GLM for claim frequency modelling

Vermeir Jellen

December 21, 2015

Contents

1 Introduction **2**

2 Claim frequency modelling **2**

2.1 Claim frequency data - Descriptive Statistics 2

2.2 GLM calibration: Age-gender interaction 7

2.3 GLM calibration: Include all covariates 8

1 Introduction

In an insurance context it is important to classify risk factors in an appropriate manner. The amount of risk that is carried by an insured depends on a number of external covariates that can each be classified in a number of categories. In such a heterogeneous portfolio, the financial result depends on the composition of risk factors inside the portfolio. Hence, we must also incorporate these risk factors into the modeling process of the claim frequency (and severity) process.

In this paper we investigate claim frequency data and build regression models to explain the number of claims that are dependent on a number of external covariates.

2 Claim frequency modelling

In this chapter we first investigate the properties of claim frequency data and the respective dependent covariates under consideration. We provide the reader with a descriptive analysis of the data and investigate possible interactions between some of the covariates. Next, GLM models are calibrated to the claim frequency data and drop-in deviance analysis and Wald tests are performed to evaluate significance of the model parameters. We obtain a parsimonious model that captures the claim frequency data and its dependency on the external covariates.

2.1 Claim frequency data - Descriptive Statistics

In this subsection we investigate a sample of insurance policies. For each individual policy, the amount of claims, the start and end date of the policy and information about a number of external covariates is reported. The exposures of the individual policies are calculated and a frequency table with the amount of claims and the total corresponding exposure is reported in table 1 below. The corresponding mean annual claim frequency for the complete dataset is 0.227.

Numer of claims	Number of policies	Total exposure
0	33386	24025.86
1	4957	4007.07
2	646	568.11
3	78	69.4
4	7	6.23
5	1	0.98
Total	39075	28677.66

Table 1: Claim amounts - frequency table

Figures 1 and 2 display the covariates and their categories together with their corresponding amount of policyholders. Note that we have manually split up the age, duration and vehicle year covariates in smaller amounts of discrete subclasses. We have done this by using their $p = [0.025, 0.10, 0.25, 0.5, 0.75, 0.9, 0.975]$ quantiles as category boundaries.

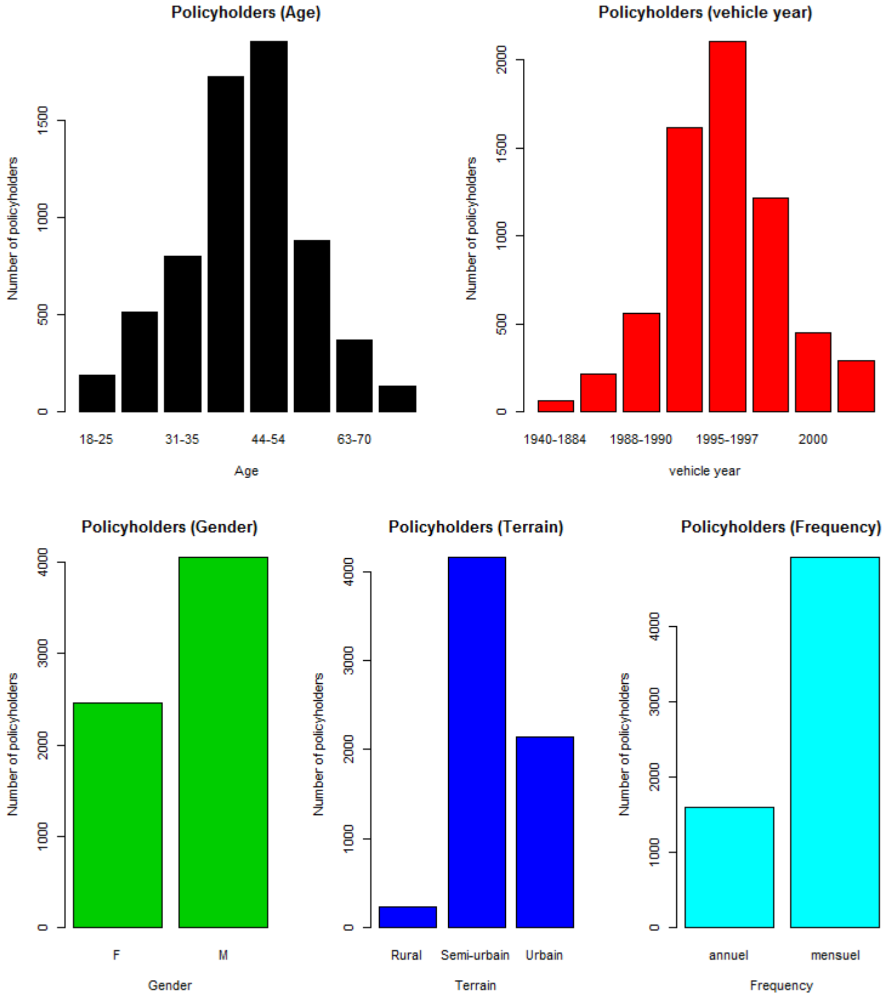

Figure 1: Covariate categories 1 - Amount of policyholders

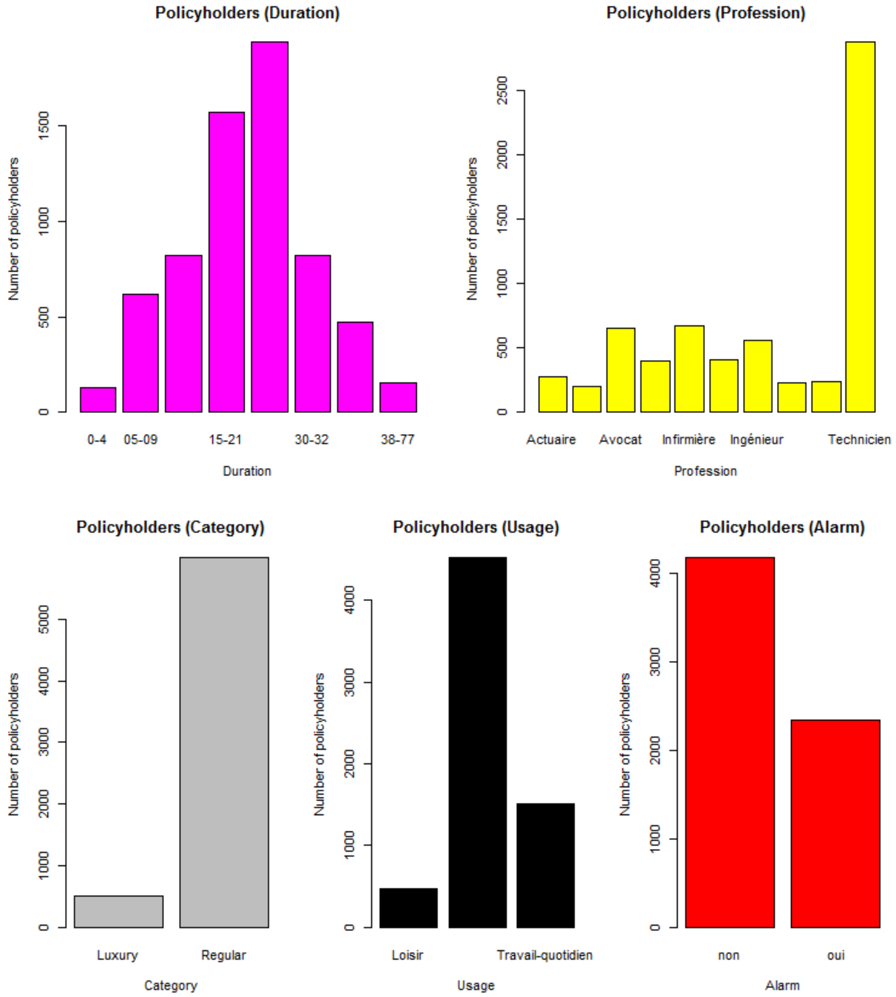

Figure 2: Covariate categories 2 - Amount of policyholders

Figures 3 and 4 display the annual mean claim frequencies for the covariate categories.

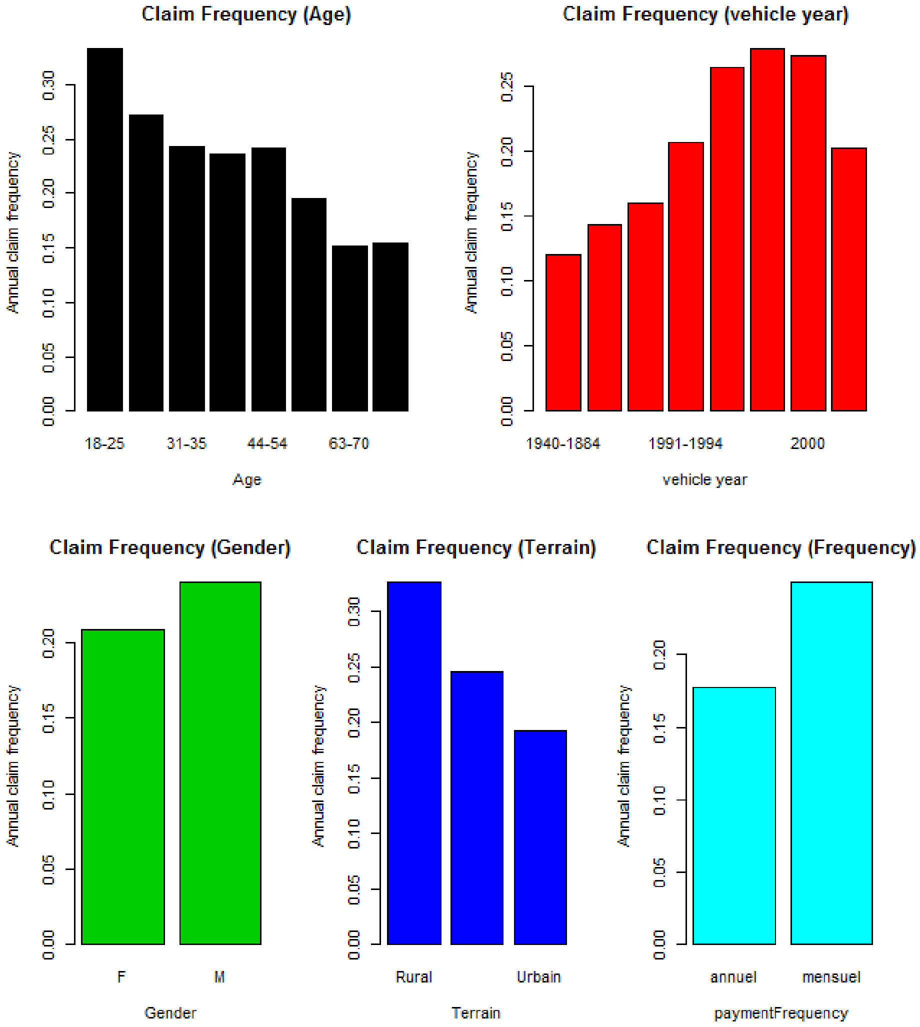

Figure 3: Covariate categories 1 - Mean annual claim frequency

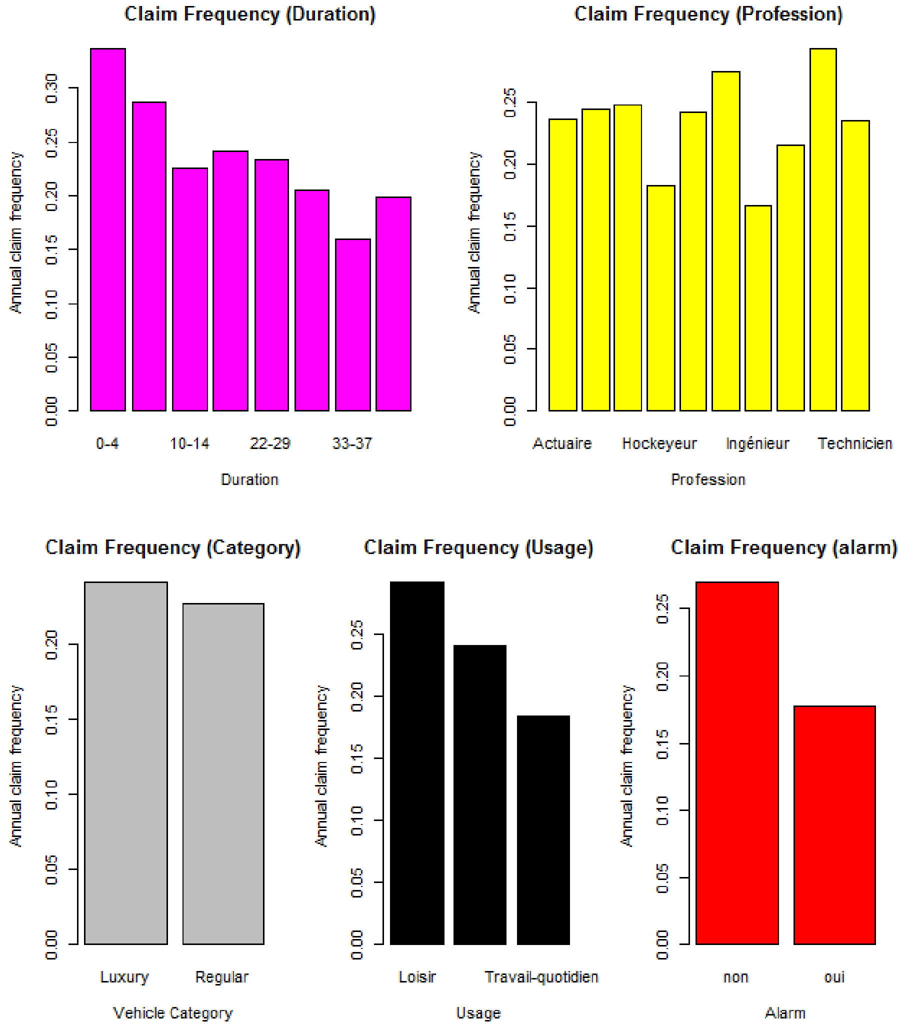

Figure 4: Covariate categories 2 - Mean annual claim frequency

A known problem with these one-way analyses is that they can be distorted by correlations. When two variables interact, the effect of one factor varies depending on the levels of another factor. For example, empirical observations have suggested that the effect of age on the average claim frequency is different between males and females.

Figure 5 illustrates the gender-age interaction for our data sample. We notice that younger men seem to posses a significantly higher risk in comparison to young females, while the difference becomes negligable at older ages. Hence, it might be necessary to take the age-gender interaction into account explicitly.

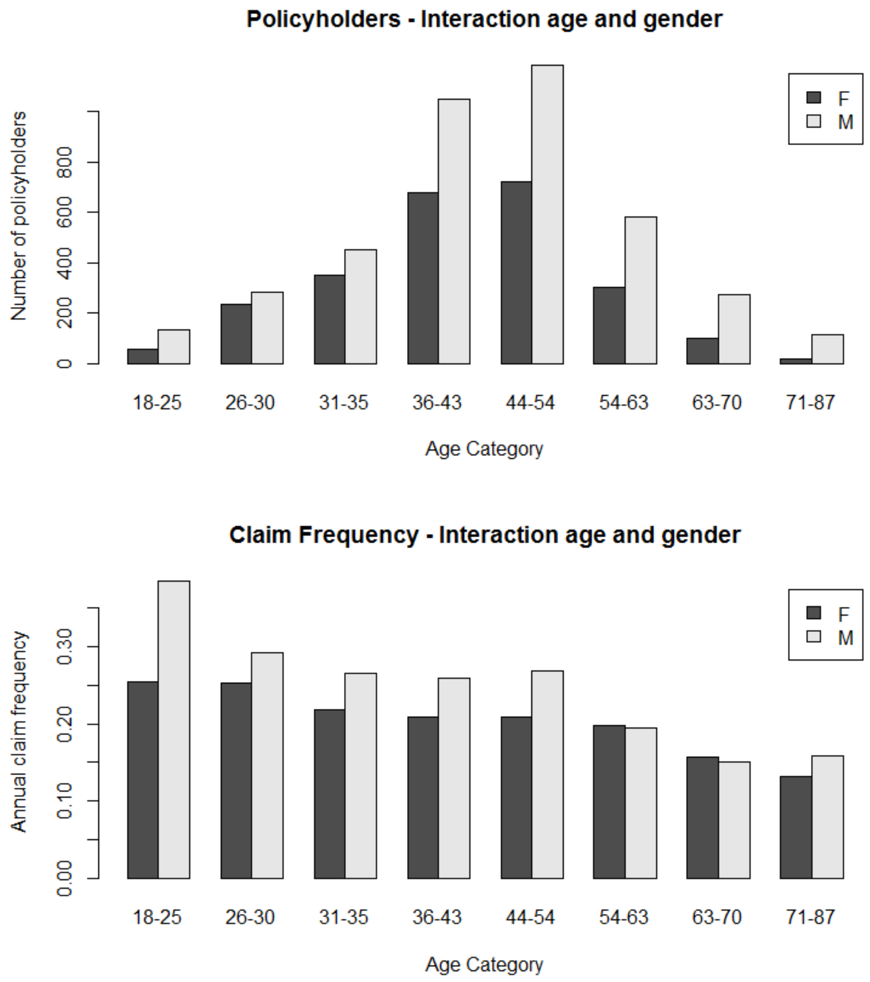

Figure 5: Interaction age and gender

2.2 GLM calibration: Age-gender interaction

In this subsection we calibrate simple GLM models to the claim count data and evaluate their likelihood via a 'Type 1' drop-in-deviance analysis. Our goal here is to determine if the age-gender interaction is a necessary (significant) component in our final model.

We start with the calibration of the trivial model, which has a deviance of 25545. We subsequently add the gender covariate to the initial model and receive a drop in deviance of 28.917 in comparison to the original model. A chi-square test shows that the likelihood of the model entails a significant improvement

($p < 0.001$) and the covariate should thus be kept in the model. In the next step we add the age category variable to the model. A chi-square test again shows that this new variable should be kept inside the model. Further inspection reveals that all age category parameters contain significant parameter values and should be retained. However, we note here that some of the parameter values might not be significantly different from eachother and categories can potentially be merged together.

Finally, we add the interaction of the age and gender categories to the model. A chi-square test shows that the likelihood of the more complex interaction model produces a significant improvement in likelihood in comparison to the previous model ($p < 0.05$). Results of the analysis are summarized in table 2.

Model	Dev.	Δ Dev.	DF	Δ DF	$1 - pchisq(\Delta Dev., \Delta DF)$
Trivial	25545		39074		
sex	25516	28.917	39073	1	< 0.0001
agecat	25328	187.695	39066	7	< 0.0001
sex:agecat	25313	15.745	39059	7	0.028

Table 2: Drop-in-deviance analysis

Hence, we conclude that we should incorporate the interactions between the age and gender covariates in order to adequately capture their correlation structure in the model. For this purpose, we have created a new dummy interaction explanatory variable that corresponds to the class divisions as they are illustrated in figure 5.

2.3 GLM calibration: Include all covariates

Our goal in this subsection is to obtain an optimal and most parsimonious claim frequency model that includes all relevant and significant explanatory variables. We start with the calibration of a GLM model that incorporates all the covariates that are displayed in figures 1 and 2. However, we add our agesex interaction category dummy variable instead of the two original age and sex variables. Lines 342-354 in the R demo calibrates the model and prints the corresponding output. We notice that the model contains many parameters that are insignificant and many more that can potentially be grouped together. Here, we perform an iterative analysis, one variable at a time.

Agesex interaction variable There are 16 agesex categories, of which the first one can be considered the 'default reference class'. Hence, there are 15 explicit agesex parameters left in the model. We start by performing a Wald test with null hypothesis stating that the values of the 9th and 10th agecategory parameters are identical. The hypothesis can not be rejected ($p > 0.05$), thus we merge the categories in such a way that they now correspond to one and the same parameter in the model. We repeat the process a few more times by merging parameters that are not significantly different from eachother, until we end up at model $g5$. This model contains 8 parameters for the agesex category, of which 7 are explicitly defined. However, only one of these parameter values is significantly different from 0 and the other ones can thus be merged with the default category. Finally, we end up with only two remaining categories for the agesex classification. More precisely, the default category corresponds to females of all ages plus males that are older than 54, while the explicit parameter corresponds to males aged 18-54.

The analysis in table 3 shows that the value of the explicit parameter in the final model is 0.197, which indicates that the insureds that fall in this group correspond to a higher risk than the default reference class. In figure 5 we indeed notice that males < 54 years old pose a higher risk in comparison to the other subcategories that correspond to the reference class.

Vehicle year category We perform a similar procedure on the vehicle year covariate and end up with three significant explanatory classes. The default class corresponds to years 1945-1991 and 2001-2003. The first explicit parameter corresponds to years 1991-1995 and has a value of 0.21, which entails added risk in comparison to the reference class. The second parameter has an even higher value of 0.36 and corresponds to years 1995-2001. The vehicle year graph in figure 1 illustrates the consistency of these results.

Terrain category The terrain subclass parameters are all significantly different from 0 and significantly different from eachother and hence can not be merged together. The semi-urban and urban class parameter values are -0.205 and -0.391 respectively, which indicates that they imply significantly less risk than the rural default class.

Duration category During the analysis we are able to merge a few categories together and end up with 4 significantly different classes. The default class corresponds to durations 0-5. The parameter values for the other three explicit variables are -0.23, -0.47 and -0.73. The second class corresponds to durations 5-10, the third class corresponds to 10-33 plus 38-70 and the fourth class corresponds to 33-38.

Vehicle and Profession category None of the parameters for these covariates are significantly different from 0 ($p > 0.05$) so they are removed from the model altogether.

Usage category All three usage class parametervalues are significant. Both the work usage classes correspond to lower risk in comparison to the leisure reference class. Parametervalues are -0.18 and -0.27 for the 'occasional work' and 'daily work' usage classes respectively.

Alarm category The alarm class is significant in the model and the explicit parameter value that corresponds to the 'yes' class implies a lower claim frequency than the 'no' reference class. The parametervalue is -0.28.

Paymentfrequency category The paymentfrequency covariate is significant in the model and the 0.264 parameter value for the 'monthly' class implies a higher claim frequency rate in comparison to the 'yearly' reference class.

Final model A summary of the modelparameters is reported in table 3 below.

Coefficient	Value	p
Intercept	-0.97268	< 0.001
agesex	0.197	< 0.001
pay. freq. monthly	0.264	< 0.001
terrain.semi.urban	-0.205	0.003
terrain.urban	-0.391	< 0.001
usage.w.occasional	-0.18	< 0.001
usage.w.dailyl	-0.27	< 0.001
alarm.yes	-0.28	< 0.001
duration.2	-0.23	0.018
duration.3	-0.47	< 0.001
duration.4	-0.73	< 0.001
vehicleyear.2	0.21	< 0.001
vehicleyear.3	0.36	< 0.001

Table 3: Final model - Parameter Values