Questions and answers – Project 2

Here are some commonly answered questions + answers from the computer exercise for Project 2, 1st of March. Thanks for some good discussions today, and good luck with the project!

Question: What is duration?

Answer: The share of the RiskYear that a customer had the insurance policy. For example, a customer with a 1 year insurance policy from 2013-07-01 ␣ 2014-06-30 will be represented by two rows in the data: one with RiskYear = 2013 and one with RiskYear = 2014, both with duration = 0.5.

Question: How to group a variable?

Answer: You want to consider two things:

1. Have a "Risk homogenous" group, meaning that you believe that the risk behavior within the group doesn't vary too much.
2. Have enough data (duration), in order to get a stable GLM estimate for that group.

Approach (1) means you usually want many, small groups, but you still need enough data (2) in each group. What is "enough" has no definite answer, but you should think about both aspects 1 and 2.

Question: How set the "base level", gamma0, in part 2 of the project.

Answer: Don't think about gamma0 in the first part of the project, when you decide your base factors. When you reach part 2 of the project, having a risk factor for each group of each variable, you want to:

1. Give each customer a "total risk factor" = risk factor_var1 * risk factor_var2 *…
2. Find a constant, gamma0, which turns the total risk factor into an actual price. The total sum of all prices for the customers that are with us in 2015 should cover all our expected claims costs plus some more (you want expected claims cost / total premium = 90%).

What will be the expected total claims cost is up for to you to estimate/guess, based on earlier years.

Question: What does offset(log(duration)) mean in the frequency GLM code?

Answer: Don't think about this, let it be there. In short, you are modelling Claims frequency = Y = X/w, where X = Total claims cost and w = duration. This means that ln(E(Y)) = ln(E(X/w)) = ln(E(X)) – ln(w). The model actually models E(X) rather than E(Y), but by adding the offset(log(duration)) we are able to consider the weight (duration) and therefore create a model for E(Y) instead.

Question: How to handle "Missing" or strange values?

Answer: One approach is to remove all of these rows from the data. BUT, this may lead to removing too many rows ␣ little data for modelling. Keep in mind that when you remove a row because one of

the variables had a strange value, you also remove all valuable info from the other variables, which may have good values. Therefore, a safer method is to put all of these “strange” values in a group itself, and let it get a factor a factor itself. This may complicate the GLM estimation a bit, but is usually better than removing the data completely.

Question: What does the factor interval in the excel mean, for the continuous variables? (such as Age, Weight…)

Answer: Instead of just putting one factor per group for the continuous variable, you might assume that there is a continuous trend (for example a linear trend), which also applies *within* the group itself. Therefore you can use the group estimate as an “average” factor for this group, and give the group a start- and an end factor instead, where the group estimate is somewhere in the middle. Remember that this is not an exact science!