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

Kaari Kuus

Claims frequency modelling with usage-based insurance data

Actuarial and Financial Engineering

Master's Thesis (30 ECTS)

Supervisors: PhD Meelis Käärik
Ervin Sääsk (Swedbank P&C Insurance AS)

Tartu 2023

CLAIMS FREQUENCY MODELLING WITH USAGE-BASED INSURANCE DATA

Master's thesis

Kaari Kuus

Abstract

Motor insurance providers can think about collecting information about drivers' behavior using telematics to get more accurate pricing. The goal of this thesis is to give an overview of usage-based insurance and the use of telematics variables in insurance modeling. The thesis is divided into three parts. The first part gives background on telematics and usage-based insurance. The second part introduces generalized linear models suitable for modeling claim frequency. In the third part, analysis was done on specific vehicles and claims, with results and possible future steps given.

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

Key Words: motor vehicle insurance, generalized linear models, telematics

KAHJUSAGEDUSE MODELLEERIMINE KASUTUSPÕHISTE KINDLUSTUSANDMETEGA

Magistritöö

Kaari Kuus

Lühikokkuvõte

Mootorsõiduki kindlustuse pakkujad võivad kaaluda telemaatika abil juhtide käitumise kohta teabe kogumist, et pakkuda täpsemat hinnastust. Lõputöö eesmärk on anda ülevaade kasutuspõhisest kindlustusest ja telemaatika andmete kasutamisest kindlustusmudelites. Lõputöö on jagatud kolme ossa. Esimene osa avab telemaatika ja kasutuspõhise kindlustuse tausta. Teises osas tutvustatakse üldistatud lineaarseid mudeleid, mis sobivad kindlustuskahjude sageduse modelleerimiseks. Kolmandas osas analüüsitakse konkreetseid sõidukeid ja kahjusid, esitatakse tulemused ja võimalikud tulevased sammud.

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

Märksõnad: mootorsõiduki kindlustus, üldistatud lineaarsed mudelid, telemaatika

Contents

Introduction

Usage-based motor insurance uses factors about vehicle use and behaviour to price insurance. It is a relatively new idea for insurance policy pricing and needs looking into for insurance companies wanting to get on the market. In this thesis usage-based solutions are researched and new variables are used in the modelling of insurance claim frequency.

The first chapter gives a setting on the topic with overview of collection of telematics data in a business setting, its opportunities, problems and use on the current market. Insight into the previous research done on the matter was also given.

The second chapter gives theoretical overview and structure of suitable generalized linear models to use for insurance claim frequency.

Third chapter is about the data used for analysis containing information about one company's fleet of vehicles. An exploration of the data set was done and described models in the second chapter are applied. Results of the models and discussion about the problems and future research was given.

The author would like to give great gratitude to supervisor Meelis Käärik for the guidance, suggestions and corrections for the thesis and being extremely responsive and forthcoming during the process. The author also thanks Ervin Sääsk for introductions and discussions about the topic and suggestions for the thesis. The author also notes peers from the master's programm Kelly Tilga and Nicholas Lupul for much needed support and constructive discussions throughout the whole master's journey.

1 Background

Motor insurance pricing refers to the process of determining the premiums that individuals or businesses must pay to obtain insurance coverage for their vehicles Traditionally, motor insurance pricing uses variables such as driver's age, vehicle age, driver experience and historical claim data. This can however discriminate certain segments like young people or people who don't use their car very often. In this thesis, in addition to such traditional factors, it is looked into how to use factors about drivers' behaviour and their use of the vehicle. Because of the technological advancements tools to measure such information are already built in a lot of new cars or are easy to install. This gives an opportunity to look into if and how such information can be put to use in insurance companies.

The goals of this thesis are the following:

- to give an overview of UBI solutions and their opportunities;
- to give an overview of telematics data and its use in insurance companies;
- to find out if there are appropriate ways to use telematics variables in insurance modelling;
- to find out if and how different factors of telematics data can affect claim size and frequency using generalized linear models.

Next, a brief overview of telematics data, usage-based insurance and the current market situation is given. Followed also with a concise overview of previous research done on the matter as a setting for this thesis.

1.1 Telematics

Telematics is a word usually used for services where telecommunication systems are applied to transmit information provided by sensors. Telematics are usually

used for vehicles and fleet management but for example also for smart buildings. With the continuing development of technology, telematics are used more and more. (Wahlström, Skog, and Händel, 2017)

For vehicles' telematics, data is information collected about vehicle use and driver behaviour using GPS devices, black-box devices or in-phone applications.

Since almost everyone owns a smartphone, the use of them to collect data can be cheaper and logistically preferable. But on the downside, these smartphone sensors are usually poorer in quality and may not have been designed to collect vehicle data. The added battery drain and the disadvantage of the smartphone not being attached to the vehicle are to be considered as well. (Wahlström, Skog, and Händel, 2017)

More precise information can be recorded with in-vehicle telematics devices. Newer cars usually have such a device installed (black-boxes) or have an OBD port to easily insert one. When the device is set it can start to record information and pass it to a connected application which could also be a phone app. These in-vehicle devices are also able to send maintenance reports to keep the car in good shape. (Habas, 2023)

Many measures can be collected and some variables are a lot easier to collect than others. For example, measuring driven distance and driving location is quite direct. Observing driver fatigue though is a lot harder since it needs sensors (a camera) to observe the driver's eyes and reaction speed. Raw telematics data is quite large and robust and won't give a lot of insight. To model using telematics data, factors are usually either:

1. taken as aggregated (e.g. driven km per year, speeding events per 100km);
2. given a scoring for each factor (e.g. a lot of speeding gives a score of 0, no speeding score 100);
3. taken a combined scoring using all factors as input.

Data collected about driving through telematic devices has already many different usages and future applications. For example, it is used in traffic planning and management in urban areas and estimation of fuel consumption for environmental impact. It is possible to track CO2 emitted from cars and in a world which needs to become environmentally friendly this kind of observation is becoming more important (England, 2021). This all indicates the importance of telematics use for governmental as well as commercial use in the private sector.

1.2 Usage-based insurance

Usage-based insurance (UBI) is a form of motor insurance solution which differs from traditional insurance as it takes into account the actual use of the vehicle.

There are two main approaches to usage-based motor insurance: pay-as-you-drive (PAYD) and pay-how-you-drive (PHYD). PAYD is calculated based on vehicle use with measures like driven distance, time, and location. PHYD takes also into account driving behaviour such as speeding, acceleration/breaking, steep cornering and driver fatigue.

UBI should give the insurer a better and more precise way to price their policies and also should motivate the driver to drive more carefully and improve road behaviour. Reimers and Shiller, 2019 showed that using UBI with a rewards system for safe driving can reduce the risks of a crash by almost 50% and reduce emissions.

UBI is achieved through the use of telematics devices that collect data and can then be used to assess a driver's risk profile to price the driver by their behaviour in traffic. It is usually marketed as a way for safe and/or non-often drivers to save money. This also gives the insurance company more accurate information about the usage of cars which traditional variables fail to give. For example, how much is the vehicle actually used (exposure) and the location it is being driven.

This system can also give the driver real-time information about their driving to

improve their behaviour and most drivers will use this information. A case study done in Tokyo used GPS, a camera and an accelerometer to track driving behaviour giving real-time alerts for risky situations. The study showed the potential of using telematics in insurance and found risk improvements at the customer level. It also detected better renewal rates in the telematics solution group which indicates great customer satisfaction with these services. (Flückiger and Carbone, 2021)

1.3 Market situation

Usually, UBI solutions are advertised as cost-friendlier for the user and more innovative and flexible than offers from traditional car insurance companies. This could make this product more desirable for consumers and is therefore a great market opportunity for an insurance company to look into.

The first UBI companies started to surface in the United States more than a decade ago with Progressive Insurance and General Motors AssuranceCompany (*Telematics/ Usage-Based Insurance* n.d.). A lot of companies are still in the experimental or research phase of implementing UBI solutions and it is still just a small amount compared to the traditional market. In 2016 there were 292 programmes of active trials in 39 countries with 15.4 million active policies. Most UBI policies were in the US, Italy and the UK and most of them followed a PHYD solution. It is also worth noting that a good percentage of these policies tend to target younger drivers. (*UBI Infographic 2016* 2016)

By far the strongest market is the US with 12 million policies as of 2022. The U.S. has also good regulations for this kind of data use and a large customer demand for innovative solutions. (Orsoni, 2022)

Europe has the second-largest UBI market. The main contributors are Italy and the United Kingdom. In Italy with 8 million policies in 2021, the success of the UBI solution is strongly based on the need to reduce theft for which black-boxes were initially installed in the cars. (Orsoni, 2022) First UBI programmes in Italy were

launched nearly 20 years ago. As of 2021, the European market is strong with 11.5 million active policies and 4.3 billion euros of premiums. It is also expected that the UBI market will continue to grow in the region reaching 47 million policies and 15 billion euros of premiums by 2030. (*European Connected Auto Insurance Study* 2022)

In the Baltic region, usage-based insurance is quite an untapped market opportunity. From traditional insurance companies operating in the Baltic market, ERGO insurance used to offer casco and traffic insurance a product named Vehicle's e-insurance. It was a PAYD solution including a monthly fee and the cost for days driving within the month. This product was discontinued and no new contracts are being issued as of 01.12.2022. (*Vehicle's e-insurance* n.d.)

In Estonia, with the emergence of companies like Bolt that offer ride-hailing services, the demand for more flexible motor insurance is a relevant topic since 2018 (Pokk, 2018). Cachet.me (Cachet Insurance Broker OÜ) which is operating in Estonia, Latvia and Poland advertises itself as the future of insurance and stands for a fair and swift way to get an insurance contract. They argue that the new world needs insurance that is personalized and based on data. They have motor insurance products for private as well as business customers which include insurance for regular and drive-hailing drivers as well as for fleets. (*Cachet: Kindlustusmaakler | Nutikad kindlustuslahendused* n.d.)

There are already several applications in smart-phones' app stores that can measure your vehicle use and behaviour and give the driver their safe driving scores and suggestions (e.g. SafeDrive, UNIGO plus). For example, Toyota owners can also already check their driving behaviour using Toyota's own MyT app which can be downloaded from the Google play store and Apple's app store. The application gives an overview of previous trips including measures for speed, breaking and driven kilometres. (*MyTdigiteenused* n.d.)

There are though also concerns about data privacy, as UBI needs the collection and

sharing of personal driving behaviour data of customers. Insurance companies also need to think about how to access telematics data - if to use an external provider or to develop their own systems to collect the data which can be an expensive and extensive process. Therefore there could first be an opportunity to use it for insuring fleets that are usually already using such devices to track telematics data for better fleet management anyway.

Since the market is there and growing, looking into UBI solutions is a great opportunity and might even be expected from the customers.

1.4 Previous research

PAYD insurance was first proposed already in 1968 by Vickrey, 1968. Many previous attempts to implement mileage-based insurance where the user reported the driven distances have not been successful since clients are not usually very precise when reporting it. Now, with the advancements in the use of telematics devices, usage-based insurance has been researched more.

Ayuso, Guillen, and Nielsen, 2019 used real-life data and implemented a frequency model with distance driven and driving behaviour variables. It was found that these factors significantly influenced the expected accidents and therefore also have an effect on price. They proposed that using telematics values with traditional variables works best in contrast to using two systems separately. They also looked into introducing telematics data for companies that are already using classical pricing. They argued that even though it is not statistically efficient, from a practical standpoint companies starting to use telematic data can introduce the data as a correction to the classical model instead of re-estimating all the variables.

Guillen et al., 2019 also looked into the use of PAYD insurance and confirmed that speeding and driving in urban areas increases the expected number of claims. They used a zero-inflated Poisson model corrected by distance and discovered a learning curve for distances driven. There is an effect on excess zeros in the number of

claims for cars with almost no distances travelled but also for drivers with very high mileage. They possibly have a lot of experience and therefore safer road behaviour. They suggest that the introduction of more telematics variables could give even better results.

Research has also shown that the use of machine learning models can give good results when using telematics data to model insurance claims. Pesantez-Narvaez, Guillen, and Alcañiz, 2019 used XGBoost to compare it to logistic regression, and Maillart, 2021 applied random forest which was utilized to improve a generalized linear model. Machine learning methods have a drawback though of losing interpretability and using a lot more time and effort in comparison to generalized linear models.

2 Generalized linear models

The following chapter is written based on (de Jong and Heller, 2008) if not referenced otherwise.

The generalized linear model (GLM) is a generalization of the classical linear model which relaxes some of its assumptions. It is widely used in modelling insurance data since the response variable doesn't have to follow a normal distribution as insurance frequency and claim size distribution have usually a long right tail. In GLM the variability of one dependent variable is explained by changes in explanatory/risk variables. In insurance risk variables such as age, vehicle body type and vehicle value are used to explain or predict average claim frequency and severity. The main questions we want to answer with GLM are:

- Which of the explanatory variables has a significant effect on the response?
- What is the conditional mean of the response variable based on explanatory variables?
- What is the prediction and prediction power of the response variable using given risk variables?

Next, a theoretical framework of GLM is given, where the following notation is used:

- Y - a response variable;
- n - number of observations;
- $X_1, \ldots, X_k$ - explanatory variables;
- $\boldsymbol{y} = (y_1, \ldots, y_n)'$ - observed values of the response variable;
- $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ - values vector of explanatory variables for i-th observation;

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ - vector of coefficients;
- $Y_i = Y|X_1 = x_{i1}, \ldots, X_k = x_{ik} = Y|\boldsymbol{x_i}$ - conditional Y given explanatory variables $\boldsymbol{x}_i$.

2.1 Structure of GLM

A random variable Y is a real number determined by chance that can take values from a specific set Ω which is called the sample space. There are discrete random variables for which Ω is a finite or countable set and continuous random variables where Ω is an interval on the real line. A discrete random variable also has a probability function which gives for each $Y \in \Omega$ a probability that the random variable takes value Y. For a continuous variable probabilities are specified using a probability density function. Here, for simplicity, we note $f(y)$ as both, density function for continuous random variable and probability function for discrete random variable. There are many distributions with known probability/density functions which can describe random variables. In modelling, we usually assume that the response variable is following a known distribution.

As said, GLM doesn't require the response variable to be from a normal distribution, only that it follows a distribution that belongs to the exponential family. The exponential family probability/density function can be expressed as:

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

or by taking a logarithm from both sides:

$$\ln f(y) = \ln c(y, \phi) + \frac{y\theta - a(\theta)}{\phi}, \tag{2}$$

where parameter θ is called the canonical parameter and parameter ϕ the dispersion parameter. Function $c(y, \phi)$ is a known function which is independent of parameter

θ and function $a(\theta)$ is a real-valued differentiable function that depends only on θ. All distributions that can have probability/density functions rewritten like this are members of the exponential family. Some well-known distributions that belong to the exponential family are Poisson, (negative) binomial, gamma function and of course, the normal distribution is also from there.

The expected value and variance of a random variable Y with distribution from the exponential family can be expressed as

$$EY = a'(\theta), \qquad DY = \phi a''(\theta), \tag{3}$$

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

When the response variable Y belongs to the exponential family the GLM gets the form:

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where $\mu_i = E(Y_i)$ is the conditional mean of the response variable for row i, $g(\mu_i)$ is the link function, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ is a vector of explanatory variables for row i and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is a vector of coefficients.

The link function is specified during the modelling process. It is called canonical when $g(\mu) = \theta$. Choosing the canonical link as a link function simplifies the estimation. Some of the commonly used links are:

- identity link $g(\mu) = \mu$, which is canonical link for the normal distribution;
- log-link $\ln \mu$, which is canonical link for Poisson distribution;
- logit-link $g(\mu) = \ln \frac{\mu}{1-\mu}$, which is canonical link for binomial distribution.

Constants that may appear in the canonical link are generally omitted.

2.1.1 Offsets

Modelling counts can also require an offset term for the exposure value. Usually offset is used as a correction for group size or for differing time intervals under observation, let's denote this number as d_i. Then instead of μ_i it is needed to find $\frac{\mu_i}{d_i}$. When, for example, using log-link as a link function then

$$g\left(\frac{\mu_i}{d_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \Rightarrow \ln\left(\frac{\mu_i}{d_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \Rightarrow \ln\mu_i = \ln d_i + \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Here variable t_i is called exposure and $\ln d_i$ is called the offset. Response variable Y_i has the expected value which directly depends on the offset:

$$\mu_i = d_i \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right).$$

2.2 Steps of modelling GLM

When the data has been gathered the modelling process follows:

1. A response variable Y and a suitable response distribution with probability/density function $f(y)$ are chosen considering the given data.
2. A link $g(\mu)$ is chosen. A canonical link can be used for this which is set for each distribution.
3. Explanatory variables $X_1, \ldots, X_k$ are chosen to model $g(\mu)$.
4. Using software, a model is fitted by estimating parameters $\boldsymbol{\beta}$ using a maximum likelihood method.
5. Estimates of parameters $\boldsymbol{\beta}$ are used to generate predictions or fitted values for response Y.

6. Model fit is examined by the divergence of predictions from actual values and other model diagnostics such as plotting the residuals, goodness-of-fit measures and overdispersion assessment.

Initial exploration of the data is important to find suggestions for usable models and distributions for the response and which variables could be used as explanatory. After the model is fitted and examined, necessary changes are made to find a better model. A lot of models and fits are usually tried to determine the best model. During the process, some explanatory variables are added, discarded, or transformed. Moreover, different distributions could be considered for dependent variables with different links.

2.2.1 Model selection

Each explanatory variable X_j added to the model improves the fit of the model, but adding unsuitable variables decreases the precision of parameter estimates. Even if the variable could be statistically significant, the trade-off between improvement of fit and loss of estimation precision may not be worthwhile. So the modelling process needs to give the best result with enough number of parameters to get a close fit and good enough parameter estimations while not overfitting.

The most known criteria which balances the fit of a model with penalty terms for added parameters is Akaike's information criterion (AIC) which is expressed as:

$$AIC = -2\ell + 2p$$

where ℓ is the log-likelihood of the model and $p = k+1$ is the number of parameters to be estimated.

Models can be compared based on AIC meaning that the lowest value of AIC indicates a better model. When using AIC, models must have been built using the same set of observations.

2.3 Count models

For modelling claim frequency methodology is based on (Guillen et al., 2019) and (de Jong and Heller, 2008).

2.3.1 The Poisson model

The Poisson distribution has the probability function:

$$f_P(y) := P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots \tag{4}$$

and a random variable that follows a Poisson distribution is denoted as $Y \sim P(\mu), \quad \mu > 0$. The expected value and variance for Y can be expressed as:

$$EY = \mu = DY.$$

So for a variable to follow Poisson distribution its mean and variance should roughly be the same. Poisson distribution works well for describing infrequently happening events.

Poisson distribution is a part of the exponential family as

$$\ln f_P(y) = \ln \frac{e^{-\mu}\mu^y}{y!} = \ln e^{-\mu} + \ln \mu^y - \ln y! = -\mu + y \ln \mu - \ln y! = \\ = -\ln y! + \frac{y \ln \mu - \mu}{1},$$

where if noted $\phi = 1$, $\theta = \ln \mu$, $a(\theta) = e^\theta$, it is in the needed form of equation (2).

When the response variable Y is a count, often the Poisson distribution is used. The Poisson regression model is expressed as:

$$Y_i \sim P(\mu_i), \quad g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where the mean μ_i is explained by explanatory variables $\boldsymbol{x}_i$ via a suitable link.

When using the canonical link for Poisson distribution we get conditional mean as:

$$\mu_i = E(Y|\boldsymbol{x}_i) = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}).$$

It is commonly assumed that the insurance claim amount follows Poisson distribution which is why usually GLM with Poisson distribution and log-link function is applied first (Tiwari, 2020).

Formula (3) for the Poisson model states that to fit it must hold that $EY_i = DY_i$ and the scale of the model is $\phi = 1$. When variance is larger than the mean, $EY_i < DY_i$, it is called overdispersion. The scale of the model can be estimated by the deviance or Pearson's χ^2-statistic as follows:

$$\hat{\phi} = \frac{D}{df}, \qquad \hat{\phi} = \frac{\chi^2}{df},$$

where deviance for the Poisson model is defined as:

$$D = 2\sum_i (y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i))$$

and Pearson's χ^2-statistic as

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

where $\hat{\mu}_i$ is the estimate of μ_i.

If the model is correct then the ratio between deviance or Pearson's statistic and degrees of freedom should roughly be 1.

2.3.2 Poisson model with exposure

When talking about risk exposure, an offset is introduced to the model to make observations comparable (see 2.1.1). So instead of count, usually claim count per unit of exposure is modelled. In this case, the exposure is going to be distance driven. Let's denote the exposure for object i by d_i. This offset is going to be included in the model as follows:

$$\mu_i = E(Y|\boldsymbol{x}_i, d_i) = d_i \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}). \tag{5}$$

According to the Poisson distribution probability function, the probability of claim count 0 can be calculated as:

$$P(Y_i = 0) = \frac{e^{-d_i\lambda_i}(d_i\lambda_i)^0}{0!} = \exp(-d_i\lambda_i),$$

which depends on distance.

2.3.3 Negative binomial model

Another possible choice for the distribution of count data is a negative binomial distribution which can be used when there is overdispersion. It is traditionally described as the number of failures in Bernoulli trials that occur until the specified number of successes. It has the probability function:

$$f_{NB}(y) := P(Y = y) = \frac{\Gamma(y + \frac{1}{\kappa})}{y!\Gamma(\frac{1}{\kappa})} \left(\frac{1}{1+\kappa\mu}\right)^{\frac{1}{\kappa}} \left(\frac{\kappa\mu}{1+\kappa\mu}\right)^{y}, \qquad y = 0, 1, 2, \ldots \tag{6}$$

A random variable which follows the negative binomial distribution is noted as $Y \sim NB(\mu, \kappa)$. The expected value and variance can be expressed as:

$$EY = \mu, \qquad DY = \mu(1 + \kappa\mu).$$

In exponential family notation (1) negative binomial probability function is in the form:

$$f_{NB}(y) = \frac{\Gamma(y+\frac{1}{\kappa})}{\Gamma(y+1)\Gamma(\frac{1}{\kappa})} \exp\left\{y \ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right) + \frac{1}{\kappa}\ln\left(\frac{1}{1+\kappa\mu}\right)\right\},$$

where $\phi = 1$, $\theta = \ln\left(\frac{\kappa\mu}{1+\kappa\mu}\right)$, $a(\theta) = -\frac{1}{\kappa}\ln\left(\frac{1}{1+\kappa\mu}\right)$.

The negative binomial model using log-link can be expressed as:

$$Y_i \sim NB(\mu_i, \kappa), \quad \ln\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Analogously as for the Poisson model, the offset term can be included in the model (see (5)). The negative binomial model using offset term and the log-link is described as:

$$Y_i \sim NB(\mu_i, \kappa), \quad \ln\mu_i = \ln d_i + \boldsymbol{x}_i'\boldsymbol{\beta}.$$

2.3.4 Zero-inflated models

This subchapter is based on (Zuur et al., 2009).

Count data can have an excessive amount of zeros. These zeros can be generated by a completely different process from the count values and those excess zeros can be modelled independently. This can occur when dealing with insurance claim frequency data. The excess zeros can happen because policyholders tend not to report small claims. An assumption is made that there are two groups of data – the so-called false zeros and the counting process which may also produce zeros.

Zero-inflated models have two parts. For the zero-inflated Poisson (ZIP) model, a Poisson model is applied for the count part and logit-model for excessive zeros.

Let's denote that the probability of response Y_i being a false zero is π_i and then the probability of it being from the counting process is $1 - \pi_i$.

Then the probability of having a zero is

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) f_P(0),$$

and the probability for y_i amount of claims is

$$P(Y_i = y_i) = (1 - \pi_i) f_P(y_i), \quad y_i = 1, 2, \ldots,$$

where the function $f_P(y_i)$ is the probability function of Poisson distribution (4).

So we have the following probability function for the ZIP model:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1 - \pi_i) e^{-\mu_i}, & y_i = 0, \\ (1 - \pi_i) \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, & y_i = 1, 2, \ldots. \end{cases} \tag{7}$$

Just like in Poisson GLM, the count model mean μ_i is modelled as

$$\mu_i = \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) = \exp\left(\beta_0 + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k\right).$$

To estimate the probabilities of false zeros a logit model is generally used:

$$\ln \frac{\pi_i}{1 - \pi_i} = \boldsymbol{x}_i^{*\prime} \boldsymbol{\beta}^*, \qquad \pi_i = \frac{\exp\left(\boldsymbol{x}_i^{*\prime} \boldsymbol{\beta}^*\right)}{1 + \exp\left(\boldsymbol{x}_i^{*\prime} \boldsymbol{\beta}^*\right)},$$

where vector $\boldsymbol{x}_i^* = (1, x_{i1}^*, \ldots, x_{ik^*}^*)'$ shows the k^* covariates and vector $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{k^*}^*)$ their coefficients for the (false/excess) zero part of the model.

The mean and variance for the ZIP model are

$$E(Y | \boldsymbol{x}_i, \boldsymbol{x}_i^*) = \mu_i (1 - \pi_i),$$
$$D(Y | \boldsymbol{x}_i, \boldsymbol{x}_i^*) = \mu_i (1 - \pi_i)(1 + \mu_i \pi_i).$$

When zero-model is logit-model, the mean of the ZIP-model becomes:

$$E(Y|\boldsymbol{x}_i, \boldsymbol{x}_i^*) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{*\prime}\boldsymbol{\beta^*})}.$$

For the zero-inflated negative binomial (ZINB) model, the Poisson distribution for the count is replaced with the negative binomial distribution. This handles the overdispersion from the non-zero counts. So the probability function for the ZINB model is a modification of the ZIP one (7):

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}, & y_i = 0, \\ (1-\pi_i) f_{NB}(y_i), & y_i = 1, 2, \ldots, \end{cases} \tag{8}$$

where function $f_{NB}(y_i)$ is the probability function for negative binomial distribution (6).

Including the offset term to zero-inflated models is analogous to adding it to Poisson and negative binomial model as described before.

3 Modelling claim frequency

In this chapter overview of the specifics of how to use telematics data and an introduction to the data sets used and modelling process are given.

3.1 Data preprocessing

Before applying any analysis data needed to be preprocessed to bring data to a needed form. All data modifications, analysis and modelling demonstrated in the next subchapters are done with `base` (R Core Team, 2020) where necessary mutations were done with `dplyr` (Wickham et al., 2023) and `reshape2` (Wickham, 2007) and plots with `ggplot2` and (Wickham, 2016) and `corrplot` (Wei and Simko, 2021).

To build an insurance model using telemetric data two data sets were used: historical insurance data about the driver's claim info and telemetric data collected with devices about the driver's behaviour and use of the car. Also, a variable for vehicle identification is needed to combine the two datasets.

3.1.1 Telematics data

Telematic data used in this thesis comes from a company that provides black-boxes to corporate clients who need an overview of their fleets. It has to be mentioned that this data was collected with drivers not aware that their driving behaviour would later be used for analysis. No driver's personal information was taken into analysis.

Telematic data is aggregated data for each trip with a vehicle. That means for each vehicle there is a certain number of rows with the trip date and aggregated driver behaviour factor values for each trip made.

Telematics data was recorded for a fleet having 65 . The Data was from time frame 02.02.2020 to 26.03.2023. Variables that were given are:

- vehicle name;
- date of the trip;
- kilometres driven;
- number of harsh acceleration events;
- number of harsh breaking events;
- number of steep cornering events;
- number of speeding events.

Data-set contained a lot of missing values which were changed into 0-s because missing values indicate that an event did not occur. This means that trips with missing values show a very smooth and proper driving with no concerning events.

String manipulation was also done to extract a vehicle registration number from the object name to get a variable for vehicle identification and to join the data sets with claim data.

Telematics data was recorded on a user basis, meaning that if one user used multiple cars it was all in one data row. This created an issue where data rows with multiple used vehicles appeared in the data set with summarized factors, but no way to distinguish between vehicles. As said, for analysis data was needed on a vehicle basis. Therefore these data rows were also removed.

To analyse vehicle-based claim frequency and severity, data needed to be summarized to bring from trip level to vehicle level. Therefore a sum of distance driven for each vehicle was summed up and factors harsh acceleration, harsh breaking, cornering and speeding were calculated for 100 kilometres driven to make these

factors comparable for all vehicles. Also a combined score using all factors was calculated using formula:

$$score = \frac{(acceleration\ events + breaking\ events + cornering\ events + speeding\ events)}{distance\ driven} \cdot 100.$$

An area variable was added to the data set as it was known which vehicles drive in which area of Estonia. Area had four different values dividing the country into northern, southern, eastern and western area.

3.1.2 Historical claims data

Another data set was extracted from an insurance companies databases containing claims information about the vehicles under study. The data set contained information about claim date and size and vehicle information. Claims with the needed vehicle registration codes and which were in the observable time frame were taken out. Claims number was summarized on the basis of vehicle. Claim numbers were then added to the summarized telematics data.

3.2 Exploratory data analysis

Most vehicles in the data did not have any claims and only 11 objects had at least one claim in total. The highest claim frequency was four claims for a vehicle (Table 1).

Table 1: Claim frequencies

number of claims	0	1	2	3	4
number of vehicles	54	8	0	2	1

The exposure variable varies from 23 070 kilometres driven to 259 927 kilometres. When looking at the sum of driven distance in groups for claim and no claim (Table 2), the mean of distance is higher for no claim group, but median of distance is actually higher for claim group.

Table 2: Description of distance driven variable

ifclaim	**n**	**min**	**max**	**mean**	**median**
0	54	23 070	259 927	105 737	94 401
1	11	25 974	194 758	100 440	97 692

As seen from Figure 1, speeding per 100 kilometres has a lot lower values as compared to other variables indicating that drivers don't violate the speed limits as often. There is one object with very high combined score of variables compared to others, where higher number of events can be seen for all variables except for speeding. Even though the distance travelled is actually lower than the mean of the data set. Objects with at least one claim seem to be dispersed throughout the whole distance driven range and can't really say that more distance travelled during the time frame would mean more probable claim event. No real pattern can be seen from telematics variables and distance travelled.

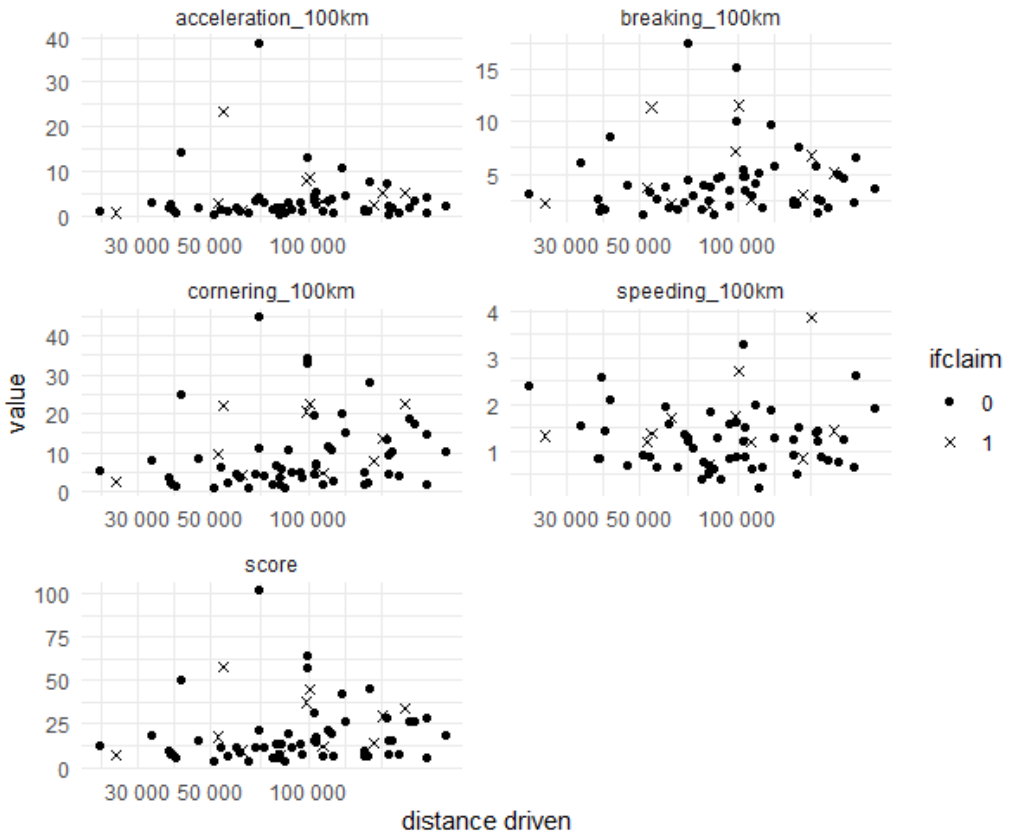

Figure 1: Visualization of telematic variables

As is seen from Figure 2, there are strong positive correlations between the telematics variables, with the exception of variable speeding. This kind of strong correlation between these variables is concerning for the modelling part.

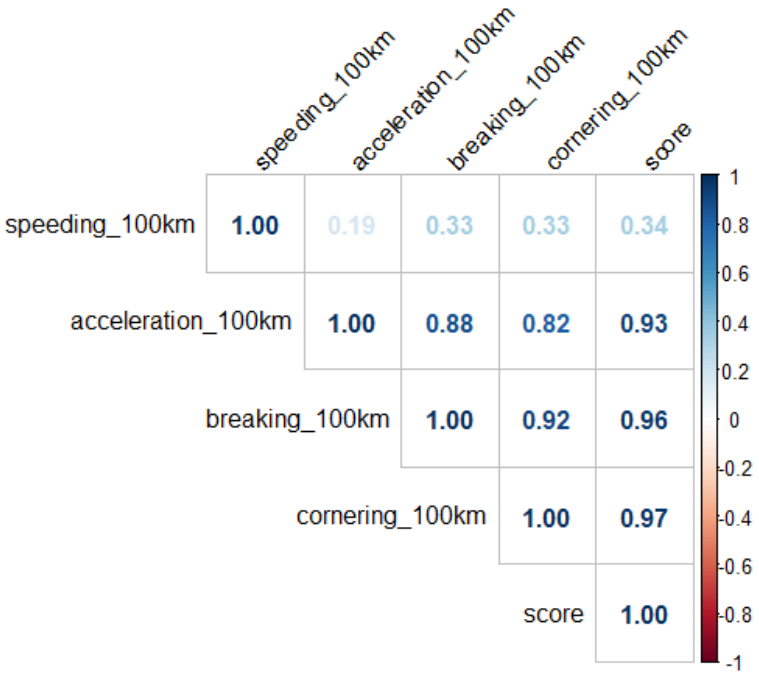

Figure 2: Correlation plot of telematics variables

3.3 Data issues

Usually, insurance data is quite big, and while the used telematic data set is very large and extensive, being on the trip level, it only contains information about 65 vehicles. Therefore for finding the claim count for vehicles after summarizing the data, there are only 65 rows to analyse and take into the model. This amount is a lot smaller than in research previously mentioned in 1.4. The data under study is also very specifically a small segment and of company cars whereas most research using telematic values is done for private customers' vehicles. Moreover, the insurance claims under study for this thesis are casco claims or motor own damage claims but for research mostly traffic-related claims are used. As is seen from (Figure 2) the correlations between telematics variables are also very high. For these reasons, there was reasonable doubt if the researched methodology would work for this data set.

3.4 Modelling claim frequency

Count models were used to build a model for claim frequency. Models tried were Poisson, Negative Binomial, ZIP and ZINB. Models were applied using packages `base` (R Core Team, 2020), `pscl` (Zeileis, Kleiber, and Jackman, 2008) and `MASS` (Venables and Ripley, 2002). The main focus was on how the driving behaviour variables could show the claim frequency. From other variables driving area was also taken into account. The offset variable showing exposure was the sum of distance driven.

First, a Poisson model with log-link and offset term as sum of distance driven was applied. All the telematic factors for 100 kilometres were taken as explanatory variables and area were taken into the first model (see listing 1). None of the driver behaviour variables showed significant effect ($\alpha = 0.05$) in the model and after backwards selection none of them still showed any significance. The effect of area was important in the model but did not statistically differ pair-wise (see listing 2). The best AIC was for the model using just area as an explanatory variable. The second best AIC was for the model which included score in addition to the area as explanatory variables (see listing 3).

Interestingly, when looking at the telematics variables' coefficients, harsh breaking per 100 kilometres gave a negative value, showing the decrease in effect for the mean response. This gave an idea to subtract breaking events in the score variable making process. However, after using it in the model it bore no significant results (see listing 4).

After trying the Poisson model the next logical step would be to try the negative binomial model. The best negative binomial model was one that used area as explanatory variable (see listing 5). The second best model was with explanatory variable score added to the area (see listing 6). Their AIC values showed better fit than of the Poisson values.

Given the nature of the data, a ZIP model was tried next with all the telematic

variables for 100 kilometres and area added to both parts of the model (see listing 7). Insignificant variables were removed from the zero model using backwards selection. Then attention was turned to the count part and the same process was followed. In the end, no variables were significant enough to be taken to the model (see listing 8). Then a model was applied using the score variable instead of four telematic variables (see listing 9). Here also the effect of score turned out to be insignificant. When area was discarded and just using score, the variable became significant in the count model part (see listing 10).

After ZIP models, a ZINB model was tried with telematics variables and area. A similar process to ZIP models was done. Unsignificant variables were removed from zero model and then from count model. No significant model was discovered. ZINB model also didn't show any significant effects of area in the model (see listing 11). The best fit gave the model which used the score variable in count part, but the effect of score wasn't itself discovered significant (see listing 12).

Model performance was compared using AIC. The smallest AIC would indicate a better model.

Table 3: AIC values for models

model	**AIC**
Poisson (area)	95.66
Poisson (area + score)	95.88
NB (area)	88.77
NB (area + score)	89.67
ZIP (score | score)	87.08
ZIP (area | 1)	86.61
ZINB (score | 1)	85.17

From Table 3, the best AIC would be given by ZINB (score|1) model, but since the effect of variable score is not statistically significant, ZIP models would be a better choice. From those the one with an explanatory variable area in the count process and just intercept in the zero process performs best. If wanting a model

with telematic variables, the ZIP model with the explanatory variable score in the count and zero process is the best.

Let's write out the two best models. For the ZIP models log link was used for the count part and logit link for the zero part. For ZIP(area | 1) model it is:

$$\begin{aligned}\hat{\mu_i} &= d_i \exp(-11.95 - 0.98I(area_i = \text{"Eastern"}) + \\ &\quad + 1.14I(area_i = \text{"Southern"}) + \\ &\quad + 1.37I(area_i = \text{"Western"})), \\ \hat{\pi}_i &= \frac{\exp(0.89)}{1 + \exp(0.89)},\end{aligned}$$

where the base level for area is "Northern". The negative coefficient in count part for when area is "Eastern" suggests an decrease in the claim frequency when compared to the base level of "Northern" whereas for other areas the positive coefficients show the increase in claim frequency compared to the base level.

The model for ZIP(score | score) we get:

$$\begin{aligned}\hat{\mu_i} &= d_i \exp(-10.48 - 0.04score_i), \\ \hat{\pi}_i &= \frac{\exp(2.49 - 0.07score_i)}{1 + \exp(2.49 - 0.07score_i)}.\end{aligned}$$

By this model, in the count part, the negative coefficient for the score would indicate that an increase in score would decrease the claim frequency. In the zero part of the model, the negative coefficient for the score indicates that a higher score implies lower odds for excess zeros.

3.5 Results and future research

For the given data set, none of the models worked very well so for an analysis done on this fleet, there can't be said that telematic variables would be good for

detecting claims.

The possible reasons for no significant generalizable results could be:

1. Insufficient data: The sample size and data quality is most possibly inadequate to detect any significant relationships as the number of object and the number of claims was very small. The statistical power to identify significant variables is limited.

2. There are problems with some of the modelling assumptions. As was seen (Figure 2), there were strong correlations between the telematic variables, which have an impact on model significance.

3. Fleet characteristics can disorder the estimations. Since vehicles can have multiple drivers the measured variables may not be suitable to use in modelling.

To look forward, a good idea would be to find more objects to take into analysis and reassess the models. Extending the subjects to personal vehicles instead of fleet cars could also be a good idea. It may be possible that selected explanatory variables are not significant in the prediction of claim frequency at all and other variables or transformations of variables should be considered. For example, the proportion of distance driven in urban areas and at night time could be some measures to think about. With more data also claim severity modelling could be done.

Conclusions

The goal of this master thesis was to give overview of opportunities for telematics data usage fin insurance pricing. Previously done research on the topic was summarized and found that usage-based insurance is successfully used already by many insurance providers and that the market has a lot of potential. It has been found that implementing usage-based insurance reduces accident rates and can be used to track green-driving. An overview of previous statistical research was given and found that telematics variables can be used in the models to predict claim frequency.

The second chapter gave an overview of suitable models which can be used for insurance claim frequency. Also an offset term was introduced where driven distance was used as an exposure measure.

The second part of the thesis could be called a case-study for one company's fleet. Data was extracted for about three years worth of time for a fleet of 65 vehicles and combined with claim information.

Data overview found 11 claims under study with the most being four claims per vehicle. Exploratory analysis showed strong correlations between telematics variables except for speeding. A combined variable score was also computed using 4 telematics variables scaled for 100 kilometres.

The biggest problem for this thesis was the lack of objects under study - the fleet only consisted of 65 vehicles. Therefore the modelling part and making statistically significant conclusions was complicated. It was found that telematics variables' are very strongly correlated with each other which is also problematic for the modelling process.

In the modelling process, Poisson, negative binomial, ZIP and ZINB models were applied with distance driven as offset. The effect of variable area was found significant in Poisson model. The best models were determined to be ZIP models on the

basis of significant effects of variables and AIC. One of them used the factor area in the count process and the other explanatory variable score in count as well as zero process. In the end a discussion of results was given and conclusions that for future research and analysis more data would be needed.

References

Ayuso, M., M. Guillen, and J.P Nielsen (2019). "Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data". In: *Transportation* 46 (3), 735–752. URL: https://doi.org/10.1007/s11116-018-9890-7.

Cachet: Kindlustusmaakler | Nutikad kindlustuslahendused (n.d.). Cachet Insurance Broker OÜ. URL: https://cachet.me/ (visited on 03/10/2023).

de Jong, P. and G.Z. Heller (2008). *Generalized linear models for insurance data*. fifth. Cambridge University Press. ISBN: 9780521879149.

England, Joanna (Nov. 4, 2021). "Are telematics the future of car insurance?" In: *IEEE Transactions on Intelligent Transportation Systems*. URL: https://insurtechdigital.com/insurtech/are-telematics-future-car-insurance.

European Connected Auto Insurance Study (2022). URL: https://www.ptolemus.com/research/european-connected-auto-insurance-study/.

Flückiger, Isabelle and Matteo Carbone (2021). *From Risk Transfer to Risk Prevention - How the Internet of Things is reshaping business models in insurance*. The Geneva Association, pp. 15–16.

Guillen, Montserrat, Jens Perch Nielsen, Mercedes Ayuso, and Ana M. Pérez-Marín (2019). "The Use of Telematics Devices to Improve Automobile Insurance Rates". In: *Risk Analysis* 39.3, pp. 662–672. DOI: https://doi.org/10.1111/risa.13172. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.13172. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13172.

Habas, Cathy (Mar. 27, 2023). *What Is a Telematics Device?* URL: https://www.safewise.com/faq/auto-safety/telematics-device/ (visited on 05/03/2023).

Maillart, Arthur (2021). "Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data". In: *European Actuarial Journal* 11 (2), pp. 579–617. ISSN: 2227-9091. DOI: 10.1007/s13385-021-00270-5.

MyTdigiteenused (n.d.). Toyota Eesti. URL: https://www.toyota.ee/owners/myt-connected-services/myt-and-multimedia (visited on 03/30/2023).

Orsoni, Damien (Oct. 7, 2022). *Usage-based insurance is growing globally but its dynamics are still regionally specific.* URL: https://www.ptolemus.com/insight/usage-based-insurance-is-growing-globally-but-its-dynamics-are-still-regionally-specific/ (visited on 02/01/2023).

Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz (2019). "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression". In: *Risks* 7.2. ISSN: 2227-9091. DOI: 10.3390/risks7020070. URL: https://www.mdpi.com/2227-9091/7/2/70.

Pokk, Priit (Aug. 2, 2018). *Eestlase hull idee nihutab kindlustuse piire.* URL: https://www.aripaev.ee/uudised/2018/08/02/eestlane-uritab-allutada-kindlustusmaailma-jagamismajanduse-reeglitele (visited on 03/27/2023).

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reimers, Imke and Benjamin R. Shiller (2019). "The Impacts of Telematics on Competition and Consumer Behavior in Insurance". In: *The Journal of Law*

and Economics 62.4, pp. 613–632. DOI: 10.1086/705119. eprint: https://doi.org/10.1086/705119. URL: https://doi.org/10.1086/705119.

Telematics/ Usage-Based Insurance (n.d.). National Association of Insurance Commissioners. URL: https://content.naic.org/cipr-topics/telematicsusage-based-insurance (visited on 02/01/2023).

Tiwari, Ajay (Mar. 13, 2020). *Modeling Insurance Claim Frequency. An illustrative guide to model insurance claim frequencies using generalized linear models in R*. URL: https://medium.com/swlh/modeling-insurance-claim-frequency-a776f3bf41dc (visited on 04/11/2023).

UBI Infographic 2016 (2016). PTOLEMUS Consulting Group. URL: http://www.ptolemus.com/content/uploads/2016/06/ptolemus-ubi-infographic-june162.png (visited on 03/05/2023).

Vehicle's e-insurance (n.d.). ERGO kindlustus. URL: https://www.ergo.ee/private-clients/vehicle-s-e-insurance-1 (visited on 09/16/2022).

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Vickrey, William (1968). "Automobile Accidents, Tort Law, Externalities, and Insurance: An Economist's Critique". In: *Law and Contemporary Problems* 33.3, pp. 464–487. ISSN: 00239186. URL: http://www.jstor.org/stable/1190938 (visited on 05/02/2023).

Wahlström, Johan, Isaac Skog, and Peter Händel (2017). "Smartphone-Based Vehicle Telematics: A Ten-Year Anniversary". In: *IEEE Transactions on Intelligent Transportation Systems* 18.10, pp. 2802–2825. DOI: 10.1109/TITS.2017.2680468.

Wei, Taiyun and Viliam Simko (2021). *R package 'corrplot': Visualization of a Correlation Matrix*. (Version 0.92). URL: https://github.com/taiyun/corrplot.

Wickham, Hadley (2007). "Reshaping Data with the reshape Package". In: *Journal of Statistical Software* 21.12, pp. 1–20. URL: http://www.jstatsoft.org/v21/i12/.

– (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=dplyr.

Zeileis, Achim, Christian Kleiber, and Simon Jackman (2008). "Regression Models for Count Data in R". In: *Journal of Statistical Software* 27.8. URL: http://www.jstatsoft.org/v27/i08/.

Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, and Graham M. Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. first. Springer New York, NY. ISBN: 978-0-387-87458-6. DOI: https://doi.org/10.1007/978-0-387-87458-6.

Appendix 1. R outputs of the models

Listing 1: Poisson base model

```
Call:
glm(formula = nbrclaims ~ acceleration_100km + breaking_100km +
    cornering_100km + speeding_100km + area, family = "poisson",
    data = data_with_claims, offset = log(sum_dist))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4357  -0.7504  -0.5509  -0.3016   3.2918

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -14.23213    0.94468 -15.066
< 2e-16 ***
acceleration_100km  0.07608    0.09075   0.838
0.40179
breaking_100km     -0.18161    0.27523  -0.660
0.50935
cornering_100km     0.04492    0.06415   0.700
0.48374
speeding_100km      0.48950    0.37021   1.322
0.18609
areaIda            -0.08795    1.16810  -0.075
0.93998
areaLõuna           1.08923    0.66465   1.639
0.10126
areaLääne           2.17306    0.75750   2.869
0.00412 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72.892  on 64  degrees of freedom
Residual deviance: 58.890  on 57  degrees of freedom
AIC: 100.14

Number of Fisher Scoring iterations: 7
```

Listing 2: Poisson model with area

```
Call:
glm(formula = nbrclaims ~ area, family = "poisson",
data = data_with_claims,
    offset = log(sum_dist))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2437  -0.8415  -0.6185  -0.3161   2.9488

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.2924     0.4082 -32.560  < 2e-16 ***
areaIda      -0.7028     1.0801  -0.651  0.51529
areaLouna     0.7291     0.6055   1.204  0.22857
areaLaane     1.7273     0.5773   2.992  0.00277 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72.892  on 64  degrees of freedom
Residual deviance: 62.415  on 61  degrees of freedom
AIC: 95.664

Number of Fisher Scoring iterations: 6
```

Listing 3: Poisson model with area and score

```
Call:
glm(formula = nbrclaims ~ area + score, family = "poisson",
    data = data_with_claims,
    offset = log(sum_dist))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3230  -0.7979  -0.6025  -0.3135   3.1919

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.88376    0.62487 -22.219  < 2e-16 ***
areaIda      -0.36417    1.12051  -0.325  0.74518
areaLõuna     0.92025    0.62741   1.467  0.14245
areaLääne     2.15271    0.67969   3.167  0.00154 **
score         0.01992    0.01358   1.467  0.14243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72.892  on 64  degrees of freedom
Residual deviance: 60.626  on 60  degrees of freedom
AIC: 95.875

Number of Fisher Scoring iterations: 6
```

Listing 4: Poisson model with area and modified score

```
Call:
glm(formula = nbrclaims ~ I(acceleration_100km + cornering_100km +
    speeding_100km - breaking_100km) + area, family = "poisson",
    data = data_with_claims, offset = log(sum_dist))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3743  -0.7877  -0.6200  -0.3098   3.2078

Coefficients:

(Intercept)
-13.85084     0.59066 -23.450  < 2e-16 ***
I(acceleration_100km+cornering_100km +
speeding_100km -breaking_100km)
0.03138     0.01993    1.574   0.11547
areaIda
-0.32093     1.12441   -0.285   0.77532
areaLõuna
0.91840     0.62452    1.471   0.14141
areaLääne
2.15833     0.67244    3.210   0.00133 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72.892  on 64  degrees of freedom
Residual deviance: 60.401  on 60  degrees of freedom
AIC: 95.65

Number of Fisher Scoring iterations: 6
```

Listing 5: Negative binomial model with area

```
glm.nb(formula = nbrclaims ~ area + offset(log(sum_dist)),
data = data_with_claims,
    init.theta = 0.3312458765, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8823  -0.6996  -0.5545  -0.2899   1.5560

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.2696     0.5470 -24.257   <2e-16 ***
areaIda      -0.8348     1.2861  -0.649   0.5163
areaLõuna     0.8996     0.8224   1.094   0.2740
areaLääne     1.5656     0.8850   1.769   0.0769 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.3312)
family taken to be 1)

    Null deviance: 37.083  on 64  degrees of freedom
Residual deviance: 31.532  on 61  degrees of freedom
AIC: 88.766

Number of Fisher Scoring iterations: 1


              Theta:  0.331
          Std. Err.:  0.217

 2 x log-likelihood:  -78.766
```

Listing 6: Negative binomial model with area and score

```
Call:
glm.nb(formula = nbrclaims ~ score + area + offset(log(sum_dist)),
    data = data_with_claims, init.theta = 0.360044076, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0381  -0.6735  -0.5334  -0.2877   1.8265

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.06639    0.84500 -16.647   <2e-16 ***
score         0.02461    0.01899   1.296   0.1951
areaIda      -0.35066    1.34069  -0.262   0.7937
areaLõuna     1.24669    0.85726   1.454   0.1459
areaLääne     2.13605    1.00324   2.129   0.0332 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.36)
family taken to be 1)

    Null deviance: 38.300  on 64  degrees of freedom
Residual deviance: 31.458  on 60  degrees of freedom
AIC: 89.673

Number of Fisher Scoring iterations: 1


              Theta:  0.360
          Std. Err.:  0.239

 2 x log-likelihood:  -77.673
```

Listing 7: ZIP base model

```
Call:
zeroinfl(formula = nbrclaims ~ acceleration_100km + breaking_100km +
    cornering_100km + speeding_100km + area |
    acceleration_100km + breaking_100km +
    cornering_100km + speeding_100km + area,
    data = data_with_claims,
    offset = log(sum_dist), dist = "poisson", link = "logit")

Pearson residuals:
      Min        1Q    Median        3Q       Max
-1.104e+00 -2.276e-01 -1.480e-08 -1.239e-08  3.114e+00

Count model coefficients (poisson with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.0598     2.2194  -3.181
0.00147 **
acceleration_100km  0.0970     0.1412   0.687  0.49197
breaking_100km      0.2865     0.3323   0.862  0.38858
cornering_100km    -0.1578     0.1091  -1.447  0.14794
speeding_100km     -1.9988     0.8962  -2.230
0.02573 *
areaIda            -3.5714     1.6684  -2.141
0.03230 *
areaLõuna          -3.2400     1.5588  -2.078
0.03767 *
areaLääne          -1.2801     1.6046  -0.798  0.42499

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        335.442    785.023   0.427    0.669
acceleration_100km   7.191    240.176   0.030    0.976
breaking_100km      20.799    127.823   0.163    0.871
cornering_100km    -15.419    157.460  -0.098    0.922
speeding_100km    -132.429    510.225  -0.260    0.795
areaIda           -194.287    475.752  -0.408    0.683
areaLõuna         -262.053    560.847  -0.467    0.640
areaLääne         -193.157    679.850  -0.284    0.776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 628
Log-likelihood: -24.07 on 16 Df

AIC: 80.13656
```

Listing 8: ZIP model with area

```
Call:
zeroinfl(formula = nbrclaims ~ +area | 1,
    data = data_with_claims, offset = log(sum_dist),
    dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5402 -0.4492 -0.3735 -0.1947  3.1800

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.9541     0.5586 -21.401   <2e-16 ***
areaIda      -0.9832     1.1684  -0.842   0.4001
areaLõuna     1.1368     0.7819   1.454   0.1459
areaLääne     1.3674     0.6986   1.957   0.0503 .

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8936     0.4448   2.009   0.0446 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 15
Log-likelihood: -38.31 on 5 Df

AIC: 86.61263
```

Listing 9: ZIP model with score and area

```
Call:
zeroinfl(formula = nbrclaims ~ score + area | 1,
    data = data_with_claims,
    offset = log(sum_dist), dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5518 -0.4469 -0.3871 -0.1955  3.2393

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.248367   1.055458 -11.605   <2e-16 ***
score         0.007235   0.022484   0.322   0.7476
areaIda      -0.815410   1.265813  -0.644   0.5195
areaLõuna     1.196014   0.781513   1.530   0.1259
areaLääne     1.574319   0.932434   1.688   0.0913 .

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8322     0.5045    1.65    0.099 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 22
Log-likelihood: -38.26 on 6 Df

AIC: 88.51213
```

Listing 10: ZIP model with score

```
Call:
zeroinfl(formula = nbrclaims ~ score | score,
    data = data_with_claims,
    offset = log(sum_dist), dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.6618 -0.3991 -0.3109 -0.2444  3.9604

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.48296    0.50613 -20.712   <2e-16 ***
score        -0.03757    0.01890  -1.988   0.0468 *

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.49119    0.78643   3.168  0.00154 **
score       -0.07330    0.03905  -1.877  0.06052 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 10
Log-likelihood: -39.54 on 4 Df

AIC: 87.07884
```

Listing 11: ZINB model with area

```
Call:
hurdle(formula = nbrclaims ~ area | 1,
data = data_with_claims, offset = log(sum_dist),
dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.4513 -0.4043 -0.3913 -0.3875  3.6303

Count model coefficients (truncated poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.9938     0.6654 -18.025   <2e-16 ***
areaIda      -9.4999   165.9730  -0.057    0.954
areaLõuna     1.0565     0.9364   1.128    0.259
areaLääne     1.3241     0.8492   1.559    0.119
Zero hurdle model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5911     0.3308   -4.81 1.51e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 375
Log-likelihood: -38.77 on 5 Df

AIC: 87.53401
```

Listing 12: ZINB model with score

```
Call:
hurdle(formula = nbrclaims ~ score | 1,
data = data_with_claims, offset = log(sum_dist),
dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.4500 -0.3971 -0.3895 -0.3874  4.0010

Count model coefficients (truncated poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.36460    0.64738 -16.010   <2e-16 ***
score        -0.05318    0.03349  -1.588    0.112
Zero hurdle model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5911     0.3308   -4.81 1.51e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 12
Log-likelihood: -39.58 on 3 Df

AIC: 85.16874
```

Non-exclusive licence to reproduce thesis and make thesis public

I, Kaari Kuus,

1. herewith 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, Claims frequency modelling with usage-based insurance data , supervised by Meelis Käärik and Ervin Sääsk.

2. I grant the University of Tartu a permit to make the work specified in p. 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 3.0, 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 p. 1 and 2.

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

Kaari Kuus
18/05/2023