Stockholms universitet

A Comparison between Methods for Performing Tariff Analysis on Auto-Insurance

Emil Erikson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:16
Matematisk statistik
September 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2023:16**
http://www.math.su.se

A Comparison between Methods for Performing Tariff Analysis on Auto-Insurance

Emil Erikson*

September 2023

Abstract

In this thesis, we study two different approaches to setting premiums for non-life insurance policies using a data set of motor third liability insurance from France. In the first approach, we use a subclass of generalized linear models, with the response distribution belonging to a class of exponential dispersion models, to analyse the frequency at which claims arrive and the average severity of a claim separately, to later use these two models to get the relatives of the premium. In the second approach, we use another subclass of GLMs called Tweedie models, specifically Tweedie models defined by having a variance function exponent $1 < p < 2$, as they are obtained from a theoretical model of the pure premium.

Models are then selected based on forward selection and backward elimination, and later also investigated through cross-validation, before we calculate the relatives of the insurance tariff or the chosen model.

The purpose of this thesis is to see how the different approaches' results differ and discuss the reasons for our findings.

We find that using a separate analysis of frequency and severity of the claims gave us more insight into what impact the different parameters have on the outcome. We also discuss a few ways of improving our analysis, among which are the fact that using the separate approach gives us more flexibility when it comes to the decision of which distributions to use.

For these reasons, we concluded that a separate analysis of the separate frequency-severity method is preferable to a Tweedie model with variance function exponent $1 < p < 2$ when deciding relatives for the pure premium.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: emilerikson@outlook.com. Supervisor: Ola Hössjer and Mohamed El Khalifi.

Sammanfattning

I den här uppsatsen studerar vi två olika metoder för att ta fram premier för sakförsäkringar. Vi använder oss av ett dataset innehållande bilförsäkringar från Frankrike.

I den första metoden använder vi oss av en underklass av generaliserade linjära modeller (GLM), med en responsfördelning som erhålls från en klass av exponentiella dispersionsmodeller (EDM), för att analysera skadefrekvensen och den genomsnittliga kostnaden för en skada separat, för att sedan ta fram relativiteten mellan premierna utifrån dessa två modeller. I den andra metoden använder vi oss av en annan underklass av GLMer som kallas Tweedie-modeller, specifikt Tweedie-modeller som definieras av en variansfunktion med exponent $1 < p < 2$, eftersom dessa beskriver en teoretisk model för premien.

En modell väljs ut baserat på framåtinkludering och bakåtelimination och dess prediktiva förmåga studeras även senare genom korsvalidering, innan vi beräknar relativiteterna för premierna för vår valda modell och för motsvarande försäkringstabell.

Syftet med denna uppsats är att se hur resultaten av de olika metoderna skiljer sig åt och diskutera orsakerna till detta.

Vi kommer fram till att den separata analysen av skadefrekvensen och det genomsnittliga skadebelopp gav oss mer insikt av hur de olika parametrarna påverkar resultatet. Vi diskuterar även vissa sätt man skulle kunnat förbättra analysen. En av dessa sätt är att en separat analys ger mer flexibilitet i hur vi kan bestämma vilka fördelningar som används. En Tweedie model med $1 < p < 2$ använder sig strikt av Poisson- och Gamma fördelningar. Om någon av dessa inte stämmer för den data som används kommer Tweedie modellen inte kunna användas, och man måste därmed använda sig av en annan metod. Däremot om vi använder den separata analysmetoden kommer vi kunna testa vilka fördelningar som passar vår data bäst, för att sedan använda den fördelning som ger oss det bästa resultatet. På grund av dessa två anledningar drar vi slutsatsen att en separat analys av skadefrekvensen och genomsnittlig skadestorlek är en bättre metod för att beräkna relativiteterna för premier i en försäkringstabell.

Acknowledgements

I want to thank my supervisors Ola Hössjer and Mohamed El Khalifi for all of the feedback and help I got during the writing of this paper, and for continuing to help me way past what was suppose to be the due date of the thesis.

Contents

1 Introduction

An insurance policy is a contract between the policyholder and the insurer, where the policyholder pays the insurer in return for economic protection. The insurer will give a payout if the policyholder suffers losses of different kinds, specified for the insurance. To be able to cover these insurance claims, the insurer charges a premium based on the probability that the policyholder suffers a loss.

There have been many different methods created for premium calculations. One of these uses generalized linear models for the claims frequency, the rate at which the insurance claims are reported, and the claim severity, the average claim cost. These two factors can then be used to calculate the premium cost.

Another method is to use a subclass of generalized linear models, called Tweedie models, to model the distribution of the premium directly.

These two methods each have their upsides and downsides and in this thesis we will perform a tariff analysis on auto-insurance data to compare the two approaches. The purpose is to investigate how the methods perform against one another and hopefully get some understanding of why we might use one over another.

2 Theory

2.1 Modeling assumptions

In order to begin discussing the theory behind pricing insurance policies, and later create a model, we start by making a few basic assumptions.

Assumption 1 (Independent policies)**.** *For the responses X_i of n different policies, $i = 1, ..., n$, it is assumed that $X_1, ..., X_n$ are independent.*

Assumption 2 (Independent time intervals)**.** *For the responses X_i of n different disjunct time intervals, $i = 1, ..., n$, it is assumed that $X_1, ..., X_n$ are independent.*

Assumption 3 (Homogeneity)**.** *For any two responses of two policies within the same tariff cell, the two responses have the same probability distribution.*

2.2 Key ratios

The theory behind non-life insurance mathematics uses rating factors, which can be seen as covariates, to divide data into tariffs. Examples of rating factors can be the age of the policyholder, and for car insurance the mileage or model of the car. Within these different tariffs we use our collected data, more specifically a response X and an exposure w, to calculate key ratios $Y = X/\omega$. An example of these variables is the *number of claims* as a response, *duration* as exposure and *claim frequency* as the key ratio. An example of a tariff can be seen in Table 1.

Table 1: Example tariff

Tariff cell i	Age	Gender	Response	Exposure	Key ratio
1	Young	Male	$X_{Y,M} = X_1$	w_1	Y_1
2	Young	Female	$X_{Y,F} = X_2$	w_2	Y_2
3	Adult	Male	$X_{A,M} = X_3$	w_3	Y_3
4	Adult	Female	$X_{A,F} = X_4$	w_4	Y_4

When pricing non-life insurance policies, the three main key ratios used are *claim frequency, claim severity* and *pure risk premium.* They are illustrated in Table 2.

The above-mentioned three ratios are of special interest due to claim frequency and claim severity multiplying together to get the pure premium of a policy. Notice however that the pure premium in a way is the only ratio that we care about in the end, as this is what roughly informs us about the premium. However, only looking at the pure premium neglects other factors that effect our price, e.g. *profit margin, cost of business* and *financial gains on investment.* Because of this, the way policies are usually priced are through the business looking at different factors, as seen above, and then setting the price of a base

Table 2: Three important key ratios for setting premiums.

Response X	Exposure w	Key ratio $Y = X/w$
Number of claims	Duration (time period)	Claim Frequency
Total claim amount	Number of claims	Claim severity
Total claim amount	Duration	Pure premium

value. The policies are then priced in relation to each other, through something called relatives, as will explained below.

A final note about how to define the expectation and variance of the responses and key ratios. Because of the modeling assumptions, if X is a sum of ω responses $Z_i,\ i = 1, ..., \omega$, then Assumptions 1-3 imply that the Z_is are independent and identically distributed (i.i.d.), so that we can define $E[Z_i] = \mu$ and $Var(Z_i) = \sigma^2$. We can then derive expressions for the expectation and variance of our response X and key ratio Y:

$$\begin{aligned} E[X] &= \omega\mu, & E[Y] &= \mu, \\ Var(X) &= \omega\sigma^2, & Var(Y) &= \frac{\sigma^2}{\omega}. \end{aligned} \tag{2.1}$$

2.3 Multiplicative model

The multiplicative model, which is the most commonly used model in insurance mathematics, is defined by expected key factors

$$\mu_{i_1,...,i_M} = \gamma_0 \gamma_{1i_1} \dots \gamma_{Mi_M}. \tag{2.2}$$

If we have M rating factors and $m_k,\ k = 1, ..., M$, classes for the kth rating factor, we can denote each tariff cell by $i = (i_1, ..., i_M)$. We can also denote the response as $X_{i_1,...,i_M}$, the exposure as $\omega_{i_1,...,i_M}$ and the key ratio as $Y_{i_1,...,i_M}$. See Table 1 for an example.

We also have $E[Y_{i_1,...,i_M}] = \mu_{i_1,...,i_M}$ where each $\mu_{i_1,...,i_M}$ is defined as above, with γ_0 a base value and $\gamma_{ki_k},\ i_k = 1, ..., m_k$ the relatives. That is, if a cell has $\gamma_{11} = 1$ and another cell has $\gamma_{12} = 2$, then the second class of the first factor changes the expected value to twice the amount of the first class. This makes it easy to adjust the tariff prices just using γ_0 as the base value and then scaling that price in different cells using the relatives. However, if one uses (2.2) without restrictions, the model will be over-parameterized. This can be seen if we multiply a relative or rating factor of a cell by a value $c \in \mathbb{R},\ c \neq 0$, and divide another rating factor of this cell by the same value c, since then the cells' expected key value will remain the same. Hence, we will need to choose a base cell $(i_1, ..., i_M)$ where

$$\gamma_{1i_1} = \gamma_{2i_2} = \dots = \gamma_{Mi_M} = 1. \tag{2.3}$$

Now all other relatives can be described in relation to the base cell. The way this is done in practice is to take the cell with the largest exposure to be the base cell.

2.4 Generalized linear models

The way we will create our expected key factors is through the use of generalized linear models (GLM). These models consist of three components; one random component, one systematic and a link function as shown in chapter 4.1.1 of [Agr13]. The random component is the distribution of a response variable Y, which has independent observations $(y_1, \dots, y_n)$ with a probability density function, or a probability mass function, of the form

$$f_{Y_i}(y_i; , \theta_i) = A(\theta_i)B(y_i)\exp\left[y_i Q(\theta_i)\right]. \tag{2.4}$$

The distribution is assumed to be part of the natural exponential family of distributions, with independent observations y_i, belonging to different cells $i = 1, ..., N$. $Q(\theta_i)$ is known as the natural parameter. An example of a distribution from the exponential family is the Poisson-distribution with the probability mass function

$$f_{Y_i}(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}. \tag{2.5}$$

This is an exponential family distribution with $A(\lambda) = e^{-\lambda}$, $B(y) = 1/y!$ and natural parameter $Q(\lambda) = \log(\lambda)$.

For a GLM the systematic component relates a vector $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)$ to the explanatory variables $x_{i,j},\ i = 1, ..., n,\ j = 1, ..., q$ through the linear transformation

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{2.6}$$

for some parameter vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)$ with unknown parameters. For a tariff with M factors we have that the number of regression parameters is

$$q = 1 + \sum_{k=1}^{M}(m_k - 1). \tag{2.7}$$

The first of these q parameters corresponds to the base cell, whereas $m_k - 1$ of the remaining parameters correspond to the non-baseline classes of factor k, for $k = 1, \dots, M$. Each row of $\boldsymbol{X}$ identifies a tariff cell, with a first component 1 corresponding to the intercept or the baseline cell, whereas the remaining $q - 1$ components are binary indicator variables that identify the non-baseline classes of the M factors.

Finally, the link function

$$g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i) \tag{2.8}$$

links the expectation of the response variable, $\mu_i = E[Y_i]$, to a linear combination η_i of the a linear combination η of the explanatory variables $x_{i1}, \ldots, x_{iq}$ that correspond to cell i. Two important link functions are the identity link, defined by $g(\mu_i) = \mu_i$ and the canonical link defined by $g(\mu_i) = Q(\theta_i)$.

2.5 Exponential dispersion models

For some applications of GLMs, one might need to use a distribution that requires more than one parameter. To be able to do this, we introduce another parameter known as the *dispersion parameter* ϕ, now referring to $\theta = Q$ as the *natural parameter*. We can now write the random component as

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]. \tag{2.9}$$

Note in particular that if the dispersion parameter is known then (2.9) reduces to (2.4), with $Q = \theta_i/a(\phi)$, $A = \exp(-b(\theta_i)/a(\phi)$ and $B = \exp(c(y_i, \phi))$. Usually when different observations have different weights, we define $a(\phi) = a_i(\phi) = \phi/\omega_i$ where ω_i is the weight or exposure of observation i, giving us

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi, \omega_i)\right]. \tag{2.10}$$

The term $b(\theta_i)$ is known as the cumulant function and it is assumed to be twice continuously differentiable, with invertible first derivative, see chapter 2.1 of [OJ10].

Some restrictions are $\phi > 0$, $\omega_i \geq 0$ and that the parameter space for θ_i must be open.

To make it easier to group policies into tariff cells, we introduce the following theorem.

Theorem 1. *Suppose we have two independent random variables Y_1 and Y_2, with distributions that belong to the same exponential dispersion model (EDM) family, with the same μ and ϕ. If these two variables have weights ω_1 and ω_2 respectively, then their ω-weighted average is*

$$Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2}$$

with weight $\omega = \omega_1 + \omega_2$. The new variable Y will then belong to the same EDM family as Y_1 and Y_2, with the same μ and ϕ.

Proof. This proof can be found in chapter 3 of [Jør97]. □

2.6 Tweedie models

2.6.1 Moment- and cumulant-generating functions

Definition 1. *The moment-generating function (MGF) of a random variable Y is*

$$\psi_Y(t) = E[e^{tY}],$$

if there exists an $\varepsilon > 0$ such that the expectation exists and is finite for any t satisfying $|t| < \varepsilon$. The cumulant-generating function (CGF) is given by

$$\Psi_Y(t) = \log \psi_Y(t) = \log E[e^{tY}].$$

For continuous EDMs, the MGF and CGF can be derived as follows:

$$\begin{aligned} E[e^{tY}] &= \int e^{ty} f_Y(y;\theta,\phi)dy = \int e^{ty} e^{\frac{y\theta-b(\theta)}{\phi/\omega}} e^c dy = \int e^{\frac{y(\theta+t\phi/\omega-b(\theta)}{\phi/\omega}} e^c dy \\ &= e^{\frac{b(\theta+t\phi/\omega)-b(\theta)}{\phi/\omega}} \cdot \int e^{\frac{y(\theta+t\phi/\omega)-b(\theta+t\phi/\omega)}{\phi/\omega}} e^c dy \overset{(*)}{=} e^{\frac{b(\theta+t\phi/\omega)-b(\theta)}{\phi/\omega}} \end{aligned} \tag{2.11}$$

which implies

$$\Psi_Y(t) = \frac{b(\theta+t\phi/\omega)-b(\theta)}{\phi/\omega}. \tag{2.12}$$

At (*), we used the fact that $\theta + t\phi/\omega$ belong to the parameter space for $|t| < \varepsilon$ for some $\varepsilon > 0$, making the integral in the previous step equal one by definition. Thus, both the moment-generating function and the cumulant-generating function exist for any exponential dispersion model for small enough t. For discrete EDMs, the MGF and CGF can be derived in a similar way.

2.6.2 Expectation and variance through generating function

It is well known that the expectation of a random variable is equal to the first moment and the variance is equal to the second central moment. In addition, we can get the nth moment though the nth derivative of the moment-generating function at $t = 0$, as seen in chapter 3.10.3 of [AB08],

$$\psi_Y^{(n)}(0) = E[Y^n]. \tag{2.13}$$

This implies

$$\Psi'_Y(t) = \frac{\psi'(t)}{\psi(t)} \Rightarrow \Psi'_Y(0) = \psi'(0) = E[Y] \tag{2.14}$$

and

$$\begin{aligned} \Psi''_Y(t) &= \frac{\psi(t)\psi''(t) - \psi'(t)^2}{\psi(t)^2} \Rightarrow \\ \Psi''_Y(0) &= \psi''(0) - \psi'(0)^2 = E[Y^2] - E[Y]^2 = Var(Y). \end{aligned} \tag{2.15}$$

Inserting (2.12) into (2.14) and (2.15) we find that for EDMs the following is true;

$$E[Y] = b'(\theta) \tag{2.16}$$

and

$$Var(Y) = b''(\theta)\phi/\omega. \tag{2.17}$$

Table 3: Examples of choices of the variance function exponent p, and the corresponding distribution of key ratios, for Tweedie models.

p	Type	Name	Key Ratio
$p = 1$	Discrete	Poisson	Claim frequency
$1 < p < 2$	Mixed, non negative	Compound Poisson	Pure premium
$p = 2$	Continuous, positive	Gamma	Claim severity

2.6.3 Variance function

If $E[Y] = \mu$, then we have that $\mu = b'(\theta)$, and since the first derivative of $b(\cdot)$ was assumed to be invertible, see subsection 2.5, we can also write $\theta = b'^{-1}(\mu)$. This makes it possible for us to express the term $b^{''}(\theta)$ of (2.17) as a function of the expectation, giving us the *variance function*

$$v(\mu) = b''\left(b'^{-1}(\mu)\right). \tag{2.18}$$

The variance of Y can then be written as $Var(Y) = v(\mu)\phi/\omega$.

Theorem 2. *Within the class of all exponential dispersion models, each EDM is characterized by its variance function.*

Proof. The proof can be found at theorem 1 in [Jør87]. □

2.6.4 Definition of Tweedie models

Tweedie models are exponential dispersion models defined by having the variance function

$$v(\mu) = \mu^p, \; p \in \mathbb{R}. \tag{2.19}$$

It can be shown that all EDMs that are scale invariant are Tweedie models. That is if we have a random variable Y with a distribution following (2.10) and a constant $c > 0$ such that Y and cY belong to the same EDM, then this EDM is a Tweedie family with $p = 2$, see theorem 4.1 in [Jør97]. A couple of different distributions for different values of p can be seen in Table 3, which is a part of a more comprehensive table, see table 2.4 in [OJ10].

2.7 Probability distributions for different key ratios

When it comes to modeling and tariff analysis, there are two main methods. We can either model the claim severity and claim frequency separately, to get the premium through multiplication, or we can model for the premium directly. The standard approach is the former alternative, see chapter 2.3.4 of [OJ10].

2.7.1 Claim frequency probability distribution

The process of our claims arriving is a stochastic process called a claims process, which can be written as $\{N(t), t \geq 0\}$, where $N(t)$ is the number of claims during the time $[0, t]$ and $N(0) = 0$. Under certain conditions, not unlike our modeling assumptions, the claims process is a Poisson process, see appendix A of [BPP84]. This makes it natural for us to assume that claims within a single policy follow a Poisson distribution, and due to Assumption 1 this is true for all policies of a given tariff cell.

Using (2.1) we get

$$X_i \sim Po(\omega_i \mu_i) \tag{2.20}$$

for the number of claims of policy i, with ω_i the time of exposure and μ_i the intensity of the underlying Poisson process. But what we actually need to price for our policies is the distribution of the claim frequency, $Y_i = X_i/\omega_i$, with probability distribution

$$f_{Y_i}(y_i) = f_{X_i}(\omega_i y_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{\omega_i y_i}}{(\omega_i y_i)!}. \tag{2.21}$$

To write this as an EDM, we first rewrite

$$f_{Y_i}(y_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{\omega_i y_i}}{(\omega_i y_i)!} = \exp\{\omega_i(y_i \log \mu_i - \mu_i) + c(y_i, \omega_i)\}, \tag{2.22}$$

with

$$c(y_i, \omega_i) = \omega_i y_i \log \omega_i - \log((\omega_i y_i)!). \tag{2.23}$$

Equation (2.22) can then be written as (2.10) with $\phi = 1$, $\theta_i = \log(\mu_i)$, $b(\theta_i) = e^{\theta_i}$, and $a_i(\phi) = \phi/\omega_i$.

2.7.2 Claim severity probability distribution

The choice of distribution for the claim severity is not as clear cut as for the frequency. Historical data shows that a distribution that is right-skewed and positive fits this type of data, some examples are Pareto, log-normal and gamma distributions. Over the years, the gamma distribution has become the standard choice when modeling claim severity so this is the distribution we will be using as well.

Suppose we have w gamma distributed variables (claims), $Z_k \sim G(\alpha, \beta)$, $k = 1, ..., w$, and let the response X be the sum of these variables (the total policy cost). It can be seen by the use of moment generating functions that $X \sim G(\omega\alpha, \beta)$, with density function

$$f_X(x) = \frac{\beta^{\omega\alpha}}{\Gamma(\omega\alpha)} x^{\omega\alpha - 1} e^{-\beta x}, \tag{2.24}$$

where $\alpha, \beta, x > 0$. The claims severity $Y = X/\omega$ then has the density function

$$f_Y(y) = \omega f_X(\omega y) = \frac{(\omega\beta)^{\omega\alpha}}{\Gamma(\omega\alpha)} y^{\omega\alpha-1} e^{-\omega\beta y}, \tag{2.25}$$

giving us $Y \sim G(\omega\alpha, \omega\beta)$. With the parameterisation $\mu = \alpha/\beta$, $\phi = 1/\alpha$ and $\theta = -1/\mu$, it follows from chapter 2.1.2 in [OJ10] that the tariff cells' claim severities Y_i can be written in EDM-form, with $b(\theta_i) = -\log(-\theta_i)$, as

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\{\frac{y_i\theta_i + \log(-\theta_i)}{\phi/\omega_i} + c(y_i, \phi, \omega_i)\}, \tag{2.26}$$

with

$$c(y, \phi, \omega) = \log(\omega y/\phi)\omega/\phi - \log(y) - \log(\Gamma(w/\phi)). \tag{2.27}$$

2.7.3 Pure premium probability distribution

As seen in Table 3, the Tweedie model for $1 < p < 2$ follows a compound Poisson distribution, which is shown in chapter 4.2.4 of [Jør97].This is the distribution of a random variable

$$X = \sum_{k=1}^{N} Z_k \tag{2.28}$$

where N is Poisson distributed whereas the Z_ks are independent with the same gamma distribution $\Gamma(\alpha, \beta)$. This is in line with subsubsection 2.7.2, where we assume gamma distributed claims. Suppose $N \sim Po(\lambda\omega)$, where λ is the claim rate and ω the time of exposure. From this it follows that the total claim X in (2.28) satisfies

$$E(X) = \lambda\omega\alpha/\beta, \tag{2.29}$$

$$Var(X) = \lambda\omega(\alpha/\beta + \alpha/\beta^2). \tag{2.30}$$

The corresponding pure premium $Y = X/\omega$ satisfies

$$E(Y) = \lambda\alpha/\beta = \mu, \tag{2.31}$$

$$Var(Y) = \lambda(\alpha/\beta + \alpha/\beta^2)/\omega = \phi\mu^{p-1}/\omega. \tag{2.32}$$

Note in particular that (2.31) and (2.32) set restrictions on how to choose μ, ϕ and p as functions of λ, α, and β.

2.8 Relatives estimation

To estimate the relatives seen in (2.2), we estimate the μ_is using equations (2.6) and (2.8). Since we are using multiplicative models for our variables, we will only be using the link function $g(\mu_i) = \log(\mu_i)$, that is

$$\log(\mu_i) = \sum_{j=1}^{q} x_{ij}\beta_j \Leftrightarrow \mu_i = \exp\left(\sum_{j=1}^{q} x_{ij}\beta_j\right) \tag{2.33}$$

In order to estimate the relatives, we will be using maximum likelihood estimation. The log-likelihood function for an EDM with n observations whose density functions follow (2.10) is

$$\ell(\theta; \{y_i\}_{i=1}^{n}, \phi) = \sum_{i=1}^{n} \left((y_i\theta_i - b(\theta_i))w_i/\phi + c(y_i, \phi, \omega_i)\right). \tag{2.34}$$

Differentiating with respect to each β_j, $j = 1, \dots, q$, and making use of equations (2.16), (2.17), (2.8), and the fact that $g(\mu_i) = \log(\mu_i)$, we obtain

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \tag{2.35}$$

$$= \sum_{i=1}^{n} \underbrace{(y_i - b'(\theta_i))w_i/\phi}_{\partial\ell/\partial\theta_i} \cdot \underbrace{b''^{-1}(\theta_i)}_{\partial\theta_i/\partial\mu_i} \underbrace{\mu_i}_{\partial\mu_i/\partial\eta_i} \underbrace{x_{ij}}_{\partial\eta_i/\partial\beta_j} \tag{2.36}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n} \frac{(y_i - \mu_i)w_i x_{ij}\mu_i}{v(\mu_i)}. \tag{2.37}$$

We know that for the claim frequency we have $\theta_i = \log(\mu_i)$ and $b(\theta_i) = b'(\theta_i) = b''(\theta_i) = \mu_i = e^{\theta_i}$, giving us $v(\mu_i) = b''(b'^{-1}(\mu_i)) = \mu_i$. We then get the ML-equation

$$\frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)w_i x_{ij} = 0, \; j = 1, ..., q. \tag{2.38}$$

For the claim severity we have $b(\theta_i) = -\log(-\theta_i)$, $\mu_i = b'(\theta_i) = -1/\theta_i$ which implies $\theta_i = -1/\mu_i$ and $b''(\theta_i) = 1/\theta_i^2$, giving us $v(\mu_i) = \mu_i^2$. We then obtain the ML-equations

$$\frac{1}{\phi}\sum_{i=1}^{n} \frac{(y_i - \mu_i)w_i x_{ij}}{\mu_i} = \frac{1}{\phi}\sum_{i=1}^{n}\left(\frac{y_i}{\mu_i} - 1\right) w_i x_{ij} = 0, \; j = 1, ..., q. \tag{2.39}$$

Lastly, for the direct approach using a Tweedie model with $1 < p < 2$, we have $v(\mu_i) = \mu_i^p$, giving us

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i-\mu_i)w_i x_{ij}}{\mu_i^{p-1}}=0,\ j=1,...,q. \tag{2.40}$$

After calculating the estimates of the β_js we can compute the relatives, by using (2.2) and (2.33), through $\gamma_j = e^{\beta_j}$.

3 Data

The data we will use is the FreMTPL data from the CASdatasets package in R, which is in turn taken from [Cha14]. This package can be installed to R through

```
install.packages(”CASdatasets”,
                 repos = ”http://cas.uqam.ca/pub/”,
                 type = ”source”
                 )
```

The FreMTPL data set contains data from a type of insurance called ”motor third party liability”, and was collected in France. The data are split into two parts, one data frame for the frequency and one for the severity. After cleaning up the variable names, the variables available can be seen in Table 4.

We are using the factors put forth in the beginning of chapter 14 of [Cha14].

A notable shortcoming of this data set is that the frequency and severity parts have very different sizes. Such is the nature of these types of data sets, as there will be a lot of policies without claims. The frequency table has about 413 000 rows, while the severity set has only about 16 000 rows. This mean that the estimates of frequencies will be based on more data than the estimates of severities.

Table 4: Variables (risk factors), variable descriptions and values (risk classes) of each variable. The descriptions are taken from the CASdatasets documentation.

Variable	Description
policy_id	The policy ID (used to link with the claims data set)
claim_nb	Number of claims during the exposure period
exposure	The period of exposure for a policy, in years
veh_power	The power of the car (ordered categorical). D-O, changed to 1-12.
car_age	The vehicle age, in years. 1 : Below 1 year 2 : 1-3 years 3 : 4-14 years 4 : 15+ years
driver_age	The driver age, in years (in France, people can drive a car at 18). 1 : Up to 22 years 2 : 23-26 years 3 : 27-42 years 4 : 43-74 years 5 : 75+ years
brand	The car brand, divided into the following groups: 1 : Renaut, Nissan and Citroen 2 : Volkswagen, Audi, Skoda and Seat 3 : Opel, General Motors and Ford 4 : Fiat 5 : Mercedes, Chrysler and BMW 6 : Japanese (except Nissan) and Korean 7 : Other
gas	The fuel of the car. 1 : Diesel 2 : Regular
density	The density of inhabitants (number of inhabitants per square-kilometer) of the city where the driver lives. 1 : 0-39 2 : 40-199 3 : 200-499 4 : 500-4499+ 5 : 4500+
region	The policy region in France (based on the 1970-2015 classification). E.g. Aquitaine, Bretagne & Ile de France. There are in total 10 regions.
claim_amount	The cost of the claim, seen as at a recent date.

4 Results

4.1 Model selection

When we are selecting which models fit our data the best, we are going to alternate between using forward selection and backward elimination. In this type of selection method, we start with the intercept model without any of the classes of our risk factors as covariates. Next we perform a first forward selection step and check much each added factor increases performance, keeping the one that increases the fit the most. The criteria for selection we are going to use is the Akaike information criterion (AIC). This criterion is defined by

$$AIC = -2\log(L) + 2p, \tag{4.1}$$

where L is the maximized likelihood function, and p is the number of parameters in the model. We have chosen to use AIC as it is well known and the standard approach for many functions inside of R. There are many other criteria we could use as well, such as the Bayesian information criterion (BIC) which has a bigger penalty for models with more parameters than AIC, or checking the p-value for different statistics for model fit or predictive ability.

After this first forward selection step, the next step is to alternate between backward elimination and forward selection, removing or adding factors to the currently chosen model. Note that this ”add or remove” is what makes this a bi-directional method, as stated above. This alternating scheme continues until the model we have is the one for which it is not possible to improve performance by adding or removing any factor. A possibility is, however, that the more factors you add to a model, the larger the goodness of fit might be. But a model with as many factors as possible is typically not the best as there is most likely overfitting, and moreover, each parameter estimate is not as informative for how much the corresponding covariate affects the outcome. For this reason, and to get a reasonable scope of the analysis, we will set our algorithm to only do 3 iterations, giving us a maximum of 3 factors, along with the intercept, in the end.

For the Tweedie model of the pure premium we will be using $p = 1.552$ in (2.19), which is the result of a maximum log-likelihood estimation algorithm from the `tweedie`-package for R. The resulting log-likelihood profile can be seen in Figure 1.

In Table 5 we can see the results of our step-wise selections. Note that some of them are not actually the model with the lowest AIC, but since a maximum number of 3 iterations was chosen these were the best among the tested models.

Now that we know which factors we are going to use, we can aggregate our data according to Theorem 1. We chose not to do this before the step-wise selection due to the number of combinations of risk factors available, if we choose 3 out of the 7 available risk factors that gives us $C(7, 3) = 35$ combinations (and therefore also 35 different aggregations needed). After aggregation we get the values seen in Table 6, where we have also added the number of parameters

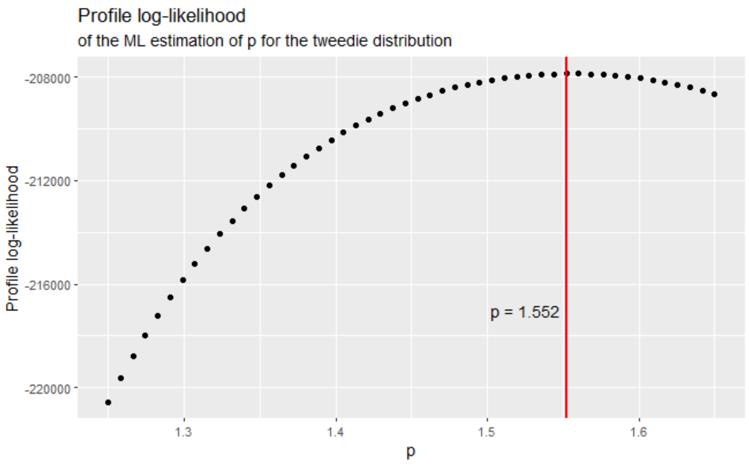

Figure 1: Profile log-likelihood for the variance function exponent p of the Tweedie distribution.

Table 5: Comparison between AIC values of the different selected models containing 3 factors. The baseline level and relativities estimates, standard errors, performance statistics and estimates of the variance function exponent p of all models can be seen in the appendix.

Risk factors in model	driver_age + density + gas	driver_age + brand + region	driver_age + region + power
Frequency	**135266**	135691	135732
Severity	305678	**305434**	305510
Tweedie Premium	484425	478618	**478033**

according to (2.7) to get some sense to how the this number might be effecting the AIC.

After the aggregation, we can see that the best model according to AIC for each response variable has changed for both the claim severity and the pure premium responses, while the frequency still has the same best model as before. To get a more clear answer to how much better the model with the lowest AIC is than the other two, we can look at two properties.

The first is the AIC differences, $\Delta_i = AIC_i - AIC_{min}$, where i is the ith model that does not have the lowest AIC. If we index our models in Table 6 as

1. driver_age + density + gas
2. driver_age + brand + region
3. driver_age + region + power

Table 6: Comparison between AIC values of the different selected models containing 3 factors, after aggregation. The baseline level and relativities estimates, standard errors, performance statistics and estimates of the variance function exponent p of all models can be seen in the appendix.

Risk factors in model	driver_age + density + gas	driver_age + brand + region	driver_age + region + power
No. of parameters	10	20	25
Frequency	**410**	1785	2494
Severity	501798	482374	**460176**
Tweedie Premium	**1373**	7182	9915

we can see that for example frequency we get $AIC_{min} = AIC_1 = 410$, $AIC_2 = 1785$ and $AIC_3 = 2494$. This gives us $\Delta_2 = 1375$ and $\Delta_3 = 2084$. We can then follow the guidelines on p.70 of [BA02], see Table 7. If we look at all of the AIC differences, , we can see that the model with the lowest The second property we can look at is the relative likelihood, $\exp(-\Delta_i/2)$. If we for choose a significance level of 0.05, as is standard, we would omit a model if its relative likelihood was below this threshold and instead choose the model with the lowest AIC. We can see all AIC differences and relative likelihoods in Table 8.

Table 7: Rules of thumb for how to determine support of models through the use of AIC differences.

Δ_i	Level of Empirical Support of Model i
0-2	Substantial
4-7	Considerably less
> 10	Essentially none

Table 8: AIC differences and relative likelihoods (RL) between models. If RL$<$ 0.05, we denote this with *.

Risk factors in model	driver_age + density + gas		driver_age + brand + region		driver_age + region + power	
	Δ_i	RL	Δ_i	RL	Δ_i	RL
Frequency	0	1	1375	*	2084	*
Severity	41622	*	22198	*	0	1
Tweedie Premium	0	1	5809	*	8542	*

It is clear that according to the rule of thumb for Δ_i and the relative likelihood that the model we should choose for each of our responses is the one with the smallest AIC.

4.2 Predictive capabilities

Another way we could have chosen models is through cross validation, which involves splitting the data into multiple subsets and then training the models on all subsets but one and then to verify how well the model can predict the actual values of the left out subset. Though we chose our models through normal selection-elimination methods, we can now also pass our models trough a cross validation algorithm to calculate some predictive values, such as the root mean squared error (RMSE), the coefficient of determination R^2 or the mean absolute error (MAE). We can see the values of these statistics for the selected models in Table 9.

We can see that the two more favourable models are one and three, i.e. the models that contain driver age, density and gas, and driver age, region and power. The model containing driver age, car brand and region is only the best when looking at the RMSE for the severity response variable. We now have an idea of which of our three models will give us the most trustworthy results, and will move on to calculate our relatives.

4.3 Relatives

To calculate the relatives, we perform the procedure discussed in subsection 2.8. In Table 10, Table 11 and Table 12 we can find the calculated relatives. In addition, the exposure and confidence intervals for the relatives can be found in the appendix. To get a better view of how the relatives differ from each other, we plot them using scatter plots.

In the scatter plots, Figure 2, Figure 3 and Figure 4, there appears to be no clear pattern that tells us the Tweedie model under- or over estimates the relatives, relative to the frequency-severity model, and therefore we cannot make a conclusion about There are cases where the Tweedie model estimates a higher and a lower relative, so we cannot make a conclusion about whether or not risk factors on both sides where the Tweedie model e

Table 9: AIC, root mean square error and mean absolute value of parameter estimates, along with the coefficient of determination R^2, for models selected via the bi-directional selection method.

Model	AIC	RMSE	R^2	MAE
	Frequency			
driver_age + density + gas	**410**	137.567	0.922	96.442
driver_age + brand + region	1785	59.326	0.822	20.849
driver_age + power + region	2494	**15.308**	**0.939**	**6.948**
	Severity			
driver_age + density + gas	501798	288688.4	**0.831**	229591.6
driver_age + brand + region	482374	**198127.2**	0.724	74943.47
driver_age + power + region	**460176**	82743.56	0.810	**38181.7**
	Pure premium			
driver_age + density + gas	**1373**	304647.1	**0.828**	249505.4
driver_age + brand + region	7182	156552.1	0.685	58362.58
driver_age + power + region	9915	**68361.05**	0.791	**28942.71**

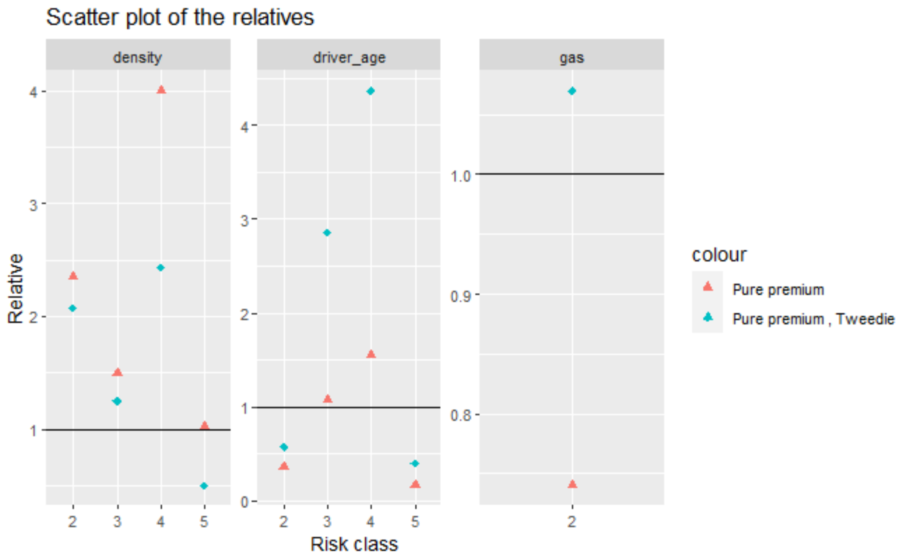

Figure 2: Scatter plot, for the two different ways of calculating premium relatives in Table 10.

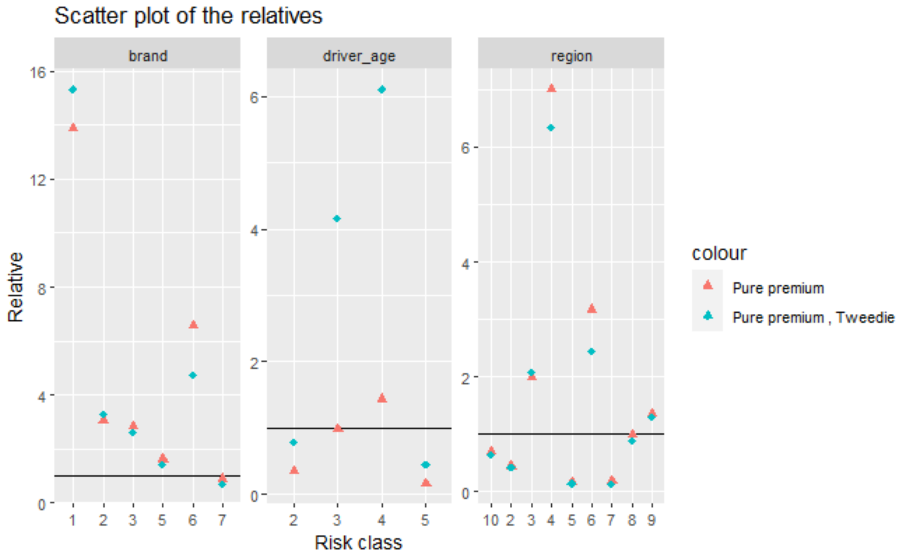

Figure 3: Scatter plot, for the two different ways of calculating premium relatives in Table 11.

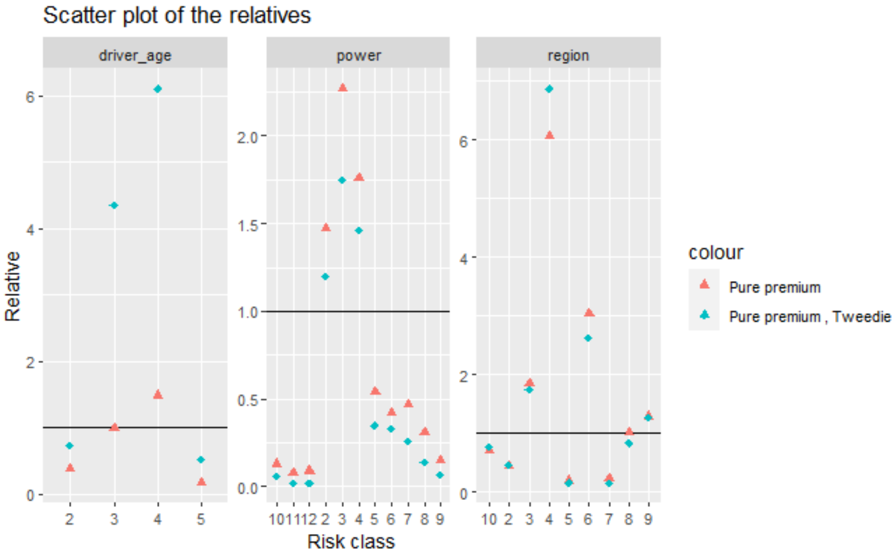

Figure 4: Scatter plot, for the two different ways of calculating premium relatives in Table 12.

Table 10: Estimated relatives for models including the risk factors driver age, density and gas.

Risk factor	Risk class	Relatives, frequency	Relatives, severity	Relatives, premium	Relatives, premium (Tweedie)
Driver age	1	1.00	1.00	1.00	1.00
	2	0.54	0.66	0.36	0.57
	3	0.34	3.16	1.07	2.86
	4	0.34	4.56	1.55	4.37
	5	0.33	0.50	0.16	0.40
Density	1	1.00	1.00	1.00	1.00
	2	1.20	1.97	2.35	2.07
	3	1.37	1.09	1.49	1.25
	4	1.69	2.37	4.00	2.43
	5	1.89	0.54	1.02	0.50
Gas	1	1.00	1.00	1.00	1.00
	2	0.81	0.90	0.74	1.07

Table 11: Estimated relatives for models including the risk factors driver age, brand and region.

Risk factor	Risk class	Relatives, frequency	Relatives, severity	Relatives, premium	Relatives, premium (Tweedie)
Driver age	1	1.00	1.00	1.00	1.00
	2	0.55	0.63	0.35	0.78
	3	0.35	2.83	0.98	4.16
	4	0.35	4.13	1.43	6.11
	5	0.32	0.46	0.15	0.45
Brand	1	0.92	15.10	13.87	15.34
	2	1.04	2.91	3.04	3.27
	3	1.06	2.67	2.83	2.60
	4	1.00	1.00	1.00	1.00
	5	1.10	1.47	1.61	1.40
	6	0.80	8.22	6.55	4.72
	7	0.99	0.89	0.88	0.69
Region	1	1.00	1.00	1.00	1.00
	2	0.89	0.49	0.44	0.43
	3	0.89	2.24	1.98	2.08
	4	0.82	8.50	7.01	6.34
	5	0.94	0.17	0.16	0.14
	6	1.22	2.61	3.17	2.45
	7	1.14	0.17	0.19	0.13
	8	1.09	0.92	0.99	0.88
	9	0.96	1.42	1.36	1.29
	10	0.94	0.73	0.69	0.64

Table 12: Estimated relatives for models including the risk factors driver age, power and region.

Risk factor	Risk class	Relatives, frequency	Relatives, severity	Relatives, premium	Relatives, premium (Tweedie)
Driver age	1	1.00	1.00	1.00	1.00
	2	0.55	0.70	0.38	0.73
	3	0.34	2.94	0.98	4.35
	4	0.33	4.46	1.48	6.11
	5	0.31	0.50	0.16	0.52
Power	1	1.00	1.00	1.00	1.00
	2	1.17	1.26	1.47	1.20
	3	1.20	1.90	2.27	1.75
	4	1.13	1.55	1.76	1.46
	5	1.20	0.45	0.54	0.35
	6	1.27	0.33	0.42	0.33
	7	1.27	0.37	0.47	0.26
	8	1.36	0.23	0.31	0.14
	9	1.21	0.12	0.15	0.07
	10	1.32	0.09	0.13	0.06
	11	1.40	0.06	0.08	0.02
	12	1.28	0.07	0.09	0.02
Region	1	1.00	1.00	1.00	1.00
	2	0.90	0.50	0.45	0.45
	3	0.91	2.01	1.84	1.74
	4	0.84	7.18	6.06	6.88
	5	0.93	0.20	0.19	0.16
	6	1.17	2.61	3.04	2.63
	7	1.13	0.20	0.23	0.15
	8	1.09	0.94	1.02	0.83
	9	0.97	1.33	1.29	1.26
	10	0.96	0.73	0.70	0.76

5 Discussion

From the analysis of this thesis, there is no clear pattern whether the frequency-severity approach leads to higher or lower estimated relatives compared to the Tweedie method. We can however conclude that using a separate analysis of the frequency and severity gives us more insight into why we get our result. For the exponential dispersion models of the frequencies and severities, we can see more clearly where the estimated relatives come from, and conduct further analysis on each of them separately. We can look closer at each of the models and see how the frequency of the claims effects each risk factor and risk class, and see how these match up with the effects that the severity model gives us. If we were to use a Tweedie model, we just get one single value of each relative and we have no real idea of where the estimated relatives comes from, except what the theory of Tweedie models tells us. Since we can more clearly see the effect the data has on the frequency and severity separately, we would also be able to tweak our values more closely and with more information backing our decisions.

Furthermore, the separate model is more flexible when it comes to the distributions used. If the data used does not fit the assumed distributions for the claims frequency or the claim severity, the Tweedie model can simply not be used. This is because a Tweedie model with a variance function exponent between one and two, $1 < p < 2$, describes gamma distributed claim sizes and a Poisson distributed number of claims. Hence, if the data differs from this, the Tweedie approach will not work. But if the frequency-severity approach is used, we could use other distributions that fit the data better if needed.

There are of course some shortcomings of the conclusions of this thesis. First of all, our choice to use exactly 3 risk factors was purely based on discussions on how extensive the analysis should be. A more comprehensive analysis that instead uses models that are the outcomes of a more elaborate model selection method might give more insight into how the two premium models preform relative to each other. However, such a comparison could become very time consuming and have an incredibly broad scope. This is why the decision was made to limit ourselves to models with a maximum of three factors in this paper. Another assumption of ours was trusting how the editor of the book that supplied the data set, see [Cha14], split the risk factors into their risk classes. We could have preformed our own analysis of the quantiles of the quantitative risk factors, or looked at real groups used by real insurance companies. We could also have studied more closely how to divide these factors into risk classes which would have given us more insight into how the different classes differ from each other.

It is worth noting that in the data set of claim costs, there are three policies that make up a considerable part of the total amount. These policies all have claim costs of above 1 million euro, one of them is even above 2 million euro. For context, the policy with the fourth largest claim has a cost of around 300 000 euro. There are many ways of handling large claims in data sets. One common way is to choose a value or threshold c, so that an amount above this value is reduced to the value. That is, for a chosen threshold c the total claim amount X_i for policy i is changed to $min(X_i, c)$. The money that is now neglected is then instead spread out among all tariff cells in the budgeting stage of setting the premium, after having calculated the relatives. This reduces the impact that these outliers have on the premiums.

Also, when we looked at our data we followed normal practice for the numeric and continuous risk factors and split them into groups, turning the risk factors into categorical data instead. There are however other approaches, such as polynomial regression or generalised additive models (GAM). This is, however, a whole other method for pricing and was not considered in this analysis.

It is also worth mentioning that Tweedie models could have been used in the analysis of the frequency and severity, with $p = 1$ for the frequency component and $p \geq 2$ for different models of severity, among them the gamma distribution. Such an estimate of p could be used to find which distribution best suited the given data. This could then be used in an GLM analysis with separate frequency and severity components, but now with information behind the distribution of the models, primarily the severity distribution.

6 Summary

In this thesis we have looked at two different models for calculating the relatives for non-life insurance pricing. We could not conclude a clear pattern of over- and under-estimation of relatives when comparing these two models, or see which method gave a more accurate result as there were no correct relatives to compare to.

We concluded, however, that the method of separate analyses of frequency and severity gave more insight into how the relatives are estimated. This could give better performance since tweaking of the model parameters could be done, with more information behind their interpretation. We also note that the separate analysis method gives more room to vary the chosen distributions if needed, which would not be possible if a Tweedie model was used for the pure premium.

Therefore, our results indicate that the method with separate models for frequency and severity is more appropriate to use for setting relatives for non-life insurance.

References

[AB08] Alm, Sven Erick and Britton, Tom. (2008). *Stokastik : sannolikhetsteori och statistikteori med tillämpningar*. Liber, Stockholm.

[Agr13] Agresti, Alan. (2013). *Categorical Data Analysis*. Wiley, Hoboken, N.J.

[BA02] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A practical information-theoretic approach*. Springer, New York, NY.

[BPP84] Beard, Robert Eric, Pentikäinen, Teivo, and Pesonen, Erkki. (1984). *Risk Theory*. Chapman & Hall, London.

[Cha14] Charpentier, Arthur, ed. (2014). *Computational Actuarial Science with R*. CRC Press, Boca Raton, FL.

[Jør87] Jørgensen, Bent. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 49, pp. 127–162.

[Jør97] Jørgensen, Bent. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

[OJ10] Ohlsson, Esbjörn and Johansson, Björn. (2010). *Non-life Insurance Pricing with Generalized Linear Models*. Springer, Berlin.

Appendix

Table 13: Covariates and their effect parameter estimates, for the frequency distribution of model 1.

Coefficients	Estimate	Standard error	Wald statistic	p-value
Intercept	-1.8618397	0.0405988	-45.859460	$< 2e - 16$ ***
driver_age2	-0.6162680	0.0460771	-13.374705	$< 2e - 16$ ***
driver_age3	-1.0796310	0.0364052	-29.655966	$< 2e - 16$ ***
driver_age4	-1.0775837	0.0354940	-30.359634	$< 2e - 16$ ***
driver_age5	-1.1070824	0.0518830	-21.338073	$< 2e - 16$ ***
density2	0.1781819	0.0269060	6.622385	$3.53e - 11$ ***
density3	0.3141370	0.0298160	10.535872	$< 2e - 16$ ***
density4	0.5242068	0.0261174	20.071204	$< 2e - 16$ ***
density5	0.6341391	0.0349586	18.139701	$< 2e - 16$ ***
gas2	-0.2057810	0.0160347	-12.833483	$< 2e - 16$ ***

Table 14: Covariates and their effect parameter estimates, for the severity distribution of model 1.

Coefficients	Estimate	Standard error	t-statistic	p-value
Intercept	12.4410731	0.2272311	54.750760	< 2e-16 ***
driver_age2	-0.4099083	0.2540214	-1.613676	0.114
driver_age3	1.1504857	0.2010273	5.723032	1.16e-6 ***
driver_age4	1.5175656	0.1955672	7.759816	1.68e-9 ***
driver_age5	-0.6956608	0.2864299	-2.428729	0.0197 *
density2	0.6763510	0.1548659	4.367333	8.66e-5 ***
density3	0.0874693	0.1714157	0.510276	0.613
density4	0.8626667	0.1500621	5.748730	1.07e-6 ***
density5	-0.6099982	0.1964770	-3.104680	0.00349 **
gas2	-0.1002027	0.0910298	-1.100769	0.278

Table 15: Covariates and their effect parameter estimates, for the premium distribution of model 1.

Coefficients	Estimate	Standard error	t-statistic	p-value
Intercept	12.4610500	0.2408989	51.7273006	< 2e-16 ***
driver_age2	-0.5556649	0.2599295	-2.1377522	0.0387 *
driver_age3	1.0495062	0.2193062	4.7855750	2.34e-5 ***
driver_age4	1.4758211	0.2118484	6.9664022	2.09e-8 ***
driver_age5	-0.9151947	0.2723902	-3.3598661	1.72e-3 **
density2	0.7270814	0.2206939	3.2945236	2.07e-3 **
density3	0.2210817	0.2321658	0.9522578	0.347
density4	0.8896791	0.2174201	4.0919815	2.01e-4 ***
density5	-0.6841354	0.2582860	-2.6487512	0.0115 *
gas2	0.0664692	0.1407176	0.4723589	0.639

Table 16: Covariates and their effect parameter estimates, for the frequency distribution of model 2.

Coefficients	Estimate	Standard error	Wald statistic	p-value
Intercept	-1.5198917	0.0580241	-26.1941531	< 2e-16 ***
driver_age2	-0.5905139	0.0460879	-12.8127635	< 2e-16 ***
driver_age3	-1.0630671	0.0364135	-29.1943086	< 2e-16 ***
driver_age4	-1.0637872	0.0356061	-29.8765739	< 2e-16 ***
driver_age5	-1.1357769	0.0519160	-21.8771973	< 2e-16 ***
region2	-0.1138904	0.0563284	-2.0219008	0.0432 *
region3	-0.1201286	0.0387788	-3.0977870	1.95e-3 **
region4	-0.1926307	0.0336564	-5.7234452	1.04e-8 ***
region5	-0.0655493	0.0741355	-0.8841828	0.377
region6	0.1954144	0.0368265	5.3063470	1.12e-7 ***
region7	0.1352030	0.0776370	1.7414763	0.0816 .
region8	0.0827503	0.0448702	1.8442143	0.0652 .
region9	-0.0455941	0.0399425	-1.1414941	0.254
region10	-0.0624802	0.0470189	-1.3288312	0.184
brand6	-0.2264919	0.0444071	-5.1003482	3.39e-7 ***
brand5	0.0925765	0.0510520	1.8133775	0.0698 .
brand3	0.0591271	0.0444938	1.3288833	0.184
brand7	-0.0099224	0.0618557	-0.1604123	0.873
brand1	-0.0847273	0.0389597	-2.1747441	0.0297 *
brand2	0.0436418	0.0456561	0.9558815	0.339

Table 17: Covariates and their effect parameter estimates, for the severity distribution of model 2.

Coefficients	Estimate	Standard error	t-statistic	p-value
Intercept	8.9150807	0.2142772	41.6053576	< 2e-16 ***
driver_age2	-0.4579453	0.1676845	-2.7309939	6.71e-3 **
driver_age3	1.0413775	0.1325708	7.8552573	8.27e-14 ***
driver_age4	1.4191450	0.1293326	10.9728311	< 2e-16 ***
driver_age5	-0.7866850	0.1894286	-4.1529357	4.35e-05 ***
region2	-0.7111431	0.2097520	-3.3904003	7.97e-4 ***
region3	0.8048754	0.1437384	5.5995834	5.07e-08 ***
region4	2.1402351	0.1241161	17.2438100	< 2e-16 ***
region5	-1.7912625	0.2768210	-6.4708324	4.26e-10 ***
region6	0.9580299	0.1351122	7.0906226	1.06e-11 ***
region7	-1.7828601	0.2854629	-6.2455064	1.54e-09 ***
region8	-0.0885632	0.1641826	-0.5394190	0.590
region9	0.3541349	0.1478592	2.3950821	0.0173 *
region10	-0.3152369	0.1748497	-1.8029014	0.0725 .
brand6	2.1059770	0.1643662	12.8127131	< 2e-16 ***
brand5	0.3848023	0.1915831	2.0085399	0.0455 *
brand3	0.9823248	0.1667747	5.8901318	1.09e-08 ***
brand7	-0.1164705	0.2312651	-0.5036233	0.615
brand1	2.7145414	0.1458695	18.6093805	< 2e-16 ***
brand2	1.0691172	0.1708918	6.2561075	1.45e-09 ***

Table 18: Covariates and their effect parameter estimates, for the premium distribution of model 2.

Coefficients	Estimate	Standard error	t-statistic	p-value
Intercept	8.7693688	0.2779426	31.5510053	< 2e-16 ***
driver_age2	-0.2490108	0.2021691	-1.2316952	0.219
driver_age3	1.4258240	0.1718230	8.2982118	2.91e-15 ***
driver_age4	1.8104441	0.1670806	10.8357553	< 2e-16 ***
driver_age5	-0.8065126	0.2181325	-3.6973523	2.56e-4 ***
region2	-0.8512350	0.2732214	-3.1155501	2.00e-3 **
region3	0.7304090	0.2277149	3.2075585	1.47e-3 **
region4	1.8470992	0.2080958	8.8761981	< 2e-16 ***
region5	-1.9761585	0.3243299	-6.0930506	3.15e-09 ***
region6	0.8944356	0.2243116	3.9874700	8.26e-05 ***
region7	-2.0050456	0.3275340	-6.1216415	2.68e-09 ***
region8	-0.1288195	0.2491033	-0.5171329	0.605
region9	0.2510923	0.2388661	1.0511847	0.294
region10	-0.4490247	0.2587970	-1.7350458	0.0837 .
brand6	1.5510746	0.2241466	6.9199113	2.44e-11 ***
brand5	0.3393384	0.2499892	1.3574122	0.176
brand3	0.9546416	0.2352181	4.0585383	6.20e-05 ***
brand7	-0.3642630	0.2705309	-1.3464747	0.179
brand1	2.7307164	0.2083870	13.1040608	< 2e-16 ***
brand2	1.1853282	0.2308157	5.1353871	4.88e-07 ***

Table 19: Covariates and their effect parameter estimates, for the frequency distribution of model 3.

Coefficients	Estimate	Standard error	Wald statistic	p-value
Intercept	-1.7067374	0.0483517	-35.2983578	< 2e-16 ***
driver_age2	-0.6037967	0.0460795	-13.1033622	< 2e-16 ***
driver_age3	-1.0924397	0.0364617	-29.9613300	< 2e-16 ***
driver_age4	-1.1030658	0.0356064	-30.9794498	< 2e-16 ***
driver_age5	-1.1662339	0.0518163	-22.5071051	< 2e-16 ***
region2	-0.1030345	0.0562396	-1.8320641	0.0669 .
region3	-0.0897298	0.0385337	-2.3286070	0.0199 *
region4	-0.1697625	0.0332530	-5.1051756	3.30e-7 ***
region5	-0.0694314	0.0741334	-0.9365740	0.349
region6	0.1529409	0.0365733	4.1817631	2.89e-5 ***
region7	0.1246016	0.0776399	1.6048658	0.109
region8	0.0831347	0.0448754	1.8525683	0.0639 .
region9	-0.0286201	0.0398050	-0.7190064	0.472
region10	-0.0424715	0.0468980	-0.9056150	0.365
power2	0.1595045	0.0271605	5.8726737	4.29e-9 ***
power3	0.1820748	0.0260455	6.9906517	2.74e-12 ***
power4	0.1234359	0.0267875	4.6079598	4.07e-6 ***
power5	0.1847364	0.0378558	4.8799969	1.06e-6 ***
power6	0.2357127	0.0425851	5.5350922	3.11e-8 ***
power7	0.2411154	0.0429562	5.6130552	1.99e-8 ***
power8	0.3082563	0.0554147	5.5627189	2.66e-8 ***
power9	0.1928840	0.0813631	2.3706584	0.0178 *
power10	0.2809620	0.1166040	2.4095404	0.0160 *
power11	0.3373327	0.1352482	2.4941740	0.0126 *
power12	0.2471917	0.1378319	1.7934293	0.0729 .

Table 20: Covariates and their effect parameter estimates, for the severity distribution of model 3.

Coefficients	Estimate	Standard error	t-statistic	p-value
Intercept	10.2843459	0.1491399	68.9577260	< 2e-16 ***
driver_age2	-0.3593096	0.1411985	-2.5447117	0.0113 *
driver_age3	1.0770742	0.1116435	9.6474431	< 2e-16 ***
driver_age4	1.4944343	0.1087076	13.7472870	< 2e-16 ***
driver_age5	-0.6951407	0.1592453	-4.3652195	1.62e-05 ***
region2	-0.6942022	0.1763428	-3.9366632	9.75e-05 ***
region3	0.7000909	0.1204522	5.8121904	1.27e-08 ***
region4	1.9706806	0.1034524	19.0491448	< 2e-16 ***
region5	-1.5987851	0.2331326	-6.8578369	2.68e-11 ***
region6	0.9604038	0.1130823	8.4929627	4.06e-16 ***
region7	-1.6031916	0.2405253	-6.6653747	8.83e-11 ***
region8	-0.0658694	0.1382117	-0.4765832	0.634
region9	0.2837995	0.1241574	2.2858045	0.0228 *
region10	-0.3107082	0.1468928	-2.1152040	0.0350 *
power2	0.2277425	0.0849438	2.6810963	0.00764 **
power3	0.6395870	0.0815771	7.8402728	4.16e-14 ***
power4	0.4394689	0.0839492	5.2349386	2.68e-07 ***
power5	-0.8073806	0.1186516	-6.8046311	3.74e-11 ***
power6	-1.1066764	0.1342437	-8.2437852	2.45e-15 ***
power7	-0.9884314	0.1349176	-7.3261859	1.33e-12 ***
power8	-1.4639336	0.1747742	-8.3761416	9.47e-16 ***
power9	-2.1143864	0.2565344	-8.2421175	2.48e-15 ***
power10	-2.3601923	0.3710203	-6.3613569	5.50e-10 ***
power11	-2.8959582	0.4189298	-6.9127525	1.90e-11 ***
power12	-2.6319173	0.4324341	-6.0862855	2.72e-09 ***

Table 21: Covariates and their effect parameter estimates, for the premium distribution of model 3.

Coefficients	Estimate	Standard error	t-statistic	p-value
Intercept	10.1989675	0.2477619	41.1643883	< 2e-16 ***
driver_age2	-0.3195782	0.1975569	-1.6176516	0.106
driver_age3	1.4707584	0.1658303	8.8690588	< 2e-16 ***
driver_age4	1.8106782	0.1619066	11.1834747	< 2e-16 ***
driver_age5	-0.6589172	0.2083369	-3.1627485	1.6552e-3 **
region2	-0.8030850	0.2613576	-3.0727445	2.23e-3 **
region3	0.5528791	0.2228148	2.4813392	0.0134 *
region4	1.9293300	0.1993471	9.6782458	< 2e-16 ***
region5	-1.8277249	0.3083633	-5.9271798	5.65e-9
region6	0.9680921	0.2143627	4.5161406	7.82e-6
region7	-1.9289788	0.3165417	-6.0939165	2.16e-9
region8	-0.1837838	0.2427057	-0.7572289	0.449
region9	0.2300665	0.2311159	0.9954595	0.320
region10	-0.2762909	0.2452118	-1.1267440	0.260
power2	0.1781650	0.1953501	0.9120291	0.362
power3	0.5600327	0.1876335	2.9847155	2.97e-3 **
power4	0.3781551	0.1910867	1.9789708	0.0484 *
power5	-1.0391105	0.2265733	-4.5862006	5.68e-6
power6	-1.1072268	0.2293667	-4.8273212	1.83e-6
power7	-1.3423545	0.2376596	-5.6482233	2.69e-8
power8	-1.9339682	0.2618092	-7.3869370	6.11e-13
power9	-2.6597546	0.3009267	-8.8385465	< 2e-16 ***
power10	-2.8339877	0.3196926	-8.8647276	< 2e-16 ***
power11	-3.8121307	0.3747088	-10.1735813	< 2e-16 ***
power12	-3.7306699	0.3798883	-9.8204395	< 2e-16 ***

Table 22: Duration and number of claims for risk factors and their risk classes. These are the exposures of the frequency, severity and pure premium key ratios.

Risk factor	Risk class	Duration	No. claims
Driver age	1	4702	1041
	2	9509	1173
	3	81519	5941
	4	125558	8942
	5	11097	740
Density	1	39158	2123
	2	73293	4867
	3	37329	2810
	4	66849	6363
	5	15755	1674
Gas	1	113397	9298
	2	118987	8539
Power	1	37907	2657
	2	44760	3553
	3	56015	4373
	4	51770	3780
	5	13955	1104
	6	9433	788
	7	9304	782
	8	4714	416
	9	2206	178
	10	977	82
	11	664	64
	12	677	60
Region	1	14372	1203
	2	6677	498
	3	27816	2033
	4	102878	6911
	5	3183	242
	6	30331	3009
	7	2405	225
	8	11545	1102
	9	21986	1736
	10	11186	878
Brand	1	18351	9750
	2	21885	1622
	3	21884	1889
	4	9547	786
	5	10567	905
	6	31352	2428
	7	5781	457

Table 23: Confidence intervals for Table 10.

		Limits							
Risk	Risk	Frequency		Severity		Premium		Premium , Tweedie	
factor	class	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Driver age	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	0.49	0.59	0.40	1.10	0.20	0.65	0.34	0.96
	3	0.32	0.37	2.08	4.65	0.66	1.70	1.85	4.40
	4	0.32	0.37	3.03	6.62	0.96	2.42	2.88	6.64
	5	0.30	0.37	0.29	0.89	0.09	0.32	0.24	0.68
Density	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.13	1.26	1.44	2.65	1.63	3.34	1.34	3.19
	3	1.29	1.45	0.77	1.53	1.00	2.22	0.79	1.97
	4	1.61	1.78	1.74	3.18	2.80	5.65	1.59	3.73
	5	1.76	2.02	0.37	0.81	0.64	1.64	0.30	0.84
Gas	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	0.79	0.84	0.75	1.09	0.59	0.92	0.81	1.41

Table 24: Confidence intervals for Table 11.

		Limits							
Risk	Risk	Frequency		Severity		Premium		Premium , Tweedie	
factor	class	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Driver age	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	0.51	0.61	0.45	0.88	0.23	0.54	0.52	1.16
	3	0.32	0.37	2.16	3.66	0.69	1.36	2.96	5.85
	4	0.32	0.37	3.17	5.31	1.02	1.97	4.38	8.53
	5	0.29	0.36	0.31	0.67	0.09	0.24	0.29	0.69
Brand	1	0.85	0.99	11.18	19.92	9.53	19.77	10.17	23.07
	2	0.96	1.14	2.06	4.06	1.97	4.64	2.07	5.15
	3	0.97	1.16	1.91	3.68	1.86	4.27	1.64	4.12
	4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5	0.99	1.21	1.01	2.14	1.00	2.60	0.86	2.29
	6	0.73	0.87	5.83	11.40	4.26	9.92	3.03	7.33
	7	0.88	1.12	0.57	1.42	0.50	1.59	0.41	1.18
Region	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	0.80	1.00	0.33	0.75	0.26	0.75	0.25	0.73
	3	0.82	0.96	1.67	2.97	1.38	2.84	1.32	3.25
	4	0.77	0.88	6.58	10.84	5.09	9.55	4.20	9.55
	5	0.81	1.08	0.10	0.30	0.08	0.32	0.07	0.26
	6	1.13	1.31	1.97	3.41	2.23	4.46	1.57	3.80
	7	0.98	1.33	0.10	0.31	0.10	0.41	0.07	0.26
	8	0.99	1.19	0.66	1.27	0.66	1.50	0.54	1.43
	9	0.88	1.03	1.06	1.91	0.94	1.97	0.80	2.05
	10	0.86	1.03	0.52	1.03	0.44	1.07	0.38	1.06

Table 25: Confidence intervals for Table 12.

		Limits							
Risk	Risk	Frequency		Severity		Premium		Premium , Tweedie	
factor	class	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Driver age	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	0.50	0.60	0.53	0.93	0.26	0.55	0.49	1.07
	3	0.31	0.36	2.34	3.65	0.73	1.32	3.12	6.06
	4	0.31	0.36	3.57	5.50	1.10	1.96	4.43	8.44
	5	0.28	0.34	0.37	0.69	0.10	0.24	0.34	0.78
Power	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.11	1.24	1.06	1.49	1.18	1.84	0.81	1.75
	3	1.14	1.26	1.61	2.23	1.84	2.81	1.21	2.53
	4	1.07	1.19	1.31	1.83	1.41	2.18	1.00	2.12
	5	1.12	1.30	0.35	0.57	0.39	0.73	0.23	0.55
	6	1.16	1.38	0.26	0.43	0.30	0.60	0.21	0.52
	7	1.17	1.38	0.29	0.49	0.34	0.68	0.16	0.42
	8	1.22	1.52	0.17	0.33	0.20	0.50	0.09	0.24
	9	1.03	1.42	0.07	0.21	0.08	0.30	0.04	0.13
	10	1.04	1.65	0.05	0.22	0.05	0.36	0.03	0.11
	11	1.06	1.81	0.03	0.14	0.03	0.26	0.01	0.05
	12	0.97	1.66	0.03	0.19	0.03	0.32	0.01	0.05
Region	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	0.81	1.01	0.35	0.72	0.29	0.72	0.27	0.76
	3	0.85	0.99	1.58	2.55	1.34	2.51	1.12	2.69
	4	0.79	0.90	5.82	8.76	4.61	7.89	4.64	10.19
	5	0.80	1.08	0.13	0.33	0.10	0.35	0.09	0.30
	6	1.09	1.25	2.08	3.26	2.26	4.08	1.73	4.01
	7	0.97	1.32	0.13	0.33	0.12	0.44	0.08	0.27
	8	1.00	1.19	0.71	1.23	0.71	1.46	0.52	1.34
	9	0.90	1.05	1.04	1.69	0.93	1.78	0.80	1.98
	10	0.87	1.05	0.55	0.98	0.48	1.03	0.47	1.23

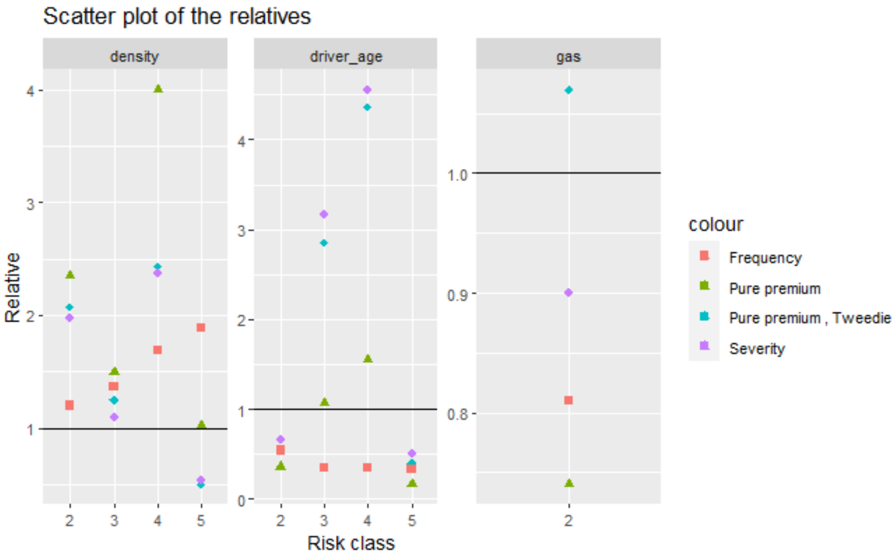

Figure 5: Scatter plot of all of the relatives in Table 10.

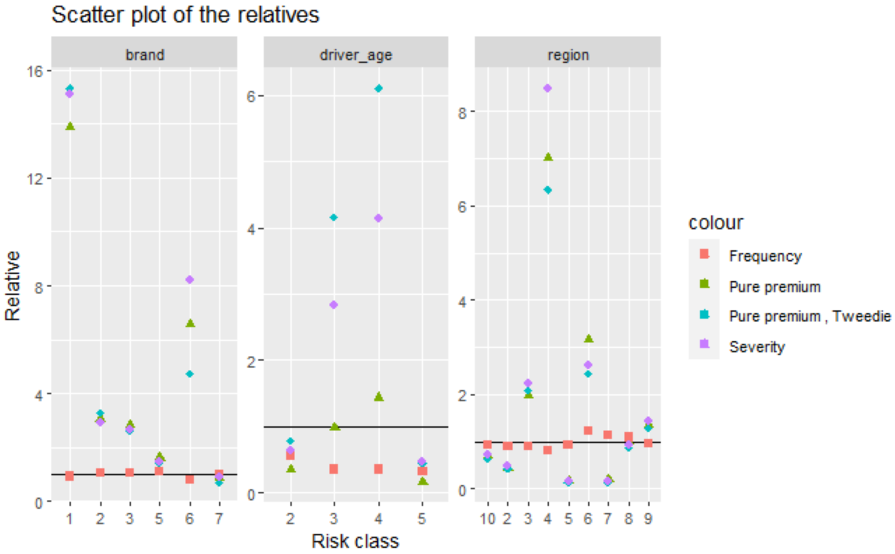

Figure 6: Scatter plot of all of the relatives in Table 11.

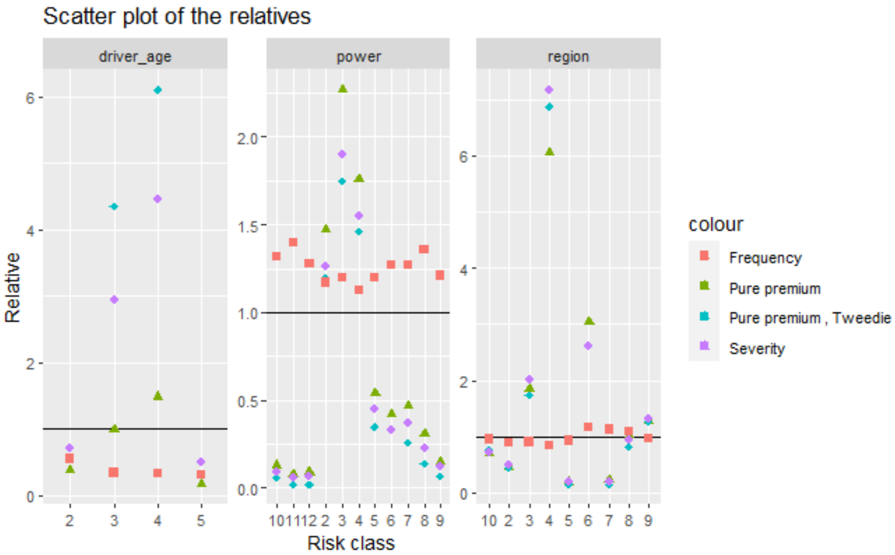

Figure 7: Scatter plot of all of the relatives in Table 12.