SF2930 Regression Analysis

Project 2

Emil Book

ebook@kth.se

940122-0034

Deepika Anantha Padmanaban

deap@kth.se

950303-4408

11.03.2019

1.1 INTRODUCTION

The aim of this project is to develop a Generalized Linear Model for finding the risk and tariff involved in the insurance contract for the tractors dataset, where the response variable is expected to follow a Poisson's distribution rather than a normal distribution. In Sweden, tractors are required by law to have a third part liability insurance. Many tractor owners complement this legally required insurance with an insurance covering vehicle damage to their own tractor.

The given dataset tractors contains the claim frequency and claim cost for 99074 tractors insurances P&C had during the period 2006-2016. The possible explanatory variables included in the dataset are non-categorical variables like weight, vehicle age and categorical variables like climate and the activity code (purpose of using the tractors). In this project, we have tried to find the best possible grouping of the non-categorical explanatory variables, to convert them into categorical and develop a simple yet well-fit GLM model to explain the risk frequency and risk severity. Then, we used the available data is set the base level to calculate the actual tariff from the obtained risks. This project includes evaluation of the model fit using Likelihoood Ratio tests, AIC and R-squared. The basic generalized linear model is of the form;

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

where, γ_0 is the base level and $\gamma_{k,i}$ is the risk associated with variable number k and variable group number i and in our case is given by two terms, namely, risk frequency and risk severity.

$$\gamma_{k,i} = risk\ frequency\ *\ risk\ severity$$

1.2 MODEL BUILDING:

REASON FOR USING GLM:

The properties of the insurance policies are;

- Policy Independence - For different insurance policies the number of claims X_1,X_2,...,X_n are independent.
- Time independence - For a policy we may divide the time of the insurance contract into different time intervals which are assumed to be independent.
- Homogeneity - Consider two different policies in the same tariff cell, having the same number of insurance years, then the number of claims X1 and X2 have the same probability distribution,

All the above properties suit a Poisson's distribution, which belongs to the exponential family, Thus, we end up using a multiplicative GLM to accomodate the exponential family instead of the normal family in a linear model.

GROUPING AND RISK DIFFERENTIATION:

Our dataset, like most real-world datasets is complex and far from perfect, with many explanatory variables involved and many values missing. The actual problem arises with continuous variables such as weight and vehicle age. These variables take a numeric value within the range 0-960 for vehicle age and 0 to 25500 for weight. Handling these variables in the model is too complicated and would lead to fitting an insurance for each combination of these values. Thus, we perform the grouping of the values of these variables to ensure that they can be converted into a categorical format and fall within a range where the number of variables in the model is manageable. But, there were main criteria involved in the grouping of variables;

- The insurance that fall under the same group must almost be similar,i.e., the groups are risk homogeneous.

- There must be enough data to represent the group formed and account for the risk involved.

For any claim, the Expected Claim Cost is;

$$Expected\ Claim\ Cost = Expected\ Claim\ Frequency\ *\ Expected\ Average\ Claim\ Cost = Risk$$

Hence, the pricing of the policy is this **Risk + some extra**.

For finding groups with almost similar risks, we tried forming a way to create temporary variables for calculating the frequency and severity approximately an tried plotting them against each continuous variable to check their spread and find commonalities between the values of the variable. The plots are as shown in figure[1] below;

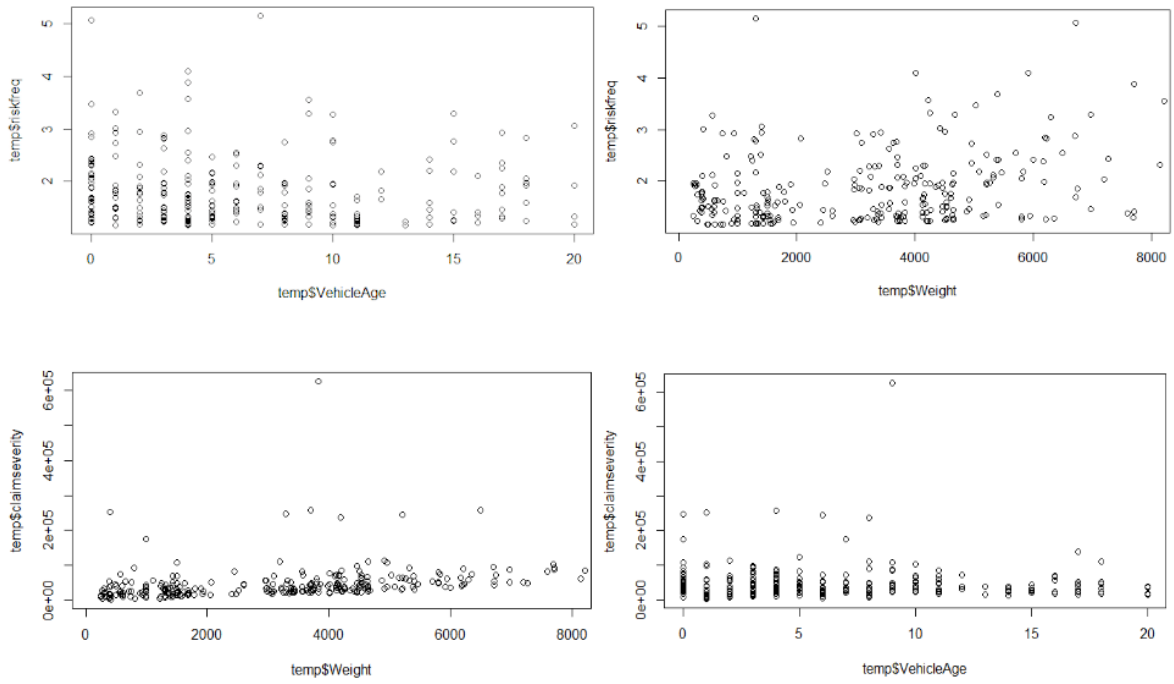

Figure 1: An approximate spread of risk severity and frequency among weight and vehicle age

From the plots, we approximately found values of the Weight and VehicleAge variables that could be grouped together and hence, converted the continuous variables into categorical variables. We used even smaller intervals to see more accurately how it was dispersed. We then iteratively grouped variables differently and analyzed to see which grouping led to the best GLM fit. We concluded that following grouping should be used figure[2 and 3]

Weight_Group	Weight
1	<2500kg
2	>=2500kg, <5000kg
3	>=5000 kg, <10000 kg
4	>=10000

Figure 2: Groupings of the variable Weight

Vehicle_Age_Group	Vehicle Age
1	<1 years
2	>=1, <5 years
3	>=5, <12 years
4	>=12 , < 20 years
5	>=20 years

Figure 3: Groupings of the variable VehicleAge

FULL MODEL FIT:

Along with the above grouped variables were the other already categorical variables such as Climate, Activity Group. We included all of these in the model, with this being the full GLM model.In the next section, we try to find the best fitting model in comparison to the full model using different model evaluation criteria such as the Likelihood Ratio Test, AIC and Adjusted - R2.

1.3 MODEL SELECTION AND TEST FOR GOOD FIT:

FULL MODEL AND REDUCED MODELS COMPARISON:

Overfitting is a serious problem that occur in regression fits, it often occurs when we have to many explanatory variables for a limited set of data points. This may result in an overly complex model, and can cause the model to be misleading. That would make our model a bad predictor for other samples of the same population, thus leading to poor generalization. Also, not variables included in the full model is necessarily explanatory of the variance in data and the Occam Razor rule also states that a simple model is always preferred over an overly complex model. Thus, we try to select variables that can explain the variation in data to the best possible extent, thus reducing the complexity.

Since, we have multiple variables, one for each group in Weight group and vehicle age group along with climate and activity code. We tried multiple models, with one variable removed at a time and did the following tests to find the significance of the variables and fit of the model. The tests that were conducted are;

Akaike Information Criteria:

The AIC of a model determines if the presence of a variable in the model leads to any improvement in the model performance.AIC is a estimator for the goodness of a model, or to compare the different models. This model penalizes the GLM function that has many explanatory variables, so it's a trade off function between goodness of a fit and the simplicity. Usually a model with the lowest AIC is considered the best. The AIC for any model is given by;

$$AIC = -2ln(L) + 2p$$

$\mathbf{R}^2_{GLM}$:

A measure of the adequacy of a regression model that has been widely used is the coefficient of multiple determination, R^2 . All subset regression models that have an R^2 not significantly different from the R^2 for the full model can be identified and preferred over the full model.

The results of the tests performed are tabulated below. Since, we have two models, one for risk frequency and one for risk severity, we calculate the best model for each of them. Figure[4] represents risk severity, while Figure[5] represents risk frequency.

Model used	Number of Variables	Likelihood ratio	R-Squared	AIC
Full Model	21	-4598.9	0.48415	9239.8
Model Without ActivityCode	11	-4634.9	0.39269	9291.8
Model Without Climate	19	-4609.1	0.45956	9256.3
Model Without Vehicle_Age_Group	17	-4618.2	0.43684	9270.4
Model Without Weight_Group	18	-4678.2	0.26201	9392.4

Figure 4: Model comparisons for Risk Severity

Model used	Number of Variables	Log -Likelihood	R-Squared	AIC
Full Model	21	-483.45	0.66977	1006.9
Model Without ActivityCode	11	-517.12	0.48055	1054.2
Model Without Climate	19	-484.31	0.6624	1004.1
Model Without Vehicle_Age_Group	17	-529.72	0.58422	1091.4
Model Without Weight_Group	18	-535.65	0.53583	1105.3

Figure 5: Model comparisons for Risk frequency

Likelihood Ratio Test:

The likelihood ratio can be used to determine whether or not the model is a good fit for the data and thus, is it necessary to add a new explanatory variable to the model. The likelihood Ratio test was conducted with the different models starting with a one variable model and comparing it with the null model/previously chosen model without any variables. We tried adding and checking if the variables included in the model led to some improvement in the Likelihood ratio. We found that the inclusion of each variable led to the improvement in the LR ratio at each step, with the LR exceeding the cut-off of significance. Thus, we preferred using the full model for the analysis.

$$LR = 2 \cdot \ln\left(\frac{\mathcal{L}(FM)}{\mathcal{L}(RM)}\right) = 2\left(\ln \mathcal{L}(FM) - \ln \mathcal{L}(RM)\right)$$
$$= 2\left(\ell(FM) - \ell(RM)\right),$$

Model	LR
Risk Frequency	104.4
Risk Severity	86.6

Figure 6: Likelihood Ratio Test

1.3 LEVELLING AND BASE LEVEL:

Having determined the different risk factors for every group, we now need to find the base level y0. To calculate the base level we first need to estimate expected total amount of cash from claims P&C needs to pay, one way to do this would be to calculate some sort of moving average. After discussing with some experts at IF, working in the field of insurances they have reasons to believe that the amount paid will be very similar 2016, which we have data on. Adding all claims costs for 2016 gives us a value of 1407425. Since we we need to have some profit margin and other overhead expenses for running a business, P&C counts that 90% target to cover these will be needed, which means that our total earnings need to be 1407425/0.9= 1563806 . Then we need to calculate the total risk factor, we do that by calculating risk factor for every tractor and sum for every insurance policy. Base Level is calculated as total expected earnings for 2016 divided by our total risk factor, which gives us a base level y0 of 98.03.

$$Total\ Active\ Insurance\ cost\ =\ 1407425$$

$$insurance\ cost\ =\ Total\ Active\ Insurance\ cost/0.9\ =\ 1407425/0.9$$

$$insurance\ cost\ =\ 1563806$$

$$insurance\ cost = \gamma_0 \prod_{k=1}^{M} \gamma_{k,i}$$

where,

γ_0 is the base level and $\gamma_{k,i}$ is the risk associated with the kth insurance for.

$$\gamma_0\ =\ 1563806/15950.89 = 98.03$$

Grouping and Risk Differentiation: The results for the risk factors after the groupings are shown below in the Figures below. After trying a large combination of different groupings, the most homogeneous risk factors for each variable groups are the ones shown in these plots.

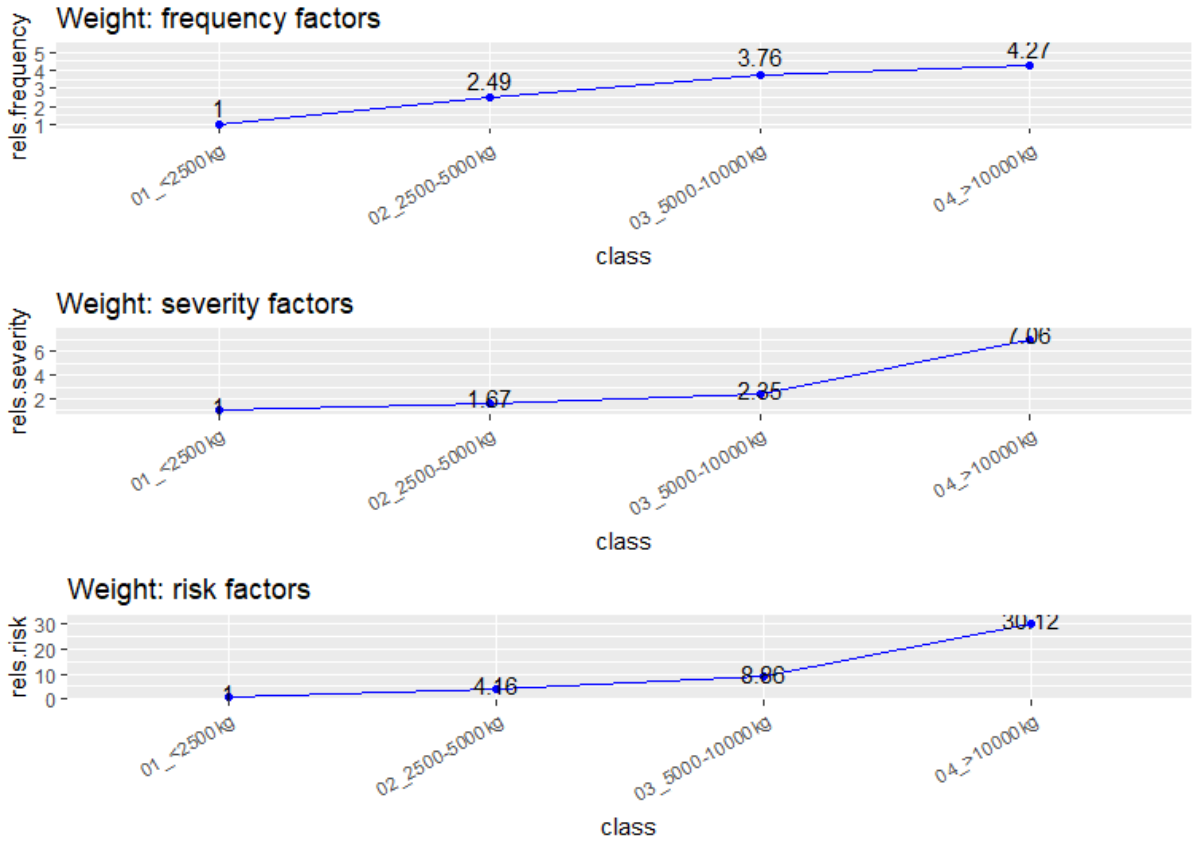

Figure 7: Risk variations among weight groups

Figure 8: Risk variations among Climate

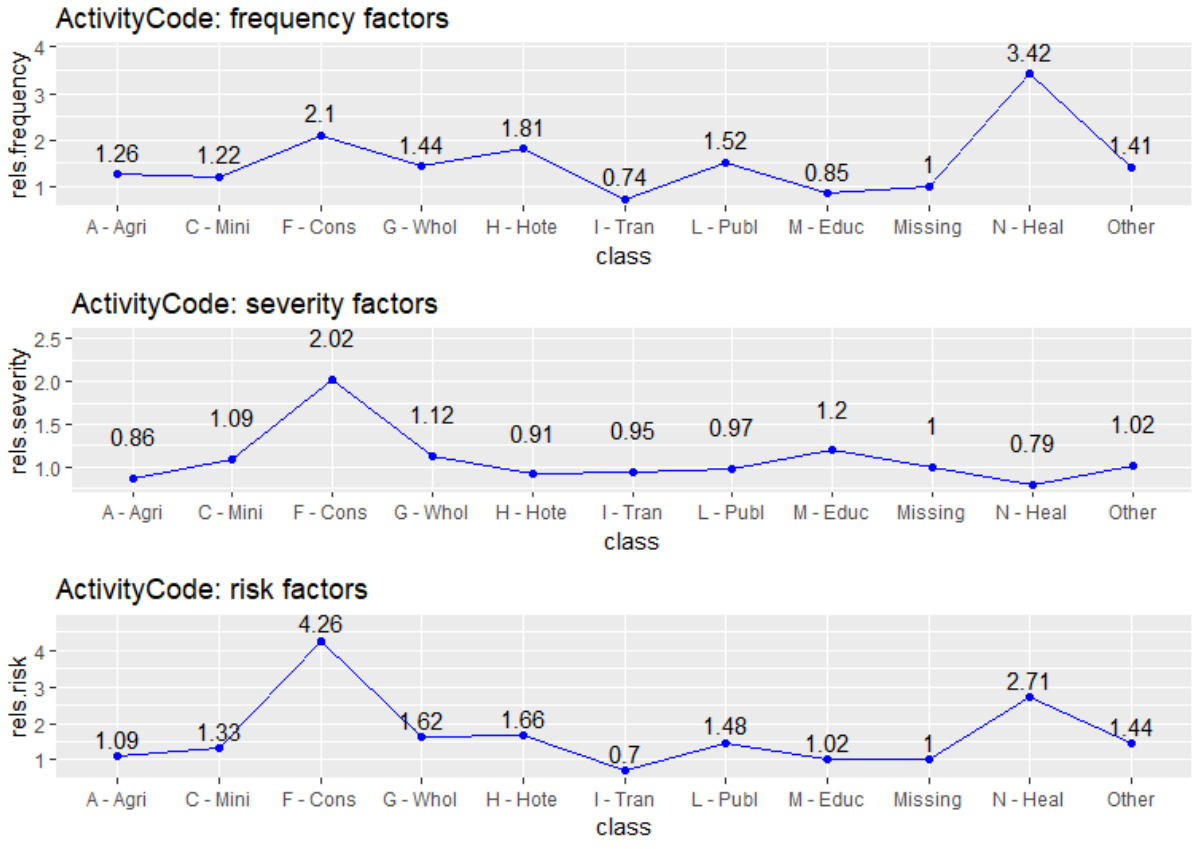

Figure 9: Risk variations among Activity Code

Figure 10: Risk variations among Vehicle Age groups

From Figure[7], it can be seen that the risk frequency and severity increases with increase in weights of the tractor, as we had expected. Thus, the overall risk also has an increasing trend with increase in weights.

From Figure[8], it can be seen that the risk frequency and severity variations with different climatic zone. As we had expected, the overall risk was high in the South region, which was almost the same as in the North, due to frequency of risk being high in the North.

From Figure[9], it can be seen that the risk frequency and severity varies with different activities. As we had expected, the risk was highest for the construction activity.

From Figure[10], it can be seen that risk frequency and severity decreases with increase in age of the tractor. Thus, the overall risk also has a decreasing trend with increase in the age. This can be attributed to the fact that there are large number of vehicles in the age group of less than 3 years, when compared to aged tractors.

1.4 CONCLUSION:

The task of this project was to develop a GLM model for calculating the risk and base level for insurances on tractors. We initially had to put in more efforts on the grouping since, it involved analyzing the groups with risk homogeneity and fitting a model each time to check if the results produced were sensible. On iterative grouping, we concluded upon the groupings as presented above. We then tried if there was a possibility of removing variables from the model. It was a futile effort, as the R-squared value had a significant reduction, with the removal of any variable. Thus, we decided against removing any variable. This was also motivated because we did not have granular grouping and each of our grouped variables had a maximum of 4 to 5 groups. Using these grouped variables we modeled the risk frequency and then the risk severity and used them for calculating the risk. We plotted them against the variables to see their trend. We also, did some model fit tests that yielded reasonable results. Finally, we concluded by finding the base level to be set up for insurance policies, which was around 98Kr and seemed reasonable.