Report 2 - Regression analysis SF2930

Group 18
Magnus Axén maaxen@kth.se 980915-7119
Henrik Fagerlund hfag@kth.se 991012-2655

March 10, 2022

1 Scenario

This is the second project working with the cancellation data provided by If insurance.

2 Introduction

In this project we are asked to do thorough analysis of how a GLM could be applied to cancellation data of various companies. We are given a large data-set, with 136893 entries. These entries contain data which is binary, discrete and continuous.

3 Grouping and risk differentiation

To start off we wish to create grouping, seeing as the data-set is vast. Provided with summary(data) in R we can get a quick rough estimate of how such a grouping could be done. When we dealt with continuous values we would plot the histogram to get a rough estimate of how the values are distributed, as for the discrete values, table() was used.

Moreover, some data was "odd" such as having negative company age or negative amounts of people working at the company. These data which were considered false or wrong were grouped into its own category and not included in the full model, seeing as there probably is something wrong with the data.

Firstly the grouping was done by pure intuition. For example, given the summary of the data we saw that,

NumberOfClaims		ClaimCost	
Min.	: 0.000000	Min.	: 0.00
1st Qu.	: 0.000000	1st Qu.	: 0.00
Median	: 0.000000	Median	: 0.00
Mean	: 0.004821	Mean	: 47.44
3rd Qu.	: 0.000000	3rd Qu.	: 0.00
Max.	:13.000000	Max.	:219908.00

Figure 1: Summary 1

Num_of_ft_Employee		CompanyAge	
Min.	: -1.0	Min.	: -4.216
1st Qu.	: 0.0	1st Qu.	: 0.000
Median	: 1.0	Median	: 5.545
Mean	: 10.5	Mean	: 11.026
3rd Qu.	: 6.0	3rd Qu.	: 17.156
Max.	:10000.0	Max.	:126.579

Figure 2: Summary 2

Here we can tell from the number of claims and claim costs that the vast majority will be 0, and so grouping of these would make sense to have one single category just as such. Moreover for the Company age and number of ft employee

we plotted histograms to get a sensible grasp of the grouping. Both in the sense of data amount and risk homogeneity.

We cross-referenced, number of claims with dangerous areas. Seeing as the dangerous areas that were not excluded had a small sample-size and all of them had 0 claims we decided not to investigate this coefficient further. We also did the same for travelling area but considered the number of claims too few to be able to trust in the categorisation.

We decided on grouping the company age via $< 5, 5 < 20, 20 <$, after looking at a histogram (can be seen in Appendix) and confirming that a sufficient amount of number of claims were existent in all of the categories. Also since the

Lastly we decided on looking at the risk year, seeing as the Corona virus came about 2020. Due to this we saw an almost double in claims, thus we decided to include risk year as a factor, but only if it was 2020 or not.

Moreover we categorized by activity by having, a Industry category including, Mining,production and distribution of electricity/power/heat, Recycling, public entities,Agricultural activities and construction. And a service category including Wholesale, Restaurants, Culture, Service companies, Education and Care. Finally a office category including Consultants, Lawyers, Property owners, Leasing and Finance/insurance, the rest were labeled as others.

Lastly we can't include to many parameter as it will hyper-tune the GLM, thus we decided to conclude here, however we did inspect other parameters such as financial rating.

We decided to do likelihood tests, where we used,

$$LR = 2log\frac{L(fullmodel)}{L(reducedmodel)}$$

here we reject $H_0 =$ *a reduced model is correct* if $LR > \chi^2_\alpha$. Our full model included the variables, NOP group, Activity group, Company age, risk year. So we took reduced models, i.e. removing one coefficient for the model one at a time and tested them against the full-model.

We did the test and saw that none of the variables should be removed, as can be seen in the table.

Removed Variable	LogLike	ChiSquared
Activity group	-278.68	151.67
Company age	-204.19	2.6964
Risk year	-220.92	36.147
NoP group	-532.64	659.59

We decided on not including number of full-time employee and number of people, seeing as there might be some co linearity between it and company age, and for model simplicity.

4 Leveling

The base level, γ_0, is heavily based on the estimation for next year expected claim cost. In retrospect, basing the 2021 year prediction on such a odd year as 2020 should be fine since 2021 also was quite unstable. The same decision would probably not have been done before 2021. When modelling the next year the pandemic should be taken into considering, depending on it's severity.

1. Claim cost

The claim costs for years 2016-2020 is plotted in the following figure,

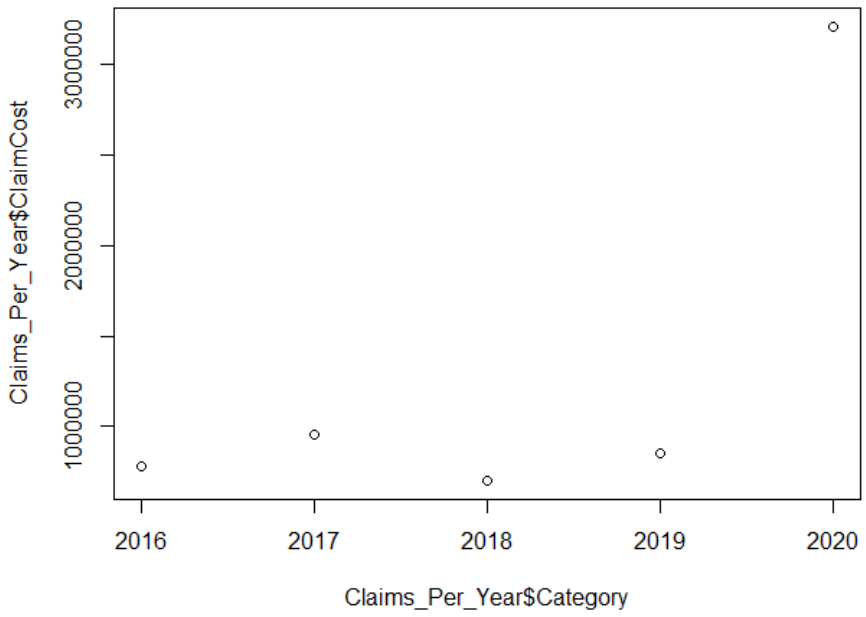

Figure 3: Claim-cost

Since the simplest possible model is used, the estimated claim cost for 2021 is predicted to be the same as 2020 which had a value of $claim\,cost = 3\,211\,117$.

2. Total sum of the companies' premium

The total sum of the companies' premium is calculated using the assumption that IF PC has a ratio target of 90%. The following formula is used,

$$SUM(Premium) = \frac{SUM(Claims)}{Target}.$$

Put in the the values and the total sum of premiums obtained was $SUM(Premium) = 3\,567\,908$.

3. Total risk for the portfolio

Each of the insurance's "total risk factor", which is the sum of all $\gamma_{k,i}$ for that insurance. This is done through iterating over all of the rating factors. These are later summarized in order to determine the total risk for the whole portfolio and the value obtained was $SUM(Total\,risk) = 136\,100.5$.

4. Base level

The base level, γ_0, can now be calculated using the following formula,

$$\gamma_0 = \frac{SUM(Premium)}{SUM(Total\,risk)}.$$

Using this, the base level obtained was $\gamma_0 = 26.21524$

5 Conclusion

The GLM model depended severely on the previous year, 2020, seeing as it was the wake of the pandemic. Noting this it will greatly affect 2021. Besides that the greater number of people would increase severity, frequency and risk, by an increasing factor with people. We note moreover that the office was the most severe, but not the most frequent, which was service which seems reasonable as it may have more frequent butt less costly complaints. Lastly by age of company we could see rather equal frequency but a larger severity in the older companies which could be since the older companies tend to be larger. This can all be seen in the last figure in the appendix.

Appendix

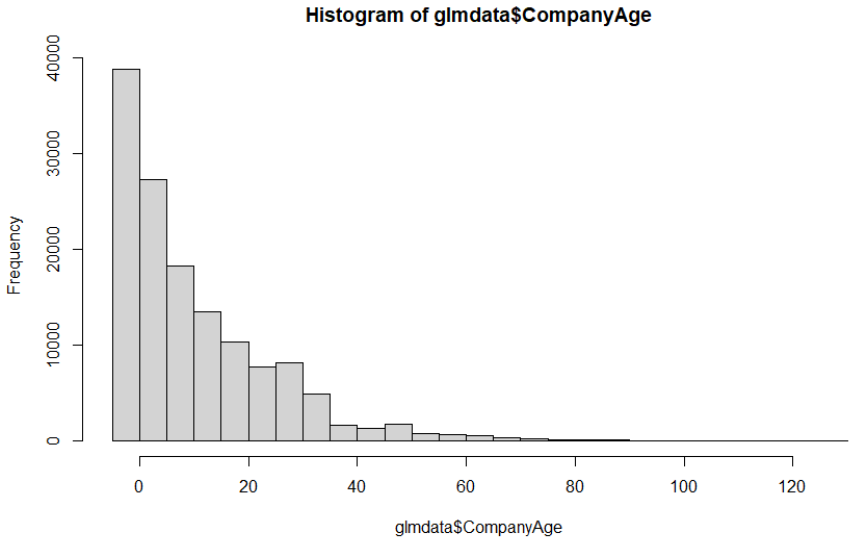

Figure 4: Histogram

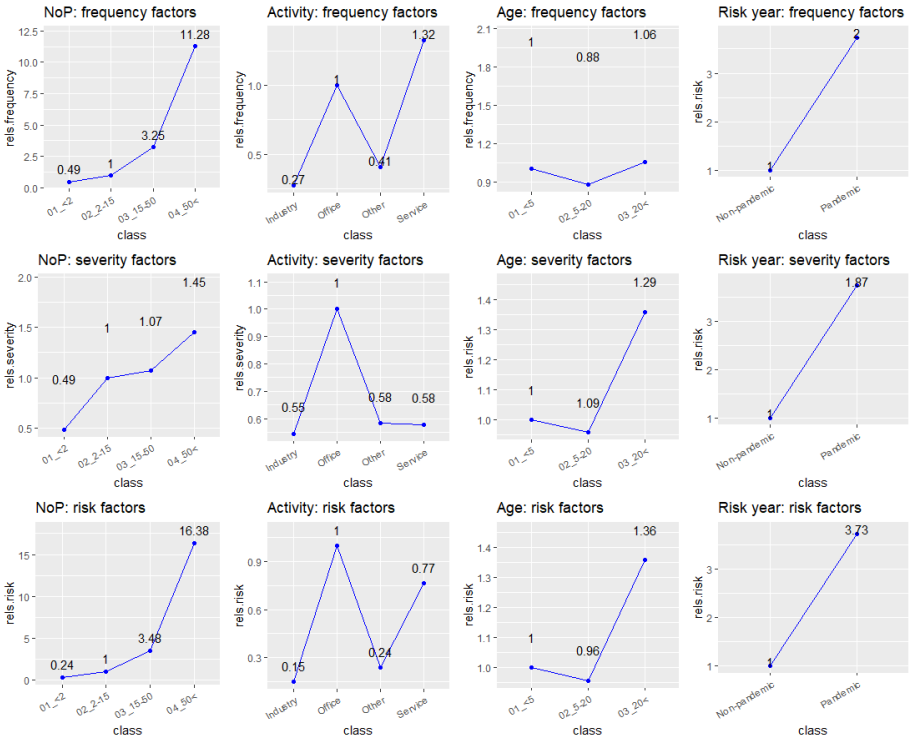

Figure 5: Frequency, Severity, Risk