Report 2 - SF2930 Regression Analysis

David Haapanen, 010813-3794 Gustav Karlander, 020317-3299

February 28, 2025

Introduction

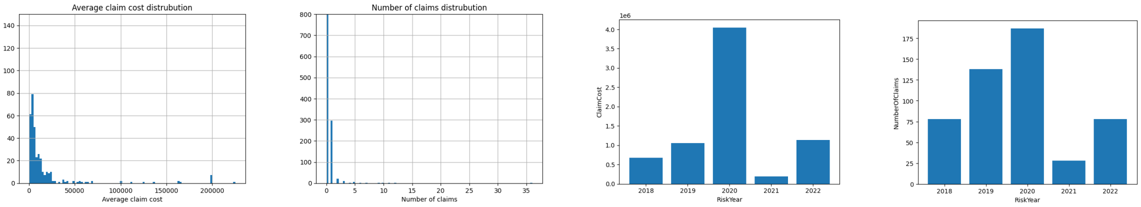
Home insurance doesn't cover travel done related to your occupation, thus companies has to pay for this type of insurance to cover their employees in case of lost luggage, delayed flight times etc. Since the past pandemic there has been an increase in interest for travel insurance, thus task is to help If insurance to create a price model that will have the given expression

$$\text{price} = \gamma_0 \prod_{k=1}^{M} \gamma_{k,i} \tag{1}$$

where γ_0 denotes the so called base premium while $\gamma_{k,i}$ are the risk factors for variables $k = 1, .., M$ and variable group i. Dependent on the company that is buying the insurance, their risk factor $\gamma_{k,i}$ will differ. To solve the task a given data set, `GLM_KTH_Data_Train.csv`, was given aswell as an template for a general linear model (GLM).

Data

The given dataset contained 149486 rows and 10 columns providing insight on companies that had payed for travel insurance from If during 2018-2022. To get an initial overview of the data histograms with bin size of 100 were made regarding the average claim cost as well as the average number of claims. Moreso both the total claim cost and number of claims per year were also evaluated. These are seen in fig. 1, which show that most of the data set contains companies that do not file claims. Furthermore it can be seen that both the average claim cost and number of claims share a power law distribution with a narrow tail. One can also see that both the cost of claims and the number of them increased every year during the pandemic up until 2021.



(a) Histogram of the average claim cost.

(b) Histogram of number of claims.

(c) Claim cost for 2018-2023.

(d) Number of claims for 2018-2023.

Figure 1: Summary of data overview

Grouping of data

Secondly the data was grouped into tariffs with the consideration of having risk homogeneity between groups (the risk shouldn't vary much within groups), each group should have a sufficiently large amount of data points and lastly that each group would have at least one claim corresponding to a claim cost. The considered groupings are listed below, where all data points that had missing values was grouped to a separate group in itself.

NoPGroup: The number of claims would probably increase for companies who insure more people. Thus the partition regarding the number of people insured was kept the same as the provided initial model.

ActivityGroup: The risk of making a claim can be argued to be more similar for certain work fields, as well as the range of claim cost. The partition regarding this was kept from the initial model, which grouped companies regarding their activity code by "Industrial", "Service" or "Other".

CompanyAgeGroup: As a company's age increases it means it has established itself more, thus companies that are more experienced most likely have more similarity between them and are less likely to claim insurance. The partition for this was New (0-10 years), new (3-10 years), established (11-25 years), mature (26-50 years), seasoned (51+).

TravelGroup: It could be speculated that more claims might be made for companies who travel far, since this would imply flights. Thus the grouping of travel area was made as international (whole world), euro (Europe) and scandic (Sweden and Nordic countries).

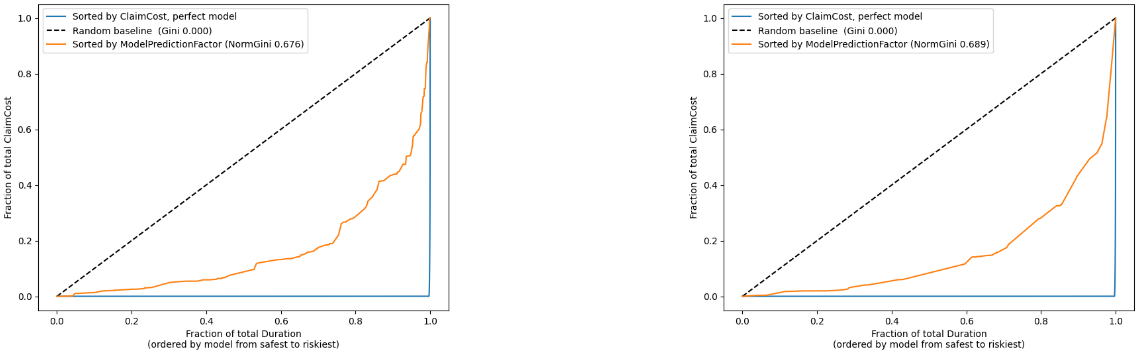
FinancialGroup: Companies with similar financial rating might depend on insurance in a similar manner. The partition consisted of "High" (AAA), "Upper-mid" (AA-A), "Mid" (BBB-BB), "Low" (B-C) and "Other" (IR, AN, missing).

GLM analysis

Both the frequency and severity model were made from the same groupings for every model that was considered. Furthermore the reference group for every combination of groupings was set to the group with the largest duration. This was 2-4 people insured for NoPGroup, ”Other” for the ActivityGroup, ”New” for CompanyAgeGroup, ”International” for TravelGroup and ”Upper-mid” for FinancialGroup. Furthermore a poisson distribution was used for the frequency model where the target value was the number of claims as target and duration as exposure measure to get the frequency. For the severity model a gamma distribution was instead used with a log link where the target was average claim cost with number of claims as weight.

Two models were developed, one which considered all groupings (full model) and one smaller model which only took into account for the groupings concerning the number of people insured by companies, the financial rating of a company aswell as travel locations (reduced model). Testing was then done for these models for comparison. This involved gini-score testing between the models and likelihood ratio testing between the frequency and severity models within both the full and reduced model.

The full model had a gini-score of 0.676 which can be seen in fig. 2a, while the reduced model had the gini-score of 0.689 illustrated in fig. 2b. Hence from this it can be said that the reduced model outperforms the full model with a larger gini-score whilst being less complex.



(a) Gini plot for full model

(b) Gini plot for reduced model

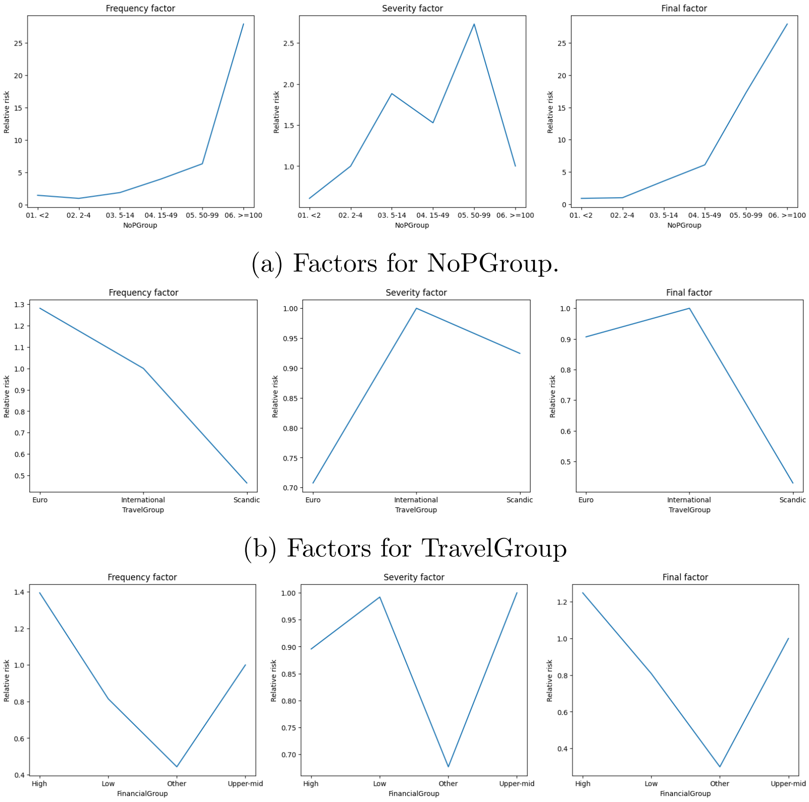
Figure 2: Comparision of gini score between models

Moreso two likelihood ratio tests were made regarding frequency and severity when considering the full model and the reduced model. It is known that using the loglikelihood function for the full model (FM) and reduced model (RM) one can get a $\chi^2(P_{diff})$ distribution where P_{diff} is the difference between the models. Hence

$$2 \cdot (l(FM) - l(RM)) \sim \chi^2(P_{diff})$$

Using a significance level of $\alpha = 0.05$ the likelihood-ratio test for the full and reduced model concerning the frequency and the severity proved that the null-hypothesis that the reduced model gives the same results as the full model was rejected (p-value $< \alpha$ for both tests).

From these tests it was decided that the final model would be the reduced model. The final factor for each grouping of the final model coming from both the corresponding frequency and severity factor is illustrated in fig. 3



(a) Factors for NoPGroup.

(b) Factors for TravelGroup

(c) Factors for FinancialGroup.

Figure 3: Summary of factors of final model

Leveling

Next the base level γ_0 in eq. (1) was determined such that the pricing for each insurance would cover its predicted claim cost (assuming a full-year duration basis). Firstly, the prediction factor for the year 2022, was determined as

$$\mathrm{PF}_{2022} = \frac{\sum \text{Predicted Claim Cost}_{2022}}{\sum \text{Duration}_{2022}} \approx 0.9120 \quad \left[\frac{\text{kr}}{\text{year}}\right]$$

The assumption was that each company that are customers for 2022 would continue be customers for the whole year of 2023. Thus PF_{2022} was multiplied with each registered customer of 2022 to compute the predicted claim cost for 2023.

$$\mathrm{PCC}_{2023} = \mathrm{PF}_{2022} \cdot \#\text{Customers}_{2022} \approx 3333941 \quad [kr]$$

Secondly the ratio target by If is given as 90%, hence

$$0.9 = \frac{\mathrm{PCC}_{2023}}{\sum Premiums} \Rightarrow \sum Premiums \approx 3704379 \quad [kr]$$

Then the total risk factor was computed by first multiplying each risk factor in accordance to eq. (1) for every insurance policy and then combining them for the whole portfolio while weighing every policy with their duration. The total risk factor was computed as $\mathrm{RF}_{tot} \approx 199251$. The base level γ_0 was then found by dividing the total claim cost for 2023 by the total risk factor. Hence

$$\gamma_0 = \frac{\mathrm{PCC}_{2023}}{\mathrm{RF}_{tot}} \approx 35.71 \quad \left[\frac{kr}{year}\right]$$

Lastly using this base level and accounting for a future risk ratio set to 85% and a safety margin set to 20% the adjusted leveling of the base level became $\gamma_0^{adj} \approx 24.28 \quad \left[\frac{kr}{year}\right]$. Which tested on unseen policies from 2023-2024 (provided by If) gave a histogram for the premium cost distribution which is seen in fig. 4

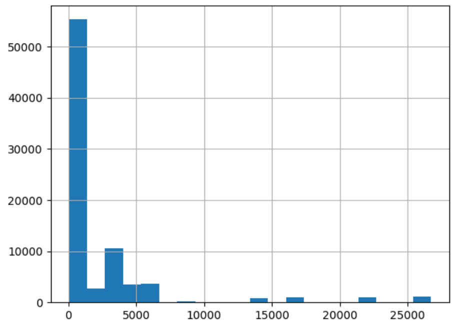


Figure 4: Histogram of premium based on unseen data from 2023-204