Report II - SF2930 Regression Analysis

Anna Gashi 1, 910223-2189 Omar Contreras 2, 930917-3996

2 March, 2020

Introduction and project goals

The general idea of the project was to perform a Generalized Linear Model (GLM) analysis to analyze the associated risk of tractors from a data file provided by the insurance company If P&G. From this, the goal of the model was to determine the associated risk of tractors belonging to different categories compared to some base line group where the relative risk factor was set to 1. Tariff factors could be determined which enabled calculation of the price of the insurance policy in order to cover expected claim costs of the tractors.

If P&G tractor insurance data

As mentioned from the project description, tractors are by law in Sweden required to have a third party liability insurance for tractor type vehicles. There are therefore a lot of options for insurance for a tractor owner to utilize, leading to pricing being a key component in being successful in a competitive insurance market. The insurance company If P&G offers vehicle damage insurance and provided a data file that contains historical data of tractors using If P&G's insurance between the years 2006-2016. The data file also contains information on each individual tractor regarding age, weight, climate that the tractor operates in, number of claims, claim cost, as well as duration which represents the share of the risk year that the tractor was insured. The goal of the project was to create one's own tractor tariff on the form:

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

In which γ_0 represents the base level price, and $\gamma_{k,i}$ for k=1,...,M represent the risk factors that correspond to variable number k and variable group number i. In other words, $\gamma_{k,i}$ will have different values for each tractor depending on its different characteristics. The final risk factors $\gamma_{l,i}$ was determined by taking the product of the claim frequency and the claim severity.

Generalized Linear Model

When the assumptions of normal distribution and constant variance from the data that's being analyzed are not satisfied, the generalized linear model (GLM) is an approach that allows non-normal response distributions. The response variable must be a member of the exponential family, such as normal, Poisson, binomial, exponential or gamma distributions. GLMs can analyze the effects of variables when some of them are categorical variables and some are continuous variables.

The probability distribution of the response variable of a GLM is:

$$Y_i \sim P(\mu_i, \sigma_i) \tag{2}$$

where P(x) comes from an exponential family of distributions.

$$g(\mu_i) = \sum_j x_{ij}\beta_j \tag{3}$$

In the last equation, x_{ij} are the independent variables, β_j are the unknown parameters, and $g(\mu_i)$ is a link function, which links the linear model to the mean. The link function used, depends on the distribution of P and β_j is determined with maximum likelihood.

Analyses and model development

Grouping and risk differentiation

In order to be able to perform the GLM model, the variables had to be grouped based on the criteria of making each group as "risk homogeneous" as possible, and also with enough data in each group to make the GLM model stable. The two variables that were subdivided into groups were the weight of the tractors as well as the age. In terms of both tractor weight and tractor age, two histograms with the original data were generated in order to get an initial idea of the distribution. Histograms were generated in order to see if there was any natural grouping in the data. This beneficial to look at since it would be likely that tractors with similar characteristics of weight and age would also have similar risk associated with them. In terms of tractor weights there appears to a natural grouping of different weights, where as in the histogram for tractor ages the frequency rapidly decreases with increased tractor age. The amount of data in each group was also researched in order to insure there was sufficient amount of tractors in each group. As seen from table 1, the grouping result was 5 tractor weight groups and 5 tractors age groups of the data.
Table 1:

Tractor weight groups	0-600kg	600-1000kg	1000-2100kg	2100-7000kg	$\geqslant$ 7000kg
Tractor age groups	0-6	6-9	9-17	17-30	$\geqslant$ 30

Following the grouping of the tractor weight and tractor age, the data set was aggregated based on the these variables in addition to climate and activity code. The aggregated data was then used in the GLM analysis.

GLM Analysis

The GLM analysis for the project was performed by using R's built in GLM function. To begin, the claim frequency of the data was modeled by setting number of claims as the dependent variable, and the variables weight group, age group, climate and activity code as the dependent variables in the GLM function. The variable duration was in the function used as an offset term, meaning that the variable was held as a constant while the other explanatory variables were evaluated. A Poisson distribution function was used in the GLM model, which is a commonly used probability distribution for discrete data.

The resulting coefficients from the GLM analysis were saved in an array, and used in the second part of the analysis involving modeling the average claim cost (claim severity in the code). Similarly to the previous step, the built in GLM function was used, with weight group, age group, climate and activity code as the explanatory variables. In this step, a Gamma distribution was utilized, which is commonly used for non-negative and positive-skewed data, such as insurance claims. Finally, the risk factors $\gamma_{k,i}$, for the aggregated values weight group, age group, climate, and activity code was calculated from computing the product of the frequency and severity factors.

Base Level Estimation

Once the risk factors corresponding to each of the variables (weight group, age group, climate, activity code) had been calculated for all the data corresponding to years between 2006-2016, one could determine the base level γ_0 corresponding to the insurances active in year 2016. The base level (or base rate) is a key component in insurance pricing and is scaled according to the risk factors associated with the particular group that the tractor belongs to. The goal was find a base level so that the ratio between the estimated claim cost and the total insurance income was 90%. The estimated claim cost was calculated by averaging the the total claim cost, and the insurance income was computed by dividing the average yearly cost by 0.9. Each risk factor was multiplied together and used in the calculating the income for each group by taking the product of the total factors with duration and an initially set base level. The total income could be computed and subtracted from the previously commuted target value, and a base rate was able to be estimated based on the idea that the difference between the total income and target was needed to be as small as possible.

Model validation-Likelihood Ratio Test

Once a model is constructed, it is of importance to be able to determine whether or not one's obtained results are significant or not, or in other words how good one's model is. In order to validate the GLM model, a Likelihood Ratio Test was performed. The likelihood ratio test has advantages over other goodness of fit test in its characteristics of sensitivity and specification. The likelihood ratio test compares a full model with a reduced model of interest, to see if the full model performs better than the reduced model. it follows the X^2 distribution and the test is significant if:

$$LR \geqslant X^2_{\alpha}(n-r) \tag{4}$$

Results

class	Rating Factor	Duration	n.claims	Rels.frequency	Rels.severity	Rels.risk
<=600kg	Weight	21995.8	79	1	1	1
600-1000kg	Weight	13753.4	124	2.9258079	0.9460941	2.76809
1000-2100kg	Weight	15672.8	219	4.5057035	1.6304167	7.34617
2100-7000kg	Weight	2228.47	48	6.2391992	2.836567	17.6979
>=7000kg	Weight	780.677	18	7.5196255	4.146397	31.1794
<=6	Age	3311.28	52	1.0975041	1.8367326	2.01582
from 6-9	Age	18618.9	219	1	1	1
from 9-17	Age	12950.6	103	0.620886	1.2950394	0.80407
from 17-30	Age	17292.3	106	0.3644374	0.955808	0.34833
>=30	Age	2258.21	8	0.2328346	1.1224586	0.26135
Middle	Climate	21991.9	188	0.9712844	0.7531184	0.73149
North	Climate	8887.61	89	1.1374939	0.7884647	0.89687
South	Climate	23551.7	211	1	1	1
A	ActivityCode	9530	106	1.171064	0.785407	0.91976
C	ActivityCode	1324.35	13	1.1339915	1.0245641	1.16185
F	ActivityCode	2504.51	44	1.9820249	1.6938198	3.35719
G	ActivityCode	1353.13	14	1.331622	1.0473592	1.39469
H	ActivityCode	1245.15	19	1.6302859	0.8691127	1.4169
I	ActivityCode	475.757	2	0.7045579	0.8313599	0.58574
L	ActivityCode	5639.04	48	1.485423	0.8980681	1.33401
M	ActivityCode	749.244	5	0.8110845	1.279293	1.03761
Missing	ActivityCode	27273.9	151	1	1	1
N	ActivityCode	2661.54	66	3.1592772	0.8261276	2.60997
Other	ActivityCode	1674.69	20	1.2971472	1.1352575	1.4726

Figure 1: Summary of glmdata3

The figure summarizes the results for the GLM analysis. The relative frequency and

relative severity factor for each class was used to calculate the risk factor for each group, which was further used in the calculations of the base level. The optimal base level for a 90% premium target representing approximately 140378 SEK, was estimated to be around 95.10 SEK, giving a total insurance income 1405078 SEK.

Model Validation

The likelihood ratio test was computed both in terms of model frequency and model severity using a built in function in R. Both models had a p-value of $2.2 \cdot 10^{-16}$, indicating that the difference was very significant for all significance levels.

Conclusion

The GLM model was used to capture the risk of different tractors and find a base price for their insurance, each tractor has different features, which were analyzed and tractors with similar risks were grouped together. By performing this grouping we were able to find a low base level of premium by fitting the model with those groupings. Risk factors modify the base level of premium increasing the cost for the riskiest tractors to ensure a profitable insurance company. The model was validated with Likelihood ratio test, confirming the significance of the model constructed.