# PREDICTIVE MODELLING OF AUTO INSURANCE CLAIMS

Ernest Tamekloe (eetamek@ilstu.edu)
Kelvin Nii Lartey-Abrahams (knlarte@ilstu.edu)

04/01/2021

# Introduction

- Premium pricing is an important task in insurance which involves finding the fair premium that covers an insurer's expected costs and expenses while providing a fair return to the insurer's investors.
- Frequency and severity of insurance claims play a major role in the pricing of the premiums.
- Frequency is the average number of claims per period. Severity is the amount paid due to a loss.
- Adequately modelling past and current data on claim experience can help insurers settle claims from existing or future portfolios.

# Objectives

The objectives of the study are as follows:

- To develop separate models for frequency and severity of claims data (Frequency-Severity method).
- develop a single model for pure premium (i.e., average claim cost involving frequency and severity).
- Perform a comparative analysis of the best Frequency-Severity method and the best Pure Premium method.
- evaluate the overall best method and make recommendations for actuaries.

# Data

- The data used for this project is taken from the Third Actuarial Pricing Game.
- There are two datasets each including 100,000 insureds for Year 0:

1. An underwriting dataset with information about insurance policies, insured drivers and their cars.
2. A claims dataset with all claims collected during year 0 to all policyholders,

- Variables to be modeled being number of claims (claims_nb) and claim amount (claim_amount).

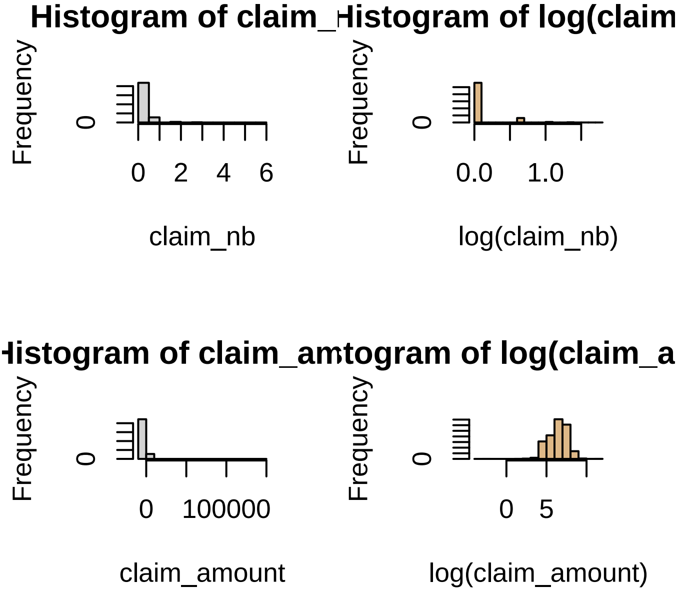
# Data ctd.

- Claims dataset was modified, because each client might make several claims with the same vehicle, which leads to duplication in id_client $*$ id_vehicle.
- To prevent duplicates, we summed up all the times and claim amounts for each occurrence of id_client $*$ id_vehicle.
- The merged dataset, which was used in conducting analysis, has $100,000$ observations with 33 variables.
- The merged dataset was split into 70%-train for building the models, and 30%-test for predictive purposes.

# Methodology

- Concept of Generalized Linear Models (GLM) was used to build models
- Frequency and Severity can be modelled separately or could be modelled at once using the pure premium approach
- Coefficient Estimates are found using MLE.
- Goodness of fit statistics such as AIC and Loglikelihood were used to compare nested models.
- For non-nested models, Root Mean Square Error with 10-fold cross-validation was used as a performance metric.

# Exploratory Data Analysis



# Exploratory Data Analysis Ctd.

Table 1: Summary Statistics for Variables

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | Number of Claims | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2 | Observation | 87346 | 11238 | 1264 | 134 | 16 | 1 | 1 |

Takeaways from histogram and summary statistics:

- From histogram, the majority of policy holders have claim counts of less than 1. The histogram for actual claim numbers is right skewed which means that most of the numbers are 0. Log-transforming claim numbers took the same shape of right skewness.
- The log transformed claim amount is normally distributed or appears to be symmetrical. Thus, a potential model for claim severity is the log transformed model.

# Exploratory Data Analysis Ctd.

- Table 1 shows the number of clients organized by claim counts. The proportion of clients having no claims is 87.35% which validates the results obtained from the histogram.
- Possible candidate models for fitting claim numbers are the zero-inflated Poisson and negative binomial models.
- the log transformed claim amount is normally distributed or appears to be symmetrical. Thus, a potential model for claim severity is the log transformed model.

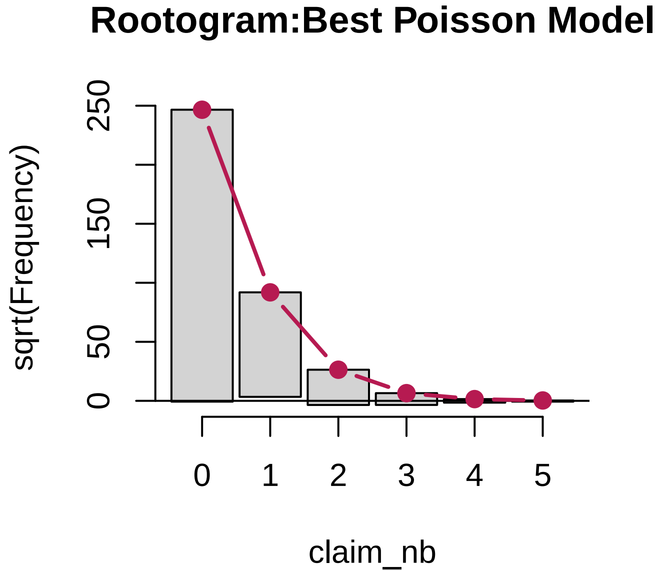
# Choosing Best Frequency Model

Table 2: 10-Fold CV for Frequency Models

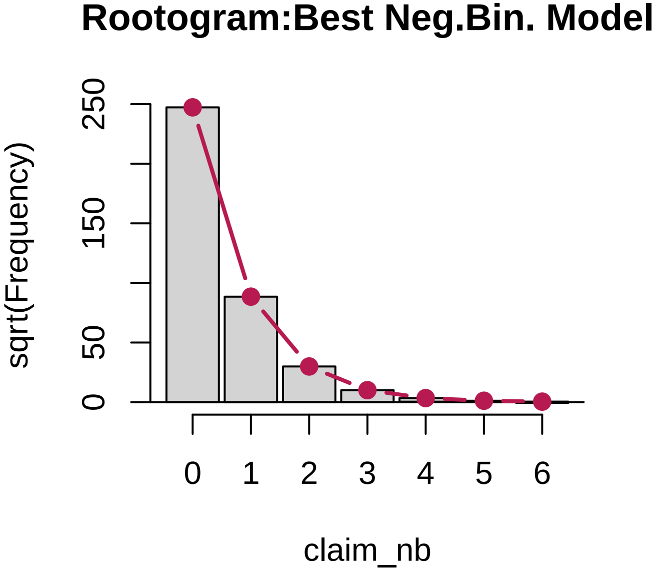
| | Best Poisson Model | Best Negative Binomial Model |
|---|---|---|
| RMSE | 0.3939 | 0.394 |

- The Poisson model was compared to the negative binomial model.
- Best Poisson model chosen based on AIC and Loglikelihood was the Poisson model with no offset.
- Best Negative model chosen based on AIC and Loglikelihood was the Negative Binomial with pol_duration as offset.
- From Table 2, there was no significant difference in RMSE of both models so further analysis were performed using rootograms and over-dispersion test to determine the best frequency model.

# Choosing Best Frequency Model ctd



# Choosing Best Frequency Model ctd



# Choosing Best Frequency Model ctd

- Over-dispersion tested the null hypothesis that the variance equals the mean which led to the rejection of that hypothesis. In this case the Negative binomial model was preferred.
- From the 2 Rootograms, we realize that the Poisson GLM sees an overprediction at claim count 1 as well as under-predictions at claim counts 2 and 3.
- Comparatively, the rootogram for the negative binomial shows a perfect agreement between expected and observed claim counts with no occurrence of over-prediction or under-prediction.
- Hence, we picked the negative binomial model with **pol_duration** offset as the best frequency model.

# Choosing Best Severity Model

Table 3: 10-Fold CV for Severity Models

| | Gamma Model | Lognormal Model |
|---|---|---|
| RMSE | 1954.885 | 2323.112 |

- The Gamma model has a smaller RMSE based on Table 3.
- Hence we settle on the Gamma model as our severity model.
- The best Frequency Severity-Model is the Negative Binomial-Gamma model which we denoted as F2-S1

# Choosing Best Pure Premium Model

Table 4: MSE for TWeedie Models

| | Model | MSE |
|---|---|---|
| 1 | Tweedie (p=0) | 982519.60 |
| 2 | Tweedie (p=1) | 982274.68 |
| 3 | Tweedie (p=2) | 982274.22 |
| 4 | DGLM Tweedie | 982407.22 |

- The results are displayed in Table 4 shows that the Tweedie model with $p = 2$ has the lowest MSE among the three models.
- To make results comparable across the board, the RMSE for the selected Tweedie model was calculated for the purpose of answering our final research question.

# Overall best model Selection

Table 5: Frequency-Severity vs Pure Premium

| | F2-S1 | Tweedie (with $p = 2$) |
|---|---|---|
| RMSE | 1955.249 | 1950.948 |

- RMSE of F2-S1 is found by adding the RMSE of the best frequency model to the RMSE of the best severity model.
- From the table, the pure premium model has the better RMSE so it is slightly preferred over the frequency-severity model.

# Discussion

- Actuaries have primarily used the Frequency-Severity approach but the pure premium model is gaining popularity. This study appropriately validates this claim.
- One challenge faced in this study was that a high proportion of zero claims suggested zero-inflated models were suitable to the data. But the models failed to converge and the standard frequency (Poisson and Neg. Bin.) models were used instead.
- Apart from auto-insurance, further studies could be conducted in other property and casualty lines and comparison of final results made.
- Models could be built using real world data and results could be compared against the results of fictitious data used.

Thank You