Regression Analysis
Project 2

Written by: Oguzhan Ugur, Shadman Ahmed,
E-mail: Oguzhanu@kth.se shadmana@kth.se

June 6, 2018

Introduction

Insurances and insurance analyzes are of great interest nowadays, due to the need of protection and safety of private as well as company assets. The price of insurances depends on risk factors, if an asset is exposed to higher risk then the insurance price will also be higher. The risk factors is determined by collecting data and applying different regression models to see correlations that are of interest. In this project a model that prices insurance for tractor vehicle damage was created in the form

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

where γ_0 is the base level and $\gamma_{k,i}, k = 1, ..., M$ are the risk numbers corresponding to variable number k and variable group number i. $\gamma_{k,i}$ will take different values for each individual tractor depending on the characteristics of the gathered data, such as in which sectors the tractor is used, which weight class it has etc. In other word the goal is to determine the risk factor for different insurances in order to predict future costs and use this to determine a correct price of the insurance.

The calculation of the risk-factor was done by performing General linear model analysis, mainly because GLM is a flexible generalization of ordinary linear regression and allows response variables that have error distribution models other than a normal distribution.

Data

The data used in this project was a csv file named Tractors.csv containing information on all tractors with vehicle damage insurance in If P&C during 2004-2014, including claims history. Each row in the file corresponds to one tractor and the columns corresponds to information about the tractor such as, Risk year(the year of the insurance period), Vehicle Age, Weight of the tractor, Climate(geographical location in Sweden), Activity code(the activity code registered on the company that owns the tractor. For each tractor there is also information regarding Duration(the share of the risk year the tractor was insured), number of claims and claims cost corresponding to the insurance period.

GLM program

Task 1: Grouping and risk differentiation

In order to perform a GLM analysis, the variables in the dataset was assigned into groups. Each group was tried to be constructed risk homogeneously meaning that the risk does not vary much within the group, with regard to the variable. The variables that was assigned into groups was weight and Vehicle age, since the other variables Activity code and climate already was assigned into groups. The weights was assigned into 5 groups and Vehicle age was assigned into 7 groups. The groups was constructed as follows:

Weight group: $<$ 1000 kg, $1000 - 2500kg$, $2500 - 5000kg$, $5000 - 7500kg$, $>= 7500kg$
VehicleAge group: $< 1year$, $1 - 3year$, $4 - 7year$, $8 - 12year$, $13 - 18$, $19 - 24year$, $>= 24year$

These grouped was optimized by looking at the risk factors. It is unnecessary to have risk factors close to each other so the groups was constructed in such a way that the risk factors was different from each other. The risk factors was calculated by multiplying the model severity with the model frequency and the glm function built in R was used to model the GLM.

Datasets that were strange, missing, or incomplete data, was handled by manipulating the dataset. This was done by deleting data on year 2003, 2015, weights less than 300 kg, vehicle age greater than 45 years and claims without claim costs.

The Likelihood Ratio test(LRT) performed on our data gave the P values written on the tables below:

Model Frequency :

	Pr(>Chi)
weight group	$2.2 \bullet 10^{-6}$
Vehicle age group	$6.485 \bullet 10^{-6}$
Climate	0.04167
Activity Code	0.06473

Model Severity :

	Pr(>Chi)
weight group	$2.817 \bullet 10^{-12}$
Vehicle age group	0.05158
Climate	0.26170
Activity Code	0.04533

It is possible to see that the p-values for weight and vehicle age are of interest since they have a p-value around $p < 0.05$ which indicates the significance level. The climate value is less than 0.05 for model frequency but much greater than 0.05 for the model severity, the conclusion drawn from this result is that the Climate maybe can be deleted from our model, this is decided by looking into the AIC(Akaike Information Criterion) and BIC(Bayesian Information Criterion) if the values of these criterion's reduces when climate is deleted then the conclusion drawn above is correct. The AIC value before Climate was deleted was 983.0565 and after it was deleted the p-value reduced to 618.3343. The BIC also got reduced when Climate was deleted which indicates that our conclusion about Climate was correct. The p-vlaue for the Activity code is also around 0.05, this indicates that it should be included in our model.

Task 2: Levelling

To determine the base level mentioned in equation (1) two steps was used. The first one was to estimate the claim cost for 2015, assuming that If P&C has a ratio target between the estimated claim cost and the total premium of 90%. The second step was to calculate the total risk factor for each insurance and then finding a base level γ_0, that turns the total risk factor into an actual price, such that the total sum of all prices for tractors that are insured in 2015 match the result obtained from the first step.

Step 1: The Claim cost was calculated by summing over all claim costs and then dividing it by 11 years 2004-2014, which gives a mean value of the claim cost over one year. The MeanClaimCost was equal to 1246045kr. The total premium was then calculated by$totPremium = \frac{MeanClaimCost}{0.9} = 1384494$kr

Step 2: The base level was calculated by multiplying the risk factors for each tractor and then summing over the risk factors. This gave $\gamma_0 = 845.85$. Which is a quite large value for the base level so the groupings could probably be constructed better.

To determine the insurance prices the graphs depicted below was analyzed the results are written on the table below for the lowest, midmost and highest price for the vehicle damage insurance.

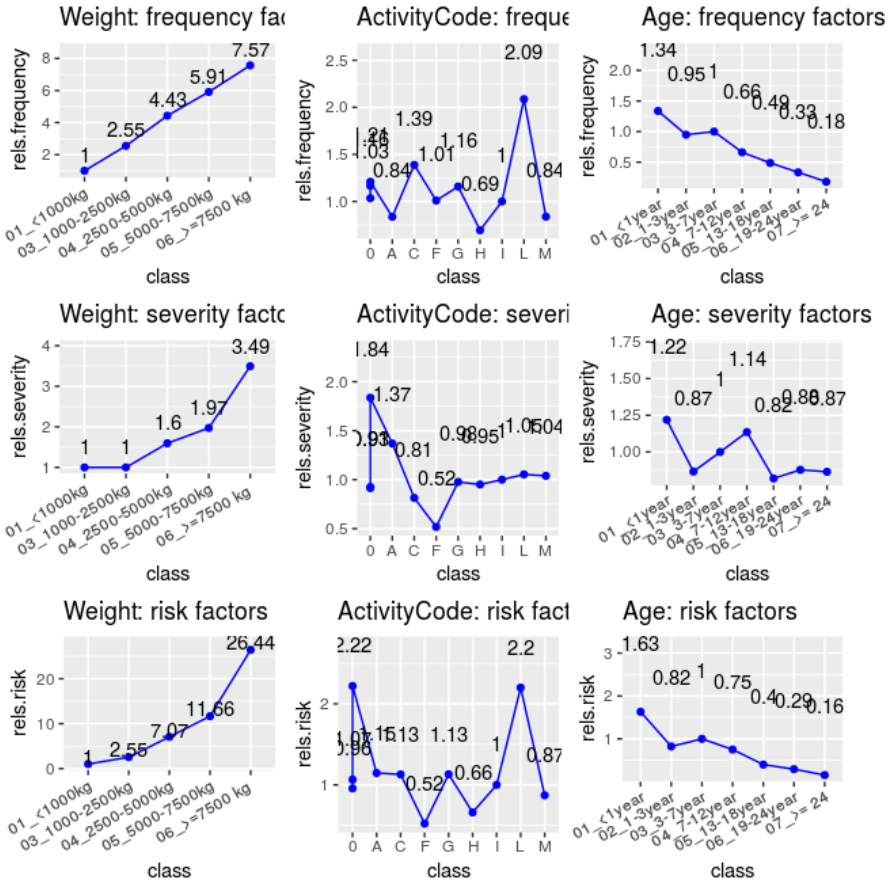

Figure 1: Plots of Model frequency, severity and risk factor for the different groups

The price of the insurance can now be calculated by price = baselevel $\bullet \prod$ riskfactors in other words equation (1) written above. The table below shows the prices calculated by using equation (1).

	Price
Lowest	70 kr per year
Midmost	2960 kr per year
Highest	80927 kr per year

The lowest and highest prices are extremums depicted at the figure 1 for the risk factor plots and doesn't probably represent large scale of the tractors. It is possible to see the the prices obtained is not reasonable since 80927kr per year is a very high insurance price and 70 kr per year is too cheap. The reason to why these prices was obtained is probably because the grouping. The grouping process is an iterative process and hard to decide properly.