SF2930 Regression analysis VT2018

Project 2

The project should be done in groups of **two**.

A computer written[1] report should be handed in no later than **2018-03-02**, by email to `dabergl@math.kth.se`. The subject of the email should be "SF2930 Project 2: Full Name 1, Full Name 2"

In addition to this, send in your resulting factors on the form given by the template TariffFactors.xlsx found on the course site by email to `dabergl@math.kth.se`, no later than **2018-02-26**. The subject of the email should be "SF2930 Project 2: Full Name 1, Full Name 2"

Introduction

A tractor is a vehicle designed to deliver a high torque at slow speeds, mostly used in agriculture or construction. In Sweden, most of these vehicles 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.

If P&C seeks your help to create a model that prices this insurance as risk correct as possible. In other words, you are going to make your own tractor tariff on 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 factors corresponding to variable number k and variable group number i. $\gamma_{k,i}$ will take different values for each individual tractor, depending on its characteristics. For example, let $k = 1$ be Vehicle age and for one particular tractor the age is 3 years old. Then, according to below table, $\gamma_1 = 0.95$.

VehicleAge group i	Risk factor $\gamma_{1,i}$
1: Age <= 1	1.00
2: Age = 2	0.98
3: Age = 3	0.95
4: Age = 4	0.90
5: Age >= 5	0.85

[1] Preferably using LaTeX

Material

1. Dataset

The file Tractors.csv contains information on all tractors with a vehicle damage insurance in If P&C during 2004-2014, including claims history. The file has one row per tractor and Risk year, as shown in the table below.

RiskYear	VehcleAge	Weight	Climate	ActivityCode	Duration	NoOfClaims	Claim cost
2010	009	3830	North	Construction	0.63	1	627 099
2008	001	400	South	Missing	0.59	1	253 850
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Here, Risk year is the year of the insurance period, Vehicle age and Weight denote the age and weight of the tractor, respectively, Climate is the geographical location in Sweden where the tractor is used and Activity code is the activity code registered on the company that owns the tractor. For each tractor, there is also information regarding Duration. This is the share of the risk year the tractor was insured. For example, a tractor with a one year insurance policy from 2013-07-01 to 2014-06-30 will be represented by two rows in the data; one with Risk year = 2013 and one with Risk year = 2014, both with Duration = 0.5. Finally, the number of claims and claims cost corresponding to the insurance period are denoted by NoOfClaims and ClaimCost.

2. GLM program

The template GLM.R contains a structure for a GLM analysis.

Tasks

1. Grouping and risk differentiation

Perform a GLM analysis to figure out how best to describe the risk for the tractors. Use the template GLM.R. The outcome should be a multiplicative GLM model, as described in Eq. 1, that model claims frequency and claim severity separately. Use the same variables and variable groups in both models, and propose the final risk factor $\gamma_{k,i}$, where the final risk factor is the product of the claim frequency and the claim severity.

In order to perform your GLM analysis, you will have to group some of the variables. Consider, for example, the tractors' weights. These cover a very wide range, as tractors can be both very small and light, and extremely big and heavy. Thus, it would be impossible to analyze each individual weight alone; it is necessary to group them. When grouping a variable, there are two things to consider:

- Make each group "Risk homogenous", meaning that you believe that the risk does not vary much within the group, with regard to the particualr variable.

- Create groups with enough data to get a stable GLM analysis for each group. What is "enough" has no clear answer, bur varies, depending among other things on how many variables you use in your analysis.

Creating good groups is usually an iterative process, so try different ways to do it!

No dataset is perfect. You will find many rows with strange, missing, or incomplete data, and need to handle this. One good strategy is to put all these values in a group of its own, letting it get its own factor in the GLM analysis.

Present and explain your choice of risk arguments, grouping of data and risk factors. How does this comply with Likelihood Ratio Test and different measures for goodness of fit discussed in this course? Compare all tests and measures.

2. Levelling

Having found the risk factors $\gamma_{k,i}$, determine the base level γ_0. Note that γ_0 is estimated automatically by the GLM program, but this base level corresponds to the total claims cost of the analysis data, not the policies that are active today. To find the correct base level, follow the following two steps:

a. Estimate the expected claim cost for 2015. Assume that If P&C has a ratio target between the estimated claim cost and the total premium of 90% – what should the total sum of tractors' premium be to accommodate this target?

b. For each insurance, calculate the "total risk factor" - i.e. the product of all risk factors $\gamma_{k,i}$, for that insurance. Then find the 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 your result from part a.

All tariffs will compete against each other – so make sure to describe the risk as well as you can!

Good luck!