main reasons for this strong correlation is the accumulation of zeros. If we exclude the zeros and consider only positive claim number amounts, correlation coefficients are quite low compared to the ones belonging to data with zero claims. The negative value of Pearson's coefficient indicates that there is a negative correlation between the claim number and the mean claim amount per policy.

Finally, data is randomly separated into two; 80% (53,887 policies) as train data to construct models, and 20% (13,471 policies) as test data to test constructed models.

3.2 Independent-GLM

In the independent model, two GLMs are constructed under the assumption of the independence of the claim frequency and severity; one for claim number N_i, in other words, frequency, and the other one for mean claim amount $\bar{Y}_i$, in other words, severity. Models are constructed in R using function *glm* from the package *stats* [16].

Poisson distribution is used for modeling randomly occurring count data in a time interval. That is why it is a convenient choice for the frequency model; it is the number of claims during the period the policy is valid, which is generally one year. Another requirement for applying Poisson distribution is the independence of obser-

vations, which is correct for claim number data, too. Namely, the number of claims caused by different and independent policies are independent of each other. Besides, Poisson distribution has only one parameter to estimate, which makes it handy to apply. Due to all these features, Poisson distribution is widely used for claim frequency modeling [4, 15].

As the accumulation at zero claims manipulates the data and affects the dependency structure of the model (see Table 3.7), it becomes very difficult to analyze the properties of non-zero claims and construct a model compatible with non-zero claims. Therefore, to better analyze the characteristics of the non-zero data and build a more consistent model, the zero-truncated Poisson distribution that excludes zeros is used for frequency GLM.

Gamma distribution is a skewed, always positive and continuous distribution. These features are extremely suitable for the claim amount data which is highly skewed and has continuous and positive values. Due to this suitability, the Gamma distribution is the most preferred distribution for the severity modeling [4, 15]. In addition, goodness of fit test is made, and the p-value is 0.15 which is greater than 0.05, thus indicates that the hypothesis is valid and the severity data is coherent with the Gamma distribution.

Another common practice is using the log link function for both of the models, as it is the link function that gives the best results, and we prefer it for both of the models [9]. Note that the log link is the canonical link of Poisson distribution, but not Gamma distribution.

The data contains policies with a duration shorter than a policy year, for example, 6 months. This feature in the data set is given with an exposure variable, that represents the duration of the policy in terms of years. For instance; exposure is equal to 1 means that policy is in force for a year, while 0.25 means a quarter of a year, etc. Since a policy with a shorter duration is less likely to cause damage, exposure has to be adjusted while constructing a model for frequency. To make this adjustment, an offset is used for the exposure variable in the frequency model by taking its regression

coefficient 1, to consider it as a fixed effect. In this way, all observations are taken into account in a congruent way. As the severity model is constructed for the mean claim amount per occurrence, $\bar{Y}_i$, the exposure amount does not affect it in a fixed proportion, that is adjusting severity GLM for the exposure by taking it as an offset is not needed.

3.2.1 Frequency GLM

If we denote exposure as e_i, $ln(e_i)$ is the offset term, and GLM for claim frequency with zero-truncated Poisson is constructed as follows:

$$N_i \sim \text{ZTP}(\mu_{iN}) \tag{3.1}$$

$$\begin{aligned} \ln(\frac{\mu_{iN}}{e_N}) &= X_{iN}^T \beta_N \\ \Rightarrow \ln(\mu_{iN}) &= \ln(e_i) + X_{iN}^T \beta_N \\ \Rightarrow \mu_{iN} &= e_i \times \exp(X_{iN}^T \beta_N) \end{aligned} \tag{3.2}$$

Firstly, we construct a model with all the covariates, called full model, as follows;

$$\begin{aligned} \ln(\mathbb{E}[N_i|X_i]) = \ln(e_i) &+ \beta^*_{N,0} + \beta^*_{N,1} vehicle\ value \\ &+ \beta^*_{N,2} vehicle\ body\ type + \beta^*_{N,3} vehicle\ age \\ &+ \beta^*_{N,4} gender + \beta^*_{N,5} area + \beta^*_{N,6} driver\ age \end{aligned} \tag{3.3}$$

As the gender, area, and vehicle age covariates have significance level above %10, they are excluded from the model, and the final model for $ln(\mathbb{E}[N_i|X_i])$ is;

$$\begin{aligned} \ln(\mu_{iN}) = \ln(e_i) &+ \beta_{N,0} + \beta_{N,1} vehicle\ value \\ &+ \beta_{N,2} vehicle\ body\ type + \beta_{N,3} driver\ age \end{aligned} \tag{3.4}$$

which implies that $\mathbb{E}[N_i|X_i]$ is equal to;

$$\begin{aligned} \mu_{iN} = e_i \times \exp\{&\beta_{N,0} + \beta_{N,1} vehicle\ value + \beta_{N,2} vehicle\ body\ type \\ &+ \beta_{N,3} driver\ age\} \end{aligned} \tag{3.5}$$

For the regression parameters, significance levels of covariates and the other coefficients of the full and main model see Appendix B.1, B.2. AIC of the full model

(Equation 3.3) is 1834.914, and the main model (Equation 3.4) is 1827.184. Since the AIC of the main model is less than the AIC of the full model, the main model is more appropriate than the full model. Moreover, the mean frequency estimated by this model is 1.05 while the real mean is 1.068. Detailed review on the results is made in the Section 3.5.

3.2.2 Severity GLM

GLM for the claim severity with Gamma distribution is constructed as follows:

$$\bar{Y_i} \sim \text{Gamma}(\mu_{iY,}, v^2) \tag{3.6}$$

$$\begin{aligned} \ln(\mu_{iY}) &= X_{iY}^T \beta_Y \\ \Rightarrow \mu_{iY} &= \exp(X_{iY}^T \beta_Y) \end{aligned} \tag{3.7}$$

As in the frequency GLM, we first construct a model with all the covariates, which we refer to as the full model, and eliminate the non-significant covariates. The full model is as follows;

$$\begin{aligned} \ln(\mathbb{E}[\bar{Y_i}|X_i]) = \beta^*_{Y,0} &+ \beta^*_{Y,1} vehicle\ value + \beta^*_{Y,2} vehicle\ body\ type \\ &+ \beta^*_{Y,3} vehicle\ age + \beta^*_{Y,4} gender + \beta^*_{Y,5} area \\ &+ \beta^*_{Y,6} driver\ age. \end{aligned} \tag{3.8}$$

After eliminating the covariates whose significance level is greater than 10%; vehicle value, vehicle body type, and vehicle age, we have what we call the main model for $ln(\mu_{iY})$ is equal to;

$$\ln(\mu_{iY}) = \beta_{Y,0} + \beta_{Y,1} gender + \beta_{Y,2} area + \beta_{Y,3} driverage \tag{3.9}$$

which implies that $\mathbb{E}[\bar{Y_i}|X_i]$ is equal to;

$$\mu_{iY} = \exp\{\beta_{Y,0} + \beta_{Y,1} gender + \beta_{Y,2} area + \beta_{Y,3} driver\ age\}. \tag{3.10}$$

For the regression parameters, significance levels of covariates and the other coefficients of the full and main model see Appendix B.3, B.4. The AIC of the full model is 63,712.24, and the main model is 63,717.26. Although the AIC of the full model is slightly lower than the main model, we select the main model with fewer variables to

Table 3.8: Correlation Coefficients for Residual Frequency and Severity

Spearman's	**Pearson's**	**Kendall's**
-0.211	-0.077	-0.17

ease the computation process, since the difference in the AIC is very small. In addition, the mean severity estimated by this model is 133.05 while the real mean severity is 124.46. The models are examined in detail in Section 3.5.

3.2.3 Compound Model

The AIC of the compound model is 65,544.44, found by summing the AICs of the frequency and severity models. The total loss per policy is 141.86 while the real total loss id 131.37. This informationis used to compare this model with dependent GLM and copula models in Section 3.5.

3.2.4 Dependency Between Residuals of Independent-GLM

To investigate the dependence in the variables, we compute the GLM residual values for the frequency and severity models. As can be seen in Table 3.8, there exists a negative correlation between the claim number and the mean claim amount as in the original data. It is clear from here that, after modeling the frequency and the severity independently we still have a dependence on the residual which is not considered by the model.

Since total claim estimates are essentially used for the premium estimates, which directly affect the sustainability of the insurance company, ignoring this dependence in the model may result in less accurate premium estimation. Therefore, this shows that we need to incorporate the dependence in the GLM model and leads to the construction of the dependent model in Section 3.3.

3.3 Dependent-GLM

In this section, we include dependency into the independent-GLM by assuming the claim mean is a function of the claim number. We will provide this by adding the claim number into severity GLM as a covariate. For the same reasons as the independent-GLM, and for ensuring the consistency between the independent-GLM and dependent-GLM, the Poisson distribution is used for the frequency, and the Gamma distribution is used for the severity again. As described in Section 2.3.5, dependency is ensured by only the severity model, therefore, there is no change in the frequency model.

3.3.1 Severity GLM

In this model, the claim number is added to severity GLM as a covariate to construct dependency between frequency and severity.

$$\bar{Y}_i \sim \text{Gamma}(\hat{\mu}_{iY}, v^2) \tag{3.11}$$

$$\begin{aligned} \ln(\hat{\mu}_{iY}) &= \hat{X}_{iY}^T \hat{\beta}_Y + N_i \beta_D \\ \Rightarrow \hat{\mu}_{iY} &= \exp(\hat{X}_{iY}^T \beta_Y + N_i \beta_D) \end{aligned} \tag{3.12}$$

As above, firstly the full model is built, and then significant covariates are chosen to build the main model;

$$\begin{aligned} \ln(\mathbb{E}[\bar{Y}_i|X_i, N_i]) = \hat{\beta}_{Y,0} &+ \beta_D N_i + \hat{\beta}_{Y,1} vehicle\ value \\ &+ \hat{\beta}_{Y,2} vehicle\ body\ type + \hat{\beta}_{Y,3} vehicle\ age \\ &+ \hat{\beta}_{Y,4} gender + \hat{\beta}_{Y,5} area + \hat{\beta}_{Y,6} driver\ age \end{aligned} \tag{3.13}$$

The same covariates are found insignificant as independent-GLM; thus the final model of $\ln(\mathbb{E}[\bar{Y}_i|X_i, N_i])$ is;

$$\ln(\hat{\mu}_{iY}) = \hat{\beta}_{Y,0} + \beta_D N_i + \hat{\beta}_{Y,1} gender + \hat{\beta}_{Y,2} area + \hat{\beta}_{Y,3} driver\ age \tag{3.14}$$

which implies that $\mathbb{E}[\bar{Y}_i|X_i, N_i]$ is equal to;

$$\hat{\mu}_{iY} = \exp\{\hat{\beta}_{Y,0} + \hat{\beta}_{Y,1} gender + \hat{\beta}_{Y,2} area + \hat{\beta}_{Y,3} driver\ age + \beta_D N_i\}. \tag{3.15}$$

In this dependent model, the significance of the number of claims, N_i, is quite high, showing that it is a highly significant covariate supporting the dependence. For the regression parameters, the significance levels of covariates, and the other coefficients of the full and main model see Appendix B.5,B.6. The AIC of the full model is 63,584.05, and the main model is 63,589.28. Although the AIC of the full model is slightly lower than the main model, we can select the simple one as the differences are very few, and using a simple model is more beneficial. The mean severity estimated by this model is 127.67 while the real mean severity is 124.46. for detailed comparison, see Section 3.5.

3.3.2 Compound Model

The AIC of the dependent-GLM is found by summing the AICs of the frequency GLM and the severity GLM, and equal to 65,416.46. The estimated total loss per policy is 136.06 while the real total loss is 131.37. This information is used to compare this model with dependent GLM and copula models in Section 3.5.

3.4 Copula Model

To build the copula-GLM, we use the independent frequency and severity GLMs constructed in Section 2.3.4. To incorporate the copula to these GLMs, we use *CopulaReg* and *VineCopula* packages in R [2, 18].

As mentioned in Section 2.4, there are two different copula families; Archimedean and elliptical. First we have to choose which copula family to use. To decide, the dependence structure between the claim number and the mean claim amount is examined (Figure 3.4) and it is observed that there is a negative relationship between them; as the number of claims per policy increases, the average claim amount decreases. The Archimedean family, namely Gumbel, Frank and Clayton copulas cannot be used with data that has a negative relationship, on the other hand, the elliptical family is a quite convenient choice for this type of data, and Gaussian copula is widely used in claim amount calculations to link claim frequency and severity [3]. In addition,

R function *BiCopSelect* from the package *VineCopula* is used to suggest the most appropriate copula function for the data. It fits 36 different copula functions to the dataset, and the results based on both AIC and BIC shows that, the Gaussian copula is the best fitting one. Therefore, the Gaussian copula is preferred to construct the copula-GLM.

The AIC of the compound model is 65474.82, the mean frequency is 1.066, and the mean severity is 127.72. Finally, the estimated claim per policy is 136.13. This information is used to compare this model with the independent and dependent GLMs.

3.5 Model Evaluation

In this section, the results of the independent-GLM, the dependent-GLM, and the copula-GLM are firstly compared both by AIC values and real data values. Secondly, the means of the frequency model, the severity model, and the compuond model results are compared with the means of the real data, which are shown in Tables 3.10, 3.11, and 3.12. As mentioned in Section 3, we split the data into two parts as 80% of it as train data to construct the models, and 20% of it as the test data to test the accuracy of the model. In this section, we use this test data to apply the models, and calculate the real data results.

Firstly, the AICs of the models are compared. As seen from the Table 3.9, the highest AIC belongs to independent-GLM, which indicates that it is the model with the lowest consistency among them. On the other hand, according to the AIC, dependent-GLM is a slightly better choice compared to the copula-GLM as the AIC of the dependent-GLM is less than the AIC of the copula-GLM (See [14, 1]. The same holds for BIC too. According to BIC, the dependent-GLM is the most accurate model while the second is copula-GLM and the least accurate one is the independent-GLM.

When we compare the mean frequency results, it is seen that the copula-GLM gives the most real-like results, which is 0.19% less than the real data. The frequency estimations for the independent-GLM and dependent-GLM are the same since they have the same frequency GLM, and it is 1.69% less than the real average frequency.

Table 3.9: AIC and BIC of Models

Model	**AIC**	**BIC**
Independent-GLM	65544.44	65681.35
Dependent-GLM	**65416.46**	**65559.60**
Copula-GLM	65474.82	65611.73

Table 3.10: Mean Estimations of Frequency Models

Frequency Model	**Value**	**% of Deviation**
Real Value	1.068	-
GLM	1.05	1.69%
Copula Model	1.066	0.19%

Table 3.11: Mean Estimations of Severity Models

Severity Model	**Value**	**% of Deviation**
Real Value	124.46	-
GLM of Independent Model	133.05	6.9%
GLM of Dependent Model	127.67	2.58%
Copula Model	127.72	2.61%

Table 3.12: Loss Estimation of Models

Compound Model	**Mean Loss**	**Total Loss**	**% of Deviation**
Real Data	131.37	1,769,655	-
Independent-GLM	141.86	1,910,993	7.99%
Dependent-GLM	136.06	1,832,897	3.57%
Copula-GLM	136.14	1,833,887	3.63%

If we look at the mean severity, the dependent-GLM and the copula-GLM give very close results, however dependent-GLM is slightly better than the copula-GLM. The dependent-GLM is 2.58% higher from the real data while the copula-GLM is 2.61% higher. On the other hand, the prediction of the independent-GLM is much higher than the actual data, which is 6.9%.

Finally, if we compare the estimated total claims of these models, again the dependent-GLM and the copula-GLM give fairly close results compared to the independent-GLM, while the dependent-GLM is slightly better than the copula-GLM. Independent-GLM is 7.99%, dependent-GLM is 3.57%, and the copula-GLM is 3.63% higher than the real data.

As all the covariates have different significance levels and coefficients in the three models, the results of the models change. For significance levels and coefficients see Appendix B. According to the results, it is remarkably obvious that the independent model deviates from the real data much more than the other models. Therefore, a dependent model should be preferred. However, as the dependent-GLM and copula-GLM give quite close results, which one performs better would vary according to the dependence structure inherent in the specific dataset to be modeled. Therefore, examining both of them on the dataset to be modeled to see which one gives better results and then proceeding with the more accurate one would be the most convenient method.

3.6 Utilization of Models

In this section, premiums of all the three models are calculated according to the expected value principle and the standard deviation principle, and then the results are compared.

Where P is the pure premium for a policy, the random variable X is the claim size, $\mathbb{E}[X]$ is the expected claim, and θ is the safety loading which is always non-negative.

Table 3.13: Premium Calculations

	Expected Premium Principle	**Standard Deviation Principle**
Independent-GLM	$141.86(1+\theta)$	$141.86+551.01\theta$
Dependent-GLM	$136.06(1+\theta)$	$136.06+520.57\theta$
Copula-GLM	$136.14(1+\theta)$	$136.14+331.28\theta$

The expected value principle for premium is expressed as:

$$P_X = (1+\theta)\mathbb{E}[X],$$

and the standard deviation principle for premium is denoted by:

$$P_X = \mathbb{E}[X] + \theta\sqrt{\mathbb{V}ar[X]}.$$

According to these information, the expected premiums calculated by the expected value principle and the standard deviation principle for the independent-GLM, the dependent-GLM and the copula-GLM is given in Table 3.13.

The ranking of the premiums does not change depending on θ in the expected value principle, on the other hand, the ranking changes depending on theta in the standard deviation principle. For instance, assuming $\theta = 0.01$, the independent-GLM has the highest price with 147.37, the second is the dependent-GLM with 141.27, and the cheapest one is the copula-GLM with 139.45. On the other hand, assuming $\theta = 0.0001$, the independent-GLM is the highest one again which is 141.92, however in this case the second one is the copula-GLM which is 136.17, and the cheapest one is the dependent-GLM with 136.11. Therefore, companies should choose the best pricing model, premium principle and the safety loading combination for them according to the features of their portfolio.

CHAPTER 4

CONCLUSION

The common practice of the insurance industry on non-life insurance pricing is modeling the claim frequency and severity separately using GLM. Assuming these two are independent, the total loss is computed by simply multiplying the estimations of severity and frequency models. Although the independence assumption simplifies the model and enables it to run in a shorter time, it ignores the dependence-sensitive features of the model and it may cause a deviation from an accurate prediction. To prevent this disadvantage, models that take dependence between the claim frequency and severity into account have been proposed. This is needed because determining accurate premiums is becoming increasingly important both for the company's reserve studies and for compliance with criteria such as Solvency and IFRS.

One of these dependent models, dependent-GLM, provides dependence by putting the claim number variable to marginal severity GLM as a covariate of that model. In other words, the amount of claims is a function that varies depending on the number of claims. Furthermore, the significance of the number of claims is quite high in the severity model showing that it is a highly significant covariate supporting the dependence.

The second dependent model that is constructed in this thesis is the copula-GLM. In this model, after constructing the marginal frequency and severity GLMs, they are linked to each other with the help of a copula function, and the new regression coefficients that take into account the dependence for the frequency and severity GLMs are obtained.

By investigating the MTPL insurance data in the application section, zero-truncated Poisson distribution for the frequency GLM, Gamma distribution for the severity GLM, and Gaussian copula for the copula-GLM are used.

The two dependent models and the classic independent-GLM are compared by applying them to MTPL data. The results show that the dependent models are noticeably more accurate than the independent model, which overestimates the total loss. On the other hand, although the dependent-GLM gives a slightly better result than the copula-GLM, no significant difference is observed between these two dependent models. In this case, it would be a more reasonable choice to use the dependent GLM due to its efficiency. As the copula model uses numerical methods, it takes a lot of time to process, on the other hand, application of the dependent GLM approach is much simpler and hence quick.

Nevertheless, the selection and creation of the dependent model to be used depends on the characteristics of the data. It should be taken into account that, insurance data shows different characteristics depending on many variables, such as country, geographical features, customer profile of the insurance company, some features of the insured property, etc. Even if all of these remain constant, the properties of the data vary over time. For instance, the customer profile of the company, or some features of the country like population may change over time. As the characteristics of these variables change, the possible risk amounts also change. Therefore, the premiums of the policies must be updated according to the changing risk amounts. The premiums of policies that have become more risky should be increased, and the premiums of policies that have become less risky should be reduced. Therefore, a pricing model that is compatible with the data today may become incompatible in the future. That's why different pricing models should be provided for different datasets, and also, these models should be tested and revised periodically to continue to fit data and give accurate results.

In this regard, pricing models that revise themselves according to the characteristics and evolutions of the data using machine learning and deep learning techniques

may be further studies.

REFERENCES

[1] D. Anderson and K. Burnham, Model selection and multi-model inference, Second. NY: Springer-Verlag, 63(2020), p. 10, 2004.

[2] S. B. Brant and I. Hobæk Haff, *copulareg: Copula Regression*, 2021, r package version 0.1.0.

[3] C. Czado, R. Kastenmeier, E. C. Brechmann, and A. Min, A mixed copula model for insurance claims and claim sizes, Scandinavian Actuarial Journal, 2012(4), pp. 278–305, 2012.

[4] P. De Jong and G. Z. Heller, *Generalized linear models for insurance data*, Cambridge University Press, 2008.

[5] X. K. Dimakos and A. F. Di Rattalma, Bayesian premium rating with latent structure, Scandinavian Actuarial Journal, 2002(3), pp. 162–184, 2002.

[6] E. W. Frees, J. Gao, and M. A. Rosenberg, Predicting the frequency and amount of health care expenditures, North American Actuarial Journal, 15(3), pp. 377–392, 2011.

[7] E. W. Frees and E. A. Valdez, Understanding relationships using copulas, North American actuarial journal, 2(1), pp. 1–25, 1998.

[8] G. Gao and J. Li, Dependence modeling of frequency-severity of insurance claims using waiting time, Insurance: Mathematics and Economics, 109, pp. 29–51, 2023.

[9] J. Garrido, C. Genest, and J. Schulz, Generalized linear models for dependent frequency and severity of insurance claims, Insurance: Mathematics and Economics, 70, pp. 205–215, 2016.

[10] S. Gschlößl and C. Czado, Spatial modelling of claim frequency and claim size in non-life insurance, Scandinavian Actuarial Journal, 2007(3), pp. 202–225, 2007.

[11] B. Jørgensen and M. C. Paes De Souza, Fitting tweedie's compound poisson model to insurance claims data, Scandinavian Actuarial Journal, 1994(1), pp. 69–93, 1994.

[12] N. Krämer, E. C. Brechmann, D. Silvestrini, and C. Czado, Total loss estimation using copula-based regression models, Insurance: Mathematics and Economics, 53(3), pp. 829–839, 2013.

[13] J.-F. Mai and M. Scherer, *Simulating copulas: stochastic models, sampling algorithms, and applications*, volume 4, World Scientific, 2012.

[14] P. A. Murtaugh, In defense of p values, Ecology, 95(3), pp. 611–617, 2014.

[15] E. Ohlsson and B. Johansson, *Non-life insurance pricing with generalized linear models*, volume 174, Springer, 2010.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024.

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

[18] U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, T. Nagler, T. Erhardt, C. Almeida, A. Min, C. Czado, M. Hofmann, et al., Package 'vinecopula', R package version, 2(5), 2015.

[19] P. Shi, X. Feng, and A. Ivantsova, Dependent frequency–severity modeling of insurance claims, Insurance: Mathematics and Economics, 64, pp. 417–428, 2015.

[20] A. Sklar, Functions de repartition and dimensions et leurs marges. l'institut de statistique de l'université de paris, Paris, France, 1959.

[21] P. X.-K. Song, M. Li, and Y. Yuan, Joint regression analysis of correlated data using gaussian copulas, Biometrics, 65(1), pp. 60–68, 2009.

[22] A. Wolny-Dominiak and M. Trzesiok, *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance.*, 2014, r package version 1.0.

[23] P. Xue-Kun Song, Multivariate dispersion models generated from gaussian copula, Scandinavian Journal of Statistics, 27(2), pp. 305–320, 2000.

APPENDIX A

SNAPSHOT OF THE DATA

A snapshot of the original data is presented below:

	X	veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat	X_OBSTAT_	veh_value_cat
1	1	1.060	0.30390144	0	0	0.0000	HBACK	3	F	CDF	2	01101 0 0 0	2
2	2	1.030	0.64887064	0	0	0.0000	HBACK	2	F	ABC	4	01101 0 0 0	2
3	3	3.260	0.56947296	0	0	0.0000	UTE	2	F	ABC	2	01101 0 0 0	5
4	4	4.140	0.31759069	0	0	0.0000	STNWG	2	F	CDF	2	01101 0 0 0	5
5	5	0.720	0.64887064	0	0	0.0000	HBACK	4	F	CDF	2	01101 0 0 0	1
6	6	2.010	0.85420945	0	0	0.0000	other	3	M	CDF	4	01101 0 0 0	4
7	7	1.600	0.85420945	0	0	0.0000	other	3	M	ABC	4	01101 0 0 0	3
8	8	1.470	0.55578371	0	0	0.0000	HBACK	2	M	ABC	6	01101 0 0 0	3
9	9	0.520	0.36139630	0	0	0.0000	HBACK	4	F	ABC	3	01101 0 0 0	1
10	10	0.380	0.52019165	0	0	0.0000	HBACK	4	F	ABC	4	01101 0 0 0	1
11	11	1.380	0.85420945	0	0	0.0000	HBACK	2	M	ABC	2	01101 0 0 0	3
12	12	1.220	0.85420945	0	0	0.0000	HBACK	3	M	CDF	4	01101 0 0 0	2
13	13	1.000	0.49281314	0	0	0.0000	HBACK	2	F	CDF	4	01101 0 0 0	2
14	14	7.040	0.31485284	0	0	0.0000	STNWG	1	M	ABC	5	01101 0 0 0	5
15	16	2.350	0.39151266	0	0	0.0000	SEDAN	2	M	CDF	4	01101 0 0 0	4
16	17	1.510	0.99383984	1	1	806.6100	SEDAN	3	F	CDF	4	01101 0 0 0	3
17	18	0.760	0.53935661	1	1	401.8055	HBACK	3	M	CDF	4	01101 0 0 0	1
18	20	0.890	0.59411362	0	0	0.0000	HBACK	3	F	CDF	3	01101 0 0 0	1
19	21	1.950	0.59411362	0	0	0.0000	HBACK	1	M	ABC	1	01101 0 0 0	4
20	22	0.390	0.53661875	0	0	0.0000	SEDAN	4	M	CDF	5	01101 0 0 0	1
21	24	1.370	0.59137577	0	0	0.0000	HBACK	1	F	ABC	1	01101 0 0 0	3

Figure A.1: Snapshot of the MTPL Dataset

APPENDIX B

COEFFICIENTS OF GLMS

In this appendix, the coefficient values of GLMs that we construct are given via the snapshots from R.

B.1 Frequency: Full Model

Table B.1: Coefficients of the Frequency GLM: Full Model

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.05090    0.54759  -1.919   0.0550 .
areaCDF       -0.14433    0.12690  -1.137   0.2554
genderM       -0.08688    0.13238  -0.656   0.5116
veh_age2       0.09244    0.19919   0.464   0.6426
veh_age3       0.03247    0.21895   0.148   0.8821
veh_age4       0.15149    0.26016   0.582   0.5604
veh_value      0.03188    0.08714   0.366   0.7145
veh_bodyHBACK -0.78413    0.43782  -1.791   0.0733 .
veh_bodyother -0.68568    0.48675  -1.409   0.1589
veh_bodySEDAN -0.24022    0.41855  -0.574   0.5660
veh_bodySTNWG -0.52987    0.42824  -1.237   0.2160
veh_bodyTRUCK -0.23962    0.52255  -0.459   0.6466
veh_bodyUTE   -0.70240    0.50223  -1.399   0.1619
driver_age2    0.01266    0.24440   0.052   0.9587
driver_age3   -0.02930    0.24007  -0.122   0.9029
driver_age4    0.03120    0.23722   0.132   0.8954
driver_age5   -0.34775    0.28468  -1.222   0.2219
driver_age6   -0.08838    0.30221  -0.292   0.7699
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Name of linear predictor: loglink(lambda)

Log-likelihood: -899.457 on 3709 degrees of freedom

Number of Fisher scoring iterations: 7
```

Although most of the variables are not significant in the 10% level or below, pricing insurance policies, past industry knowledge and expert opinion are also quite important. In light of general knowledge of the insurance industry and common pricing

studies, we decide to exclude area, gender, and vehicle age from the full frequency model. The AIC of the model is 1834.914 (Table B.1, Figure B.1).

```
> AIC(ztpoissonfull)
[1] 1834.914
```

Figure B.1: AIC of the Frequency GLM: Full Model

B.2 Frequency: Main Model

Table B.2: AIC of the Frequency GLM: Main Model

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.00342    0.46209  -2.171   0.0299 *
veh_value      0.00372    0.06491   0.057   0.9543
veh_bodyHBACK -0.81576    0.42637  -1.913   0.0557 .
veh_bodyother -0.70038    0.48651  -1.440   0.1500
veh_bodySEDAN -0.27234    0.41294  -0.660   0.5096
veh_bodySTNWG -0.54719    0.42717  -1.281   0.2002
veh_bodyTRUCK -0.31136    0.52024  -0.598   0.5495
veh_bodyUTE   -0.74437    0.50126  -1.485   0.1375
driver_age2    0.02352    0.24428   0.096   0.9233
driver_age3   -0.01925    0.23960  -0.080   0.9360
driver_age4    0.03715    0.23690   0.157   0.8754
driver_age5   -0.34169    0.28440  -1.201   0.2296
driver_age6   -0.08636    0.30193  -0.286   0.7749
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Name of linear predictor: loglink(lambda)

Log-likelihood: -900.5918 on 3714 degrees of freedom

Number of Fisher scoring iterations: 7
```

```
> AIC(ztpoisson)
[1] 1827.184
```

Figure B.2: AIC of the Frequency GLM: Main Model

What remains in the main frequency model is the vehicle value, vehicle body type and driver age. The AIC of the model is 1827.184. (See Table B.2, Figure B.2).

B.3 Severity: Full Model

In the full severity model, two of the four vehicle age categories, vehicle value, and all the vehicle body categories have more than 10% significance level, hence they are excluded from the model. The AIC of the model is 63,712.24. (See Table B.3).

Table B.3: Coefficients of the Severity GLM: Full Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.67682    0.30038  25.557  < 2e-16 ***
genderM        0.17453    0.06299   2.771  0.00562 **
areaCDF        0.12058    0.05999   2.010  0.04451 *
driver_age2   -0.25758    0.11332  -2.273  0.02308 *
driver_age3   -0.32095    0.11024  -2.911  0.00362 **
driver_age4   -0.32610    0.11015  -2.960  0.00309 **
driver_age5   -0.39502    0.12365  -3.195  0.00141 **
driver_age6   -0.30309    0.14190  -2.136  0.03275 *
veh_age2       0.11552    0.09336   1.237  0.21601
veh_age3       0.16825    0.10242   1.643  0.10052
veh_age4       0.22554    0.12337   1.828  0.06760 .
veh_value      0.03112    0.04368   0.713  0.47615
veh_bodyHBACK -0.10623    0.25462  -0.417  0.67655
veh_bodyother -0.15854    0.27291  -0.581  0.56134
veh_bodySEDAN -0.28615    0.25110  -1.140  0.25454
veh_bodySTNWG -0.21528    0.25308  -0.851  0.39504
veh_bodyTRUCK -0.15350    0.30617  -0.501  0.61614
veh_bodyUTE   -0.22204    0.27514  -0.807  0.41973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.301248)

    Null deviance: 6102.8  on 3726  degrees of freedom
Residual deviance: 5971.7  on 3709  degrees of freedom
AIC: 63712

Number of Fisher Scoring iterations: 8
```

B.4 Severity: Main Model

In the main severity model, gender, area and driver age are remained, and all the categories have significance levels between 0% and 10%. The AIC of the model is 63,717.26. (See Table B.4).

Table B.4: Coefficients of the Severity GLM: Main Model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.67978    0.09882  77.717  < 2e-16 ***
genderM      0.18538    0.06028   3.075 0.002117 **
areaCDF      0.11331    0.05944   1.906 0.056683 .
driver_age2 -0.25531    0.11228  -2.274 0.023033 *
driver_age3 -0.31822    0.10903  -2.919 0.003537 **
driver_age4 -0.33443    0.10881  -3.073 0.002132 **
driver_age5 -0.40917    0.12228  -3.346 0.000827 ***
driver_age6 -0.33015    0.14017  -2.355 0.018558 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.274253)

    Null deviance: 6102.8  on 3726  degrees of freedom
Residual deviance: 6004.4  on 3719  degrees of freedom
AIC: 63717

Number of Fisher Scoring iterations: 6
```

B.5 Conditional Severity: Full Model

In the full conditional severity model, one of the four vehicle age categories, vehicle value, and all the vehicle body categories have more than 10% significance level, hence they are excluded from the model. The AIC of the model is 63,584.05. (See Table B.5).

Table B.5: Coefficients of the Conditional Severity GLM: Full Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.65898    0.31419  27.560  < 2e-16 ***
numclaims     -0.89959    0.10796  -8.333  < 2e-16 ***
genderM        0.17410    0.06104   2.852 0.004365 **
areaCDF        0.11193    0.05814   1.925 0.054279 .
driver_age2   -0.25897    0.10981  -2.358 0.018403 *
driver_age3   -0.31857    0.10682  -2.982 0.002879 **
driver_age4   -0.33242    0.10674  -3.114 0.001857 **
driver_age5   -0.40873    0.11982  -3.411 0.000654 ***
driver_age6   -0.30558    0.13750  -2.222 0.026313 *
veh_age2       0.12141    0.09046   1.342 0.179650
veh_age3       0.17916    0.09925   1.805 0.071121 .
veh_age4       0.23604    0.11954   1.975 0.048398 *
veh_value      0.03199    0.04232   0.756 0.449754
veh_bodyHBACK -0.16534    0.24679  -0.670 0.502917
veh_bodyother -0.21855    0.26449  -0.826 0.408673
veh_bodySEDAN -0.32257    0.24331  -1.326 0.185007
veh_bodySTNWG -0.26239    0.24526  -1.070 0.284769
veh_bodyTRUCK -0.18044    0.29667  -0.608 0.543080
veh_bodyUTE   -0.26778    0.26666  -1.004 0.315331
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.099489)

    Null deviance: 6102.8  on 3726  degrees of freedom
Residual deviance: 5803.9  on 3708  degrees of freedom
AIC: 63584

Number of Fisher Scoring iterations: 7
```

B.6 Conditional Severity: Main Model

In the main conditional severity model, number of claims, area, gender and driver age are remained, and all the categories have significance levels are between 0% and 10%. The AIC of the model is 63,589.28. (See Table B.6).

Table B.6: Coefficients of the Conditional Severity GLM: Main Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.61909    0.14903  57.836  < 2e-16 ***
numclaims   -0.89794    0.10717  -8.379  < 2e-16 ***
areaCDF      0.10510    0.05752   1.827 0.067762 .
genderM      0.18669    0.05833   3.201 0.001383 **
driver_age2 -0.25316    0.10865  -2.330 0.019856 *
driver_age3 -0.31293    0.10550  -2.966 0.003036 **
driver_age4 -0.33791    0.10530  -3.209 0.001343 **
driver_age5 -0.41888    0.11832  -3.540 0.000405 ***
driver_age6 -0.33015    0.13564  -2.434 0.014979 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.065655)

    Null deviance: 6102.8  on 3726  degrees of freedom
Residual deviance: 5836.1  on 3718  degrees of freedom
AIC: 63589

Number of Fisher Scoring iterations: 6
```

B.7 Frequency after Copula Application

The regression coefficients for the frequency model obtained from the copula-GLM model are shown in the Table B.7.

Table B.7: Coefficients of the Frequency Copula-GLM

Name	Type	Value
alpha0	double [8]	7.707 0.187 0.110 -0.254 -0.318 -0.337 ...
R(Intercept)	double [1]	7.706691
RgenderM	double [1]	0.1867579
RareaCDF	double [1]	0.1097192
Ragecat2	double [1]	-0.253818
Ragecat3	double [1]	-0.3179274
Ragecat4	double [1]	-0.3367526
Ragecat5	double [1]	-0.417932
Ragecat6	double [1]	-0.3324036

B.8 Severity after Copula Application

The regression coefficients for the severity model obtained from the copula-GLM model are shown in the Table B.8.

Table B.8: Coefficients of the Severity Copula-GLM

Name	Type	Value
beta0	double [13]	-1.6913 0.0223 -0.7448 -0.5514 -0.1963 -0.4818 ...
S(Intercept)	double [1]	-1.691277
Sveh_value	double [1]	0.02230208
Sveh_bodyHBACK	double [1]	-0.7448034
Sveh_bodyother	double [1]	-0.5514142
Sveh_bodySEDAN	double [1]	-0.1962924
Sveh_bodySTNWG	double [1]	-0.4817711
Sveh_bodyTRUCK	double [1]	-0.1375811
Sveh_bodyUTE	double [1]	-0.6544267
Sagecat2	double [1]	0.122266
Sagecat3	double [1]	0.06561149
Sagecat4	double [1]	0.1231702
Sagecat5	double [1]	-0.2357592
Sagecat6	double [1]	0.05243853