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

Minni-Marii Paarmets

Claims frequency modeling and model comparison for travel insurance

Actuarial and Financial Engineering

Master's Thesis (30 ECTS)

Supervisor: Assoc. prof. Raul Kangro

Tartu 2025

CLAIMS FREQUENCY MODELING AND MODEL COMPARISON FOR TRAVEL INSURANCE

Master thesis

Minni-Marii Paarmets

Abstract

Travel insurance is a widely used insurance product covering various risks that might occur while traveling. The aim of this thesis is to model the frequency of travel insurance claims using generalized linear models and different types of metrics to determine which type of model performs best. The thesis is split into three parts, where the first gives an overview of the background, the data, and travel insurance. The second part describes the theory of the models and model comparison techniques. The third part details the analysis of the models and their comparisons.

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

Key Words: Travel insurance, generalized linear models, generalized additive models, k-fold cross-validation .

REISIKINDLUSTUSE KAHJUDE SAGEDUSE MODELLEERIMINE JA MUDELITE VÕRDLUS

Magistritöö

Minni-Marii Paarmets

Lühikokkuvõte

Reisikindlustus on laialdaselt kasutatav kindlustustoode, mis katab erinevaid reisimisega seonduvaid riske. Selle lõputöö eesmärk on modelleerida reisikindlustuse kahjude sagedust kasutades üldistatud lineaarseid mudeleid ning võrrelda saadud mudeleid mitmete erinevate mõõdikutega. Lõputöö on jaotatud

kolmeks erinevaks osaks, millest esimene osa annab ülevaate töö taustast, andmestikust ning reisikindlustusest. Teine osa kirjeldab mudelite teooriat ja mudelite võrdlemise metoodikaid ning kolmas osa sisaldab mudelite analüüsi ja võrdlust.

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

Märksõnad: Reisikindlustus, üldistatud lineaarsed mudelid, üldistatud aditiivsed mudelid, ristvalideerimine.

Contents

Introduction **5**

1 Background **6**

2 Generalized linear models **8**

2.1 Poisson regression 8

2.2 Negative binomial regression 10

2.3 Zero-inflated model 11

2.4 Zero-altered model 13

2.5 Generalized additive model 14

2.6 Lasso regularization 16

3 Model comparison methods **17**

3.1 Likelihood ratio test 17

3.2 Akaike and Bayesian information criterion 18

3.3 k-fold cross-validation 19

4 Model analysis **22**

4.1 Poisson and negative binomial regression models 22

4.2 ZIP and ZINB regression models 25

4.3 ZAP and ZANB regression models 29

4.4 GAM 31

4.5 Lasso regularization 36

4.6 Results 37

Summary **39**

Bibliography **40**

Appendix 1. R model outputs **43**

Introduction

In insurance, product prices are usually calculated using statistical models that predict the claim frequency, defined as the expected number of claims per exposure unit, and claim severity, meaning the average cost per claim. The premium is obtained by multiplying these two estimates (Ohlsson and Johansson, 2010, p. 4). Generalized linear models are most commonly applied to estimate both the frequency and severity of insurance products.

Whereas most previous theses have focused on motor vehicle or casco insurance, which normally have policies that last 1 year, this thesis models the frequency of travel insurance claims where policy duration varies with trip length. The aim of this thesis is to develop several different types of models, evaluate their performance with various types of metrics to see which type of model performs the best.

The first chapter gives an overview of travel insurance and explains the concepts of claim frequency and severity. It also includes a description of the used dataset. The second chapter introduces the theory of the different types of models that will be used and gives a brief description of the distributions used in the models. The third chapter describes the theoretical background of different methods used for model comparison. Lastly, the fourth chapter discusses the model selection process, presents the final models, and model comparison metrics. Additionally, it gives final results and an overview of how each model performed.

1 Background

The main method in insurance pricing is frequency and severity modeling. Frequency refers to how often a claim occurs within a specific period and is usually expressed as a rate. Severity represents the average cost of a claim when it occurs, which can vary depending on the insured product or the risk incurred (Ohlsson and Johansson, 2010, p. 4).

In travel insurance, frequency models the likelihood of events such as trip cancellations, lost luggage, flight delays, or medical emergencies. The severity of these claims can vary greatly depending on different factors like destination or length of the trip. Unlike other insurance products, travel insurance is short-term and covers only the duration of your trip. Travel insurance policies are also quite customizable, as it is often possible to select specific coverage amounts for things like luggage or specify the reason for the trip. In addition, travel insurance is more affected by seasonal trends, holidays, and external factors like geopolitical situations or weather conditions compared to other insurance products (Actuaries, 2018).

The data set is from Swedbank P&C Insurance AS. Data mining was performed in the Oracle SQL database, where preliminary data cleaning was performed. Additional data cleaning, analysis, and risk grouping were done using the R software. The data set covers the period from 1 September 2021 to 31 August 2024 and includes information about the insurance company's policies, customers in Estonia, Latvia, and Lithuania, and any claims that occurred. A subset of this dataset was created, focusing only on policies that were part of a single risk group. From this point onward, only this subset will be referenced in this thesis.

The subset contains 374,806 rows, with only around 2,400 representing policies that had claims. The maximum number of claims per risk group was 4, while the average number of claims was close to zero due to the low frequency of claims. The earned period, measured in days, ranged from 1 to 365, with a median of 8 days.

In the modeling process, the earned period is converted into years and used as a measure of exposure, representing the time an insured person is covered. In the data set, in addition to exposure and number of claims, there are seven numeric, eight nominal, one categorical, and four binary variables, making a total of 22 variables.

2 Generalized linear models

Unless said otherwise, this chapter is written based on (McCullagh and Nelder, 1989). Generalized linear models (GLMs) are used when the dependent variable's conditional distribution is not normally distributed. Instead of assuming that the dependent variable is normally distributed given the independent variables, GLMs can handle different types of data by using distributions from the exponential family, such as the Poisson and negative binomial distributions for count data. In linear regression, the mean of the response is modeled directly as a linear combination of the predictors. In contrast, GLMs use a link function, which connects the mean of the response to the linear predictor.

2.1 Poisson regression

Poisson regression is a generalized linear model used to model count data, where the response variable represents the number of occurrences of an event over a fixed period or unit of measurement. It assumes that the response variable follows a Poisson distribution.

A discrete random variable Y is said to follow a Poisson distribution, denoted as $Y \sim \text{Po}(\mu)$, if it has the probability mass function:

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y \in \{0, 1, 2, \dots\}, \quad \mu > 0$$

In a Poisson model, the conditional mean and variance of the response variable are equal:

$$E(Y|\mathbf{x}) = \text{Var}(Y|\mathbf{x}) = \mu$$

Since count data is non-negative and often follows a right-skewed distribution, standard linear models are not appropriate for modeling it. Instead, Poisson regression models the relationship between predictor variables and the expected count μ_i,

using a log link function:

$$g(\mu_i) = \ln(\mu_i), \quad \ln \mu_i = \mathbf{x}_i^T \beta$$

This ensures that μ_i remains positive, because taking the exponent of both sides gives:

$$\mu_i = \exp(\mathbf{x}_i^T \beta) \tag{1}$$

Expanding this, we get:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdots \exp(\beta_k x_{ik})$$

It is reasonable to assume that if the other variables remain constant, then the expected number of claims is proportional to the length of the exposure. Therefore, estimating the quantity $\frac{\mu}{t}$ is logical. As a result, it is necessary to include an additional term, called offset, in the model, which accounts for the time period. In this case:

$$\ln\left(\frac{\mu}{t}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

or equivalently,

$$\eta = \ln(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \ln(t),$$

where $\ln(t)$ is the offset (Hilbe, 2011, pp. 134–135).

Consider the vector y_i, where each value is an independent random variable with a distribution having a mean of μ_i, then the likelihood function for parameter estimation for Poisson regression is defined as:

$$L(\boldsymbol{\mu}, \mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

From this, we derive the log-likelihood function:

$$\ell(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{n} (-\mu_i + y_i \ln \mu_i - \ln(y_i!))$$

2.2 Negative binomial regression

This subchapter is written based on (Hilbe, 2011, pp. 185–219).

Count data often exhibit overdispersion, where the variance exceeds the mean. In such cases, alternative models like negative binomial regression are used, as they generalize Poisson regression by relaxing its assumption that the variance equals the mean.

A discrete random variable Y follows a negative binomial distribution, denoted as $Y \sim \mathrm{NB}(k, \pi)$, if it has the probability mass function:

$$p(y; k, \pi) = \frac{\Gamma(k+y)}{y!\,\Gamma(k)} \pi^k (1-\pi)^y, \quad y \in \{0, 1, 2, \dots\}, \quad k > 0, \quad 0 < \pi < 1.$$

In negative binomial regression, the log-link function is used to relate the expected value μ_i to the linear predictor η_i:

$$\eta_i = g(\mu_i) = \ln(\mu_i), \quad \mu_i = h(\eta_i) = \exp(\eta_i).$$

Similarly to the Poisson regression:

$$\mu_i = \exp\left(\ln(t_i) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\right) \tag{2}$$

where $\ln(t)$ is known as the offset.

Let $\alpha = 1/k$. The likelihood function for parameter estimation for negative binomial regression is defined as:

$$L(\boldsymbol{\mu}, \mathbf{y}, \alpha) = \prod_{i=1}^{n} \frac{\Gamma(y_i + 1/\alpha)}{y_i!\,\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

where α is called the dispersion parameter. From the likelihood function, it's possible to derive the log-likelihood function:

$$\ell(\boldsymbol{\mu}, \mathbf{y}, \alpha) = \sum_{i=1}^{n} \Bigg[\log \Gamma(y_i + 1/\alpha) - \log y_i! - \log \Gamma(1/\alpha) \\ + y_i \log(\alpha \mu_i) - \left(y_i + \frac{1}{\alpha} \right) \log(1 + \alpha \mu_i) \Bigg]$$

2.3 Zero-inflated model

This subchapter is written based on (Cameron and Trivedi, 2013, pp. 140–141). In cases where the data contains an excessive number of zeros, it is possible to model them separately using zero-inflated models. These models assume that zeros come from a separate process, typically modeled using a Bernoulli distribution, while the counts (including some of the zeros) are modeled using either Poisson or negative binomial regression.

Consider first the case when Poisson regression is used for estimating the count process (1). We also define a process that generates excess zeros:

$$\ln \frac{\pi_i}{1 - \pi_i} = Z_i^T \gamma, \quad \pi_i = \frac{\exp\left(Z_i^T \gamma\right)}{1 + \exp\left(Z_i^T \gamma\right)}, \tag{3}$$

where $Z_i = (1, Z_{i1}, Z_{i2}, \ldots, Z_{ik})$ and $\gamma = (\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_k)$

Zero-inflated Poisson (ZIP) model:

$$\mathbf{P}(Y_i = 0) = \pi_i + (1 - \pi_i) \exp\left(-\mu_i\right), \\ \mathbf{P}(Y_i = y_i) = (1 - \pi_i) \frac{\exp\left(-\mu_i\right) \left[\mu_i\right]^{y_i}}{y_i!}, \quad y_i = 1, 2, \ldots$$

The likelihood function for parameter estimation for ZIP model is defined as:

$$L(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}) = \prod_{i:y_i=0} \left[\pi_i + (1 - \pi_i) e^{-\mu_i}\right] \cdot \prod_{i:y_i>0} (1 - \pi_i) \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

From this, we derive the log-likelihood function:

$$\ell(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \Big\{ \mathbf{1}_{\{y_i=0\}} \ln\big[\pi_i + (1-\pi_i)e^{-\mu_i}\big] + \mathbf{1}_{\{y_i>0\}} \big[\ln(1-\pi_i) + y_i \ln \mu_i - \mu_i - \ln(y_i!)\big] \Big\}.$$

Alternatively, we can estimate the counts using a negative binomial regression with a counting process (2) and a zero-generating process (3). Then the zero-inflated negative binomial (ZINB) model:

$$\mathbf{P}(Y_i = 0) = \pi_i + (1-\pi_i)\,(1+\alpha\mu_i)^{-1/\alpha},$$

$$\mathbf{P}(Y_i = y_i) = (1-\pi_i)\,\frac{\Gamma\big(y_i + 1/\alpha\big)}{y_i!\,\Gamma(1/\alpha)}\,(\alpha\mu_i)^{y_i}\,(1+\alpha\mu_i)^{-\,(y_i+1/\alpha)}, \quad y_i = 1, 2, \ldots$$

where α is called a dispersion parameter of the Negative Binomial distribution.

The likelihood function for parameter estimation for the ZINB model is defined as:

$$L(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}, \alpha) = \prod_{i:y_i=0} \Big[\pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-1/\alpha}\Big] \cdot \prod_{i:y_i>0} \Big[(1-\pi_i)\,\frac{\Gamma(y_i+1/\alpha)}{y_i!\,\Gamma(1/\alpha)}(\alpha\mu_i)^{y_i}(1+\alpha\mu_i)^{-\,\left(y_i+1/\alpha\right)}\Big].$$

From this, we derive the log-likelihood function:

$$\ell(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}, \alpha) = \sum_{i:y_i=0} \log\Big[\pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-1/\alpha}\Big] + \sum_{i:y_i>0} \Big\{ \log(1-\pi_i) + \log\Gamma(y_i+1/\alpha) - \log y_i! - \log\Gamma(1/\alpha) + y_i \log(\alpha\mu_i) - (y_i+1/\alpha)\log(1+\alpha\mu_i) \Big\}.$$

2.4 Zero-altered model

This subchapter is written based on (Cameron and Trivedi, 2013, pp. 136–139). Contrary to zero-inflated models, zero-altered (hurdle) models do not include zeros in their count process. Rather than assuming that zeros come from two distinct processes as in ZI, hurdle models separate the modeling of zero vs. positive counts.

A hurdle model consists of two components:

1. A binary model that determines whether the observation is zero or takes a positive value.
2. A truncated count model (Poisson or negative binomial) that models the positive counts only if it has crossed the hurdle.

For the zero-altered Poisson model π_i and μ_i are respectively defined as (3) and (1). The zero-altered Poisson (ZAP) model given by:

$$\begin{aligned}
\mathbf{P}(Y_i = 0) &= \pi_i, \\
\mathbf{P}(Y_i = y_i) &= [1 - \pi_i] \cdot \frac{\exp(-\mu_i)\,\mu_i^{y_i}}{(1 - e^{-\mu_i})y_i!}, \quad y_i = 1, 2, \ldots
\end{aligned}$$

The likelihood function for parameter estimation for the ZAP model is defined as:

$$L(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}) = \prod_{i:y_i=0} \pi_i \cdot \prod_{i:y_i>0} (1 - \pi_i) \frac{\mu_i^{y_i} e^{-\mu_i}/y_i!}{1 - e^{-\mu_i}}$$

From this, we derive the log-likelihood function:

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}) = \sum_{i:y_i=0} \ln(\pi_i) + \sum_{i:y_i>0} \Big[&\ln(1 - \pi_i) + y_i \ln(\mu_i) \\
&- \mu_i - \ln(y_i!) - \ln\big(1 - e^{-\mu_i}\big) \Big]
\end{aligned}$$

Analogously, for zero-altered negative binomial model counts are estimated using a negative binomial regression with counting process (2) and a zero-generating

process (3). Thus, ZANB model is given by:

$$\mathbf{P}(Y_i = 0) = \pi_i,$$
$$\mathbf{P}(Y_i = 0) = (1 - \pi_i)\,\frac{\frac{\Gamma(y_i+1/\alpha)}{y_i!\,\Gamma(1/\alpha)}\,(\alpha\mu_i)^{y_i}\,(1+\alpha\mu_i)^{-\,(y_i+1/\alpha)}}{1-(1+\alpha\mu_i)^{-1/\alpha}} \quad , \quad y_i = 1, 2, \ldots$$

where α is called a dispersion parameter of the negative binomial distribution.

The likelihood function for parameter estimation for the ZANB model is defined as:

$$L(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}, \alpha) = \prod_{i:y_i=0} \pi_i \cdot \prod_{i:y_i>0} \Bigg[(1-\pi_i)\,\frac{\Gamma(y_i+1/\alpha)}{y_i!\,\Gamma(1/\alpha)} (\alpha\mu_i)^{y_i} \\ (1+\alpha\mu_i)^{-\left(y_i+1/\alpha\right)} \Big/ \left[1-(1+\alpha\mu_i)^{-1/\alpha}\right] \Bigg].$$

From this, we derive the log-likelihood function:

$$\ell(\boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\pi}, \alpha) = \sum_{i:y_i=0} \ln \pi_i \; + \sum_{i:y_i>0} \Big\{ \ln(1-\pi_i) + \ln\Gamma(y_i + 1/\alpha) \\ - \ln y_i! - \ln\Gamma(1/\alpha) + \; y_i \ln(\alpha\mu_i) - \; (y_i + 1/\alpha) \ln(1+\alpha\mu_i) \\ - \ln\left[1-(1+\alpha\mu_i)^{-1/\alpha}\right] \Big\}.$$

2.5 Generalized additive model

This subchapter is written based on (Wood, 2017, pp. 161–182), unless stated otherwise.

A generalized additive model (GAM) is an extension of a generalized linear model that allows the linear predictor to be a sum of smooth, potentially non-linear functions of the predictor variables. The general form of a GAM can be written as

$$g(\mu_i) = \eta_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip})$$

where $\mu_i = \mathbb{E}[Y_i]$ is the expected value of the response variable Y_i for observation i, and it is assumed to follow a distribution of the exponential family. The functions in the model can include multiple variables, meaning that x_{ij}, where $j = 1, \ldots, p$, can also be a multidimensional set of variables.

Here, the link function $g(\cdot)$ connects the expected value μ_i to the additive predictor η_i. In this case, if the log-link function is used, it relates the expected value μ_i to the additive predictor η_i as follows:

$$\eta_i = g(\mu_i) = \ln(\mu_i), \quad \mu_i = h(\eta_i) = \exp(\eta_i)$$

The functions $f_j(x_j)$ are smooth functions that describe the potentially nonlinear relationship between the predictor variable x_j and the response. These functions are not specified in parametric form. Instead, they are estimated from the data using nonparametric techniques such as smoothing splines.

The idea of smoothing splines is to fit a smooth curve to a set of data by using flexible piecewise polynomials. To prevent overfitting, an additional term is added to the loss function that penalizes rapid changes in the fitted function $f(x)$. This ensures that the curve remains smooth while still fitting the data well. We want the average loss from our predictions to be small, but on the other hand, we want it to be smooth, so the approach is to find a function g that minimizes

$$\sum_{i=1}^{n} L(f(x_i), y_i) + \lambda \int \left(f''(t)\right)^2 dt.$$

where $\lambda \geq 0$ is a tuning parameter and $L(f(x_i), y_i)$ is the loss function. In case of a linear regression model $L(f(x_i), y_i) = (y_i - f(x_i))^2$. The smoothing function is penalized by an additional integral to control its smoothness. Each predictor may have its smoothing parameter, which helps prevent overfitting while allowing for flexible, non-linear relationships between the predictors and the response variable. The term $\lambda \int (f''(t))^2$ is called the penalty term, which penalizes the variability in

f (James et al., 2021, pp. 301–302).

2.6 Lasso regularization

This subchapter is written based on (James et al., 2021, pp. 241–242).

Lasso regression, when applied to Poisson models, is a regularization technique that improves the prediction accuracy and interpretability of generalized linear models by penalizing large coefficients. The penalty term added to the loss function is proportional to the sum of the absolute values of the model coefficients, which tends to shrink some coefficients exactly to zero, performing variable selection. When lasso regularization is used for Poisson regression, then the coefficients are estimated by minimizing the following function (Hastie, Friedman, and Tibshirani, 2025):

$$\sum_{i=1}^{N}[y_i \ln \mu_i - \mu_i - \ln(y_i!)] + \lambda \sum_{j=1}^{p} |\beta_j|$$

It uses an ℓ_1 penalty, which refers to the ℓ_1-norm of the coefficient vector. The ℓ_1-norm of a coefficient β is given by $\|\beta\|_1 = \sum_j |\beta_j|$.

3 Model comparison methods

Models can be compared in two different ways: by assessing their fit on training data and by evaluating their performance on unseen data. In this thesis, Aikaike Information Criterion and Bayesian Information Criterion are used to compare the model fit on the training set, and the likelihood ratio test is performed to compare strictly nested models. To evaluate performance on unseen data, 5-fold cross-validation is done, where root-mean-square error, mean absolute error, and log-likelihood are computed.

3.1 Likelihood ratio test

The likelihood ratio test is used to compare the fit of a reduced model to the fit of the other model by comparing the maximum likelihood values. Let's define models $M_1 = M(X_1)$ and $M_2 = M(X_2)$, where M_1 is nested in M_2 and X is the design matrix. Additionally, let's define their maximum values of likelihood:

$$L(\hat{\boldsymbol{\beta}}_{M_1}; \mathbf{y}) = \max_{\boldsymbol{\beta} \in M_1} L(\boldsymbol{\beta}; \mathbf{y}), \quad L(\hat{\boldsymbol{\beta}}_{M_2}; \mathbf{y}) = \max_{\boldsymbol{\beta} \in M_2} L(\boldsymbol{\beta}; \mathbf{y})$$

Then, the likelihood ratio statistic is:

$$\lambda^* = \frac{L(\hat{\boldsymbol{\beta}}_{M_1}; \mathbf{y})}{L(\hat{\boldsymbol{\beta}}_{M_2}; \mathbf{y})},$$

which lies between 0 and 1 (Käärik, 2023).

The idea is to test H_0 against H_1, which are defined as:

$$H_0 : \ \theta \in \Theta_0, \quad H_1 : \ \theta \in \Theta \setminus \Theta_0,$$

where Θ is the full parameter space and $\Theta_0 \subset \Theta$ is the restriction under the null hypothesis. Values of λ^* close to zero indicate that the constrained model explains

the data less well than the full model, suggesting that the restriction in H_0 may not be valid (Casella and Berger, 2002, pp. 373–375).

3.2 Akaike and Bayesian information criterion

The Akaike information criterion (AIC) is a model selection criterion, which estimates the quality of a statistical model for a given dataset by balancing goodness-of-fit and model complexity. It is defined as:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

where

- k - the number of parameters in the model,
- $\hat{L}$ - the maximum value of the likelihood function.

AIC rewards models that fit the data well, but includes a penalty for each additional parameter to discourage overfitting. AIC is comparable only if the models are fitted on the same dataset (Burnham and Anderson, 2002, pp. 60–64).

The Bayesian Information Criterion (BIC) includes a sample size-dependent penalty. It is defined as:

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

where

- n - the number of observations,
- k - the number of parameters in the model,
- $\hat{L}$ - the maximum value of the likelihood function.

Like AIC, BIC penalizes model complexity, but more strongly as n increases, making it more conservative. Both criteria use the likelihood to assess model fit, but differ in their penalty terms. AIC uses a constant penalty and tends to favor more complex models, making it suitable for prediction. BIC's penalty grows with sample size, often favoring simpler models. Similarly to AIC, BIC can only be compared across models when they have been fitted to the same dataset (Burnham and Anderson, 2002, pp. 271–293).

3.3 k-fold cross-validation

This subchapter is written based on (Hastie, Tibshirani, and Friedman, 2009, pp. 241–247), unless it is stated otherwise.
In statistical models, one of the main goals is to assess how the chosen model performs on new data. Usually, it is measured by expected loss, which means evaluating $\mathbb{E}(L(f(X), Y))$, where (X, Y) is a random sample of the population and f is the prediction function for the fitted model.

To evaluate this, a model is fitted to a data subset, and an expected loss is calculated by using the remaining data as an independent sample from the population. In this case, two problems arise. First, the model isn't fitted using all of the available data, resulting in worse results. Secondly, if a big part of the data is left out from model estimation, then the expected loss calculated from the remaining data can be quite imprecise.

Alternatively, instead of looking at a specific model with fixed coefficients, it is possible to compare the models by their expected loss when each model is fitted on one random sample and then evaluated on another random sample. In other words, the prediction function is also treated as a random function.

The goal is to estimate

$$\mathbb{E}[L(f(X, data_n), Y)],$$

where the expected value is taken over every n-point random samples $data_n$ drawn from a given distribution, and (X, Y) is a random vector from the same population distribution.

To approximate this, k tries are done, where each time a model is trained on $\frac{k-1}{k}$ of the data and then the loss is evaluated on the remaining $\frac{1}{k}$. Thus, instead of the target quantity

$$H_n = \mathbb{E}[L(f(X, data_n), Y)],$$

this is estimated:

$$H_k = \mathbb{E}[L(f(X, data_{n\times(k-1)/k}), Y)].$$

The typical choices for k are 5 or 10, since it's clear that this estimator becomes more biased as k decreases, since each model is trained on fewer observations. Conversely, for large k, the different training sets overlap, differing only by a small fraction of points, which results in the estimates having high variance depending on the particular sample.

Usually, the loss from prediction errors is measured by mean squared error, but it can be replaced with any other metric. In this case, the root mean squared error (RMSE) is used (James et al., 2013, pp. 203–205) and defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2},$$

where:

- y_i is the true response value for the i-th observation in the test set,
- $\hat{f}(x_i)$ is the predicted value produced by the model for the corresponding predictor x_i,
- n is the number of observations in the test set.

Additionally, mean absolute error (MAE) and log-likelihood values are estimated.

MAE is the average of the absolute differences between observed and predicted values (Willmott and Matsuura, 2005):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4 Model analysis

Modeling was done in the R software using the `MASS` package (Venables and Ripley, 2002) for the Poisson and negative binomial regression, the `pscl` package (Jackman et al., 2024) for zero-inflated and zero-altered models, the `mgcv` package (Wood, 2025) for GAMs, and the `tidymodels` (Kuhn and Wickham, 2025) and `offsetreg` (Heaphy, 2025) packages for lasso regularization.

First, a model is estimated by removing variables with statistically insignificant effects one by one, starting with a variable that has the highest p-value. In this process, AIC values are also considered, as a model with the lowest AIC is preferred. The likelihood ratio test was used to compare nested models, for example, when interaction terms were added. After the final model is estimated, AIC and BIC values are calculated. Next, 5-fold cross-validation is performed, where for each fold, RMSE, MAE, and log-likelihood values are calculated. The average of these results over the five folds is then computed to compare the models. The statistically significant effect of the variables is set as $p < 0.1$, because the models performed better at this level compared to the stricter $p < 0.05$ criterion. The model outputs can be seen in Appendix 1.

4.1 Poisson and negative binomial regression models

A Poisson regression model was fitted to the full dataset. After removing variables with statistically insignificant effects, 12 variables and an offset term remained. Next, another Poisson regression model was estimated, this time including interaction terms to assess whether it would perform better. A likelihood ratio test was then performed, having a p-value below 0.05. This indicates that the model with interaction terms provides a significantly better fit.

After adding the interaction terms, the effect of variable X_3 is no longer statistically significant. However, it won't be removed, as its inclusion results in a lower AIC.

Final Poisson regression model formula:

$$
\begin{aligned}
\mu_i = &\exp(-188.340 - 0.661 \cdot I(X_{1i} = 2) - 0.214 \cdot I(X_{1i} = 3) + 0.093 \cdot X_{2i} \\
&+ 0.001 \cdot X_{3i} - 0.695 \cdot I(X_{4i} = \text{yes}) + 0.119 \cdot I(X_{5i} = 1) + 0.330 \cdot I(X_{5i} = 2) \\
&+ 0.640 \cdot I(X_{5i} = 3) - 0.073 \cdot I(X_{6i} = 1) - 0.286 \cdot I(X_{6i} = 2) - 0.855 \cdot I(X_{6i} = 3) \\
&+ 0.488 \cdot X_{7i} + 0.486 \cdot I(X_{8i} = 1) - 0.064 \cdot I(X_{8i} = 2) - 0.075 \cdot I(X_{8i} = 3) \\
&- 0.443 \cdot I(X_{9i} = 1) - 0.122 \cdot I(X_{9i} = 2) + 0.157 \cdot X_{10i} - 0.004 \cdot X_{11i} \\
&- 0.292 \cdot I(X_{12i} = 2) - 0.088 \cdot I(X_{12i} = 3) - 0.146 \cdot I(X_{12i} = 4) \\
&- 0.001 \cdot I(X_{12i} = 5) - 0.342 \cdot I(X_{12i} = 6) - 0.520 \cdot I(X_{12i} = 7) \\
&- 0.303 \cdot I(X_{12i} = 8) - 0.058 \cdot I(X_{12i} = 9) + 0.098 \cdot I(X_{12i} = 10) \\
&+ 0.263 \cdot I(X_{12i} = 11) + 0.292 \cdot I(X_{12i} = 12) - 0.002 \cdot (X_{10i} \cdot X_{11i}) \\
&+ 0.004 \cdot (X_{11i} \cdot I(X_{1i} = 2)) - 0.008 \cdot (X_{11i} \cdot I(X_{1i} = 3)) + ln(\text{exposure}_i))
\end{aligned}
$$

Even though there was no overdispersion in the previous model, a negative binomial model was estimated for comparison. Similar to the Poisson regression model, two versions were estimated: one without interaction terms and one with them. The likelihood ratio test had a p-value lower than 0.05, allowing us to conclude once again that the model with interaction terms provides a better fit. Best negative binomial regression model formula:

$$
\begin{aligned}
\mu_i = &\exp(-183.759 - 0.661 \cdot I(X_{1i} = 2) - 0.196 \cdot I(X_{1i} = 3) + 0.090 \cdot X_{2i} \\
&+ 0.003 \cdot X_{3i} - 0.755 \cdot I(X_{4i} = \text{yes}) + 0.131 \cdot I(X_{5i} = 1) + 0.333 \cdot I(X_{5i} = 2) \\
&+ 0.637 \cdot I(X_{5i} = 3) - 0.063 \cdot I(X_{6i} = 1) - 0.281 \cdot I(X_{6i} = 2) - 0.877 \cdot I(X_{6i} = 3) \\
&+ 0.519 \cdot X_{7i} + 0.459 \cdot I(X_{8i} = 1) - 0.055 \cdot I(X_{8i} = 2) - 0.067 \cdot I(X_{8i} = 3) \\
&- 0.453 \cdot I(X_{9i} = 1) - 0.143 \cdot I(X_{9i} = 2) + 0.170 \cdot X_{10i} - 0.004 \cdot X_{11i} \\
&- 0.301 \cdot I(X_{12i} = 2) - 0.099 \cdot I(X_{12i} = 3) - 0.152 \cdot I(X_{12i} = 4) \\
&- 0.007 \cdot I(X_{12i} = 5) - 0.350 \cdot I(X_{12i} = 6) - 0.530 \cdot I(X_{12i} = 7) \\
&- 0.322 \cdot I(X_{12i} = 8) - 0.070 \cdot I(X_{12i} = 9) + 0.097 \cdot I(X_{12i} = 10) \\
&+ 0.253 \cdot I(X_{12i} = 11) + 0.272 \cdot I(X_{12i} = 12) - 0.002 \cdot (X_{10i} \cdot X_{11i}) \\
&+ 0.004 \cdot (X_{11i} \cdot I(X_{1i} = 2)) - 0.008 \cdot (X_{11i} \cdot I(X_{1i} = 3)) + ln(\text{exposure}_i))
\end{aligned}
$$

Using 5-fold cross-validation, the RMSE, MAE, and log-likelihood were estimated. The AIC and BIC were also calculated for both models.

Table 1: Poisson and NB regression model comparison

Metric	**Poisson**	**negative binomial**
RMSE	0.08296	0.08304
MAE	0.01273	0.01274
Log-likelihood	-2715.03	-2700.44
AIC	27145.41	27006.22
BIC	27513.77	27385.41

Both RMSE and MAE are slightly higher for the negative binomial regression model. However, the difference is so small that it is practically negligible. On the other hand, log-likelihood is higher, while AIC and BIC are both smaller for the negative binomial regression model, with a notably larger difference (Table 1).

4.2 ZIP and ZINB regression models

The zero-inflated model was estimated in two ways. In the first model, only variables with statistically significant effects were retained in the count process, while the zero-inflation process included only an intercept. In the second model, the zero-inflation process was also estimated using variables with statistically significant effects. The likelihood ratio test comparing the two models had a p-value of 0.8714, indicating no statistically significant difference between the two models. Interaction terms were added to the model where the zero-inflation process included only an intercept, and 5-fold cross-validation was performed to see if the model with interaction terms performs better.

Table 2: Comparison of the ZIP models

Metric	**ZIP model**	**ZIP model with interaction terms**
RMSE	0.08298	0.08297
MAE	0.01272	0.01272
Log-likelihood	-2704.98	-2703.36
AIC	27047.76	27028.19
BIC	27383.62	27396.55

The ZIP model with interaction terms has slightly lower RMSE and higher log-likelihood than the simpler model. It also has a lower AIC but a higher BIC, which is expected because the BIC penalizes additional parameters more heavily (Table 2). The likelihood ratio test comparing the ZIP model with its version including interaction terms had a p-value below 0.05. Therefore, it can be concluded that including interaction terms improves model performance. Variable X_{11} doesn't have a statistically significant effect, but the model with it performed better, thus it isn't removed.

Final ZIP count model:

$$\begin{aligned}
\mu_i = &\exp(0.543 - 0.659 \cdot I(X_{1i} = 2) + 0.206 \cdot I(X_{1i} = 3) \\
&+ 0.002 \cdot X_{3i} - 0.763 \cdot I(X_{4i} = \text{yes}) + 0.122 \cdot I(X_{5i} = 1) + 0.311 \cdot I(X_{5i} = 2) \\
&+ 0.618 \cdot I(X_{5i} = 3) - 0.054 \cdot I(X_{6i} = 1) - 0.262 \cdot I(X_{6i} = 2) - 0.850 \cdot I(X_{6i} = 3) \\
&+ 0.456 \cdot X_{7i} + 0.443 \cdot I(X_{8i} = 1) + 0.053 \cdot I(X_{8i} = 2) - 0.064 \cdot I(X_{8i} = 3) \\
&- 0.457 \cdot I(X_{9i} = 1) - 0.194 \cdot I(X_{9i} = 2) + 0.166 \cdot X_{10i} - 0.004 \cdot X_{11i} \\
&- 0.313 \cdot I(X_{12i} = 2) - 0.105 \cdot I(X_{12i} = 3) - 0.157 \cdot I(X_{12i} = 4) \\
&- 0.014 \cdot I(X_{12i} = 5) - 0.360 \cdot I(X_{12i} = 6) - 0.544 \cdot I(X_{12i} = 7) \\
&- 0.336 \cdot I(X_{12i} = 8) - 0.154 \cdot I(X_{12i} = 9) + 0.008 \cdot I(X_{12i} = 10) \\
&+ 0.155 \cdot I(X_{12i} = 11) + 0.139 \cdot I(X_{12i} = 12) - 0.002 \cdot (X_{10i} \cdot X_{11i}) \\
&+ 0.004 \cdot (X_{11i} \cdot I(X_{1i} = 2)) - 0.008 \cdot (X_{11i} \cdot I(X_{1i} = 3)) + ln(\text{exposure}_i)),
\end{aligned}$$

ZIP zero-inflation model:

$$\hat{\pi}_i = \frac{\exp(1.150)}{1 + \exp(1.150)}.$$

As before, the ZIP model didn't exhibit any overdispersion, but still, a ZINB model is estimated to see if it performs better. Similar comparisons were made for ZINB models. According to the likelihood ratio test, the ZINB models with variables in the zero-inflation process performed better. 5-fold cross-validation was performed, and the AIC and BIC values were calculated for the ZINB models with interaction terms.

Table 3: Comparison of ZINB models with interaction terms

Metric	**ZINB model**	**ZINB model with interaction terms**
RMSE	0.08310	0.08302
MAE	0.01274	0.01272
Log-likelihood	-2701.54	-2678.26
AIC	27017.69	26762.1
BIC	27396.89	27347.15

According to the 5-fold cross-validation results and the estimated AIC and BIC values (Table 3), the ZINB model with interaction terms performs better. The likelihood ratio test also produced a p-value below 0.05, thus indicating that the ZINB model with added interaction terms and significant variables in the zero-inflation process is preferable. The ZINB model had the same MAE as the ZIP model and even though the RMSE was better for the ZIP model, the log-likelihood, AIC, and BIC values are better for the ZINB model. Although variables X_3 and X_{11} weren't statistically significant, they weren't removed because the model had better 5-fold cross-validation results and AIC values with them.

Final ZINB count model:

$$
\begin{aligned}
\mu_i = &\exp(-0.793 - 0.705 \cdot I(X_{1i} = 2) + 0.157 \cdot I(X_{1i} = 3) \\
&- 0.002 \cdot X_{3i} - 0.651 \cdot I(X_{4i} = \text{yes}) + 0.146 \cdot I(X_{5i} = 1) + 0.225 \cdot I(X_{5i} = 2) \\
&+ 0.483 \cdot I(X_{5i} = 3) - 0.069 \cdot I(X_{6i} = 1) - 0.247 \cdot I(X_{6i} = 2) - 0.781 \cdot I(X_{6i} = 3) \\
&+ 0.642 \cdot X_{7i} + 0.418 \cdot I(X_{8i} = 1) + 0.051 \cdot I(X_{8i} = 2) - 0.085 \cdot I(X_{8i} = 3) \\
&- 0.404 \cdot I(X_{9i} = 1) - 0.143 \cdot I(X_{9i} = 2) + 0.188 \cdot X_{10i} - 0.003 \cdot X_{11i} \\
&- 0.300 \cdot I(X_{12i} = 2) - 0.025 \cdot I(X_{12i} = 3) - 0.149 \cdot I(X_{12i} = 4) \\
&+ 0.120 \cdot I(X_{12i} = 5) - 0.373 \cdot I(X_{12i} = 6) - 0.585 \cdot I(X_{12i} = 7) \\
&- 0.292 \cdot I(X_{12i} = 8) - 0.130 \cdot I(X_{12i} = 9) + 0.137 \cdot I(X_{12i} = 10) \\
&+ 0.378 \cdot I(X_{12i} = 11) + 0.264 \cdot I(X_{12i} = 12) - 0.002 \cdot (X_{10i} \cdot X_{11i}) \\
&+ 0.004 \cdot (X_{11i} \cdot I(X_{1i} = 2)) - 0.007 \cdot (X_{11i} \cdot I(X_{1i} = 3)) + ln(\text{exposure}_i))
\end{aligned}
$$

The ZINB zero-inflation model:

$$\hat{\pi}_i = \frac{W_i}{1 + W_i},$$

where:

$$
\begin{aligned}
W_i = \exp(&9.696 - 0.933 \cdot I(X_{1i} = 2) - 0.421 \cdot I(X_{1i} = 3) - 1.153 \cdot X_{3i} + 1.208 \cdot X_{7i} \\
&+ 0.118 \cdot X_{10i} + 0.016 \cdot X_{11i} + 0.027 \cdot I(X_{12i} = 2) + 0.827 \cdot I(X_{12i} = 3) \\
&- 0.032 \cdot I(X_{12i} = 4) + 1.345 \cdot I(X_{12i} = 5) - 0.271 \cdot I(X_{12i} = 6) \\
&- 0.650 \cdot I(X_{12i} = 7) + 0.538 \cdot I(X_{12i} = 8) + 0.489 \cdot I(X_{12i} = 9) \\
&+ 1.558 \cdot I(X_{12i} = 10) + 1.840 \cdot I(X_{12i} = 11) + 0.663 \cdot I(X_{12i} = 12) \\
&- 1.1752 \cdot X_{13i} - 0.687 \cdot I(X_{14i} = 1)) + ln(\text{exposure}_i))
\end{aligned}
$$

4.3 ZAP and ZANB regression models

Since zero-altered models handle all zeros separately, one of the estimated ZAP models used only an intercept in the count process, while the zero process was modeled using variables with a statistically significant effect. After performing a likelihood ratio test and comparing AIC and BIC values, the alternative model, where the count process was also modeled using variables, proved to perform better. Additionally, a comparison was made to evaluate whether adding interaction terms improves model performance. 5-fold cross-validation was performed, and AIC and BIC values were estimated.

Table 4: Comparison of ZAP models

Metric	**ZAP model**	**ZAP model with interaction terms**
RMSE	0.08294	0.08293
MAE	0.01272	0.01271
Log-likelihood	-2704.11	-2702.75
AIC	27044.81	27025.31
BIC	27391.51	27404.51

The likelihood ratio test shows that the ZAP model with interaction terms performs better. The 5-fold cross-validation results are quite similar between the two models, although BIC is lower for the ZAP model without interaction terms, all of the other metrics are better for the other model (Table 4).
Final ZAP count model:

$$\mu_i = \exp(0.720 - 0.006 \cdot X_{3i} + 0.331 \cdot X_{7i} + ln(\text{exposure}_i))$$

The ZAP zero-inflation model:

$$\hat{\pi}_i = \frac{Z_i}{1 + Z_i},$$

where

$$
\begin{aligned}
Z_i = \exp(\, & -0.858 - 0.662 \cdot I(X_{1i} = 2) + 0.185 \cdot I(X_{1i} = 3) - 0.645 \cdot I(X_{4i} = \text{yes}) \\
& + 0.108 \cdot I(X_{5i} = 1) + 0.306 \cdot I(X_{5i} = 2) + 0.604 \cdot I(X_{5i} = 3) \\
& - 0.080 \cdot I(X_{6i} = 1) - 0.285 \cdot I(X_{6i} = 2) - 0.853 \cdot I(X_{6i} = 3) \\
& + 0.506 \cdot X_{7i} + 0.465 \cdot I(X_{8i} = 1) + 0.055 \cdot I(X_{8i} = 2) - 0.100 \cdot I(X_{8i} = 3) \\
& - 0.438 \cdot I(X_{9i} = 1) - 0.136 \cdot I(X_{9i} = 2) + 0.164 \cdot X_{10i} - 0.004 \cdot X_{11i} \\
& - 0.291 \cdot I(X_{12i} = 2) - 0.095 \cdot I(X_{12i} = 3) - 0.145 \cdot I(X_{12i} = 4) \\
& - 0.035 \cdot I(X_{12i} = 5) - 0.344 \cdot I(X_{12i} = 6) - 0.544 \cdot I(X_{12i} = 7) \\
& - 0.347 \cdot I(X_{12i} = 8) - 0.162 \cdot I(X_{12i} = 9) - 0.009 \cdot I(X_{12i} = 10) \\
& + 0.127 \cdot I(X_{12i} = 11) + 0.200 \cdot I(X_{12i} = 12) - 0.002 \cdot (X_{10i} \cdot X_{11i}) \\
& + 0.004 \cdot (X_{11i} \cdot I(X_{1i} = 2)) - 0.007 \cdot (X_{11i} \cdot I(X_{1i} = 3)) + ln(\text{exposure}_i)).
\end{aligned}
$$

A ZANB model is also fitted to see if it fits the data better. Analogous comparisons were conducted for the ZANB models. Similarly to other models, the ZANB model with interaction terms and variables in the count process was better according to the likelihood ratio test. ZANB model formula:

Final ZANB model:

$$\mu_i = \exp(-0.681 - 0.006 \cdot X_{3i} + 0.376 \cdot X_{7i} + ln(\text{exposure}_i))$$

$$\hat{\pi}_i = \frac{Z_i}{1 + Z_i},$$

where

$$
\begin{aligned}
Z_i = \exp(\, & -0.858 - 0.662 \cdot I(X_{1i} = 2) + 0.185 \cdot I(X_{1i} = 3) - 0.645 \cdot I(X_{4i} = \text{yes}) \\
& + 0.108 \cdot I(X_{5i} = 1) + 0.306 \cdot I(X_{5i} = 2) + 0.604 \cdot I(X_{5i} = 3) \\
& - 0.080 \cdot I(X_{6i} = 1) - 0.285 \cdot I(X_{6,i} = 2) - 0.853 \cdot I(X_{6,i} = 3) \\
& + 0.506 \cdot X_{7i} + 0.465 \cdot I(X_{8i} = 1) + 0.055 \cdot I(X_{8i} = 2) - 0.100 \cdot I(X_{8i} = 3) \\
& - 0.438 \cdot I(X_{9i} = 1) - 0.136 \cdot I(X_{9i} = 2) + 0.164 \cdot X_{10i} - 0.004 \cdot X_{11i} \\
& - 0.291 \cdot I(X_{12,i} = 2) - 0.095 \cdot I(X_{12,i} = 3) - 0.145 \cdot I(X_{12,i} = 4) \\
& - 0.035 \cdot I(X_{12,i} = 5) - 0.344 \cdot I(X_{12,i} = 6) - 0.544 \cdot I(X_{12,i} = 7) \\
& - 0.347 \cdot I(X_{12i} = 8) - 0.163 \cdot I(X_{12i} = 9) - 0.010 \cdot I(X_{12i} = 10) \\
& + 0.127 \cdot I(X_{12i} = 11) + 0.197 \cdot I(X_{12i} = 12) - 0.002 \cdot (X_{10i} \cdot X_{11i}) \\
& + 0.004 \cdot (X_{11i} \cdot I(X_{1i} = 2)) - 0.007 \cdot (X_{11i} \cdot I(X_{1i} = 3) + ln(\text{exposure}_i)).
\end{aligned}
$$

This model had an AIC of 27,025.14 and a BIC of 27,415.17. The 5-fold cross-validation yielded a RMSE of 0.08308, a MAE of 0.01272, and a log-likelihood value of -2705.28. Surprisingly, the ZANB model performed worse than the ZAP model according to 5-fold cross-validation and AIC results.

4.4 GAM

For GAMs, the effect of numerical variables on the value of the linear predictor are described by smooth functions. Since the numeric variables have skewed distributions, where there are a few extreme values, a natural logarithm was taken from these variables. This helps with the interpretability of the models.
First, a GAM model with a log link and Poisson distribution was estimated. As with the other models, the initial model didn't include interaction terms, and another model with interaction terms was estimated. As expected, according to the likelihood ratio and AIC, adding interaction terms results in a better-performing

model. Since variable X_2 didn't have enough different numeric values, it will not be estimated as a smoothing function.
GAM model with Poisson distribution:

$$\begin{aligned}\mu_i = \exp(&-200.610 - 0.506 \cdot I(X_{1i} = 2) - 0.049 \cdot I(X_{1i} = 3) + 0.098 \cdot X_{2i} \\ &- 0.483 \cdot I(X_{4i} = \text{yes}) + 0.138 \cdot I(X_{5i} = 1) + 0.266 \cdot I(X_{5i} = 2) + 0.577 \cdot I(X_{5i} = 3) \\ &- 0.048 \cdot I(X_{6i} = 1) - 0.197 \cdot I(X_{6i} = 2) - 0.699 \cdot I(X_{6i} = 3) + 0.425 \cdot I(X_{8i} = 1) \\ &+ 0.012 \cdot I(X_{8i} = 2) - 0.261 \cdot I(X_{8i} = 3) - 0.393 \cdot I(X_{9i} = 1) + 0.050 \cdot I(X_{9i} = 2) \\ &- 0.246 \cdot I(X_{12i} = 2) - 0.070 \cdot I(X_{12i} = 3) - 0.145 \cdot I(X_{12i} = 4) - 0.057 \cdot I(X_{12i} = 5) \\ &- 0.325 \cdot I(X_{12i} = 6) - 0.486 \cdot I(X_{12i} = 7) - 0.265 \cdot I(X_{12i} = 8) - 0.102 \cdot I(X_{12i} = 9) \\ &+ 0.084 \cdot I(X_{12i} = 10) + 0.292 \cdot I(X_{12i} = 11) + 0.349 \cdot I(X_{12i} = 12) + 0.281 \cdot I(X_{14i} = 1) \\ &+ f_1(\ln(X_{7i})) + f_2(\ln(X_{10i}), \ln(X_{11i})) + f_3(\ln(X_{3i})|I(X_{13i} = \text{E})) \\ &+ f_4(\ln(X_{3i})|I(X_{13i} = \text{W})) + \ln(\text{exposure}_i))\end{aligned}$$

To understand the model behavior, it is useful to graph different smoothing terms, for example $f_3(\ln(X_{3i})|I(X_{13i} = \text{E}))$ in the GAM with Poisson distribution (graph 1). This is a smooth effect of variable $\ln(X_3)$, where smoothing curves were created for each separate level of variable X_{13}. The y-axis represents the logarithm of the expected count for the response variable. The line on the graph shows how the logarithm of the expected count changes as $\ln(X_3)$ varies, while $X_{13} = \text{E}$. As seen in the graph, when $\ln(X_3)$ has values between 2 and 5, the logarithm of the expected count is the highest, and it decreases when $\ln(X_3)$ has values between 0 and 1 or when it is larger than 5. The dotted lines represent the 95% confidence intervals, and from the graph it is possible to see that the smoothing term is the most precise near $\ln(X_3) = 2$.

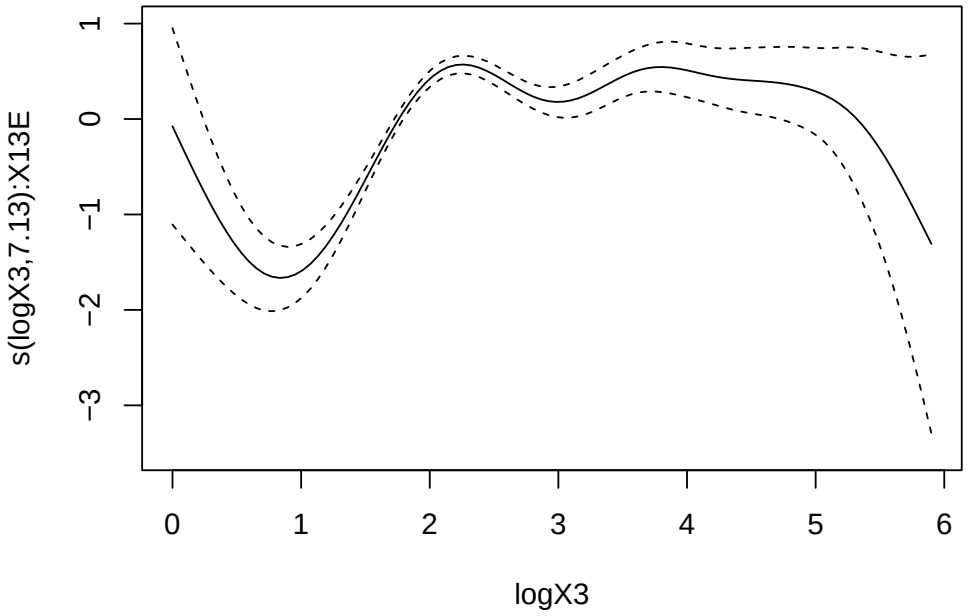

Figure 1: Estimated effect of $\ln(X_3)$ on the mean number of claims when $X_{13} = \text{E}$.

Even though the GAM with Poisson distribution didn't exhibit overdispersion, a GAM with a negative binomial distribution was fitted to see if it performs better. Final model formula:

$$
\begin{aligned}
\mu_i = \exp(&-198.163 - 0.507 \cdot I(X_{1i} = 2) - 0.051 \cdot I(X_{1i} = 3) + 0.097 \cdot X_{2i} \\
&- 0.512 I(X_{4i} = \text{yes}) + 0.144 \cdot I(X_{5i} = 1) + 0.263 \cdot I(X_{5i} = 2) + 0.578 \cdot I(X_{5i} = 3) \\
&- 0.047 \cdot I(X_{6i} = 1) - 0.195 \cdot I(X_{6i} = 2) - 0.713 \cdot I(X_{6i} = 3) + 0.408 \cdot I(X_{8i} = 1) \\
&+ 0.019 \cdot I(X_{8i} = 2) - 0.237 \cdot I(X_{8i} = 3) - 0.408 \cdot I(X_{9i} = 1) + 0.011 \cdot I(X_{9i} = 2) \\
&- 0.267 \cdot I(X_{12i} = 2) - 0.089 \cdot I(X_{12i} = 3) - 0.163 \cdot I(X_{12i} = 4) - 0.074 \cdot I(X_{12i} = 5) \\
&- 0.345 \cdot I(X_{12i} = 6) - 0.509 \cdot I(X_{12i} = 7) - 0.295 \cdot I(X_{12i} = 8) - 0.121 \cdot I(X_{12i} = 9) \\
&+ 0.065 \cdot I(X_{12i} = 10) + 0.281 \cdot I(X_{12i} = 11) + 0.330 \cdot I(X_{12i} = 12) + 0.292 \cdot I(X_{14i} = 1) \\
&+ f_1(\ln(X_{7i})) + f_2(\ln(X_{10i}), \ln(X_{11i})) + f_3(\ln(X_{3i}) | I(X_{13i} = \text{E})) \\
&+ f_4(\ln(X_{3i}) | I(X_{13i} = \text{W})) + \ln(\text{exposure}_i))
\end{aligned}
$$

Another example of the smoothing function $f_3(\ln(X_{10i}), \ln(X_{11i}))$ is from GAM with a negative binomial regression, where the smoothing effect has been estimated from an interaction term. Since smoothing effects plots from interaction terms can be complicated and hard to interpret, the data was split into three parts based on $\ln(X_{10})$ values. On the graph $\ln(X_{10}) = 1.5$, which is the mean of the variable values in between $(1.18, 2.35]$. Here, it can be seen how the value of $\ln(X_{11})$ affects the average number of claims when $\ln(X_{10}) = 1.5$. The points on the graph correspond to the values of $\ln(X_{11})$, which are in the data set together with $\ln(X_{10})$ values in the interval $(1.18, 2.35]$ (graph 2). When $\ln(X_{11})$ values increase, the expected count decreases, although the expected count increases when $\ln(X_{11})$ values are near 3 and over 4.

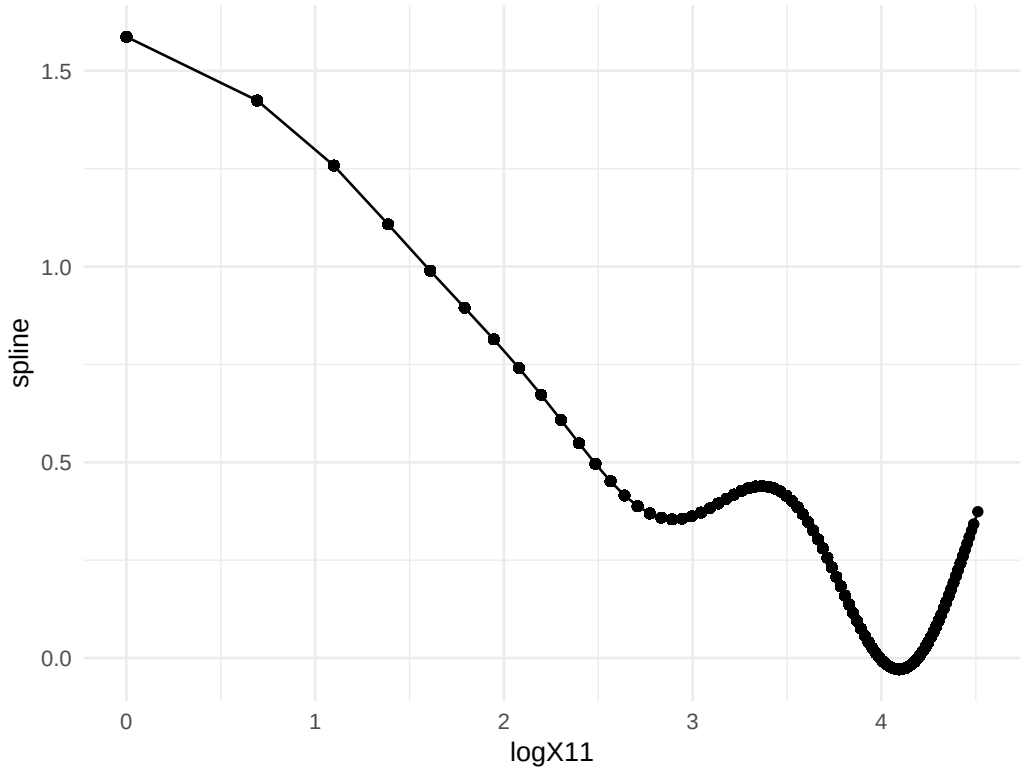

Figure 2: Estimated effect of $\ln(X_{11}$ on the mean number of claims when $ln(X_{10}) = 1.5$.

5-fold cross-validation was performed for GAM with Poisson and negative binomial distributions. Also, as before, AIC and BIC values were estimated for the models. While MAE is slightly lower for the model with Poisson distribution and RMSE has the same value for both of models, all of the other metrics have better values for the model with negative binomial distribution (Table 5).

Table 5: Comparison of GAMs

Metric	GAM Poisson	GAM NB
RMSE	0.08279	0.08279
MAE	0.01268	0.01269
Log-likelihood	-2694.55	-2651.84
AIC	26627.43	26522.15
BIC	27294.99	27190.75

4.5 Lasso regularization

Lasso regularization was applied to a Poisson regression using `tidymodels` package. Unlike other models estimated before, it performs 5-fold cross-validation and estimates the model at the same time. The 5-fold cross-validation results for each penalty value can be seen in Table 6. As the penalty decreases, RMSE, MAE, and log-likelihood values for the model get better. The lowest log-likelihood value is when the penalty is 0, and in that case, RMSE and MAE remain the same as for the penalty of 0.0002.

Table 6: Metric estimates with penalties

Metric Penalty	0.0002	0.0004	0.0006	0.0008	0.001
RMSE	0.0829	0.0829	0.0829	0.0829	0.0830
MAE	0.0127	0.0127	0.0128	0.0128	0.0128
Log-likelihood	-2722.99	-2731.97	-2740.27	-2748.31	-2757.24

From this, it is possible to conclude that in the case of the current data set and the form of the model, there is no benefit from Lasso regularization, and the unpenalized Poisson regression fit is optimal.

4.6 Results

For all models, AIC and BIC values were calculated, after which a 5-fold cross-validation was conducted. While RMSE and MAE yielded quite similar results for most of the models, log-likelihood, AIC, and BIC values had notably larger differences in some cases.

Table 7: Results

Model	RMSE	MAE	Log-likelihood	AIC	BIC
Poisson	0.08296	0.01273	-2715.03	27145.41	27513.77
NB	0.08304	0.01274	-2700.44	27006.22	27385.41
ZIP	0.08297	0.01272	-2703.36	27028.19	27396.55
ZINB	0.08302	0.01272	-2678.26	26762.10	27347.15
ZAP	0.08293	0.01271	-2702.75	27025.31	27404.51
ZANB	0.08308	0.01272	-2705.28	27025.14	27415.17
GAM Poisson	0.08279	0.01268	-2694.55	26627.43	27294.99
GAM NB	0.08279	0.01269	-2651.84	26522.15	27190.75

The ZANB model had the highest RMSE value, while the negative binomial model had the highest MAE value. Surprisingly, the Poisson regression had the poorest log-likelihood, AIC, and BIC values. Overall, the models with the Poisson distribution yielded better RMSE and MAE values, whereas the models with the negative binomial distribution yielded better log-likelihood, AIC, and BIC values with the exception of ZA models.

Although the overall differences in RMSE and MAE are small, the results varied considerably more across individual folds. This indicates that, for the given dataset, performing 5-fold cross-validation was valuable. As can be seen from the table, the GAMs outperformed all of the other models. Even though MAE is better for the GAM with Poisson distribution, log-likelihood, AIC, and BIC values are better for

the GAM with negative binomial distribution (Table 7).

Summary

The aim of this thesis was to estimate the frequency of travel insurance claims using generalized linear models and compare them with different metrics. First, the models were estimated by removing the variables with statistically insignificant effects one by one, while AIC was taken into account. After the models were estimated, the AIC and BIC were calculated as the first metrics for model comparison. Except for ZA models, both of these metrics seemed to favor models with a negative binomial distribution.

With 5-fold cross-validation, additional metrics were calculated, and an average of them over 5 folds was taken. As seen from the results, the metrics RMSE and MAE did not differ much between the models, but generally, the models with a Poisson distribution proved to have better RMSE and MAE results. In addition to these metrics, the log-likelihood was calculated using 5-fold cross-validation. In contrast to RMSE and MAE results, log-likelihood estimates were better for the models with negative binomial distribution, and the differences were also bigger.

In the end, it appears that generalized additive models performed better compared to other models according to every metric. As seen from the graphs, GAMs were able to model the non-linear relationships that other GLMs weren't able to capture, making them more precise. As the differences in accuracy are so small, the final conclusion is that the generalized additive model with negative binomial distribution is the most suitable for this data set.

Bibliography

Actuaries, American Academy of (2018). *Travel Insurance Monograph*. URL: https://www.actuary.org/sites/default/files/files/publications/TravelInsuranceMonograph_09052018.pdf (visited on 04/27/2025).

Burnham, Kenneth P. and David R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd. New York: Springer.

Cameron, A. Colin and Pravin K. Trivedi (2013). *Regression Analysis of Count Data*. 2nd. New York: Cambridge University Press.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. 2nd. Duxbury.

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani (2025). *An Introduction to glmnet*. Version 4.1-9. URL: https://glmnet.stanford.edu/articles/glmnet.html (visited on 05/16/2025).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York, NY: Springer. ISBN: 978-0-387-84858-7.

Heaphy, Matt (Mar. 2, 2025). *offsetreg: An Extension of 'Tidymodels' Supporting Offset Terms*. Version 1.1.1. URL: https://cran.r-project.org/web/packages/offsetreg/index.html (visited on 04/29/2025).

Hilbe, Joseph M. (2011). *Negative Binomial Regression*. 2nd. New York: Cambridge University Press.

Jackman, Simon, Alex Tahk, Achim Zeileis, Christina Maimone, James Fearon, and Zoe Meers (Jan. 31, 2024). *pscl: Political Science Computational Laboratory*. Version 1.5.9. URL: https://cran.r-project.org/web/packages/pscl/index.html (visited on 04/29/2025).

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. 1st ed. Springer Texts in Statistics. New York: Springer.

– (2021). *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. Springer Texts in Statistics. Springer. ISBN: 978-1-0716-1418-1.

Kuhn, Max and Hadley Wickham (Feb. 21, 2025). *tidymodels: Easily Install and Load the 'Tidymodels' Packages*. Version 1.3.0. URL: https://cran.r-project.org/web/packages/tidymodels/index.html (visited on 04/29/2025).

Käärik, Meelis (2023). *Generalized Linear Models. Lecture 3. Hypothesis testing*. URL: https://courses.ms.ut.ee/MTMS.01.011/2024_spring/uploads/Main/GLM_slides_3_hypothesis_gof.pdf (visited on 05/20/2025).

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. 2nd. London: Chapman & Hall/CRC.

Ohlsson, Esbjörn and Björn Johansson (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin, Heidelberg: Springer.

Venables, William N. and Brian D. Ripley (2002). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. URL: https://cran.r-project.org/web/packages/MASS/index.html (visited on 04/29/2025).

Willmott, C.J. and K. Matsuura (Dec. 2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". In: *Climate Research* 30.1, pp. 79–82. DOI: 10.3354/cr030079. URL: https://www.int-res.com/abstracts/cr/v30/n1/p79-82/.

Wood, Simon N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. CRC Press.

Wood, Simon N. (Apr. 4, 2025). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. Version 1.9-3. URL: https://cran.r-project.org/web/packages/mgcv/index.html (visited on 04/29/2025).

Appendix 1. R model outputs

Poisson regression

```
Call:
glm(formula = nrclaims ~ offset(log(expyrs)) + X1 + X2 + X3 +
    X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X11 * X10 +
    X1 * X11, family = poisson(link = "log"), data = data)
Coefficients:
               Estimate   Std. Error z value             Pr(>|z|)
(Intercept) -188.3396052   52.5234587  -3.586             0.000336 ***
X12           -0.6609054    0.1390603  -4.753  0.00000200763516854 ***
X13            0.2141066    0.0906515   2.362             0.018183 *
X2             0.0926633    0.0259612   3.569             0.000358 ***
X3             0.0011759    0.0007565   1.554             0.120077
X4yes         -0.6946504    0.1974707  -3.518             0.000435 ***
X51            0.1192557    0.0621052   1.920             0.054830 .
X52            0.3304623    0.0736058   4.490  0.00000713490823607 ***
X53            0.6397066    0.1083342   5.905  0.00000000352782574 ***
X61           -0.0730659    0.0588549  -1.241             0.214436
X62           -0.2861855    0.0862056  -3.320             0.000901 ***
X63           -0.8549019    0.1966300  -4.348  0.00001375298369272 ***
X7             0.4882723    0.0337144  14.483 < 0.0000000000000002 ***
X81            0.4862646    0.1121108   4.337  0.00001442065049519 ***
X82            0.0640137    0.0680482   0.941             0.346853
X83           -0.0745873    0.1583627  -0.471             0.637648
X91           -0.4430785    0.0567371  -7.809  0.00000000000000575 ***
X92           -0.1217404    0.1914129  -0.636             0.524770
X10            0.1566881    0.0204887   7.648  0.00000000000002048 ***
```

```
X11          -0.0041445    0.0021507  -1.927             0.053974 .
X122         -0.2918305    0.1091082  -2.675             0.007480 **
X123         -0.0875963    0.1035951  -0.846             0.397796
X124         -0.1457257    0.1010976  -1.441             0.149462
X125         -0.0008935    0.0973966  -0.009             0.992681
X126         -0.3417009    0.0999586  -3.418             0.000630 ***
X127         -0.5200970    0.1029811  -5.050  0.00000044085899384 ***
X128         -0.3030341    0.1024425  -2.958             0.003096 **
X129         -0.0577551    0.1048847  -0.551             0.581872
X1210         0.0976102    0.1006209   0.970             0.332007
X1211         0.2631006    0.1092382   2.409             0.016018 *
X1212         0.2915935    0.1050708   2.775             0.005517 **
X10:X11      -0.0021608    0.0006631  -3.259             0.001120 **
X12:X11       0.0037519    0.0034337   1.093             0.274533
X13:X11      -0.0078319    0.0024209  -3.235             0.001216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 23061  on 374805  degrees of freedom
Residual deviance: 22358  on 374772  degrees of freedom
AIC: 27145
Number of Fisher Scoring iterations: 8
```

Negative Binomial regression

```
Call:glm.nb(formula = nrclaims ~ offset(log(expyrs)) + X1 + X2 + X3 +
    X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X11 * X10 +
    X1 * X11, data = data, link = log, init.theta = 0.2610105696)
```

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-183.7589406	54.1006631	-3.397	0.000682	***
X12	-0.6608241	0.1416784	-4.664	0.00000309736539326	***
X13	0.1959927	0.0936275	2.093	0.036320	*
X2	0.0903806	0.0267407	3.380	0.000725	***
X3	0.0025399	0.0008920	2.847	0.004406	**
X4yes	-0.7546563	0.2091015	-3.609	0.000307	***
X51	0.1313122	0.0639987	2.052	0.040190	*
X52	0.3329536	0.0760425	4.379	0.00001194880592898	***
X53	0.6367903	0.1128230	5.644	0.00000001659936011	***
X61	-0.0626762	0.0607306	-1.032	0.302055	
X62	-0.2810326	0.0887447	-3.167	0.001542	**
X63	-0.8773644	0.2036396	-4.308	0.00001644267231964	***
X7	0.5192161	0.0473676	10.961	< 0.0000000000000002	***
X81	0.4592396	0.1168048	3.932	0.00008435302254028	***
X82	0.0552011	0.0695035	0.794	0.427067	
X83	-0.0670980	0.1615784	-0.415	0.677947	
X91	-0.4529418	0.0585954	-7.730	0.00000000000001076	***
X92	-0.1433401	0.1960034	-0.731	0.464587	
X10	0.1695660	0.0215258	7.877	0.00000000000000334	***
X11	-0.0038308	0.0022282	-1.719	0.085570	.
X122	-0.3008964	0.1124139	-2.677	0.007435	**
X123	-0.0994220	0.1069185	-0.930	0.352430	
X124	-0.1519226	0.1040595	-1.460	0.144301	
X125	-0.0065493	0.1002429	-0.065	0.947908	
X126	-0.3503894	0.1029255	-3.404	0.000663	***
X127	-0.5297379	0.1056674	-5.013	0.00000053515604198	***
X128	-0.3219152	0.1055365	-3.050	0.002286	**
X129	-0.0695347	0.1081539	-0.643	0.520274	

```
X1210          0.0965575    0.1035975   0.932            0.351313
X1211          0.2530274    0.1133883   2.232            0.025647 *
X1212          0.2720650    0.1093004   2.489            0.012805 *
X10:X11       -0.0023360    0.0006834  -3.418            0.000630 ***
X12:X11        0.0036289    0.0035035   1.036            0.300293
X13:X11       -0.0072109    0.0024893  -2.897            0.003771 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Negative Binomial(0.261) family taken to be 1)
    Null deviance: 18226  on 374805  degrees of freedom
Residual deviance: 17559  on 374772  degrees of freedom
AIC: 27006
Number of Fisher Scoring iterations: 1
              Theta:  0.2610
          Std. Err.:  0.0361
2 x log-likelihood:  -26936.2150
```

ZIP

```
Call:zeroinfl(formula = nrclaims ~ offset(log(expyrs)) + X1 + X3 +
X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X11:X10 + X1:X11 | 1,
data = data, dist = "poisson")
Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.50564 -0.08557 -0.06762 -0.05202 48.36218
Count model coefficients (poisson with log link):
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)  0.5433071  0.1786558   3.041             0.002357 **
X12         -0.6591868  0.1408533  -4.680  0.00000286940413571 ***
```

```
X13          0.2055624  0.0923950   2.225             0.026093 *
X3           0.0018344  0.0010655   1.722             0.085131 .
X4yes       -0.7630682  0.2082414  -3.664             0.000248 ***
X51          0.1216871  0.0636970   1.910             0.056081 .
X52          0.3113173  0.0751180   4.144  0.00003407368217488 ***
X53          0.6181387  0.1112495   5.556  0.00000002755112143 ***
X61         -0.0541465  0.0603068  -0.898             0.369265
X62         -0.2619761  0.0879457  -2.979             0.002893 **
X63         -0.8495165  0.1990736  -4.267  0.00001978101562408 ***
X7           0.4560377  0.0421255  10.826 < 0.0000000000000002 ***
X81          0.4426145  0.1154083   3.835             0.000125 ***
X82          0.0526490  0.0693742   0.759             0.447905
X83         -0.0639360  0.1624045  -0.394             0.693814
X91         -0.4567416  0.0582504  -7.841  0.00000000000000447 ***
X92         -0.1940676  0.1960820  -0.990             0.322308
X10          0.1663530  0.0229062   7.262  0.00000000000038042 ***
X11         -0.0035838  0.0022242  -1.611             0.107129
X122        -0.3132636  0.1119402  -2.798             0.005134 **
X123        -0.1053703  0.1065168  -0.989             0.322547
X124        -0.1569779  0.1038254  -1.512             0.130549
X125        -0.0136135  0.0999881  -0.136             0.891702
X126        -0.3595489  0.1025936  -3.505             0.000457 ***
X127        -0.5441595  0.1053476  -5.165  0.00000023996444519 ***
X128        -0.3357011  0.1049414  -3.199             0.001379 **
X129        -0.1536786  0.1052256  -1.460             0.144162
X1210        0.0078719  0.1005599   0.078             0.937605
X1211        0.1550414  0.1098829   1.411             0.158254
X1212        0.1394734  0.1055069   1.322             0.186190
X10:X11     -0.0023596  0.0006816  -3.462             0.000536 ***
```

```
X12:X11      0.0035174  0.0034752   1.012             0.311465
X13:X11     -0.0073462  0.0024599  -2.986             0.002823 **

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)   1.1499     0.1344   8.558 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Number of iterations in BFGS optimization: 51
Log-likelihood: -1.348e+04 on 34 Df
```

ZINB

```
Call: zeroinfl(formula = nrclaims ~ offset(log(expyrs)) + X1 + X3 +
X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X11:X10 + X1:X11 |
offset(log(expyrs)) + X1 + X3 + X7 + X10 + X11 + X12 + X13 + X14,
data = data, dist = "negbin")
Pearson residuals:
      Min        1Q    Median        3Q       Max
 -0.49667  -0.08949  -0.06692  -0.03624 127.80698
Count model coefficients (negbin with log link):
              Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.7933112  0.1606534  -4.938 0.000000789160871473 ***
X12         -0.7046553  0.1453085  -4.849 0.000001238521167033 ***
X13          0.1569122  0.0986030   1.591             0.111530
X3          -0.0015032  0.0011036  -1.362             0.173155
X4yes       -0.6505019  0.2051440  -3.171             0.001519 **
X51          0.1458738  0.0643994   2.265             0.023504 *
X52          0.2250240  0.0761723   2.954             0.003135 **
```

```
X53          0.4828304  0.1125995   4.288 0.000018026217626892 ***
X61         -0.0690326  0.0610000  -1.132             0.257768
X62         -0.2473372  0.0888074  -2.785             0.005351 **
X63         -0.7808733  0.2015432  -3.874             0.000107 ***
X7           0.6419208  0.0602418  10.656 < 0.0000000000000002 ***
X81          0.4181682  0.1165688   3.587             0.000334 ***
X82          0.0508435  0.0696261   0.730             0.465246
X83         -0.0849165  0.1618304  -0.525             0.599774
X91         -0.4042484  0.0590833  -6.842 0.000000000007808931 ***
X92         -0.1435821  0.1959894  -0.733             0.463802
X10          0.1884268  0.0288787   6.525 0.000000000068108999 ***
X11         -0.0032638  0.0024222  -1.347             0.177837
X122        -0.2997160  0.1239237  -2.419             0.015582 *
X123        -0.0245466  0.1191943  -0.206             0.836839
X124        -0.1487206  0.1162751  -1.279             0.200882
X125         0.1203358  0.1168203   1.030             0.302967
X126        -0.3729268  0.1156464  -3.225             0.001261 **
X127        -0.5853344  0.1166495  -5.018 0.000000522424485750 ***
X128        -0.2923636  0.1195849  -2.445             0.014492 *
X129        -0.1303375  0.1169491  -1.114             0.265073
X1210        0.1370159  0.1177884   1.163             0.244733
X1211        0.3576796  0.1238780   2.887             0.003885 **
X1212        0.2641776  0.1155189   2.287             0.022203 *
X10:X11     -0.0022632  0.0007645  -2.960             0.003073 **
X12:X11      0.0037182  0.0034708   1.071             0.284035
X13:X11     -0.0073360  0.0024566  -2.986             0.002824 **
Log(theta)  -1.1820027  0.1467884  -8.052 0.000000000000000812 ***

Zero-inflation model coefficients (binomial with logit link):
```

```
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)  9.69640    0.74600  12.998 < 0.0000000000000002 ***
X12         -0.93284    0.46867  -1.990                0.0465 *
X13         -0.42158    0.37072  -1.137                0.2555
X3          -1.15333    0.08319 -13.864 < 0.0000000000000002 ***
X7           1.20826    0.25224   4.790           0.00000167 ***
X10          0.11777    0.06686   1.761                0.0782 .
X11          0.01628    0.00674   2.415                0.0157 *
X122         0.02732    0.86508   0.032                0.9748
X123         0.82734    0.77104   1.073                0.2833
X124        -0.03171    0.80939  -0.039                0.9687
X125         1.34499    0.74074   1.816                0.0694 .
X126        -0.27145    0.94826  -0.286                0.7747
X127        -0.65007    1.00168  -0.649                0.5164
X128         0.53770    0.84841   0.634                0.5262
X129         0.48884    0.83504   0.585                0.5583
X1210        1.55766    0.74020   2.104                0.0353 *
X1211        1.84002    0.72534   2.537                0.0112 *
X1212        0.66323    0.72672   0.913                0.3614
X13W        -1.75155    0.96971  -1.806                0.0709 .
X141        -0.68650    0.28046  -2.448                0.0144 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta = 0.3067
Number of iterations in BFGS optimization: 172
Log-likelihood: -1.333e+04 on 54 Df
```

ZAP

```
Call:hurdle(formula = nrclaims ~ offset(log(expyrs)) + X3 + X7 |
offset(log(expyrs)) + X1 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
X12 + X11:X10 + X1:X11, data = data, dist = "poisson")
Pearson residuals:
     Min       1Q   Median       3Q      Max
-1.06813 -0.08488 -0.06716 -0.05185 48.59907
Count model coefficients (truncated poisson with log link):
             Estimate Std. Error z value   Pr(>|z|)
(Intercept)  0.719531   0.138860   5.182 0.00000022 ***
X3          -0.006427   0.002715  -2.367     0.0179 *
X7           0.330907   0.082193   4.026 0.00005674 ***

Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.8583333  0.1435454  -5.980  0.00000000223787697 ***
X12         -0.6620067  0.1415683  -4.676  0.00000292191457854 ***
X13          0.1849171  0.0932164   1.984             0.047285 *
X4yes       -0.6445785  0.2003142  -3.218             0.001292 **
X51          0.1083334  0.0634363   1.708             0.087683 .
X52          0.3056336  0.0753421   4.057  0.00004979030147686 ***
X53          0.6044217  0.1117634   5.408  0.00000006371590779 ***
X61         -0.0803865  0.0600865  -1.338             0.180946
X62         -0.2846970  0.0879818  -3.236             0.001213 **
X63         -0.8530877  0.2009799  -4.245  0.00002189427582736 ***
X7           0.5059478  0.0459335  11.015 < 0.0000000000000002 ***
X81          0.4649722  0.1158373   4.014  0.00005969582326576 ***
X82          0.0551656  0.0692752   0.796             0.425843
```

```
X83         -0.1001635  0.1636530  -0.612             0.540506
X91         -0.4384435  0.0584749  -7.498  0.00000000000006481 ***
X92         -0.1363700  0.1968688  -0.693             0.488501
X10          0.1644718  0.0209864   7.837  0.00000000000000461 ***
X11         -0.0037316  0.0021789  -1.713             0.086775 .
X122        -0.2905336  0.1113884  -2.608             0.009100 **
X123        -0.0952217  0.1060312  -0.898             0.369157
X124        -0.1450130  0.1031574  -1.406             0.159800
X125        -0.0345996  0.0998895  -0.346             0.729058
X126        -0.3442643  0.1019865  -3.376             0.000737 ***
X127        -0.5436056  0.1051810  -5.168  0.00000023625097340 ***
X128        -0.3473171  0.1051817  -3.302             0.000960 ***
X129        -0.1620850  0.1052097  -1.541             0.123417
X1210       -0.0096500  0.1004097  -0.096             0.923436
X1211        0.1270944  0.1104658   1.151             0.249925
X1212        0.1996666  0.1045108   1.910             0.056070 .
X10:X11     -0.0023998  0.0006489  -3.698             0.000217 ***
X12:X11      0.0038499  0.0035040   1.099             0.271892
X13:X11     -0.0068895  0.0024858  -2.772             0.005580 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Number of iterations in BFGS optimization: 14
Log-likelihood: -1.348e+04 on 35 Df
```

ZANB

```
Call:hurdle(formula = nrclaims ~ offset(log(expyrs)) + X3 + X7 |
offset(log(expyrs)) + X1 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
X12 + X11:X10 + X1:X11, data = data, dist = "negbin")
```

```
Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.77018 -0.08486 -0.06715 -0.05185 48.59954
Count model coefficients (truncated negbin with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.681279   2.580415  -0.264  0.79177
X3          -0.005529   0.003231  -1.711  0.08705 .
X7           0.375688   0.117029   3.210  0.00133 **
Log(theta)  -1.114279   3.422379  -0.326  0.74474

Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.8583333  0.1435454  -5.980  0.00000000223787697 ***
X12         -0.6620067  0.1415683  -4.676  0.00000292191457854 ***
X13          0.1849171  0.0932164   1.984             0.047285 *
X4yes       -0.6445785  0.2003142  -3.218             0.001292 **
X51          0.1083334  0.0634363   1.708             0.087683 .
X52          0.3056336  0.0753421   4.057  0.00004979030147686 ***
X53          0.6044217  0.1117634   5.408  0.00000006371590779 ***
X61         -0.0803865  0.0600865  -1.338             0.180946
X62         -0.2846970  0.0879818  -3.236             0.001213 **
X63         -0.8530877  0.2009799  -4.245  0.00002189427582736 ***
X7           0.5059478  0.0459335  11.015 < 0.0000000000000002 ***
X81          0.4649722  0.1158373   4.014  0.00005969582326576 ***
X82          0.0551656  0.0692752   0.796             0.425843
X83         -0.1001635  0.1636530  -0.612             0.540506
X91         -0.4384435  0.0584749  -7.498  0.00000000000006481 ***
X92         -0.1363700  0.1968688  -0.693             0.488501
X10          0.1644718  0.0209864   7.837  0.00000000000000461 ***
```

```
X11          -0.0037316  0.0021789  -1.713             0.086775 .
X122         -0.2905336  0.1113884  -2.608             0.009100 **
X123         -0.0952217  0.1060312  -0.898             0.369157
X124         -0.1450130  0.1031574  -1.406             0.159800
X125         -0.0345996  0.0998895  -0.346             0.729058
X126         -0.3442643  0.1019865  -3.376             0.000737 ***
X127         -0.5436056  0.1051810  -5.168  0.00000023625097340 ***
X128         -0.3473171  0.1051817  -3.302             0.000960 ***
X129         -0.1620850  0.1052097  -1.541             0.123417
X1210        -0.0096500  0.1004097  -0.096             0.923436
X1211         0.1270944  0.1104658   1.151             0.249925
X1212         0.1996666  0.1045108   1.910             0.056070 .
X10:X11      -0.0023998  0.0006489  -3.698             0.000217 ***
X12:X11       0.0038499  0.0035040   1.099             0.271892
X13:X11      -0.0068895  0.0024858  -2.772             0.005580 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta: count = 0.3282
Number of iterations in BFGS optimization: 20
Log-likelihood: -1.348e+04 on 36 Df
```

GAM with Poisson distribution

```
Family: poisson
Link function: log
Formula:
nrclaims ~ offset(log(expyrs)) + X1 + X2 + X4 + X5 + X6 + s(logX7, k = 5) +
X8 + X9 + X12 + X14 + s(logX10, logX11) + s(logX3, by = X13)
Parametric coefficients:
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-200.61000	52.72920	-3.805	0.000142	***
X12	-0.50562	0.06964	-7.261	3.86e-13	***
X13	-0.04945	0.04881	-1.013	0.310986	
X2	0.09847	0.02606	3.778	0.000158	***
X4yes	-0.48309	0.19886	-2.429	0.015131	*
X51	0.13848	0.06299	2.198	0.027916	*
X52	0.26611	0.07501	3.548	0.000389	***
X53	0.57738	0.11089	5.207	1.92e-07	***
X61	-0.04806	0.05977	-0.804	0.421398	
X62	-0.19714	0.08748	-2.254	0.024221	*
X63	-0.69871	0.19761	-3.536	0.000406	***
X81	0.42502	0.11197	3.796	0.000147	***
X82	0.01164	0.06843	0.170	0.864971	
X83	-0.26141	0.15953	-1.639	0.101303	
X91	-0.39326	0.05733	-6.859	6.92e-12	***
X92	0.04971	0.19270	0.258	0.796418	
X122	-0.24597	0.10926	-2.251	0.024375	*
X123	-0.07024	0.10371	-0.677	0.498277	
X124	-0.14470	0.10166	-1.423	0.154632	
X125	-0.05740	0.09825	-0.584	0.559096	
X126	-0.32477	0.10070	-3.225	0.001260	**
X127	-0.48567	0.10389	-4.675	2.94e-06	***
X128	-0.26468	0.10336	-2.561	0.010444	*
X129	-0.10178	0.10554	-0.964	0.334887	
X1210	0.08445	0.10127	0.834	0.404325	
X1211	0.29231	0.10952	2.669	0.007607	**
X1212	0.34867	0.10545	3.306	0.000945	***
X141	0.28146	0.05651	4.981	6.32e-07	***

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Approximate significance of smooth terms:
                    edf Ref.df Chi.sq p-value
s(logX7)          1.930  2.304 149.82  <2e-16 ***
s(logX10,logX11) 20.593 23.475 492.37  <2e-16 ***
s(logX3):X13E     7.130  7.884 172.09  <2e-16 ***
s(logX3):X13W     3.963  4.894  40.87  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) =  0.0134   Deviance explained = 5.54%
UBRE = -0.94155  Scale est. = 1         n = 374806
```

GAM with Negative Binomial distribution

```
Family: Negative Binomial(0.314)
Link function: log
Formula:
nrclaims ~ offset(log(expyrs)) + X1 + X2 + X4 + X5 + X6 + s(logX7, k = 5) +
X8 + X9 + X12 + X14 + s(logX10, logX11) + s(logX3, by = X13)
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -198.16278   54.25560  -3.652 0.000260 ***
X12           -0.50661    0.07107  -7.128 1.02e-12 ***
X13           -0.05085    0.05026  -1.012 0.311685
X2             0.09727    0.02682   3.627 0.000286 ***
X4yes         -0.51164    0.20747  -2.466 0.013660 *
X51            0.14366    0.06482   2.216 0.026685 *
X52            0.26347    0.07730   3.408 0.000653 ***
```

```
X53            0.57777    0.11502   5.023 5.08e-07 ***
X61           -0.04690    0.06154  -0.762 0.446058
X62           -0.19532    0.08983  -2.174 0.029677 *
X63           -0.71338    0.20380  -3.500 0.000464 ***
X81            0.40773    0.11640   3.503 0.000460 ***
X82            0.01852    0.06988   0.265 0.790981
X83           -0.23680    0.16294  -1.453 0.146139
X91           -0.40841    0.05918  -6.902 5.14e-12 ***
X92            0.01122    0.19754   0.057 0.954710
X122          -0.26700    0.11261  -2.371 0.017743 *
X123          -0.08932    0.10711  -0.834 0.404333
X124          -0.16314    0.10468  -1.559 0.119099
X125          -0.07380    0.10112  -0.730 0.465500
X126          -0.34486    0.10359  -3.329 0.000872 ***
X127          -0.50899    0.10650  -4.779 1.76e-06 ***
X128          -0.29503    0.10637  -2.774 0.005543 **
X129          -0.12115    0.10868  -1.115 0.264954
X1210          0.06519    0.10430   0.625 0.531955
X1211          0.28066    0.11351   2.473 0.013415 *
X1212          0.33035    0.10965   3.013 0.002590 **
X141           0.29173    0.05756   5.068 4.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Approximate significance of smooth terms:
                   edf Ref.df Chi.sq p-value
s(logX7)         1.906  2.271 107.54 < 2e-16 ***
s(logX10,logX11) 17.328 20.554 445.04 < 2e-16 ***
s(logX3):X13E    7.094  7.843 169.79 < 2e-16 ***
s(logX3):X13W    3.564  4.433  38.24 6.2e-07 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0128   Deviance explained = 6.47%

-REML =  13309  Scale est. = 1         n = 374806
```

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

I, Minni-Marii Paarmets,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis, Claims frequency modeling and model comparison for travel insurance supervised by Raul Kangro.

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

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

Minni-Marii Paarmets
21.05.2025