Project 2

Klara Zimmerman and Ville Wassberg

March 2024

In this lab, we look at data from the file "Cancellation.csv", which holds information about travel insurance claims from different companies. We will perform a GLM analysis to find a model that calculates the price of this insurance based on a number of variables.

Preprocessing and grouping data

For each variable considered for the GLM model, we decided how to handle any missing data, and whether to group the data into categories. The results of this is presented below.

Activity Code

- This data consists of categorical values that tell us what type of work the company does. We decided to group these categories together into parts we thought would have similar risks, while making sure that each category has at least one claim with some claim cost.
- Through trial and error, we finally decided upon the groups Industry, Service, Engineering, Government, Missing and Other (see figure 1).
- If information in this category is missing, we find it difficult to assign a category, since this probably has a large influence on the risk. These are therefore placed in their own category, labled "missing".

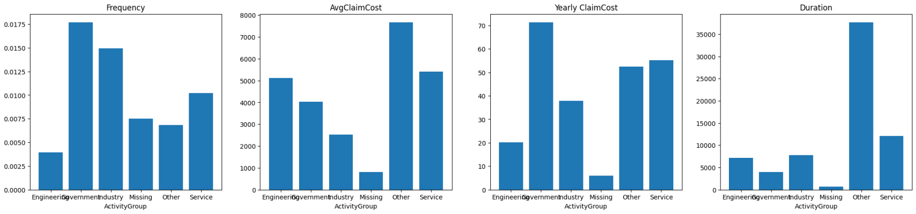


Figure 1: Grouping of "ActivityCode".

Company Age

- This data has numerical values. In order to group together company ages, we considered different spans that seemed relevant. A company that has only been registered for three years or less might have a different risk than a more established company. Likewise for companies that have been registered for a very long time. We considered companies registered for over 25 years to all belong to the same category, since not much changes after such a long time. In between 3 and 25 years we grouped the data with 5 year intervals.

- This resulted in groups that don't vary too much, but are large enough for there to be lots of data, and there is at least one claim with some claim cost in each interval (see figure 2).
- To fill in missing values, we decided to assume the mean of the category.

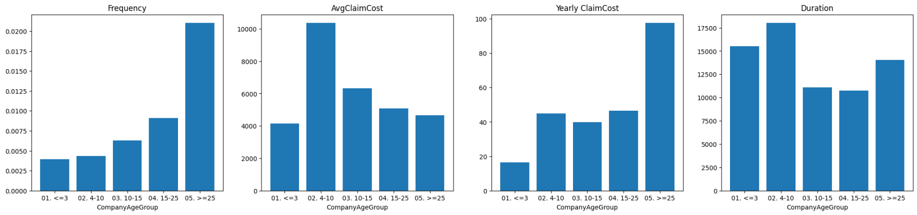


Figure 2: Grouping of "CompanyAge".

Dangerous Areas

- This data is categorical, and consists of only two different categories, hence we did not do any further grouping.
- For missing values, we chose the most common value.

Financial Rating

- This data has categorical values, and we found the categories fit to model without further groupings.
- For missing data, we chose the most common value.

Number of Persons

- This data has numerical values. We chose to select the first interval to companies 3 persons or less, as there are many companies like this, and we believe them to be a homogeneous group. Next, we had intervals 4-10, 11-40 and 41-100 to represent medium sized companies which we did not think seemed fit to all be in the same category. For the larger companies the number of people varies a lot, so we divided them into intervals 101-500, 501-1500 and >1500. These are shown in figure 3.
- For missing values, we assumed the mean of the category, and rounded this to a whole number.

Travelling Area

- This data has categorical values, with only 4 categories. We did not perform any grouping.
- For missing values, we assigned the most common area.

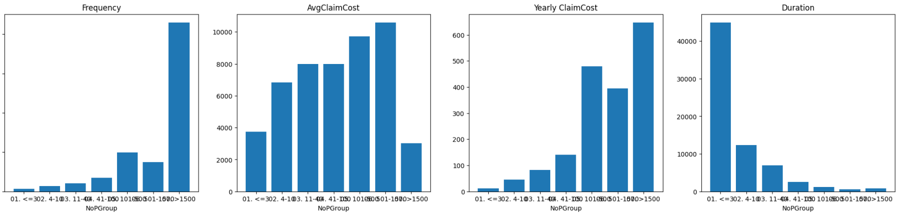


Figure 3: Grouping of "CompanyAge".

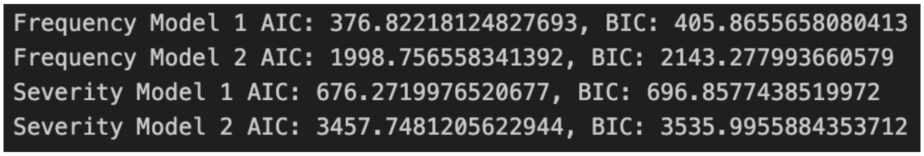
Different Models

Using the groupings defined above, we now want to decide which parameters to use in our model. We thus created one model with only three parameters, and one using all 6 parameters.

- Model 1
  - parameters: Travelling Area, Number of Persons and Activity Code.
  - motivation: We considered these variable the most relevant, and therefore want to see if a simple model performs better than a larger model.
- Model 2
  - parameters: Company age, Number of Persons, Activity Code, Dangerous Area, Financial Rating and Travelling Area.
  - motivation: We wanted to include all parameters to see if this model benefits from all the available information.

GLM Analysis

We now create a frequency model and a severity model for both model 1 and model 2. In order to compare how they perform, we first look at the AIC and BIC values. These are presented in figure 4. As we can see, model 1 better minimizes the AIC and BIC values for both frequency and severity. This seems reasonable, as these tests penalize models with many variables.



```
Frequency Model 1 AIC: 376.82218124827693, BIC: 405.8655658080413
Frequency Model 2 AIC: 1998.756558341392, BIC: 2143.277993660579
Severity Model 1 AIC: 676.2719976520677, BIC: 696.8577438519972
Severity Model 2 AIC: 3457.7481205622944, BIC: 3535.9955884353712
```

Figure 4: Aic and Bic values of Frequency and Severity models, for model 1 and model 2.

We produce plots that show the frequency factors and severity factors of each variable, and we combine them for a final factor model for each variable. Here, we check that the confidence intervals look reasonable for the variables and their groupings. We have omitted these plots in the report.

Model Validation

We want to analyze these two models further, which can be done by a gini validation. The gini test looks at how well the model assesses the different tariff cells against each other. In order to compare the two models, we create a gini plot using the gini score and a Lorenz curve. The results for both model 1 and model 2 are plotted in figure 5.

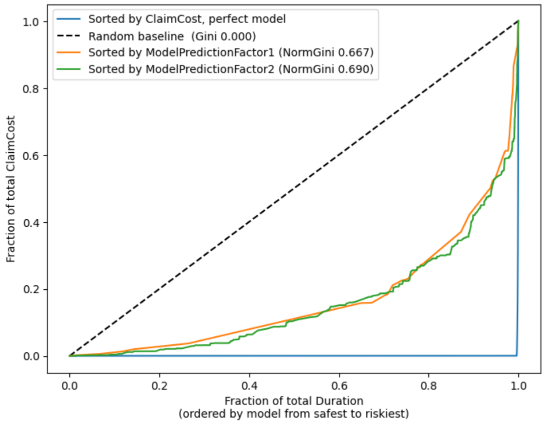


Figure 5: Gini plot comparing model 1 (ModelPredictionFactor1) and model 2 (ModelPredictionFactor2).

As we see in figure 5, the Model 2 is generally slightly closer to the "Perfect model", and has a slightly larger gini value than model 1. This could be explained by the larger number of parameters in model 2. However, this score only looks at the relative risk between different tariff cells in the model.

Next, we look at the risk ratio tests. Here we consider the magnitude, where a good model has more even risk ratios between groups. In these tests, model 2 seems to be the most even model, concerning most of the variables. For variables "DangerousAreas", "TravellingAreas" and "FinancialRating", only model 2 is trained, hence in these plots model 1 can be viewed as a random model. Clearly, both model 1 and model 2 perform better than the completely random model, for all variables.

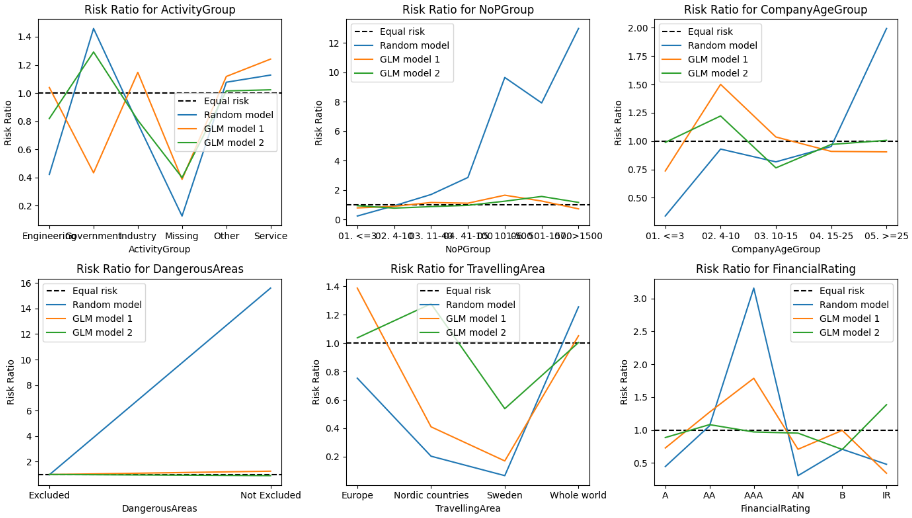


Figure 6: Risk ratio plots comparing model 1, model 2 and a random model.

Based in the plots in figure 6, we choose to continue with model 2. In order to get the price we need to set the base level which the factors are to be multiplied with such that the price for each insurance per year covers the predicted claim cost. If we assume all companies that are registered 2023 will be registered 2024 and that no companies are cancelling their insurance during 2023, we can get the sum of claim cost of the members in 2024.

If we assume the following year is expected to have the same customers as the current year with no additions nor loss, we can estimate the total claim cost of the upcoming year as the total claim cost of the current:

$$\sum_{RiskYear=2023} ClaimCost_i = 278014, \text{ for each individual i where riskYear is 2023.}$$

The total sum of the companies' premium for a price/cost ratio to be 90% must therefore, in this model, be 308904. Finally, we determine the corresponding total risk for the portfolio described by

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

where $\gamma_{i,k}$ is the corresponding risk factor of variable number k and group number i. Computing the product $\prod_k^M \gamma_{i,k} = 23467.8$ lets us now determine the base value γ_0, which in this case, thus, is 13.163. This can be verified by checking the product $23467.8 \times 13.163 = 308904$. Below are tables representing each variable group, with their respective frequency, severity and final risk factors.

Tables

ActivityGroup	**Frequency Factor**	**Severity Factor**	**Final risk Factor**
Engineering	0.273646	1.081792	0.296028
Government	0.487297	0.852186	0.415628
Industry	0.68592	0.450717	0.309157
Missing	0.376816	0.329899	0.124345
Other	1.00000	1.00000	1.00000
Service	1.214429	0.704554	0.855631

Table 1: Factors for ActivityGroup

NoP Group	**Frequency Factor**	**Severity Factor**	**Final risk Factor**
< 3	1.000000	1.000000	1.000000
4-10	1.648004	2.291544	3.776474
11-40	2.165974	2.498124	5.410874
41-100	3.671302	2.368871	8.696840
101-500	9.692224	2.510415	24.331507
501-1500	13.928838	2.562840	35.639729
> 1500	102.511825	1.013151	103.859945

Table 2: Factors for Number of People Group

CompanyAgeGroup	**Frequency Factor**	**Severity Factor**	**Final risk Factor**
< 3	1.476049	0.525375	0.775480
4-10	1.000000	1.000000	1.000000
10-15	1.269027	0.829398	1.053213
15-25	1.273140	0.475297	0.605120
> 25	1.335276	0.575382	0.768294

Table 3: Factors for Company Age Group

Dangerous Areas	**Frequency Factor**	**Severity Factor**	**Final Factor**
Excluded	1.000000	1.000000	1.000000
Not Excluded	2.913577	0.823305	2.398763

Table 4: Factors for Dangerous Areas

Travelling Area	**Frequency Factor**	**Severity Factor**	**Final Factor**
Europe	0.733714	1.609479	1.180897
Nordic countries	0.181989	1.648453	0.311870
Sweden	0.188596	1.696817	0.315766
Whole world	1.000000	1.000000	1.000000

Table 5: Factors for Travelling Area

Financial Rating	**Frequency Factor**	**Severity Factor**	**Final Factor**
A	0.646793	1.107464	0.716300
AA	1.000000	1.000000	1.000000
AAA	1.218168	1.328298	1.618900
AN	0.761439	0.974568	0.742074
B	0.705588	1.666966	1.171691
IR	0.275573	0.790139	0.217741

Table 6: Factors for Financial Rating