University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Nicholas Lupul

Claims Severity Modelling on the Basis of Publicly Available Vehicle Insurance Data

Actuarial and Financial Engineering

Master's Thesis (30 ECTS)

Supervisor: Assoc. Prof. Meelis Käärik

Tartu 2023

CLAIMS SEVERITY MODELLING ON THE BASIS OF PUBLICLY AVAILABLE VEHICLE INSURANCE DATA

Master's thesis

Nicholas Lupul

Abstract

The purpose of this master's thesis is to establish a baseline for what data potential new vehicle insurance providers should collect in order to establish models for claim size. An additional goal was to gather publicly available insurance data sets suitable for modelling claim size and provide an overview of these data sets. The first chapter is used to refresh the reader's knowledge of generalized linear models. In the second chapter the framework around which the analysis was conducted is introduced and the procedure for the analysis is detailed. The data sets used in the analysis are introduced in chapter three. Chapter four is used to present top models for each data set. The final chapter compares and highlights similarities between top models across the various data sets.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: motor vehicle insurance, generalized linear models, generalized additive models, generalized additive models for location scale and shape, R (programming language).

KAHJUTASU SUURUSE MODELLEERIMINE AVALIKULT KÄTTESAADAVATE MOOTORSÕIDUKITE KINDLUSTUSE ANDMETE PÕHJAL

Magistritöö

Nicholas Lupul

Lühikokkuvõte

Magistritöö eesmärk on määrata baasandmed, mida uued potentsiaalsed sõidukikindlustuse pakkujad peaksid kahju suuruse hindamise mudelite jaoks koguma. Eesmärk oli ka leida selle analüüsi jaoks avalikult kättesaadavaid kindlustusandmestikke ja anda neist ülevaade. Töö esimeses peatükis antakse ülevaade üldistatud lineaarsetest mudelitest. Teises peatükis tehakse analüüsi meetodite tutvustus. Kolmandas peatükis tutvustatakse analüüsis kasutatud andmeid. Mudeleid, mis sobisid igale andmestikule kõige paremini, esitletakse neljandas peatükis. Viimases peatükis tuuakse välja parimate mudelite võrdlused erinevate andmestike lõikes.[1]

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: mootorsõiduki kindlustus, üldistatud lineaarsed mudelid, üldistatud aditiivsed mudelid.

[1]The author is thankful to Kelly Tilga for providing translations from English to Estonian.

Contents

Introduction

In the case that a new company is starting out in the vehicle insurance industry, the prospective company will not have any previous data to base their claim size models. Furthermore, they may not have a clear understanding of what information should be gathered from customers in order to accurately model claim size, which will then be used to help make pricing models. The purpose of this master's thesis is to gather various vehicle insurance data sets in order to model claim size for each set individually. These models will then be compared to find common predictor variables across models, common trends among predictors, as well as potentially common distributions for claim size modelling.

The first chapter will serve the purpose of refreshing the reader's background knowledge of generalized linear models. In the second chapter the framework for the modelling process, generalized additive models for location, scale, and shape will be introduced. Additionally in this chapter the distributions considered in the modelling process will be discussed along with the step by step procedure taken during analysis. The data gathering process and a description of data sets is given in the third chapter. Chapter four discusses the model results for each data set, drawing attention to predictors with statistically significant coefficients in each model. The fifth chapter is used to discuss commonalities between models across the various data sets.

The author would like to offer his gratitude to the supervisor Meelis Käärik for his guidance, direction, and advice over the course of the thesis writing process. Lastly, the author is grateful to his peers, notably Artur Tuttar, Kaari Kuus, and Kelly Tilga, who continually provided constructive criticism, motivation, and support throughout the duration of the masters study program.

1 Generalized linear models

When studying the relationship between a response variable and predictor variables often the first approach is to apply a basic linear regression model. In order to trust the results of linear regression the normality assumption must hold, that is the residuals must follow a normal distribution. Failing to meet this assumption can come from a non linear relationship between the response and predictor variables, making a linear model inappropriate. In many real world situations, this assumption does not hold. For example insurance claim size cannot be assumed to be normally distributed, rather claim size is often assumed to be right skewed.

Building on basic linear models one will arrive at generalized linear models or GLMs which do not share the same normality assumptions discussed previously. GLMs are widely used across scientific fields due to the ease of interpretation and applicability.

1.1 GLM structure

We refer to (De Jong and Heller, 2008) for the results of this subsection.

Generalized linear models are another tool used to evaluate the relationship between a response variable Y and a number of predictor variables X_j for $j = 1, 2, ..., J$. GLMs accomplish this by estimating the transformed conditional mean of the response variable by a linear predictor. The transformation of the conditional mean is achieved via a link function, which will be discussed in Section 1.2. A restriction of GLMs is the assumption that the distribution of the response variable is a member of the exponential family of distributions. In order for a continuous distribution to be a member of the exponential family it must have a probability density function in the case of continuous distributions or probability mass function for discrete distributions that can be expressed in the following form:

$$f_Y(y) = c(y, \phi) \exp\left(\frac{y\theta - a(\theta)}{\phi}\right) \tag{1}$$

with θ referred to as the canonical parameter and ϕ the dispersion parameter. Some commonly known members of the exponential family include binomial, negative binomial, normal, Poisson, and gamma distributions.

It can also be shown that if a random variable Y has a distribution from the exponential family:

$$E(Y) = a'(\theta), \qquad Var(Y) = \phi a''(\theta)$$

where $a'(\theta)$ and $a''(\theta)$ are the first and second derivatives with respect to θ of $a(\theta)$.

1.2 Link functions

We refer to (De Jong and Heller, 2008) and (Lindsey, 1997) for the results of this subsection.

The link function is often expressed in the following form:

$$g(\mu_i) = \boldsymbol{x}_i{}^T \boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{iJ})^T$ is the vector of values of predictor variables for the ith observation, $\boldsymbol{\beta}$=$(\beta_0, \beta_1, \beta_2, ..., \beta_J)^T$ is the vector of coefficients, and $\mu_i = \mathrm{E}(Y|X_1 = x_{i1}, X_2 = x_{i2}, ..., X_J = x_{iJ})$ is the conditional mean of the response variable Y. Link functions are specified during the modelling process and there are different options to choose from.

The link function describes the relationship between the conditional mean of the ith observation and its linear predictor. During the modelling process, the link function is specified from a list of monotonic and differentiable functions. In the case that $g(\mu) = \theta$ the link function is referred to as the canonical link function. The choice of a canonical link function simplifies the computation of estimates, however advancements in computing technology allow the choice of more complex link functions. Generally when the canonical form of θ includes some constants, these constants are not included in the canonical link function. This can be seen in the case of the inverse Gaussian distribution where the canonical link is $\frac{1}{\mu^2}$ even though $\theta = -\frac{1}{2\mu^2}$. Some well known link functions and their names are given in Table 1.

Table 1: Well known link functions.

Name	Link function
log link	$g(\mu) = \ln \mu$
identity link	$g(\mu) = \mu$
inverse link	$g(\mu) = \frac{1}{\mu}$
squared inverse link	$g(\mu) = \frac{1}{\mu^2}$

1.3 Modelling with GLMs

We refer to (De Jong and Heller, 2008) for the results of this subsection.

The process for building a GLM is as follows:

1) An appropriate distribution for the response variable Y, from the exponential family should be chosen. The choice of distribution will determine $a(\theta)$.

2) A link function $g(\mu)$ must be decided upon as well. For simplicity the canonical link can be chosen but this does not have to be the case.

3) A choice of which predictor variables $X_1, X_2, ..., X_J$ will be used in the model must be made.

4) A model can then be fit on the data by estimating β. If the dispersion parameter ϕ is unknown it should also be estimated. Most commonly these estimations are done using maximum likelihood estimation.

5) Upon fitting the model, the model can be assessed for goodness of fit. One way to assess this is check the model's performance on a test data set. Additionally at this step it can be checked if the coefficients of any predictor variables are perhaps not statistically significant and therefore can be eliminated from the model.

1.4 Maximum likelihood estimation

We refer to (Hardin and Hilbe, 2018) for the results of this subsection.

Maximum likelihood estimation is a commonly used approach to parameter estimations within GLMs. As introduced in Section 1.1, the probability density/mass function for the response variable Y of a GLM can be seen in Equation 1. Since all y_i are assumed to be independent observations of the response variable Y defined by parameter θ_i, the joint density/mass function of the sample is simply the product of the probability density/mass functions at each individual y_i. When this joint density/mass function is expressed as a function of the canonical and dispersion parameters given y_i, it is referred to as the likelihood function and takes the form:

$$L(\boldsymbol{\theta}, \phi|\boldsymbol{y}) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - a(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)^T$ being the vector of exponential family canonical parameters for the sample and $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ being the response vector for the sample. The intent is to find values for $\boldsymbol{\theta}$ and ϕ that will maximize the likelihood function. The typical process for maximizing a function is to take the first derivative of the function with respect to the parameter of interest, set this derivative equal to zero and solve for the parameter of interest. This process will lead you to an estimate for the parameter. However, as the likelihood function is a product it can quickly become hard to derive. This can be remedied by exploiting the properties of logarithms.

By taking the natural logarithm of the likelihood function we will arrive at the log likelihood function:

$$l(\boldsymbol{\theta}, \phi|\boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - a(\theta_i)}{\phi} + c(y_i, \phi) \right\}.$$

Additionally, θ_i can be expressed using $\boldsymbol{\beta}$ in the following form: $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\boldsymbol{x}_i{}^T\boldsymbol{\beta}))$, where $g^{-1}(.)$ is the inverse of the link function. In order to find estimates for β_j, $j = 0, 1, ..., J$, the next step is to take the derivative of the aforementioned log likelihood function with respect to the model parameters. For a more detailed breakdown of the estimation process, refer to (Hardin and Hilbe, 2018).

2 Modelling with GAMLSS

2.1 Generalized additive models

The following subsection refers to (Hastie and Tibshirani, 1986).

Generalized additive models, or GAMs, extend upon the concept of GLMs. The linear predictor in GLMs is replaced by an additive predictor. As discussed in the previous section, for GLM we have:

$$g(\mathrm{E}(Y|X_1 = x_{i1}, X_2 = x_{i2}, ..., X_J = x_{iJ})) = \beta_0 + \beta_1 x_{i1} + ... + \beta_J x_{iJ}.$$

While in GAMs we modify this definition for the conditional expectation with:

$$g(\mathrm{E}(Y|X_1 = x_{i1}, X_2 = x_{i2}, ..., X_J = x_{iJ})) = s_0 + \sum_{j=1}^{J} s_j(x_{ij})$$

where $s_j(.)$ represent smooth functions which are standardized such that $\mathrm{E}\big(s_j(X_j)\big) = 0$. The R package Generalized Additive Models for Location, Scale and Shape or `GAMLSS`, was used in the modelling process.

2.2 GAMLSS

We refer to (Stasinopoulos, Rigby, and Bastiani, 2018) for the results of this subsection.

Since its inception in 2005, GAMLSS has been used for scientific research in varied fields, including but not limited to: actuarial science, biology, economics, finance and medicine. Although the smoothing capabilities of GAMLSS were not explored in this analysis, the justification for using GAMLSS rather than GLM is in the flexibility of the GAMLSS framework. While GLMs are restricted to exponential family distributions, GAMs applied through the `GAMLSS` package in this case offer a wider variety of distributions for the response variable. Additionally, by using a package with a wider scope of capabilities it allows this research to easily be extended upon in future works.

As was previously mentioned, GAMLSS which have the capability to handle linear, non-linear, or smooth non-parametric relationships between response and predictors, can be used as an alternative to GLMs. Often viewed as an extension of GLMs or GAMs, GAMLSS offers the user more flexibility in modelling. Unlike GLMs and GAMs, the assumed distribution used in GAMLSS is not restricted to the exponential family of distributions, in contrast it can be any parametric distribution. Furthermore, GAMLSS allows all distribution parameters to be modelled using linear or smooth functions of the predictor variables. This means that depending on the predictor variables, a change in location, scale, and shape of the response variable is allowed.

Due to the fact that parameter modelling can include the use of non-parametric smoothing functions of predictor variables, and the requirement of an assumed parametric distribution for the response variable, GAMLSS are referred to as semi-parametric regression-type models. There is an extensive range of distributions available for modelling the response variable, with a comprehensive guide available in (Rigby et al., 2017). GAMLSS models are conditional on up to 4 parameters μ_i being the location parameter, σ_i the scale parameter, ν_i and τ_i the shape parameters. Therefore under the GAMLSS framework, for independent observations $i = 1, 2, ..., n$, the probability density function of Y_i is as follows $f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$. While this model framework can be generalized to distributions with greater number of parameters, the package used to implement GAMLSS does not include distributions with more than 4 parameters.

The (theoretical) observations of response variable $Y_1, Y_2, ..., Y_n$ are independent with:

$$Y_i \sim D(\mu_i, \sigma_i, \nu_i, \tau_i).$$

For $k = 1, 2, 3, 4$ allow $g_k(.)$ to be a known monotonic link function relating a distribution to a predictor $\boldsymbol{\eta}_k$, where:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} s_{1j}(\boldsymbol{x}_j) \tag{3}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} s_{2j}(\boldsymbol{x}_j) \tag{4}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} s_{3j}(\boldsymbol{x}_j) \tag{5}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} s_{4j}(\boldsymbol{x}_j) \tag{6}$$

with $\boldsymbol{X}_k$ being a known design matrix, $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, ..., \beta_{kJ'_k})^T$ is a vector of coefficients, $\boldsymbol{\mu}$=$(\mu_1, \mu_2, ..., \mu_n)^T$, $\boldsymbol{\sigma}$=$(\sigma_1, \sigma_2, ..., \sigma_n)^T$, $\boldsymbol{\nu}$=$(\nu_1, \nu_2, ..., \nu_n)^T$, $\boldsymbol{\tau}$=$(\tau_1, \tau_2, ..., \tau_n)^T$, s_{kj} a smooth non-parametric function of variable X_{kj} and the $\boldsymbol{x}_j$'s are vectors of length n, for $j = 1, 2, ..., J_k$. Note that the first column of $\boldsymbol{X}_k$ and the first element of $\boldsymbol{\beta}_k$ correspond to the intercept term. Thus `GAMLSS` models permit distribution parameters to be modelled as linear, $\boldsymbol{X}_k\boldsymbol{\beta}_k$, or smooth term functions $s_{kj}(\boldsymbol{x}_j)$ for $k = 1, 2, 3, 4$.

As was mentioned at the onset of this subsection, for the scope of this analysis smoothing terms were not considered. In the modelling process, `GAMLSS` was used to fit individual models with 10 different assumed two parameter distributions for the response variable, claim size. Under these conditions, Equations 3 and 4 simplify to:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2$$

.

2.3 Candidate distributions

The following section provides an overview of the main properties of distributions that will be used as candidates to model claim size, Y, throughout the analysis as described in (Rigby et al., 2017). Going forward, it is important to note that μ and σ represent the shape and scale parameters respectively. It is not always the case that $\mu = \mathrm{E}(Y)$ in this parameterization. Similarly, it is not correct to view σ as the standard deviation of the response variable.

2.3.1 Gamma

The Gamma distribution is a 2 parameter member of the exponential family. It is continuous and right skewed making it a candidate to model right skewed data such as vehicle insurance claim size. From this point on, the gamma distribution will be referred to as `GA`. The parameterization of the probability density function used in the `GA` function is:

$$f_Y(y|\mu,\sigma) = \frac{y^{(\sigma^{-2}-1)}\exp\{\frac{-y}{\sigma^2\mu}\}}{(\sigma^2\mu)^{\sigma^{-2}}\Gamma(\sigma^{-2})}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the gamma distribution are given below:

- $\mathrm{E}(Y) = \mu$;
- $\mathrm{Var}(Y) = \sigma^2\mu^2$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = \frac{\gamma(\sigma^{-2}, y\mu^{-1}\sigma^{-2})}{\Gamma(\sigma^{-2})}$$

where $\gamma(a, x) = \int_0^x t^{a-1}e^{-t}dt$.

The two different link functions that were used to fit two separate models for each data set are given below.

Identity link: $g(\mu) = \mu$ and $g(\sigma) = \sigma$.

Log link: $g(\mu) = \ln\mu$ and $g(\sigma) = \ln\sigma$.

2.3.2 Inverse gamma

The inverse gamma distribution is another 2 parameter, continuous, right skewed member of the exponential family. From this point on, the inverse gamma distribution will be referred to as `IGAMMA`. The parameterization of the probability density function used in the `IGAMMA` function is:

$$f_Y(y|\mu, \sigma) = \frac{\mu^{\sigma^{-2}}(\sigma^{-2}+1)^{\sigma^{-2}}y^{-(\sigma^{-2}+1)}}{\Gamma(\sigma^{-2})}\exp\left\{-\frac{\mu(\sigma^{-2}+1)}{y}\right\}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the inverse gamma distribution are given below:

- $\mathrm{E}(Y) = \frac{(1+\sigma^2)\mu}{(1-\sigma^2)}$ if $\sigma^2 < 1$, $\mathrm{E}(Y) = \infty$ if $\sigma^2 \geq 1$;
- $\mathrm{Var}(Y) = \frac{(1+\sigma^2)^2\mu^2\sigma^2}{(1-\sigma^2)^2(1-2\sigma^2)}$ if $\sigma^2 < 1/2$, $\mathrm{Var}(Y) = \infty$ if $\sigma^2 \geq 1/2$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = \frac{\Gamma\left(\sigma^{-2}, \frac{\mu(\sigma^{-2}+1)}{y}\right)}{\Gamma(\sigma^{-2})}$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}dt$.

The link functions used in the modelling process were the `GAMLSS` package default log link for both μ and σ: $g(\mu) = \ln\mu$ and $g(\sigma) = \ln\sigma$.

2.3.3 Inverse Gaussian

The inverse Gaussian distribution is a 2 parameter, continuous, right skewed member of the exponential family. From this point on, the inverse Gaussian distribution will be referred to as

IG. The parameterization of the probability density function used in the IG function is:

$$f_Y(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left\{-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right\}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the inverse Gaussian distribution are given below:

- $\mathrm{E}(Y) = \mu$;
- $\mathrm{Var}(Y) = \sigma^2\mu^3$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = \Phi\Big[(\sigma^2 y)^{-1/2}\Big(\frac{y}{\mu}-1\Big)\Big] + e^{2(\mu\sigma^2)^{-1}}\Phi\Big[(\sigma^2 y)^{-1/2}\Big(\frac{y}{\mu}+1\Big)\Big].$$

The link functions used in the modelling process were the GAMLSS package default log link for both μ and σ: $g(\mu) = \ln\mu$ and $g(\sigma) = \ln\sigma$.

2.3.4 Lognormal

The lognormal distribution is a 2 parameter, continuous, right skewed member of the exponential family. From this point on, the lognormal distribution will be referred to as LOGNO. The parameterization of the probability density function used in the LOGNO function is:

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

for $0 < y < \infty$, $-\infty < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the lognormal distribution are given below:

- $\mathrm{E}(Y) = e^{\mu+\sigma^2/2}$;
- $\mathrm{Var}(Y) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = \Phi\Big(\frac{\ln y - \mu}{\sigma}\Big).$$

The link functions used in the modelling process were the GAMLSS package default identity link for μ_i and log link for σ: $g(\mu) = \ln\mu$ and $g(\sigma) = \ln\sigma$.

2.3.5 Weibull

The Weibull distribution is also a 2 parameter continuous, right skewed distribution. It was named after Swedish engineer Waloddi Weibull and is classified as an extreme value distribution as it is often used to describe uncommon or rare occurrences. It has applications in modelling real world extreme events such as floods, droughts and drastic insurance losses (*WeibullDistribution* 2016). Three different version of the Weibull distribution available in `GAMLSS` were used. From this point on, the Weibull distributions will be referred to as `WEI`, `WEI2`, and `WEI3`. The parameterization of the probability density function used in the `WEI` function is presented in the following form:

$$f_Y(y|\mu,\sigma) = \frac{\sigma y^{\sigma-1}}{\mu^\sigma} \exp\left\{-\left(\frac{y}{\mu}\right)^\sigma\right\}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the `WEI` distribution are given below:

- $\text{E}(Y) = \mu[\Gamma(\sigma^{-1}+1)]$;
- $\text{Var}(Y) = \mu^2[\Gamma(2\sigma^{-1}+1) - (\Gamma(\sigma^{-1}+1))^2]$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = 1 - \exp\left\{-\left(\frac{y}{\mu}\right)^\sigma\right\}.$$

The parameterization of the probability density function used in the `WEI2` function is presented in the following form:

$$f_Y(y|\mu,\sigma) = \sigma\mu y^{(\sigma-1)} \exp\{-\mu y^\sigma\}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the `WEI2` distribution are given below:

- $\text{E}(Y) = \mu^{(\frac{-1}{\sigma})}[\Gamma(\sigma^{-1}+1)]$;
- $\text{Var}(Y) = \mu^{(\frac{-2}{\sigma})}[\Gamma(2\sigma^{-1}+1) - (\Gamma(\sigma^{-1}+1))^2]$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = 1 - \exp\{-\mu y^\sigma\}.$$

The parameterization of the probability density function used in the WEI3 function is presented in the following form:

$$f_Y(y|\mu,\sigma) = \frac{\sigma y^{\sigma-1}}{\beta^\sigma} \exp\left\{-\left(\frac{y}{\beta}\right)^\sigma\right\}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$, where $\beta = \frac{\mu}{\Gamma(\frac{1}{\sigma}+1)}$.

The mean and variance of the WEI3 distribution are given below:

- $\mathrm{E}(Y) = \mu$;
- $\mathrm{Var}(Y) = \beta^2[\Gamma(2\sigma^{-1}+1) - (\Gamma(\sigma^{-1}+1))^2]$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = 1 - \exp\left\{-\left(\frac{y}{\beta}\right)^\sigma\right\}.$$

The link functions used in the modelling process for all of the Weibull models were the GAMLSS package default log link for both μ and σ: $g(\mu) = \ln\mu$ and $g(\sigma) = \ln\sigma$.

2.3.6 Pareto

The Pareto distribution, named after Vilfredo Pareto has historically been used to model the distribution of wealth or income (*ParetoDistribution* 2016). As with the other distributions discussed it is right skewed and continuous. Various different forms of the Pareto distribution exist, the Pareto Type 2 distribution PARETO2 and Pareto Type 2 original distribution PARETO2o were used as modelling candidates. From this point on, the Pareto distributions will be referred to as PARETO2, and PARETO2o.

The parameterization of the probability density function used in the two parameter PARETO2 function is:

$$f_Y(y|\mu,\sigma) = \frac{\sigma^{-1}\mu^{\sigma^{-1}}}{(y+\mu)^{\sigma^{-1}+1}}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the PARETO2 distribution are given below:

- $\mathrm{E}(Y) = \frac{\mu\sigma}{(1-\sigma)}$ for $\sigma < 1$, $\mathrm{E}(Y) = \infty$ for $\sigma \geq 1$;
- $\mathrm{Var}(Y) = \frac{\mu^2\sigma^2}{(1-\sigma)^2(1-2\sigma)}$ for $\sigma < 1/2$, $\mathrm{Var}(Y) = \infty$ for $\sigma \geq 1/2$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = 1 - \frac{\mu^{\sigma^{-1}}}{(y+\mu)^{\sigma^{-1}}}.$$

The parameterization of the probability density function used in the two parameter original `PARETO2o` function is presented in the following form:

$$f_Y(y|\mu,\sigma) = \frac{\sigma\mu^{\sigma}}{(y+\mu)^{\sigma+1}}$$

for $0 < y < \infty$, $0 < \mu < \infty$, and $0 < \sigma < \infty$.

The mean and variance of the `PARETO2o` distribution are given below:

- $\mathrm{E}(Y) = \frac{\mu}{(\sigma-1)}$ for $\sigma > 1$, $\mathrm{E}(Y) = \infty$ for $\sigma \leq 1$;
- $\mathrm{Var}(Y) = \frac{\sigma\mu^2}{(\sigma-1)^2(\sigma-2)}$ for $\sigma > 2$, $\mathrm{Var}(Y) = \infty$ for $\sigma \leq 2$.

The cumulative distribution function is given by the following equation:

$$F_Y(y) = 1 - \frac{\mu^{\sigma}}{(y+\mu)^{\sigma}}.$$

The link functions used in the modelling process for all of the Pareto models was the `GAMLSS` package default log link for both μ and σ: $g(\mu) = \ln\mu$ and $g(\sigma) = \ln\sigma$.

2.4 Modelling process

The modelling process followed throughout the analysis was as described in (Ramires, 2021). Step wise selection was performed through the use of the `GAMLSS` function `stepGAICALL.A`, the model selection process implemented in this function is described below.

1) Build a model for μ using a forward step wise approach.

2) Given the model for μ build a model for σ using a forward step wise approach.

3) Given the models for μ and σ build a model for ν using a forward step wise approach.

4) Given the models for μ, σ, and ν build a model for τ using a forward step wise approach.

5) Given the models for μ, σ, ν, and τ check if all terms for ν are needed via backward step wise selection.

6) Given the models for μ, σ, ν, and τ check if all terms for σ are needed via backward step wise selection.

7) Given the models for μ, σ, ν, and τ check if all terms for μ are needed via backward step wise selection.

In our case we are only dealing with 2 parameter distributions which will shorten this process substantially to:

1) Build a model for μ using a forward step wise approach.

2) Given the model for μ build a model for σ using a forward step wise approach.

3) Given the models for μ, and σ check if all terms for μ are needed via backward step wise selection.

2.4.1 Model selection

The following subsection refers to (Akaike, 1998).

The criterion for variable selection at each stage of the modelling process was Akaike information criterion, or AIC. AIC is given by the following formula: $\text{AIC} = -2\hat{l}(\theta) + 2df$ where $\hat{l}(\theta)$ is the fitted log-likelihood, and df is the degrees of freedom of the model. In the case of a GAMLSS model with 2 parameters and no smoothing terms, the degrees of freedom can be calculated using the formula $df = J_1 + J_2 + 2$ with J_1 being the number of predictors in the μ model and J_2 being the number of predictors in the σ model.

2.4.2 Manual adjustments

The top 2 models for each data set, by AIC, were then studied. In some cases these models included terms that were not statistically significant at a significance level of 5%. Although including the predictor variables reduced the AIC, it was concluded that these variables should be removed one by one until coefficients of all predictor variables are statistically significant. In the case of multi level factor variables, if the coefficients of all levels were not significant, the likelihood ratio test was preformed in to determine if the predictor should remain in the model. The process for manually removing variables that were not significant is given below.

1) Study the summary of the model given by `stepGAICALL.A`. If coefficients of all predictor variables are significant, this is the final model.

2) In the situation where not all coefficients of predictor variables are significant, first look at the σ model and remove a variable with coefficients that were not statistically significant. If the

σ model has no such terms, look at the μ model and remove a variable with coefficients that were not statistically significant. It is important to only remove one term at a time, as removing one may cause the significance level of other terms to change.

3) Repeat steps 1 and 2 until all coefficients of predictor variables are significant.

The model achieved after this process was taken as the final model and was used in comparisons across data sets. While the AIC of the manually adjusted model is higher than before the adjustments, the increase was slight.

To see the modelling process, including manual adjustments, as performed in R refer to Listings 1, 2, 3, and 4.

3 Data gathering

Data gathering proved to be a difficult and time consuming task. Vehicle insurance companies are not motivated to share their data with the public as it could allow for competitors to potentially train more effective models which would negatively effect the profit and performance of the company with public data. For this reason, many of the publicly available vehicle insurance data sets are old, simulated or both. In one particular case with data supplied by American insurance company Allstate, all variables were masked so that it was not known what they were measuring. This type of data cannot be used for the purpose of this research as the intent is to find relationships between predictors across different data sets.

3.1 Data sources

Data sets were eventually found from a variety of sources. First, R packages that contained insurance related data sets were found and assessed to see if they could be used. The package `insuranceData` (Wolny-Dominiak and Trzesiok, 2014) provided four data sets: *AutoClaims*, *AutoCollision*, *dataCar* and *dataOhlsson*. The package `CASdatasets` (Dutang and Charpentier, 2020) supplied the data sets *norauto* and *swautoins*. Kaggle competitions were searched where one useful data sets was found. It was also the case that many of the data sets were duplicated across different sources with varied names. For example, *dataCar* from `insuranceData` is the same as *ausprivauto0405* from `CASdatasets`. This confirms the previous claim that finding numerous data sets relating to vehicle insurance is not an easy task.

It is also important to note that these data sets are taken from various different time frames and different countries. Additionally, the currency of claim amount is not consistent across data sets. This is not a large issue for this analysis as the subject of interest is finding commonalities between models for claim size across various data sets. Since the goal is not to find a model for claim size in one specific region at a given time, the effect of currency exchange as well as inflation does not have to be accounted for. It should be noted however, that in subsequent models due to differences in currencies and denomination of claim size across data sets, the magnitude of coefficients cannot be directly compared across models for different data sets.

3.2 Data set description

Table 2 can be used to see a quick summary of the data sets as well as links to their sources. Additionally, opening the following link will direct the reader to a web page with data set summaries as well as links to download the data sets. The code for this web page was written with help from ChatGPT (OpenAI, 2021). ChatGPT was used to create a template that was then edited by the author, the prompt given to and template supplied by the chat bot can be seen in Listing 5.

Table 2: Data set summary.

Data set	Source	Observations	Variable count
AutoClaims	insuranceData (R package)	6773	5
AutoCollision	insuranceData (R package)	32	4
dataCar	insuranceData (R package)	67856	11
dataOhlsson	insuranceData (R package)	64548	9
norauto	CASdatasets (R package) pg.74	183999	7
swautoins	CASdatasets (R package) pg.90	2182	7
Claims	Macquarie University Website	2746	33
Kaggle5	Kaggle	8630	22

AutoClaims includes 6773 observations of 5 variables from an American provider of automobile insurance.

AutoCollision has 32 observations of 4 variables which is a sample of a 1999 data set on automobile policies in the United Kingdom. It is worth noting that the sample size is very small in this data set.

dataCar contains 67856 observations of 11 variables of one year insurance policies from 2004 and 2005.

dataOhlsson consists of 64548 entries of 9 variables from a study on motorcycle insurance by Swedish insurer Wasa from 1994 to 1998. While this data set deals with motorcycles rather than automobiles, it is included in the analysis due to the difficulty finding data. Furthermore, it would not be unreasonable to assume motorcycle insurance claims follow similar trends to automobile insurance claims.

norauto contains 183999 observations of 7 variables from a Norwegian auto insurer.

swautoins is 2182 entries of 7 variables from 1977 gathered by the Swedish Committee on the Analysis of Risk Premium.

Claims had 10303 observations of 33 variables. This data set was originally found on the website of Macquarie University a school in Sydney, Australia in 2022. Unfortunately since the data gathering phase the data set has been removed. While an attempt to locate the data from other sources was made, it was not found. As the data had already been downloaded, it was able to be used for analysis.

The *Kaggle5* data set is 8630 observations of 22 variables. The exact source of the data is unclear as it is not stated on Kaggle. While it was decided to include this dataset in the data gathering phase, it will not be included in the analysis as it is unclear if the data is simulated and some variables are unclear.

3.3 Data preparation

Prior to analysis, some data preparation was necessary. As the analysis is focused on insurance claim size modelling, it was necessary to only use the observations with claim size greater than zero. Data for policies that had no claim are not of interest, these were removed. Additionally, many of the sets had variables that were of the wrong type when initially loaded into R. This was often the case with factor variables. All data sets were sifted through to correctly relabel factor variables if needed. Some data cleaning was needed in the *Claims* data set. Variables that were not useful for modelling were removed. These removed variables were primarily relating to ID numbers and dates.

4 Modelling claim size

In this section models with various distributions for claim size as the response variable Y, were fit to the 7 data sets described in Section 3.2 in an attempt to model vehicle insurance claim size. The distributions applied were inverse gamma, log normal, gamma, inverse Gaussian, 2 forms of the Pareto distribution, and 3 forms of the Weibull distribution. Details about the distributions can be found in Section 2.3. It is well known that vehicle insurance claim sizes are right or positively skewed. This arises from the fact that many of the claims are of low cost, perhaps minor damages, or damages to a less valuable vehicle. While occasionally major damages occur or very expensive vehicles are heavily damaged or written off causing the distribution of claim sizes to be right skewed. Therefore the choice of 2 parameter right skewed distributions as candidates is justified.

The purpose of this analysis was to see if there would be similarities across models for the different data sets. These similarities could come in the form of distribution of response variable Y or common predictors across various data sets and models. Also of interest was the direction of significant coefficients of predictors and if this direction was consistent across models with similar predictors.

4.1 Model fitting

The analysis was performed in R software (version 4.2.1) using the integrated development environment RStudio (version 1.4.1717) for Mac. The R packages and their corresponding references are listed here: `base` (R Core Team, 2022), `GAMLSS` (Rigby and Stasinopoulos, 2005), `lmtest` (Zeileis and Hothorn, 2002), and `readr` (Wickham, Hester, and Bryan, 2022).

The process for modelling each of the data sets is detailed below.

1. If the data contained policies without claims, remove these observations leaving just the observations with claim size greater than 0.

2. If the data contained variables read into R as the wrong type, reclassify them. Most commonly this was the case with factor variables being classified as numeric.

3. A histogram of claim size Y was viewed in order to confirm the right skewed property of the response variable.

4. A base model with no predictors was fit on the pre processed data set for one of the candidate distributions using the `GAMLSS` function.

5. The model obtained in step 4) was taken as the base model to be used in the function `stepGAICAll.A`. Maximal scopes for predictors of μ and σ were set to be the same.

6. A summary was then printed to view the resulting model.

7. Steps 4), 5), and 6) were repeated for all candidate distributions.

8. Comparison of candidate models was made via AIC, models with the lowest and second lowest AIC were brought to the next step.

9. Iterative manual removal of variables in which the coefficient of their predictor was not statistically significant was performed as described in Section 2.4.2.

10. The predictors and their terms for the final two models were recorded to compare against the other data sets.

4.2 Top models

In the following section the top 2 models for each data set will be summarized in a table as well as described more in depth.

4.2.1 Summary

Table 3 summarizes the distributions found to be the best candidate models for each data set. More information about the candidate distributions can be found in earlier sections as follows: gamma (`GA` in Section 2.3.1), inverse gamma (`IGAMMA` in Section 2.3.2), inverse Gaussian (`IG` in Section 2.3.3), lognormal (`LOGNO` in Section 2.3.4), Weibull (`WEI`, `WEI2`, `WEI3` in Section 2.3.5), and Pareto (`PARETO2` and `PARETO2o` in Section 2.3.6).

Table 3: Top models summary.

Data set	Top candidate distribution	Second candidate distribution
AutoClaims	LOGNO	PARETO2,PARETO2o
AutoCollision	GA[1]	IG
dataCar	IG	IGAMMA
dataOhlsson	LOGNO	WEI
norauto	PARETO2,PARETO2o	GA[2]
swautoins	GA[1]	WEI
Claims	LOGNO	IG

[1] Identity link function for μ and σ.

[2] Log link function for μ and σ.

4.2.2 *AutoClaims*

Prior to manual adjustments, the best performing model, by AIC, for *AutoClaims* was found by fitting a lognormal model with default link functions. The default link function for μ is the identity link, the default link function for σ is the log link. During the manual adjustment phase it was found that the effect of gender in the μ model was not statistically significant and therefore this term was dropped. No further manual adjustments were needed. To see the summary of the top model refer to Listing 6 in the Appendix.

The model had 2 terms for μ: state and age of driver. State is not particularly helpful for this analysis as the actual states are unknown, rather they are represented by integers. Furthermore, region based variables are not of much use when data from various countries is being compared. It is interesting to note that, while keeping other predictor values unchanged, as the age of driver increased by one year, μ decreased by approximately 0.003. A squared age term was also allowed into the model but it was dropped by the `stepGAICAll.A` function.

There are two terms for σ: age and gender. The baseline gender is female. Unlike the model for μ, The age of driver coefficient is positive, implying that as a driver's age increases, so does σ. Additionally, male drivers see a decreased σ when compared to female drivers.

As seen in Section 2.3.4, the mean of the lognormal distribution can be calculated as $\text{E}(Y) = e^{\mu+\sigma^2/2}$. Taking this formula into account along with the models for μ and σ it can be concluded that as driver age increases, so does the average insurance claim size. Furthermore, female drivers can expect a lower average claim size when compared to males.

The second best performing model was found by fitting either Pareto model with default log link functions for both μ and σ. PARETO2o and PARETO2 gave nearly identical models for μ with coefficients differing after the third decimal place. The effect of the only predictor was statistically significant and therefore no manual adjustments were needed. To see the summaries of the secondary models refer to Listing 7 (PARETO2o) and Listing 8 (PARETO2) in the Appendix.

The model for μ only consists of the term state. The σ model is empty with only the intercept term. As was previously discussed the state predictor is of little use in this analysis. The PARETO2 model suggested a constant σ of $\exp(-1.55604) \approx 0.21$ while the PARETO2o model suggested a constant σ of $\exp(1.55621) \approx 4.74$. This difference is to be expected as σ in the probability density function of PARETO2o is essentially replaced by σ^{-1} in the probability density function of PARETO2, as seen in Section 2.3.6.

For this model nothing of use can be gathered about the average claim size as no information is provided about the factor variable for state.

4.2.3 *AutoCollision*

The top model for *AutoCollision* was provided by the gamma distribution with identity link functions for both μ and σ. During the manual adjustment phase it was found that the effect of the age factor variable was not statistically significant. However upon removal of this variable, the model was not able to be fit. Errors persisted as σ was being estimated as negative, which is not possible. This is a drawback of using the identity link function for σ, as the range is not restricted to $[0, \infty)$ like it is when using the log link function. As there were no other variables with coefficients that were not statistically significant, variables could not be dropped in a different order. Due to this and the fact that the gamma with identity link functions model had the best AIC by a large margin, it was decided to keep this model despite it including the not statistically significant coefficient for the age variable. To see the summary of the top model refer to Listing 9 in the Appendix.

The μ model has terms vehicle use as a factor with 4 levels, and age group as a factor variable with 8 levels. The baseline category for vehicle use is business, while the baseline category for age group is A which is unknown. The age factor is not very useful as it is not specified what the ranges are or even if the factors are in descending or ascending order. While keeping other predictor values unchanged, drivers primarily using their vehicles for short drives or for pleasure can expect a decrease in μ by about 138 and 137 pounds sterling respectively. In this case as seen in Section 2.3.1 the expected value of Y is equal to μ which allows a simple interpretation of this model. While disregarding age groups, policy holders using their vehicles for business can expect the highest claim size, followed by, long drivers, then pleasure drivers, and lastly short drivers can expect the lowest average claim size.

The σ model just has the age group term. As mentioned earlier the factor age variable does not provide much information as its definition is unclear. The only conclusion we can draw from the σ model is that the σ is not assumed to be constant.

The second best model in terms of AIC was the inverse Gaussian model with default log link functions for both μ and σ. During the manual adjustment phase it was found that coefficients of all predictors were statistically significant and therefore no manual adjustments were needed. To see the summary of the secondary model refer to Listing 10 in the Appendix.

The μ model has terms for vehicle use as a factor with 4 levels, and age group as a factor with 8 levels. Similarly to the gamma model, while keeping other predictor values unchanged, a driver using their vehicle for pleasure or for short drives would expect μ to reduce, by factors of $\exp(0.57431) \approx 1.78$ and $\exp(0.49018) \approx 1.63$ respectively. Once again the interpretation is simplified because $\mathrm{E}(Y) = \mu$ in this case. Therefore all observations stated for μ can be extended to the average claim size.

The σ model just has a term for intercept. Therefore it is implied by the model that the σ is $\exp(-4.954) \approx 0.007$.

4.2.4 *dataCar*

The top model for *dataCar* was provided by the inverse Gaussian distribution with default log link functions for both μ and σ. During the manual adjustment phase it was found that coefficients of all predictors were statistically significant and therefore no manual adjustments were needed. To see the summary of the top model refer to Listing 11 in the Appendix.

The model for μ has terms for exposure, area as a factor with 6 levels, age of driver as a factor with 6 levels and gender. The baseline category for area is A which is unknown. The baseline category for age of driver is the youngest group. The baseline category for gender is female. According to the model, as exposure, a numeric variable with a possible range of $[0, 1]$, increases the μ decreases, assuming other predictors remain unchanged. The gender coefficient implies that male drivers can expect μ to increase by a factor of $\exp(0.12644) \approx 1.13$ when compared to female drivers, assuming other predictors remain unchanged. Once again the interpretation is simplified because $\mathrm{E}(Y) = \mu$ in this case. Therefore all observations stated for μ can be extended to the average claim size.

The model for σ has terms for exposure, and area as a factor with 6 levels. While keeping other predictor values unchanged, as exposure increases from 0 to 1, σ changes by a factor of $\exp(-0.104349) \approx 0.90$.

The second best model for *dataCar* was found from fitting the inverse gamma distribution with default log link functions for both μ and σ. During the manual adjustment phase it was found that coefficients of all predictors were statistically significant and therefore no manual adjustments were needed. To see the summary of the secondary model refer to Listing 12 in the Appendix.

The model for μ has terms vehicle value, area as a factor with 6 levels, and vehicle age as a factor with 4 levels. The baseline category for vehicle age is the youngest category. While keeping other predictor values unchanged, when the vehicle value is increased by one the μ also increases by a factor of $\exp(0.063013) \approx 1.07$. While keeping other predictor values unchanged, as a vehicle increases in age the μ also increases, in the case that a vehicle falls into the oldest age category the μ rises by a factor of $\exp(0.199839) \approx 1.22$.

The model for σ has terms for exposure, vehicle value, policy holder age as a factor with 6 levels, and gender. The baseline category for policy holder age is the youngest category. The baseline category for gender is female. While keeping other predictor values unchanged, as exposure increases σ changes by a factor of $\exp(-0.15617) \approx 0.86$. As vehicle value increases by one, σ changes by a factor of $\exp(-0.02737) \approx 0.97$ in the case that other predictors remain unchanged. Additionally male drivers can expect an increase in σ. In general as the drivers age increased σ decreased, however in the oldest age category the decrease was less than that of the second oldest category.

For the inverse gamma distribution, as seen in Section 2.3.2, $\mathrm{E}(Y) = \frac{(1+\sigma^2)\mu}{(1-\sigma^2)}$ for $0 < \sigma^2 < 1$ and $\mathrm{E}(Y) = \infty$ for $\sigma^2 \geq 1$. While holding other predictors unchanged, an increase in vehicle value, or vehicle age will result in an increase in average claim size. Additionally, increases in exposure or vehicle value while keeping the other predictors the same will lead to an increased average claim size. Male policy holders can expect a higher average claim size, as compared to females, when all other predictors remain the same.

4.2.5 *dataOhlsson*

The best model for *dataOhlsson* came from the lognormal model with default link functions, identity link for μ and log link for σ. During the manual adjustment phase it was found that coefficients of all predictors were statistically significant and therefore no manual adjustments were needed. To see the summary of the top model refer to Listing 13 in the Appendix.

The model for μ has terms for vehicle age, vehicle owner age, and an area factor variable with 7 levels. The baseline category for area is 1 which is unknown. While keeping other predictor values unchanged, as vehicle age increases, μ decreases. Similarly, as owner age increases, μ decreases. As was the case with the previous models, the area factor variable is not of interest.

The σ model has terms for area as a factor variable with 7 levels and MC class as a factor variable with 7 levels. MC class is a classification based on a ratio calculated as 100 times the engine power over the sum of the weight of the vehicle and average driver in kilograms. The baseline categories for both of these factors is unknown. As the categories are unknown, these variables are of little interest.

For lognormal, $\mathrm{E}(Y) = e^{\mu+\sigma^2/2}$, additionally since σ is only influenced by the area variable, some observations about average claim size can be made. Average claim size will see a decrease as age of policy holder increases, while other predictors remain unchanged. The same conclusion can be made for an increase in age of vehicle.

The next best model for *dataOhlsson* comes from the Weibull model with default log link functions for both μ and σ. During the manual adjustment phase it was found that coefficients of all predictors were statistically significant and therefore no manual adjustments were needed. To see the summary of the secondary model refer to Listing 14 in the Appendix.

The μ model has terms for owner age, an area factor variable with 7 levels, and vehicle age. While keeping other predictor values unchanged, as owner age or vehicle age increase, the model suggests a decrease in μ.

The σ model has terms for MC class as a factor variable with 7 levels, vehicle age, and owner age. While keeping other predictor values unchanged, as owner age or vehicle age increases, σ decreases.

The formula for average claim size for the `WEI` is $\mathrm{E}(Y) = \mu[\Gamma(\sigma^{-1} + 1)]$ as found in Section 2.3.5. Referring to this formula and the fact that $\sigma > 0$ for this distribution, it can be observed that policy holder age increases, while other predictors remain unchanged, average claim size decreases. The same conclusion can be made for an increase in vehicle age.

4.2.6 *norauto*

The best performing model for *norauto* was a tie between both Pareto distributions with default log link functions for both μ and σ. However, the models for μ and σ are intercept only models.

The similarity between intercept only Pareto models is to be expected. From studying the density functions presented in Section 2.3.6, it can be seen that when σ in the PARETO2 probability density function is replaced with σ^{-1} the probability density function is the same as that of the PARETO2o distribution. The PARETO2 model suggested a constant μ of $\exp(12.4147) \approx 246400$ and a constant σ of $\exp(-2.4411) \approx 0.09$. The PARETO2o model suggested a constant μ of $\exp(12.4149) \approx 246450$ and a constant σ of $\exp(2.4413) \approx 11.49$. As expected, μ for both Pareto distributions are similar while for PARETO2, $\sigma^{-1} = \exp(-2.4411)^{-1} \approx 11.49$ which is equal to σ for PARETO2o. To see the summaries of the top models refer to Listing 15 in the Appendix.

Due to the Pareto models only consisting of intercepts, the average claim size is assumed to be constant. The average claim size for as calculated using the formulas for expectation found in Section 2.3.6 is approximately 23498 for both PARETO2 and PARETO2o models.

The next best model was provided by the gamma distribution with log link functions for both μ and σ. During the manual adjustment phase it was found that coefficients for distance limit and gender variables in the μ model were not statistically significant and therefore these terms were dropped. Prior to dropping the factor variable for distance limit, the likelihood ratio test was performed to compare the model including the predictor with the model excluding it. The p-value of the test was greater than the significance level of 5% allowing it to be concluded that this term can be removed. Similarly, coefficients for gender and if the driver is young variables in the σ model were not statistically significant and therefore these terms were dropped. To see the summary of the secondary model refer to Listing 16 in the Appendix.

After these adjustments, the models for both μ and σ were both intercept only models. The intercept only models suggested a constant μ of $\exp(10.06851) \approx 23590$ and a constant σ of $\exp(-0.061112) \approx 0.94$.

In the case of gamma models Section 2.3.1 gives us $\mathrm{E}(Y) = \mu$, therefore the average claim size suggested by this model is approximately 23590. This is very close to what was observed as the average claim size for the Pareto models.

4.2.7 *swautoins*

The best model for *swautoins* is given by the gamma distribution with identity link functions for both μ and σ. No manual adjustments were necessary as all coefficients for predictor variables were statistically significant. To see the summary of the top model refer to Listing 17 in the Appendix.

The model for μ has terms for the number of policy holder years, bonus class as a factor with 7 levels, and vehicle brand as a factor with 9 levels. The baseline categories for brand and bonus class are not specified. While keeping other predictor values unchanged, an increase by one of the variable for policy holder years resulted in an increase of μ by about 224. All levels of the bonus class aside from the baseline reduced μ to varying degrees. The brand of the vehicle had varying effects on μ, however as the actual brands are not given, little insight is gained from this.

The model for σ has terms for the number of policy holder years, bonus class as a factor with 7 levels, vehicle brand as a factor with 9 levels, kilometers driven as a factor with 5 levels and an area variable as a factor with 7 levels. The baseline categories of these factor variables is unknown. Once again brand and the location variable do not provide much insight. While keeping other predictor values unchanged, as the policy holder years increased, so did σ. Bonus class and kilometers driven factor variables are not labelled and therefore provide little insight.

Due to the simple relationship between μ and $\text{E}(Y)$ it can be concluded that for this model, while other predictors are held unchanged, as number of policy holder years increases, so does the average claim size. While the factor variable levels are not specified, the highest claim sizes are expected for brand 9 and bonus class 1.

The second best model is provided by the Weibull distribution with default log link functions for both μ and σ. No manual adjustments were necessary as all coefficients for predictor variables were statistically significant. To see the summary of the secondary model refer to Listings 18 and 19 in the Appendix.

The model for μ has terms for bonus class as a factor with 7 levels, vehicle brand as a factor with 9 levels, kilometers driven as a factor with 5 levels and an area variable as a factor with 7 levels. All of these factor variables have varied effects for different levels, and since the levels are not explained it is difficult to interpret.

The σ consists of terms for bonus class as a factor with 7 levels, vehicle brand as a factor with 9 levels, kilometers driven as a factor variable with 5 levels and an area variable as a factor with 7 levels. All of the brands aside from the baseline reduced σ. All bonus classes aside from the baseline increased σ. As previously stated, with the levels not being given for these factor variables few conclusions can be drawn.

Due to the fact that all predictors are factor variables with unknown levels, it is difficult to conclude anything about average claim size from this model.

4.2.8 *Claims*

The top model for *Claims* is given by the lognormal distribution with identity link function for μ and log link function for σ. During the manual adjustment phase it was found that coefficients for if the policy holder is married variable and motor vehicle record points in the μ model were not statistically significant and therefore these term was dropped. Additionally the gender variable for σ was dropped as its coefficient was not statistically significant. No further manual adjustments were needed. To see the summary of the top model refer to Listing 20 in the Appendix.

The μ model only had a term for value of vehicle. As vehicle value increased, so did μ.

The σ model contains terms for level of education as a factor with 5 levels, how many years the policy holder has lived in the same location, occupation as a factor with 9 levels, and vehicle value. The baseline level for education is below high school. The baseline occupation is blank. Effects of factor variables on σ varied as the levels varied. While keeping other predictor values unchanged, an increase in the number of years a policy holder has lived at their address decreases σ. An increase in vehicle value increased the σ, when other predictors are not changed.

As seen in Section 2.3.4, $\mathrm{E}(Y) = e^{\mu+\sigma^2/2}$ for the lognormal distribution. Therefore, increases in vehicle value, while other predictors remain unchanged, will lead to increases in average claim size. Additionally, additional years lived at the same address will lead to increases in claim size, if other predictors are not changed.

The next best model for *Claims* is provided by the inverse Gaussian model with default log link functions for both μ and σ. During the manual adjustment phase it was found that the coefficient for travel time in the σ model was not statistically significant and therefore this term was dropped. To see the summary of the secondary model refer to Listing 21 in the Appendix.

The μ model contains terms for vehicle value, if the policy holder is married, and how many years the policy holder has lived in the same location. As was seen with the previous model, while keeping other predictor values unchanged, as the value of the vehicle increases, so does μ. Married policy holders can expect a reduced μ when compared to non married customers. Living in the same home for multiple years also reduced μ.

The model for σ contains terms for occupation as a factor with 9 levels, vehicle type as a factor with 6 levels, education as a factor with 5 levels, how many years the policy holder has lived in the same location, home value, gender, an indicator if the vehicle is red, if the owners license has been revoked in the past 5 years, and MVR points. MVR points are motor vehicle record points and are assigned to drivers for motor vehicle offences. The baseline category for vehicle type was panel truck. Negative coefficients for σ were seen for red vehicles, owners with revoked licenses, and increases in MVR points, or years at the same address. Positive coefficients for σ were seen for male drivers and increases in home value. Additionally all categories aside from the baseline for vehicle type saw increases in σ.

The inverse Gaussian model allows easy interpretation and connection between μ and $\mathrm{E}(Y)$. While keeping other predictor values unchanged, as vehicle value increases, so does average claim size. Additionally, policy holders who are married can expect a lower average claim size when compared to their single counterparts. Lastly, as the number of years living at the same address increases, the average claim size decreases.

5 Results

The following subsections will be used to discuss common predictors, trends and distributions across the previously introduced models.

5.1 Common distributions

As was discussed in Section 2, 10 different 2 parameter, continuous, right skewed distributions were used in the modelling process. Of these candidate distributions the most commonly seen distribution among top models was lognormal. It was the distribution assumed in the top models for *AutoClaims*, *dataOhlsson* and *Claims* data sets. The next most frequent distribution was gamma with identity link functions for μ and σ, which was the distribution assumed in the top model for *AutoCollision* and *swautoins*. The inverse Gaussian distribution appeared once as the distribution assumed in the top model for *dataCar* and twice as the distribution in secondary models for *AutoCollision* and *Claims*. The Pareto distribution appeared in the top model for *norauto* and the secondary model for *AutoClaims*. Weibull as given by `WEI` was seen for *swautoins* secondary model and *dataOhlsson* secondary model. Inverse gamma appeared for *dataCar* secondary model only. The gamma distribution with log link functions was seen as in the secondary model for *norauto*. A summary of these results can be seen in Table 3.

5.2 Common predictors and trends

Predictor variables that were seen in multiple models were location, age of policy holder, gender of policy holder, vehicle value, and vehicle age.

5.2.1 Common predictors for μ

Table 4 can be used to see a summary of commonly seen predictors for μ. Predictor variables are ordered by frequency in data sets. Columns 3 and 4 show the percentage of top or secondary models the predictor was seen in.

Table 4: Commonly seen predictors for μ

Predictor	Data sets in	Top models in	Secondary models in
Policy holder age	6	4 (66.$\overline{6}$%)	2 (33.$\overline{3}$%)
Location	4	3 (75%)	4 (100%)
Vehicle age	3	1 (33.$\overline{3}$%)	2 (66.$\overline{6}$%)
Vehicle value	2	1 (50%)	2 (100%)

Policy holder age was seen as a predictor in top models for *AutoClaims*, *AutoCollision*, *dataCar*, and *dataOhlsson*. Secondary models for *AutoCollision* and *dataOhlsson* also included the predictor policy holder age.

Location was seen as a predictor in the top models for *AutoClaims*, *dataCar*, and *dataOhlsson*. Additionally location was also a predictor for μ in the second best models for *AutoClaims*, *dataCar*, *dataOhlsson*, and *swautoins*.

Vehicle age was a predictor in the top model for *dataOhlsson* and the second best models for *dataCar* and *dataOhlsson*.

Vehicle value was a seen as a predictor in both models for *Claims* and the secondary model for *dataCar*.

5.2.2 Common trends for μ

Table 5 summarizes the effects of numeric predictor variables for μ in top models. Positive and negative relationships are shown as counts and percentages as well as models where the coefficient of the predictor was not statistically significant. The percentages are calculated as the number of top models with the given relationship divided by the number of data sets the predictor was present in. In the case of policy holder age, the total does not sum to 100% because the variable is a factor in *AutoCollision* and *dataCar*. Similarly Table 6 summarizes effects for factor variables for μ in top models. With multi level factor variables the direction of relationship is not relevant however the presence of an effect is of interest.

Table 5: Effects of numeric predictors for μ in top models.

Predictor	Positive effect	Negative effect	No effect
Policy holder age[1]	0 (0%)	2 ($33.\overline{3}$%)	2 ($33.\overline{3}$%)
Vehicle age	0 (0%)	1 ($33.\overline{3}$%)	2 ($66.\overline{6}$%)
Vehicle value	1 (50%)	0 (0%)	1 (50%)
Exposure	0 (0%)	1 (50%)	1 (50%)
Policy holder years	1 (100%)	0 (0%)	0 (0%)

[1] Variable was a factor for *AutoCollision*.

Table 6: Effects of factor predictors for μ in top models.

Predictor	Effect	No effect
Location	3 (75%)	1 (25%)
Gender	1 (25%)	3 (75%)
Bonus malus	1 ($33.\overline{3}$%)	2 ($66.\overline{6}$%)
Vehicle use	1 (50%)	1 (50%)
Brand	1 (100%)	0 (0%)

While it is important to note that the magnitude of model coefficients are not comparable across data sets, general trends can still be assessed. Magnitude is of no use in comparison as data sets vary in country of origin, currency, and time period.

Policy holder age was seen to have a negative relationship with μ for *AutoClaims* and both *dataOhlsson* models. Unfortunately the age variable in *AutoCollision* and *dataCar* are factor variables and therefore the trend is not comparable. Prior to analysis it was hypothesized that policy holder age would increase average claim size at either extreme, for this a squared age term was allowed into the scope of all models where age was a numeric variable in the data set. However, this term was dropped by the step function during the modelling process in all cases.

Vehicle age was not seen to have a consistent trend. For both *dataOhlsson* models the relationship with μ was negative, while for *dataCar* secondary model this relationship was positive.

Other numeric variables only appeared in one model for μ and therefore cannot be used to establish trends.

Unfortunately, trends across factor variable location cannot be analyzed. As these factors are not described in the data set sources, it is not known what regions the locations are referring to. If regions were known, it would be possible to research these regions in order to compare characteristics such as population density across the different data sets to see if trends exist.

5.2.3 Common predictors for σ

Table 7 can be used to see a summary of commonly seen predictors for σ. Predictor variables are ordered by frequency in data sets. Columns 3 and 4 show the percentage of top or secondary models the predictor was seen in.

Table 7: Commonly seen predictors for σ

Predictor	Data sets in	Top models in	Secondary models in
Policy holder age	6	2 ($33.\overline{3}$%)	2 ($33.\overline{3}$%)
Location	5	3 (60%)	1 (20%)
Gender	4	1 (25%)	2 (50%)
Vehicle value	2	1 (50%)	1 (50%)
Bonus malus	2	1 (50%)	1 (50%)

Policy holder age was a commonly seen predictor for σ. It was present in the top models for *AutoClaims* and *AutoCollision*. This term was also in the second best models for *dataOhlsson* and *dataCar*.

Location was a predictor in both models for *swautoins*, and top models for *dataCar* and *dataOhlsson*.

Gender was present in top models for *AutoClaims*. Additionally gender was a predictor in the second best models for *Claims* and *dataCar*.

Vehicle value was seen as a predictor in the top model for *Claims* and the secondary model for *dataCar*.

Bonus malus class was a predictor in both models for *swautoins*.

5.2.4 Common trends for σ

Table 8 summarizes the effects of numeric predictor variables for σ in top models. As was seen earlier, for policy holder age, the total does not sum to 100% because the variable is a factor in *AutoCollision* and *dataCar*. Similarly Table 9 summarizes effects for factor variables for σ in top models.

Table 8: Effects of common numeric predictors for σ top models.

Predictor	Positive effect	Negative effect	No effect
Policy holder age[1]	1 ($16.\overline{6}$%)	0 (0%)	4 ($66.\overline{6}$%)
Exposure	0 (0%)	1 (50%)	1 (50%)
Vehicle value	1 (50%)	0 (0%)	1 (50%)
Policy holder years	1 (100%)	0 (0%)	0 (0%)
Years at same address	0 (0%)	1 (100%)	0 (0%)

[1] Variable was a factor for *AutoCollision*.

Table 9: Effects of factor predictors for σ in top models.

Predictor	Effect	No effect
Location	3 (75%)	1 (25%)
Gender	1 (25%)	3 (75%)
Bonus malus	1 (50%)	1 (50%)
Vehicle type	1 (50%)	1 (50%)
Distance driven	1 (50%)	1 (50%)
Education level	1 (100%)	0 (0%)
Brand	1 (100%)	0 (0%)
Occupation	1 (100%)	0 (0%)
MC class	1 (100%)	0 (0%)

As was mentioned before, only trends will be assessed and compared across data sets. Magnitude of coefficients is not of interest. In general there were not many similar trends for σ when comparing across data sets.

An increase in policy holder age caused an increase in σ for *AutoClaims*. However for *dataOhlsson* the opposite was true, older policy holders saw a decrease in σ. Unfortunately the age variables in *AutoCollision* and *dataCar* are factor variables with no description and therefore the trend is not comparable.

Gender did not have a consistent effect across data sets either. In *AutoClaims*, males saw a decrease in σ. In *Claims* and *dataCar*, males saw an increase in σ.

Due to the unknown nature of the location variable, trends cannot be analyzed in a useful manner. Additionally, variables that only were present in one data set could not be used to establish trends.

Conclusion

The goal of this master's thesis was to establish a baseline for what data potential new vehicle insurance providers should collect in order to establish models for claim size. It was hypothesized that perhaps similarities in best assumed response variable distribution, or common trends for predictor variables could be established across various data sets. An additional goal was to gather publicly available insurance data sets suitable for modelling claim size and provide an overview of these data sets.

The first chapter was used to refresh the reader's knowledge of generalized linear models. In the second chapter the framework for the modelling process, generalized additive models for location scale and shape, was introduced. Furthermore, the process of model fitting applied in the analysis was discussed. Chapter three revolves around the data gathering process and introduces the data sets that are to be used in the analysis. Chapter four focuses on the results of each data sets top models and their statistically significant predictors. The fifth chapter is used to discuss similarities in distributions, significant predictors, or trends in predictor coefficients between top models of various data sets.

While the most commonly assumed distribution of top models was lognormal, in 3 of 7 top models, it would be in the best interest of potential new vehicle insurance providers to choose candidate distributions based on their particular data. Depending on the heaviness of tails, another distribution may fit better. Predictors that were present in multiple final models included: policy holder age, location, vehicle age, vehicle value, and gender. These variables can be collected from customers seeking policies rather easily. While this should not be taken as the complete set of necessary predictors to build models for average claim size it provides a starting point. Lastly the trends in predictor coefficients can be also be taken into account.

The biggest limitation to this thesis was access to relevant data. None of the data sets are particularly current which could prove to be an issue when modelling insurance claim size in today's market. Additionally, with the data sets ranging across many different countries, currencies and time periods it is difficult to say how relevant these conclusions will prove to be. It would be valuable to future research if possible to find more current data sets including more shared predictor variables. Particularly with technological advances in recent years it is important for new insurance companies to be making decisions based on the most up to date data available. With current data, researchers would be able to search for a model to accurately predict claim size rather than only focusing on common distributions, predictors, and trends.

Initially the the author intended to use the `GLM` package for modelling, however issues with convergence and difficulty of modelling some distributions arose. The author chose to pivot to using GAMLSS as it provided more flexibility in candidate distributions. While the smoothing term capabilities of GAMLSS were not explored in this work, in the future it would be interesting to see if applying smoothing terms to the models would help to improve them.

References (with BibLaTeX)

Akaike, Hirotogu (1998). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Statistical Science*. Accessed: 17-January-2023, pp. 199–213. URL: https://gwern.net/doc/statistics/decision/1998-akaike.pdf.

De Jong, P. and G.Z. Heller (2008). *Generalized linear models for insurance data*. UK: Cambridge. ISBN: 978-0-521-87914-9.

Dutang, Christophe and Arthur Charpentier (2020). *CASdatasets: Insurance datasets*. R package version 1.0-11.

Hardin, James W. and Joseph M. Hilbe (2018). *Generalized linear models and extensions*. 4th ed. Texas, USA: StataPress. ISBN: 978-1-59718-225-6.

Hastie, Trevor and Robert Tibshirani (1986). "Generalized Additive Models". In: *Statistical Science* 1.3, pp. 297–310.

Lindsey, James K. (1997). *Applying generalized linear models*. New York, USA: Springer-Verlag. ISBN: 0-287-98218-3.

OpenAI (2021). *GPT-3.5*. https://openai.com/research/.

ParetoDistribution (2016). Accessed: 17-January-2023. Wolfram Research. URL: https://reference.wolfram.com/language/ref/ParetoDistribution.html.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramires, Thiago G. (2021). "Validation of Stepwise-Based Procedure in GAMLSS". In: *Journal of Data Science* 19.1, pp. 96–110. DOI: 10.6339/21-JDS1003.

Rigby, R. A. and D. M. Stasinopoulos (2005). "Generalized additive models for location, scale and shape,(with discussion)". In: *Applied Statistics* 54, pp. 507–554.

Rigby, Robert A., Mikis Stasinopoulos, Gillian Heller, and Fernanda De Bastiani (2017). *Distributions for modelling location, scale and shape: using GAMLSS in R*.

Stasinopoulos, Mikis D., Robert A. Rigby, and Fernanda De Bastiani (2018). "GAMLSS: A distributional regression approach". In: *Statistical Modelling* 18, pp. 248–273.

WeibullDistribution (2016). Accessed: 17-January-2023. Wolfram Research. URL: https://reference.wolfram.com/language/ref/WeibullDistribution.html.

Wickham, Hadley, Jim Hester, and Jennifer Bryan (2022). *readr: Read Rectangular Text Data*. R package version 2.1.2. URL: https://CRAN.R-project.org/package=readr.

Wolny-Dominiak, Alicja and Michal Trzesiok (2014). *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance.* R package version 1.0. URL: https://CRAN.R-project.org/package=insuranceData.

Zeileis, Achim and Torsten Hothorn (2002). "Diagnostic Checking in Regression Relationships". In: *R News* 2.3, pp. 7–10. URL: https://CRAN.R-project.org/doc/Rnews/.

A Appendix

Listing 1: Modelling process

```
## #first initialize empty model
## > M_AutoClaims_LN<-gamlss(PAID~1,family= LOGNO(),
## data=na.omit(AutoClaims),method = RS(2000))
## GAMLSS-RS iteration 1: Global Deviance = 114370.2
## GAMLSS-RS iteration 2: Global Deviance = 114370.2
## > #build model through step function
## > M_AutoClaims_LN<-stepGAICAll.A(M_AutoClaims_LN, scope = list(lower=~1
## +                   , upper= ~ STATE + CLASS + GENDER +AGE +AGE^2 ),
## +                   sigma.scope =list(lower=~1 , upper= ~ STATE + CLASS
## +                                     + GENDER +AGE +AGE^2 ))
## ---------------------------------------------------
## Distribution parameter:  mu
## Start:  AIC= 114374.2
##  PAID ~ 1
##
##          Df    AIC
## + STATE  12 114343
## + AGE     1 114367
## + CLASS  17 114374
## <none>      114374
## + GENDER  1 114374
##
## Step:  AIC= 114343.1
##  PAID ~ STATE
##
##          Df    AIC
## + AGE     1 114340
## + GENDER  1 114343
## <none>      114343
## + CLASS  17 114344
##
## Step:  AIC= 114340
##  PAID ~ STATE + AGE
##
##          Df    AIC
## + GENDER  1 114340
## <none>      114340
## + CLASS  17 114343
##
## Step:  AIC= 114339.6
##  PAID ~ STATE + AGE + GENDER
##
##         Df    AIC
## <none>     114340
## + CLASS 17 114343
## ---------------------------------------------------
```

Listing 2: Modelling process continued

```
## ## Distribution parameter:  sigma
## Start:  AIC= 114339.6
##  ~1
##
##          Df    AIC
## + AGE     1 114325
## + GENDER  1 114337
## <none>      114340
## + CLASS  17 114353
## + STATE  12 114353
##
## Step:  AIC= 114325
##  ~AGE
##
##          Df    AIC
## + GENDER  1 114323
## <none>      114325
## - AGE     1 114340
## + STATE  12 114340
## + CLASS  17 114345
##
## Step:  AIC= 114322.7
##  ~AGE + GENDER
##
##          Df    AIC
## <none>      114323
## - GENDER  1 114325
## - AGE     1 114337
## + STATE  12 114338
## + CLASS  17 114343
## ---------------------------------------------------
## Distribution parameter:  mu
## Start:  AIC= 114322.7
##  PAID ~ STATE + AGE + GENDER
##
##          Df    AIC
## <none>      114323
## - GENDER  1 114323
## - AGE     1 114327
## - STATE  12 114351
## ---------------------------------------------------
```

Listing 3: Modelling process continued

```
## > summary(M_AutoClaims_LN)
## ******************************************************************
## Family:  c("LOGNO", "Log Normal")
##
## Call:  gamlss(formula = PAID ~ STATE + AGE + GENDER, sigma.formula = ~AGE +
##     GENDER, family = LOGNO(), data = na.omit(AutoClaims),
##     method = RS(2000), trace = FALSE)
##
## Fitting method: RS(2000)
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.998436   0.117565  59.528  < 2e-16 ***
## STATESTATE 02  0.112336   0.088648   1.267 0.205125
## STATESTATE 03  0.010454   0.100501   0.104 0.917157
## STATESTATE 04  0.050649   0.092406   0.548 0.583631
## STATESTATE 06  0.289416   0.093072   3.110 0.001881 **
## STATESTATE 07  0.097088   0.105168   0.923 0.355953
## STATESTATE 10  0.211190   0.104501   2.021 0.043325 *
## STATESTATE 11  0.174264   0.364210   0.478 0.632330
## STATESTATE 12  0.394973   0.106711   3.701 0.000216 ***
## STATESTATE 13  0.207940   0.110335   1.885 0.059522 .
## STATESTATE 14  0.073876   0.116495   0.634 0.525999
## STATESTATE 15  0.079967   0.085856   0.931 0.351674
## STATESTATE 17  0.225107   0.095765   2.351 0.018771 *
## AGE           -0.003086   0.001252  -2.464 0.013747 *
## GENDERM        0.041812   0.026899   1.554 0.120137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1228610  0.0529648  -2.320   0.0204 *
## AGE          0.0032585  0.0007981   4.083  4.5e-05 ***
## GENDERM     -0.0366512  0.0177214  -2.068   0.0387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  6773
## Degrees of Freedom for the fit:  18
##       Residual Deg. of Freedom:  6755
##                       at cycle:  3
##
## Global Deviance:     114286.7
##             AIC:     114322.7
##             SBC:     114445.5
## ******************************************************************
## > #coefficient for gender not significant, remove gender from mu model manually
```

Listing 4: Modelling process continued

```
M_AutoClaims_LN2<-gamlss(PAID~STATE +AGE,sigma.formula = ~GENDER +AGE ,
                     family= LOGNO(),data=na.omit(AutoClaims),method = RS(2000))
## GAMLSS-RS iteration 1: Global Deviance = 114289.4
## GAMLSS-RS iteration 2: Global Deviance = 114289.2
## GAMLSS-RS iteration 3: Global Deviance = 114289.2

summary(M_AutoClaims_LN2)

## *******************************************************************
## Family:  c("LOGNO", "Log Normal")
##
## Call:  gamlss(formula = PAID ~ STATE + AGE, sigma.formula = ~GENDER +
##     AGE, family = LOGNO(), data = na.omit(AutoClaims),      method = RS(2000))
##
##
## Fitting method: RS(2000)
##
## -------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.029984   0.115823  60.696  < 2e-16 ***
## STATESTATE 02  0.108909   0.088634   1.229 0.219207
## STATESTATE 03  0.007062   0.100493   0.070 0.943981
## STATESTATE 04  0.048655   0.092412   0.526 0.598558
## STATESTATE 06  0.286056   0.093062   3.074 0.002122 **
## STATESTATE 07  0.092136   0.105137   0.876 0.380877
## STATESTATE 10  0.206291   0.104471   1.975 0.048352 *
## STATESTATE 11  0.163963   0.364235   0.450 0.652611
## STATESTATE 12  0.391812   0.106709   3.672 0.000243 ***
## STATESTATE 13  0.205345   0.110342   1.861 0.062789 .
## STATESTATE 14  0.075120   0.116508   0.645 0.519100
## STATESTATE 15  0.077109   0.085849   0.898 0.369113
## STATESTATE 17  0.219721   0.095718   2.296 0.021734 *
## AGE           -0.003115   0.001252  -2.487 0.012894 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1219554  0.0529457  -2.303   0.0213 *
## GENDERM     -0.0368841  0.0177271  -2.081   0.0375 *
## AGE          0.0032493  0.0007978   4.073 4.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
## No. of observations in the fit:  6773
## Degrees of Freedom for the fit:  17
##       Residual Deg. of Freedom:  6756
##                       at cycle:  3
##
## Global Deviance:     114289.2
##             AIC:     114323.2
##             SBC:     114439.1
## *******************************************************************

#model is final
```

Listing 5: ChatGPT prompt and answer

```
Prompt: can you provide me with code to make an html file with 2 tables, one will have 4
 columns, the first column with a link to download a data set, the second column with a
 link to another website. The second table will have 10 columns
Answer:
<!DOCTYPE html>
<html>
<head>
	<title>My Tables</title>
</head>
<body>
	<h2>Table 1</h2>
	<table>
		<thead>
			<tr>
				<th>Data Set</th>
				<th>Website</th>
			</tr>
		</thead>
		<tbody>
			<tr>
				<td><a href="https://example.com/data-set.zip">Download</a></td>
				<td><a href="https://example.com">Visit</a></td>
			</tr>
		</tbody>
	</table>
	<h2>Table 2</h2>
	<table>
		<thead>
			<tr>
				<th>Column 1</th>
				<th>Column 2</th>
				<th>Column 3</th>
				<th>Column 4</th>
				<th>Column 5</th>
				<th>Column 6</th>
				<th>Column 7</th>
				<th>Column 8</th>
				<th>Column 9</th>
				<th>Column 10</th>
			</tr>
		</thead>
		<tbody>
			<tr>
				<td>Value 1</td>
				<td>Value 2</td>
				<td>Value 3</td>
				<td>Value 4</td>
				<td>Value 5</td>
				<td>Value 6</td>
				<td>Value 7</td>
				<td>Value 8</td>
				<td>Value 9</td>
				<td>Value 10</td>
			</tr>
			<tr>
				<td>Value 1</td>
				<td>Value 2</td>
				<td>Value 3</td>
				<td>Value 4</td>
				<td>Value 5</td>
				<td>Value 6</td>
				<td>Value 7</td>
				<td>Value 8</td>
				<td>Value 9</td>
				<td>Value 10</td>
			</tr>
			<!-- Add more rows as needed -->
		</tbody>
	</table>
</body>
</html>
```

Listing 6: AutoClaims top model

```
## ******************************************************************
## Family:  c("LOGNO", "Log Normal")
##
## Call:  gamlss(formula = PAID ~ STATE + AGE, sigma.formula = ~GENDER +
##     AGE, family = LOGNO(), data = na.omit(AutoClaims),      method = RS(1000))
##
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.029984   0.115823  60.696  < 2e-16 ***
## STATESTATE 02  0.108909   0.088634   1.229 0.219207
## STATESTATE 03  0.007062   0.100493   0.070 0.943981
## STATESTATE 04  0.048655   0.092412   0.526 0.598558
## STATESTATE 06  0.286056   0.093062   3.074 0.002122 **
## STATESTATE 07  0.092136   0.105137   0.876 0.380877
## STATESTATE 10  0.206291   0.104471   1.975 0.048352 *
## STATESTATE 11  0.163963   0.364235   0.450 0.652611
## STATESTATE 12  0.391812   0.106709   3.672 0.000243 ***
## STATESTATE 13  0.205345   0.110342   1.861 0.062789 .
## STATESTATE 14  0.075120   0.116508   0.645 0.519100
## STATESTATE 15  0.077109   0.085849   0.898 0.369113
## STATESTATE 17  0.219721   0.095718   2.296 0.021734 *
## AGE           -0.003115   0.001252  -2.487 0.012894 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1219554  0.0529457  -2.303   0.0213 *
## GENDERM     -0.0368841  0.0177271  -2.081   0.0375 *
## AGE          0.0032493  0.0007978   4.073 4.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  6773
## Degrees of Freedom for the fit:  17
##       Residual Deg. of Freedom:  6756
##                       at cycle:  3
##
## Global Deviance:     114289.2
##             AIC:     114323.2
##             SBC:     114439.1
## ******************************************************************
```

Listing 7: AutoClaims second models

```
## ******************************************************************
## Family:  c("PARETO2o", "Pareto Type 2")
##
## Call:  gamlss(formula = PAID ~ STATE, family = PARETO2o(),
##     data = na.omit(AutoClaims), method = RS(500), trace = FALSE,
##     sigma.formula = ~1)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.703878   0.115616  75.283  < 2e-16 ***
## STATESTATE 02 0.101128   0.098074   1.031 0.302512
## STATESTATE 03 0.028689   0.112233   0.256 0.798251
## STATESTATE 04 0.040793   0.102445   0.398 0.690504
## STATESTATE 06 0.292547   0.102994   2.840 0.004519 **
## STATESTATE 07 0.106486   0.116906   0.911 0.362399
## STATESTATE 10 0.187214   0.115767   1.617 0.105891
## STATESTATE 11 0.157173   0.396610   0.396 0.691904
## STATESTATE 12 0.420719   0.118615   3.547 0.000392 ***
## STATESTATE 13 0.206892   0.122978   1.682 0.092546 .
## STATESTATE 14 0.003675   0.128758   0.029 0.977230
## STATESTATE 15 0.090899   0.094950   0.957 0.338434
## STATESTATE 17 0.213807   0.105912   2.019 0.043556 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.55621    0.06164   25.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  6773
## Degrees of Freedom for the fit:  14
##       Residual Deg. of Freedom:  6759
##                       at cycle:  114
##
## Global Deviance:     114955.5
##             AIC:     114983.5
##             SBC:     115079
## ******************************************************************
```

Listing 8: AutoClaims second models continued

```
## ******************************************************************
## Family:  c("PARETO2", "Pareto Type 2")
##
## Call:  gamlss(formula = PAID ~ STATE, family = PARETO2(),
##     data = na.omit(AutoClaims), method = RS(500), trace = FALSE,
##     sigma.formula = ~1)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.703642   0.115611  75.284  < 2e-16 ***
## STATESTATE 02 0.101145   0.098075   1.031 0.302439
## STATESTATE 03 0.028782   0.112235   0.256 0.797619
## STATESTATE 04 0.040823   0.102446   0.398 0.690287
## STATESTATE 06 0.292552   0.102995   2.840 0.004518 **
## STATESTATE 07 0.106527   0.116908   0.911 0.362221
## STATESTATE 10 0.187215   0.115768   1.617 0.105892
## STATESTATE 11 0.157126   0.396609   0.396 0.691989
## STATESTATE 12 0.420741   0.118617   3.547 0.000392 ***
## STATESTATE 13 0.206911   0.122979   1.682 0.092521 .
## STATESTATE 14 0.003674   0.128759   0.029 0.977239
## STATESTATE 15 0.090924   0.094951   0.958 0.338307
## STATESTATE 17 0.213819   0.105913   2.019 0.043547 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.55604    0.06163  -25.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  6773
## Degrees of Freedom for the fit:  14
##       Residual Deg. of Freedom:  6759
##                       at cycle:  113
##
## Global Deviance:     114955.5
##             AIC:     114983.5
##             SBC:     115079
## ******************************************************************
```

Listing 9: AutoCollision top model

```
## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:  gamlss(formula = Severity ~ Vehicle_Use + Age, sigma.formula = ~Age,
##     family = GA(mu.link = "identity", sigma.link = "identity"),
##     data = na.omit(AutoCollision), method = RS(500),      trace = FALSE)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            442.886     67.805   6.532 1.91e-05 ***
## Vehicle_UseDriveLong   -82.174      3.120 -26.339 1.15e-12 ***
## Vehicle_UseDriveShort -138.290      3.702 -37.358 1.29e-14 ***
## Vehicle_UsePleasure   -136.662      3.341 -40.906 4.01e-15 ***
## AgeB                   -56.197     70.090  -0.802    0.437
## AgeC                   -63.278     68.207  -0.928    0.370
## AgeD                   -74.376     67.812  -1.097    0.293
## AgeE                  -132.886     69.206  -1.920    0.077 .
## AgeF                  -103.790     68.022  -1.526    0.151
## AgeG                  -101.955     67.801  -1.504    0.157
## AgeH                  -107.813     67.841  -1.589    0.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  identity
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3848     0.1329   2.896   0.0125 *
## AgeB         -0.2585     0.1402  -1.843   0.0882 .
## AgeC         -0.3297     0.1345  -2.451   0.0291 *
## AgeD         -0.3699     0.1329  -2.783   0.0155 *
## AgeE         -0.2459     0.1418  -1.735   0.1065
## AgeF         -0.3361     0.1339  -2.510   0.0261 *
## AgeG         -0.3720     0.1329  -2.799   0.0150 *
## AgeH         -0.3609     0.1330  -2.713   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  32
## Degrees of Freedom for the fit:  19
##       Residual Deg. of Freedom:  13
##                       at cycle:  5
##
## Global Deviance:     261.9652
##             AIC:     299.9652
##             SBC:     327.8142
## ******************************************************************
```

Listing 10: AutoCollision second model

```
## ****************************************************************
## Family:  c("IG", "Inverse Gaussian")
##
## Call:
## gamlss(formula = Severity ~ Vehicle_Use + Age, family = IG(mu.link = "log"),
##     data = na.omit(AutoCollision), method = RS(500),
##     trace = FALSE, sigma.formula = ~1)
##
## Fitting method: RS(500)
##
## -------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.17742    0.07214  85.635  < 2e-16 ***
## Vehicle_UseDriveLong  -0.33338    0.06446  -5.172 4.63e-05 ***
## Vehicle_UseDriveShort -0.49018    0.06214  -7.888 1.45e-07 ***
## Vehicle_UsePleasure   -0.57431    0.06105  -9.407 8.75e-09 ***
## AgeB                  -0.14741    0.08912  -1.654 0.113730
## AgeC                  -0.16307    0.08886  -1.835 0.081381 .
## AgeD                  -0.20784    0.08773  -2.369 0.028004 *
## AgeE                  -0.47304    0.08352  -5.664 1.52e-05 ***
## AgeF                  -0.32976    0.08521  -3.870 0.000953 ***
## AgeG                  -0.32048    0.08554  -3.747 0.001271 **
## AgeH                  -0.34643    0.08506  -4.073 0.000594 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.954      0.125  -39.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
## No. of observations in the fit:  32
## Degrees of Freedom for the fit:  12
##       Residual Deg. of Freedom:  20
##                       at cycle:  2
##
## Global Deviance:     308.6616
##             AIC:     332.6616
##             SBC:     350.2504
## ****************************************************************
```

Listing 11: dataCar top model

```
## ******************************************************************
## Family:  c("IG", "Inverse Gaussian")
##
## Call:  gamlss(formula = claimcst0 ~ exposure + area + gender +
##     agecat, sigma.formula = ~area + exposure, family = IG(mu.link = "log"),
##     data = na.omit(dataCar0), method = RS(500), trace = FALSE)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.10767    0.11019  73.577  < 2e-16 ***
## exposure    -0.69839    0.09052  -7.715 1.47e-14 ***
## areaB       -0.06529    0.07104  -0.919  0.35810
## areaC        0.04417    0.06659   0.663  0.50717
## areaD       -0.09611    0.08149  -1.179  0.23826
## areaE        0.09754    0.09301   1.049  0.29440
## areaF        0.31745    0.11062   2.870  0.00413 **
## genderM      0.12644    0.04875   2.593  0.00953 **
## agecat2     -0.13971    0.09714  -1.438  0.15042
## agecat3     -0.20079    0.09433  -2.129  0.03334 *
## agecat4     -0.18957    0.09397  -2.017  0.04371 *
## agecat5     -0.29959    0.10104  -2.965  0.00304 **
## agecat6     -0.25990    0.11371  -2.286  0.02232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.212586   0.032764 -98.052  < 2e-16 ***
## areaB       -0.001123   0.031342  -0.036  0.97141
## areaC        0.005424   0.028593   0.190  0.84956
## areaD       -0.095466   0.038488  -2.480  0.01316 *
## areaE       -0.114614   0.042008  -2.728  0.00639 **
## areaF       -0.141529   0.047434  -2.984  0.00286 **
## exposure    -0.104349   0.039533  -2.640  0.00833 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  4624
## Degrees of Freedom for the fit:  20
##       Residual Deg. of Freedom:  4604
##                       at cycle:  3
##
## Global Deviance:     77042.6
##             AIC:     77082.6
##             SBC:     77211.38
## ******************************************************************
```

Listing 12: dataCar second model

```
## ******************************************************************
## Family:  c("IGAMMA", "Inverse Gamma")
##
## Call:  gamlss(formula = claimcst0 ~ area + veh_age + veh_value,
##     sigma.formula = ~exposure + veh_value + agecat +
##         gender, family = IGAMMA(mu.link = "log"), data = na.omit(dataCar0),
##     method = RS(500), trace = FALSE)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.364474   0.059940  89.497  < 2e-16 ***
## areaB       -0.021439   0.042420  -0.505 0.613293
## areaC       -0.005279   0.038644  -0.137 0.891340
## areaD        0.083720   0.051919   1.613 0.106920
## areaE        0.165044   0.056905   2.900 0.003745 **
## areaF        0.249249   0.064620   3.857 0.000116 ***
## veh_age2     0.075369   0.043116   1.748 0.080521 .
## veh_age3     0.138205   0.045472   3.039 0.002384 **
## veh_age4     0.199839   0.051386   3.889 0.000102 ***
## veh_value    0.063013   0.015863   3.972 7.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.17193    0.03259   5.276 1.38e-07 ***
## exposure    -0.15617    0.02967  -5.263 1.48e-07 ***
## veh_value   -0.02737    0.00766  -3.574 0.000356 ***
## agecat2     -0.07452    0.02938  -2.537 0.011227 *
## agecat3     -0.08812    0.02860  -3.081 0.002073 **
## agecat4     -0.09330    0.02861  -3.262 0.001116 **
## agecat5     -0.12220    0.03206  -3.812 0.000139 ***
## agecat6     -0.09638    0.03664  -2.630 0.008564 **
## genderM      0.03399    0.01599   2.126 0.033546 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  4624
## Degrees of Freedom for the fit:  19
##       Residual Deg. of Freedom:  4605
##                       at cycle:  9
##
## Global Deviance:     77080.5
##             AIC:     77118.5
##             SBC:     77240.84
## ******************************************************************
```

Listing 13: dataOhlsson top model

```
## ******************************************************************
## Family:  c("LOGNO", "Log Normal")
##
## Call:
## gamlss(formula = skadkost ~ fordald + agarald + zon, sigma.formula = ~zon +
##     mcklass, family = LOGNO(), data = dataOhlsson0, method = RS(500),
##     trace = FALSE)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.519227   0.186206  56.492  < 2e-16 ***
## fordald     -0.055959   0.008296  -6.745 3.35e-11 ***
## agarald     -0.019790   0.004431  -4.467 9.35e-06 ***
## zon2        -0.047432   0.163587  -0.290  0.77195
## zon3        -0.427746   0.167333  -2.556  0.01080 *
## zon4        -0.509012   0.158325  -3.215  0.00137 **
## zon5        -0.701595   0.512981  -1.368  0.17188
## zon6        -0.207637   0.226699  -0.916  0.36005
## zon7        -3.000209   0.132083 -22.715  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.33553    0.11505   2.916  0.00366 **
## zon2          0.16385    0.07824   2.094  0.03663 *
## zon3          0.06217    0.08596   0.723  0.46982
## zon4          0.17774    0.07534   2.359  0.01861 *
## zon5          0.13391    0.24428   0.548  0.58377
## zon6         -0.46363    0.18165  -2.552  0.01093 *
## zon7        -37.00292    0.71036 -52.090  < 2e-16 ***
## mcklass2     -0.27830    0.14137  -1.969  0.04942 *
## mcklass3      0.12373    0.12002   1.031  0.30294
## mcklass4     -0.01680    0.12881  -0.130  0.89625
## mcklass5     -0.03534    0.12150  -0.291  0.77125
## mcklass6     -0.08723    0.11948  -0.730  0.46563
## mcklass7     -0.26458    0.33631  -0.787  0.43173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  670
## Degrees of Freedom for the fit:  22
##       Residual Deg. of Freedom:  648
##                       at cycle:  3
##
## Global Deviance:     14561.74
##             AIC:     14605.74
##             SBC:     14704.9
## ******************************************************************
```

Listing 14: dataOhlsson second model

```
## ******************************************************************
## Family:  c("WEI", "Weibull")
##
## Call:  gamlss(formula = skadkost ~ fordald + zon + agarald,
##     sigma.formula = ~mcklass + agarald + fordald, family = WEI(),
##     data = dataOhlsson0, method = RS(500), trace = FALSE)
##
## Fitting method: RS(500)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.112791   0.178211  62.357  < 2e-16 ***
## fordald     -0.069008   0.009841  -7.013 5.87e-12 ***
## zon2         0.005653   0.139322   0.041 0.967650
## zon3        -0.376045   0.146377  -2.569 0.010420 *
## zon4        -0.331553   0.139452  -2.378 0.017716 *
## zon5        -0.564672   0.405583  -1.392 0.164322
## zon6        -0.417778   0.309765  -1.349 0.177904
## zon7        -3.783509   1.626967  -2.325 0.020351 *
## agarald     -0.015198   0.004519  -3.363 0.000816 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.063857   0.152582  -0.419  0.67571
## mcklass2     0.250669   0.144947   1.729  0.08421 .
## mcklass3    -0.039625   0.127741  -0.310  0.75651
## mcklass4     0.171626   0.140999   1.217  0.22396
## mcklass5     0.243487   0.130158   1.871  0.06183 .
## mcklass6     0.289156   0.126871   2.279  0.02298 *
## mcklass7     0.262232   0.323313   0.811  0.41762
## agarald     -0.007696   0.002488  -3.093  0.00207 **
## fordald     -0.009161   0.004341  -2.110  0.03522 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  670
## Degrees of Freedom for the fit:  18
##       Residual Deg. of Freedom:  652
##                       at cycle:  9
##
## Global Deviance:     14626.69
##             AIC:     14662.69
##             SBC:     14743.82
## ******************************************************************
```

Listing 15: norauto top models

```
## ******************************************************************
## Family:  c("PARETO2", "Pareto Type 2")
##
## Call:  gamlss(formula = ClaimAmount ~ 1, family = PARETO2(),
##     data = na.omit(norauto0), method = RS(1000), sigma.formula = ~1)
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4147     0.1123   110.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.4411     0.1039   -23.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  8444
## Degrees of Freedom for the fit:  2
##       Residual Deg. of Freedom:  8442
##                       at cycle:  302
##
## Global Deviance:     186791.9
##             AIC:     186795.9
##             SBC:     186810
## ******************************************************************

## ******************************************************************
## Family:  c("PARETO2o", "Pareto Type 2")
##
## Call:  gamlss(formula = ClaimAmount ~ 1, family = PARETO2o(),
##     data = na.omit(norauto0), method = RS(1000), sigma.formula = ~1)
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4149     0.1123   110.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4413     0.1039    23.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  8444
## Degrees of Freedom for the fit:  2
##       Residual Deg. of Freedom:  8442
##                       at cycle:  303
##
## Global Deviance:     186791.9
##             AIC:     186795.9
##             SBC:     186810
## ******************************************************************
```

Listing 16: norauto second model

```
## ******************************************************************
## Family:  c("GA", "Gamma")
##
## Call:  gamlss(formula = ClaimAmount ~ 1, sigma.formula = ~1,
##     family = GA(), data = na.omit(norauto0), method = RS(1000))
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.06851    0.01024   983.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.061112   0.006848  -8.924   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  8444
## Degrees of Freedom for the fit:  2
##       Residual Deg. of Freedom:  8442
##                       at cycle:  2
##
## Global Deviance:     186848
##             AIC:     186852
##             SBC:     186866
## ******************************************************************
```

Listing 17: swautoins top model

```
## ****************************************************************
## Family:  c("GA", "Gamma")
##
## Call:
## gamlss(formula = Payment ~ Insured + Bonus + Make, sigma.formula = ~Make +
##     Zone + Bonus + Insured + Kilometres, family = GA(mu.link = "identity",
##     sigma.link = "identity"), data = na.omit(swautoins0), method = RS(1000),
##     trace = FALSE)
##
## Fitting method: RS(1000)
##
## -------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38824.187   3974.227   9.769  < 2e-16 ***
## Insured        223.883      0.408 548.751  < 2e-16 ***
## Bonus2       -5049.774   2602.679  -1.940   0.0525 .
## Bonus3      -11374.628   2268.783  -5.014 5.88e-07 ***
## Bonus4      -10420.415   2315.730  -4.500 7.24e-06 ***
## Bonus5      -11502.837   2305.717  -4.989 6.67e-07 ***
## Bonus6      -11000.326   2495.817  -4.408 1.11e-05 ***
## Bonus7      -12649.706   2425.421  -5.215 2.05e-07 ***
## Make2       -19543.720   3738.842  -5.227 1.92e-07 ***
## Make3       -22414.233   3688.441  -6.077 1.50e-09 ***
## Make4       -28539.337   3516.645  -8.116 8.92e-16 ***
## Make5       -17792.650   3796.151  -4.687 2.98e-06 ***
## Make6       -24112.271   3737.434  -6.452 1.42e-10 ***
## Make7       -23014.258   3608.152  -6.378 2.27e-10 ***
## Make8       -19329.729   3736.339  -5.173 2.56e-07 ***
## Make9       116524.180   7764.905  15.007  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
```

Listing 18: swautoins second model

```
## ******************************************************************
## Family:  c("WEI", "Weibull")
##
## Call:  gamlss(formula = Payment ~ Make + Bonus + Zone + Kilometres,
##     sigma.formula = ~Make + Zone + Kilometres + Bonus,
##     family = WEI(), data = na.omit(swautoins0), method = RS(1000),
##     trace = FALSE)
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.06041    0.07363 177.389  < 2e-16 ***
## Make2       -1.41833    0.06499 -21.825  < 2e-16 ***
## Make3       -1.70246    0.07894 -21.568  < 2e-16 ***
## Make4       -2.11159    0.11198 -18.858  < 2e-16 ***
## Make5       -1.42201    0.05613 -25.334  < 2e-16 ***
## Make6       -0.96306    0.05312 -18.128  < 2e-16 ***
## Make7       -1.70920    0.06018 -28.402  < 2e-16 ***
## Make8       -2.05047    0.09489 -21.609  < 2e-16 ***
## Make9        1.86261    0.03318  56.132  < 2e-16 ***
## Bonus2      -0.49248    0.06109  -8.061 1.39e-15 ***
## Bonus3      -0.77502    0.06181 -12.540  < 2e-16 ***
## Bonus4      -0.96499    0.06472 -14.909  < 2e-16 ***
## Bonus5      -0.84603    0.06522 -12.972  < 2e-16 ***
## Bonus6      -0.22201    0.06452  -3.441 0.000594 ***
## Bonus7       1.16459    0.05955  19.556  < 2e-16 ***
## Zone2       -0.07705    0.04197  -1.836 0.066532 .
## Zone3       -0.10088    0.04130  -2.443 0.014669 *
## Zone4        0.47430    0.04255  11.148  < 2e-16 ***
## Zone5       -1.23681    0.05450 -22.693  < 2e-16 ***
## Zone6       -0.63436    0.04872 -13.022  < 2e-16 ***
## Zone7       -3.54197    0.10122 -34.994  < 2e-16 ***
## Kilometres2  0.32498    0.03683   8.823  < 2e-16 ***
## Kilometres3 -0.10803    0.04124  -2.620 0.008882 **
## Kilometres4 -1.02727    0.04931 -20.833  < 2e-16 ***
## Kilometres5 -1.22543    0.05270 -23.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
```

Listing 19: swautoins second model continued

```
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45856    0.09215   4.976 7.12e-07 ***
## Make2       -0.72235    0.07665  -9.424  < 2e-16 ***
## Make3       -0.90454    0.08049 -11.238  < 2e-16 ***
## Make4       -1.13032    0.08589 -13.161  < 2e-16 ***
## Make5       -0.55757    0.07529  -7.406 2.02e-13 ***
## Make6       -0.43834    0.07516  -5.832 6.52e-09 ***
## Make7       -0.68485    0.07677  -8.920  < 2e-16 ***
## Make8       -1.08693    0.08006 -13.577  < 2e-16 ***
## Make9        0.45635    0.08126   5.616 2.27e-08 ***
## Zone2        0.05743    0.06275   0.915 0.360267
## Zone3        0.13613    0.06341   2.147 0.031938 *
## Zone4        0.15072    0.06469   2.330 0.019928 *
## Zone5       -0.29748    0.06786  -4.384 1.24e-05 ***
## Zone6       -0.06782    0.06624  -1.024 0.306036
## Zone7       -0.94276    0.08582 -10.986  < 2e-16 ***
## Kilometres2  0.33124    0.05793   5.718 1.27e-08 ***
## Kilometres3  0.15982    0.06470   2.470 0.013597 *
## Kilometres4 -0.22172    0.06576  -3.372 0.000763 ***
## Kilometres5 -0.32425    0.06625  -4.895 1.08e-06 ***
## Bonus2       0.28350    0.07169   3.955 7.97e-05 ***
## Bonus3       0.33986    0.07256   4.684 3.04e-06 ***
## Bonus4       0.37481    0.07311   5.126 3.28e-07 ***
## Bonus5       0.38017    0.07309   5.201 2.21e-07 ***
## Bonus6       0.50198    0.07281   6.894 7.53e-12 ***
## Bonus7       0.83613    0.07499  11.150  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -------------------------------------------------------------------
## No. of observations in the fit:  1797
## Degrees of Freedom for the fit:  50
##       Residual Deg. of Freedom:  1747
##                       at cycle:  12
##
## Global Deviance:     42804.58
##             AIC:     42904.58
##             SBC:     43179.27
## *******************************************************************
```

Listing 20: Claims top model

```
## ******************************************************************
## Family:  c("LOGNO", "Log Normal")
##
## Call:  gamlss(formula = CLM_AMT ~ BLUEBOOK, sigma.formula = ~MAX_EDUC +
##     JOBCLASS + SAMEHOME + BLUEBOOK, family = LOGNO(), data = na.omit(claims0),
##     method = RS(1000))
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.101e+00  3.182e-02 254.619  < 2e-16 ***
## BLUEBOOK    1.169e-05  2.023e-06   5.777 8.67e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -3.045e-01  1.145e-01  -2.660 0.007873 **
## MAX_EDUCBachelors     1.783e-01  5.605e-02   3.181 0.001490 **
## MAX_EDUCHigh School  -3.842e-02  4.379e-02  -0.877 0.380410
## MAX_EDUCMasters       2.037e-02  1.016e-01   0.200 0.841197
## MAX_EDUCPhD           2.320e-01  1.165e-01   1.992 0.046468 *
## JOBCLASSBlue Collar   1.194e-01  1.048e-01   1.139 0.254748
## JOBCLASSClerical     -1.943e-02  1.083e-01  -0.179 0.857648
## JOBCLASSDoctor       -4.237e-01  1.528e-01  -2.772 0.005614 **
## JOBCLASSHome Maker   -1.522e-02  1.054e-01  -0.144 0.885135
## JOBCLASSLawyer        1.277e-01  8.664e-02   1.474 0.140514
## JOBCLASSManager      -1.232e-01  9.809e-02  -1.256 0.209157
## JOBCLASSProfessional -2.166e-03  1.041e-01  -0.021 0.983400
## JOBCLASSStudent       8.415e-04  1.116e-01   0.008 0.993987
## SAMEHOME             -1.481e-02  3.995e-03  -3.707 0.000215 ***
## BLUEBOOK              5.466e-06  1.983e-06   2.757 0.005883 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  2189
## Degrees of Freedom for the fit:  17
##       Residual Deg. of Freedom:  2172
##                       at cycle:  3
##
## Global Deviance:     41269.02
##             AIC:     41303.02
##             SBC:     41399.77
## ******************************************************************
```

Listing 21: Claims second model

```
## ******************************************************************
## Family:  c("IG", "Inverse Gaussian")
##
## Call:
## gamlss(formula = CLM_AMT ~ BLUEBOOK + SAMEHOME + MARRIED, sigma.formula = ~JOBCLASS +
##     CAR_TYPE + MAX_EDUC + HOME_VAL + GENDER + RED_CAR + SAMEHOME +
##     MVR_PTS + REVOLKED, family = IG(), data = na.omit(claims0),
##     method = RS(1000))
##
## Fitting method: RS(1000)
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.399e+00  4.944e-02 169.884   <2e-16 ***
## BLUEBOOK     2.109e-05  2.528e-06   8.345   <2e-16 ***
## SAMEHOME    -8.884e-03  3.512e-03  -2.530   0.0115 *
## MARRIEDYes  -8.859e-02  3.950e-02  -2.243   0.0250 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -4.535e+00  1.262e-01 -35.919  < 2e-16 ***
## JOBCLASSBlue Collar   1.164e-01  1.062e-01   1.096  0.27330
## JOBCLASSClerical     -3.987e-02  1.099e-01  -0.363  0.71681
## JOBCLASSDoctor       -3.866e-01  1.554e-01  -2.487  0.01295 *
## JOBCLASSHome Maker    7.077e-03  1.084e-01   0.065  0.94798
## JOBCLASSLawyer        2.083e-01  9.073e-02   2.296  0.02179 *
## JOBCLASSManager      -9.476e-02  9.979e-02  -0.950  0.34244
## JOBCLASSProfessional -1.005e-01  1.050e-01  -0.958  0.33825
## JOBCLASSStudent       3.737e-02  1.134e-01   0.329  0.74182
## CAR_TYPEPickup        2.559e-01  6.561e-02   3.899 9.93e-05 ***
## CAR_TYPESedan         1.591e-01  6.908e-02   2.303  0.02137 *
## CAR_TYPESports Car    2.251e-01  8.215e-02   2.740  0.00620 **
## CAR_TYPESUV           1.278e-01  7.641e-02   1.673  0.09455 .
## CAR_TYPEVan           3.486e-02  7.328e-02   0.476  0.63435
## MAX_EDUCBachelors     1.659e-01  5.605e-02   2.961  0.00310 **
## MAX_EDUCHigh School  -6.911e-03  4.388e-02  -0.158  0.87486
## MAX_EDUCMasters      -7.618e-02  1.020e-01  -0.747  0.45512
## MAX_EDUCPhD           1.059e-01  1.170e-01   0.905  0.36581
## HOME_VAL              4.146e-07  1.432e-07   2.895  0.00383 **
## GENDERM               2.204e-01  5.379e-02   4.097 4.34e-05 ***
## RED_CARyes           -1.326e-01  4.485e-02  -2.956  0.00315 **
## SAMEHOME             -1.126e-02  4.004e-03  -2.812  0.00497 **
## MVR_PTS              -1.209e-02  5.893e-03  -2.052  0.04029 *
## REVOLKEDYes          -7.824e-02  3.751e-02  -2.086  0.03714 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  2189
## Degrees of Freedom for the fit:  28
##       Residual Deg. of Freedom:  2161
##                       at cycle:  4
##
## Global Deviance:     41508.31
##             AIC:     41564.31
##             SBC:     41723.66
## ******************************************************************
```

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Nicholas Lupul,
(*author's name*)

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Claims Severity Modelling on the Basis of Publicly Available Vehicle Insurance Data,
(*title of thesis*)

supervised by Assoc. Prof. Meelis Käärik.
(*supervisor's name*)

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Nicholas Lupul
16/05/2023