PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 428

Wrocław Conference in Finance: Contemporary Trends and Challenges

Publishing House of Wrocław University of Economics
Wrocław 2016

Copy-editing: Marta Karaś
Layout: Barbara Łopusiewicz
Proof-reading: Barbara Cibis
Typesetting: Małgorzata Czupryńska
Cover design: Beata Dębska

Information on submitting and reviewing papers is available on websites
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
e- ISSN 2392-0041

ISBN 978-83-7695-583-4

The original version: printed

Publication may be ordered in Publishing House
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław

tel./fax 71 36-80-602; e-mail: econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Printing: TOTEM

Contents

Introduction

One of the fastest growing areas in the economic sciences is broadly defined area of finance, with particular emphasis on the financial markets, financial institutions and risk management. Real world challenges stimulate the development of new theories and methods. A large part of the theoretical research concerns the analysis of the risk of not only economic entities, but also households.

The first Wrocław Conference in Finance WROFIN was held in Wrocław between 22nd and 24th of September 2015. The participants of the conference were the leading representatives of academia, practitioners at corporate finance, financial and insurance markets. The conference is a continuation of the two long-standing conferences: INVEST (Financial Investments and Insurance) and ZAFIN (Financial Management – Theory and Practice).

The Conference constitutes a vibrant forum for presenting scientific ideas and results of new research in the areas of investment theory, financial markets, banking, corporate finance, insurance and risk management. Much emphasis is put on practical issues within the fields of finance and insurance. The conference was organized by Finance Management Institute of the Wrocław University of Economics. Scientific Committee of the conference consisted of prof. Diarmuid Bradley, prof. dr hab. Jan Czekaj, prof. dr hab. Andrzej Gospodarowicz, prof. dr hab. Krzysztof Jajuga, prof. dr hab. Adam Kopiński, prof. dr. Hermann Locarek-Junge, prof. dr hab. Monika Marcinkowska, prof. dr hab. Paweł Miłobędzki, prof. dr hab. Jan Monkiewicz, prof. dr Lucjan T. Orłowski, prof. dr hab. Stanisław Owsiak, prof. dr hab. Wanda Ronka-Chmielowiec, prof. dr hab. Jerzy Różański, prof. dr hab. Andrzej Sławiński, dr hab. Tomasz Słoński, prof. Karsten Staehr, prof. dr hab. Jerzy Węcławski, prof. dr hab. Małgorzata Zaleska and prof. dr hab. Dariusz Zarzecki. The Committee on Financial Sciences of Polish Academy of Sciences held the patronage of content and the Rector of the University of Economics in Wroclaw, Prof. Andrzej Gospodarowicz, held the honorary patronage.

The conference was attended by about 120 persons representing the academic, financial and insurance sector, including several people from abroad. During the conference 45 papers on finance and insurance, all in English, were presented. There were also 26 posters.

This publication contains 27 articles. They are listed in alphabetical order. The editors of the book on behalf of the authors and themselves express their deep gratitude to the reviewers of articles – Professors: Jacek Batóg, Joanna Bruzda, Katarzyna Byrka-Kita, Jerzy Dzieża, Teresa Famulska, Piotr Fiszeder, Jerzy Gajdka, Marek Gruszczyński, Magdalena Jerzemowska, Jarosław Kubiak, Tadeusz Kufel, Jacek Li-

sowski, Sebastian Majewski, Agnieszka Majewska, Monika Marcinkowska, Paweł Miłobędzki, Paweł Niedziółka, Tomasz Panek, Mateusz Pipień, Izabela Pruchnicka-Grabias, Wiesława Przybylska-Kapuścińska, Jan Sobiech, Jadwiga Suchecka, Włodzimierz Szkutnik, Mirosław Szreder, Małgorzata Tarczyńska-Łuniewska, Waldemar Tarczyński, Tadeusz Trzaskalik, Tomasz Wiśniewski, Ryszard Węgrzyn, Anna Zamojska, Piotr Zielonka – for comments, which helped to give the publication a better shape.

Wanda Ronka-Chmielowiec, Krzysztof Jajuga

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 428 • 2016

Wrocław Conference in Finance: Contemporary Trends and Challenges ISSN 1899-3192
e-ISSN 2392-0041

Ján Gogola, Viera Pacáková
University of Pardubice
e-mails: jan.gogola@upce.cz; viera.pacakova@upce.cz

FITTING FREQUENCY OF CLAIMS BY GENERALIZED LINEAR MODELS[1]

MODELOWANIE CZĘSTOTLIWOŚCI ROSZCZEŃ ZA POMOCĄ UOGÓLNIONYCH MODELI LINIOWYCH

DOI: 10.15611/pn.2016.428.06
JEL Classification: C10, C38, C53, G22

Abstract: The article deals with the issue of creating homogenous tariff classes of non-life insurance and setting claim frequency of each tariff class. We use Generalized Linear Models (GLM) for the purpose of finding significant risk factors and also to determine the estimated claim frequency of the individual tariff classes. The theoretical part is completed by the application on a typical heterogeneous portfolio of the Motor Third Party Liability (MTPL). All calculations are performed using the R environment.

Keywords: Generalized Linear Models, Poisson regression, R language, frequency claims, risk classification.

Streszczenie: Artykuł omawia kwestię tworzenia jednorodnych klas taryfowych ubezpieczeń majątkowych i ustalenie częstotliwości roszczeń każdej klasy taryfowej. W celu znalezienia istotnych czynników ryzyka, a także w celu określenia szacunkowej częstości roszczeń poszczególnych grup taryfowych, wykorzystany został uogólniony model liniowy (GLM). Część teoretyczna została uzupełniona zastosowaniem tego modelu dla typowego niejednorodnego portfela ubezpieczeń OC. Wszystkie obliczenia przeprowadzono z użyciem środowiska programowania R.

Słowa kluczowe: uogólnione modele liniowe, regresja Poissona, język R, częstotliwość roszczeń, klasyfikacja ryzyka.

1. Introduction

In a homogeneous portfolio, where all policyholders have the same risk level, there is no reason to let the amount of premium vary. In practice, most portfolios are

[1] RNDr. Ján Gogola, Ph.D, prof. RNDr. and Viera Pacáková, Ph.D, both at the University of Pardubice (CZE), Institute of Mathematics and Quantitative Methods.

heterogeneous. They mix individuals with different risk levels. There is a need of risk classification [Denuit, Charpentier 2005].

Nowadays, it has become extremely difficult for insurance companies to maintain cross subsidies between different risk categories in a competitive setting. Therefore, actuaries have to design a tariff structure that will fairly distribute the burden of claims among policyholders. We apply *Generalized Linear Models* (GLM) to achieve risk classification [Cox 1972]. Ratemaking (or risk classification) is essentially about classifying policies according to their risk characteristics.

The classification variables are called *a priori variables*, as their values can be determined before the policyholder starts to be covered by the insurance company. In the Motor Third Party Liability (MTPL) insurance, they include the age, gender and occupation of the policyholder, the type and use of their car, etc. These observable characteristics are typically seen as non-random covariates. Even with all the covariates included in price lists, substantial risk differentials remain amongst individual drivers (due to hidden characteristics like temper and skill, aggressiveness behind the wheel, knowledge of the highway, etc.)

2. Modelling claims in insurance

The pure premium is the amount the insurance company should charge in order to be able to indemnify all the claims, without loss nor profit. The computation of the pure premium [Jee 1989] relies on a statistical model incorporating all the available information about the risk. The aim of ratemaking is to evaluate as possible the pure premium for each policyholder.

Usually, the total claims S generated by a policy of the portfolio is not the modelling target. Instead, the different components of S are modelled separately, such as: frequency claims, standard claims costs, cost of large claims, etc. This allows for a better understanding of the price list as the risk factors influencing each component of S are isolated.

The total claim amount S_i generated by policyholder i can generally be decomposed as:

$$S_i = \sum_{k=1}^{N_i} C_{ik} + J_i \cdot L_i, \tag{1}$$

where: N_i is the number of standard claims filed by policyholder i, C_{ik} is the cost of the k-th standard claim filed by policyholder i, J_i indicates whether the policy i produced a large claim (at least), L_i is the cost of this large claims, if any.

If insurance data is subdivided into risk classes determined by many a priori variables, actuaries work with figures which are small in exposure and claim numbers (it is even possible that no observations are available for a particular combination of

the rating factors). Hence, simple average will be suspect and a regression model is required.

2.1. Generalized Linear Models (GLM)

Generalized Linear Models (GLM) [Nelder, McCullagh 1989] are ideally suited to the analysis of non-normal data which insurance analysts typically encounter. The GLM are used to assess and quantify the relationship between a *response variable* (or dependent variable) and a set of possible *explanatory variables* (or independent variables). GLM is important in insurance applications as:

- the assumption of normality is often not applicable, for example claim counts, claim sizes or claim occurrences on a single policy do not obey the Gaussian distribution,
- the relationship between outcomes and explanatory variables is often multiplicative rather than additive.

With the GLM, the variability in one variable is explained by the changes in one or more other variables. The variable being explained (claim count, claim cost, etc.) is called the *response variable*. The variables that are doing the explaining are the *explanatory variables*, also called in insurance *risk factors* or *risk characteristics*.

GLM describe the connection between the response and the explanatory variables. The explanatory variables may be, and often are, related. A question arises which explanatory variables are predictive of the response, and what is the appropriate scale for their inclusion in the model?

2.2. Frequency model – Poisson regression for claim counts

Our explanatory variables are assumed to be categorical. A categorical variable with k levels separates the portfolio into k classes. It can be coded via k–1 binary variables being all zero for the reference level. The *linear predictor* (or *score*) for each class is given by the linear combination:

$$\beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_p \cdot x_{ip} = \boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}_{\boldsymbol{i}}, \tag{2}$$

where: $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector containing the parameters, β_0 is called the intercept, $\boldsymbol{X}_{\boldsymbol{i}} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector containing the explanatory variables (or observable characteristics) for the i-th policyholder.

The annual expected claim frequency (in a Poisson model with *log* link function) for each class is given by:

$$\exp\left(\boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}\right) = \exp\left(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_p \cdot x_p\right), \tag{3}$$

The intercept β_0 representing the risk associated with the reference class (for which $x_1 = x_2 = \cdots = x_p = 0$) and $\exp(\beta_0)$ is the annual expected claim number for a policy in the reference class.

When all explanatory variables are categorical, each policyholder is represented by a vector with components equal to '0' or '1'. The annual expected number of claims is then equal to:

$$\exp\left(\boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}_{\boldsymbol{i}}\right) = \exp(\beta_0) \cdot \prod_{j=1}^{p} \exp(\beta_j \cdot x_{ij}) = \exp(\beta_0) \cdot \prod_{j=1; x_{ij}=1}^{p} \exp(\beta_j), \tag{4}$$

where: $\exp(\beta_j)$ models the effect of the *j*-th ratemaking variable.

If $\beta_j > 0$ then $\exp(\beta_j)$ increases the annual expected claim number. The $\exp(\beta_j)$ is the multiplicative effect on the annual expected claim number (frequency) due to the covariate associated with β_j, while holding the other explanatory variables constant.

On contrary, if $\beta_j < 0$ then $\exp(\beta_j)$ decreases the annual expected claim number. Denote as *n* the *number of policies* in the portfolio. Let N_i be the *number of claims* filed by policyholder *i*, *i* = 1, 2, …, *n*.

The time in which each policyholder is covered during a year is not a constant. It depends on the time the policyholder entered into or left the insurance. This various time periods need to be explicitly incorporated in the Poisson regression model [Lambert 1992; Ter Berg 1980]. So let, d_i, be the length (duration) of the coverage period (or *exposure to risk*).

Poisson regression for independent counts N_i; *i* = 1, 2, …, *n*, is based on

$$N_i \sim Po\left(d_i \cdot \exp(\boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}_{\boldsymbol{i}})\right), \; i = 1, 2, \ldots, n. \tag{5}$$

Then we can express the regression equation as:

$$\log \lambda_i = \log d_i + \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_{ij}, \tag{6}$$

where: $\log d_i$ is called the offset.

The offset takes a specific value for each policyholder, and there is no parameter associated with it to estimate. The predicted expected number of claims for the *i*-th policyholder is:

$$\hat{\lambda}_i = d_i \cdot \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \cdot x_{ij}\right). \tag{7}$$

Prediction in this sense means guessing the expected number of claims (i. e. the response) from the values of the explanatory variables in an observation. We are

estimating the parameters β_i by the maximum likelihood approach. The goodness of fit is measured by *deviance* (the smaller, the better).

The deviance can also be used to compare the fit of two models by taking the difference in the deviances. The difference in the deviance, between the more complex model (*full model*) D_F and the deviance of the simpler model (*reduced model*) D_R with some parameters dropped out, can also be used to test the null hypothesis that the additional parameters in the full model are equal to zero. Let us test the null hypothesis:

$$H_0\text{: } \boldsymbol{\beta} = \boldsymbol{\beta_0} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_q)^T,$$
$$H_1\text{: } \boldsymbol{\beta} = \boldsymbol{\beta_1} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^T.$$

That is, H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$.

The difference in the deviances is a χ^2 distributed variable (under the null hypothesis that the additional regression parameters are equal to zero) with degrees of freedom equal to the difference in the number of regression parameters between the full and the reduced model, or equivalently the number of additional parameters in the full model:

$$D_R - D_F \sim \chi^2_{p-q}. \tag{8}$$

We can formally test the hypothesis that the additional parameters in the full model are zero, by using the difference of the deviances in a formal statistical test. H_0 is rejected if $D_R - D_F$ is "too large", that is if $D_R - D_F > \chi^2_{p-q}(1-\alpha)$. If the χ^2 is statistically significant, then we accept the full model. If it is not significant, we accept the reduced model. The goal of our regression analysis is to find a set of explanatory variables that have high explanatory power as measured through goodness of fit.

3. Results

Our data set is based on one-year vehicle insurance policies of the Motor TPL (Third Party Liability) portfolio. There are 163 657 policies, of which 18 345 produced at least one claim. The analysis is performed with the help of the *GLM* procedure of R language [R Core Team 2015] and R packages "car" [Fox, Weisberg 2011], "epicalc" [Chongsuvivatwong 2012], "gmodels" [Warnes 2013].

First we eliminate atypical extreme losses. We have chosen a threshold of 100 000 units of currency. The reference class is composed of the modalities of the variables with the largest risk-exposure. Then we test whether a particular category is significant or not. We start for a model incorporating all the available information and then exclude the irrelevant explanatory variables.

Table 1. Description of variables with their modalities

Variable	Description with modalities
duree	Length of the coverage period (or exposure to risk)
nbrtotc	Number of claims
chargtot	Total claim amount
agecar	Age of the vehicle: 0-1, 2-5, 6-10, >10
sexp	Sex of the driver: Male or Female
fuelc	Type of fuel: Petrol or Gasoil
split	Split of the premium: Monthly, Once, Thrice, Twice
usec	Use of the vehicle: Private or Professional
fleetc	Vehicle belonging to a fleet: Yes or No
sportc	Sports car: Yes or No
coverp	Coverage: MTPL, MPTL+, MPTL+++
powerc	Power of the vehicle: <66 kW, 66-110 kW, >110 kW

Source: Authors' own study.

The *p*-value tests the relevance of the variable. The limit of 5% is usually used to decide on this relevance.

```
> Anova(GLM_Analysis, test.statistic="Wald",type=3,singular.ok=TRUE)
          Analysis of Deviance Table (Type III tests)

          Response: Data[["nbrtotc"]]
                                Df      Chisq Pr(>Chisq)
          (Intercept)            1 17239.0084  < 2.2e-16 ***
          Data[["sexp"]]         1    41.1735  1.393e-10 ***
          Data[["usec"]]         1     0.0347   0.852202
          Data[["fleetc"]]       1     8.7443   0.003106 **
          Data[["sportc"]]       1     6.5602   0.010429 *
          Data[["coverp"]]       2    97.1003  < 2.2e-16 ***
          Data[["fuelc"]]        1   186.5374  < 2.2e-16 ***
          Data[["agecar"]]       3    59.9103  6.143e-13 ***
          Data[["powerc"]]       2    36.3138  1.302e-08 ***
          Data[["split"]]        3   708.0591  < 2.2e-16 ***
          Residuals         163624
          ---
          Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We remove "usec" (*p*-value equals to 0.8522)

```
> Anova(GLM_Analysis, test.statistic="Wald",type=3,singular.ok=TRUE)
           Analysis of Deviance Table (Type III tests)

           Response: Data[["nbrtotc"]]
                                Df      Chisq Pr(>Chisq)
           (Intercept)           1 17275.4526  < 2.2e-16 ***
           Data[["sexp"]]        1    41.1499  1.410e-10 ***
           Data[["fleetc"]]      1     8.7150   0.003156 **
           Data[["sportc"]]      1     6.5481   0.010499 *
           Data[["coverp"]]      2    97.1279  < 2.2e-16 ***
           Data[["fuelc"]]       1   187.7628  < 2.2e-16 ***
           Data[["agecar"]]      3    59.8762  6.247e-13 ***
           Data[["powerc"]]      2    36.8955  9.733e-09 ***
           Data[["split"]]       3   708.1961  < 2.2e-16 ***
           Residuals        163625
```

With the help of Anova analysis, we observe that all the variables are significant. However, all the modalities do not need to be significant. Therefore, we try to gather the modalities using the `fit.contrast` procedure.

We first look at the confidence interval of the predictions.

```
> confint.default(GLM_Analysis,level = 0.95)
                                    2.5 %      97.5 %
(Intercept)                   -2.22603398 -2.16062057
Data[["sexp"]]B/Female         0.07105812  0.13358376
Data[["fleetc"]]B/Yes         -0.21186027 -0.04279222
Data[["sportc"]]B/Yes          0.04224610  0.31872533
Data[["coverp"]]B/MTPL+       -0.18584780 -0.11862913
Data[["coverp"]]B/MTPL+++     -0.21461854 -0.11864791
Data[["fuelc"]]B/Gasoil        0.17831654  0.23784190
Data[["agecar"]]B/0-1          0.13056022  0.26211704
Data[["agecar"]]B/2-5         -0.08943254 -0.01866191
Data[["agecar"]]C/>10         -0.03098823  0.04221339
Data[["powerc"]]B/66-110       0.05943024  0.12516634
Data[["powerc"]]C/>110         0.08711041  0.36379976
Data[["split"]]B/Thrice        0.47387455  0.56768884
Data[["split"]]B/Twice         0.19440914  0.25986695
Data[["split"]]C/Monthly       0.37205106  0.45539893
```

The process of gathering can be summarized as follows:

- Variable coverp: try to gather MTPL+ and MTPL+++ as the predicted value of the first one is in the confidence interval of the other one.
- Variable split: No overlapping of the confidence intervals.
- Variable powerc: 60-110 and >110 are overlapping.
- Variable agecar: some modalities are not significant; this means that we try to gather them:

```
    > fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,1,0,0))
                                 Estimate Std. Error  z value    Pr(>|z|)
Data[["agecar"]] c=( -1 1 0 0 ) 0.1963386 0.03356103 5.850197 4.90992e-09
> fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,0,1,0))
                                 Estimate Std. Error   z value    Pr(>|z|)
Data[["agecar"]] c=( -1 0 1 0 ) -0.0540472 0.01805407 -2.993632 0.00275678
> fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,0,0,1))
                                   Estimate Std. Error   z value  Pr(>|z|)
Data[["agecar"]] c=( -1 0 0 1 ) 0.005612582 0.01867423 0.3005523 0.7637559
> fit.contrast(GLM_Analysis,Data[["agecar"]],c(0,-1,1,0))
                                Estimate Std. Error   z value     Pr(>|z|)
Data[["agecar"]] c=( 0 -1 1 0 ) -0.2503858 0.0329191 -7.606106 2.82477e-14
> fit.contrast(GLM_Analysis,Data[["agecar"]],c(0,-1,0,1))
                                Estimate Std. Error   z value     Pr(>|z|)
Data[["agecar"]] c=( 0 -1 0 1 ) -0.190726 0.03567375 -5.346398 8.97220e-08
> fit.contrast(GLM_Analysis,Data[["agecar"]],c(0,0,-1,1))
                                 Estimate Std. Error  z value    Pr(>|z|)
Data[["agecar"]] c=( 0 0 -1 1 ) 0.0596598 0.02165873 2.754538 0.005877503
```

The modalities ">10" and "6-10" of "agecar" should be gathered (as ">5") as the *p*-value equals 0.7637. We repeat the previous process with other variables. Consequently we gather modalities "B/MTPL+" and "B/MTPL+++" of "coverp" (as "More") and the modalities "B/66-110" and "C/>110" of "powerc" (as ">=66"):

```
> Anova(GLM_Analysis, test.statistic="Wald",type=3,singular.ok=TRUE)
     Analysis of Deviance Table (Type III tests)

     Response: Data[["nbrtotc"]]
                          Df      Chisq Pr(>Chisq)
     (Intercept)           1 22033.5719  < 2.2e-16 ***
     Data[["sexp"]]        1    40.8282  1.662e-10 ***
     Data[["fleetc"]]      1     8.8546   0.002923 **
     Data[["sportc"]]      1    10.1966   0.001407 **
     Data[["coverp"]]      1   100.9727  < 2.2e-16 ***
     Data[["fuelc"]]       1   185.9554  < 2.2e-16 ***
     Data[["agecar"]]      2    59.3518  1.294e-13 ***
     Data[["powerc"]]      1    32.8338  1.004e-08 ***
     Data[["split"]]       3   710.4404  < 2.2e-16 ***
     Residuals        163628
     ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let us look at the summary of our fitting:

```
> summary(GLM_Analysis)
Call:
          glm(formula = Data[["nbrtotc"]] ~ off
     set(log(Data[["duree"]])) + Data[["sexp"]] + Data[["fleetc"]] +
     Data[["sportc"]] + Data[["coverp"]] + Data[["fuelc"]] + Data
     [["agecar"]] + Data[["powerc"]] + Data[["split"]], family =
     poisson(link = log))

Coefficients:
                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)               -2.18917    0.01475 -148.437  < 2e-16 ***
Data[["sexp"]]B/Female     0.10189    0.01595    6.390 1.66e-10 ***
Data[["fleetc"]]B/Yes     -0.12821    0.04309   -2.976  0.00292 **
Data[["sportc"]]B/Yes      0.21621    0.06771    3.193  0.00141 **
Data[["coverp"]]B/More    -0.15676    0.01560  -10.049  < 2e-16 ***
Data[["fuelc"]]B/Gasoil    0.20631    0.01513   13.637  < 2e-16 ***
Data[["agecar"]]B/0-1      0.18940    0.03214    5.893 3.79e-09 ***
Data[["agecar"]]B/2-5     -0.05822    0.01677   -3.470  0.00052 ***
Data[["powerc"]]B/>=66     0.09483    0.01655    5.730 1.00e-08 ***
Data[["split"]]B/Thrice    0.52045    0.02388   21.792  < 2e-16 ***
Data[["split"]]B/Twice     0.22683    0.01668   13.602  < 2e-16 ***
Data[["split"]]C/Monthly   0.41265    0.02116   19.497  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If we look at the summary of the GLM analysis we can see that there is no need of more gathering. All the variables and modalities are significant.

Table 2. Estimated parameters and their relativities

		Frequency	
Variable	Modality	Estimate $\vec{\beta}$	Relativities
Intercept		–2.18917	11.2%
sexp	Female	0.10189	110.7%
fleetc	Yes	–0.12821	88.0%
sportc	Yes	0.21621	124.1%
coverp	More	–0.15676	85.5%
fuelc	Gasoil	0.20631	122.9%
agecar	0-1	0.1894	120.9%
	2-5	–0.05822	94.3%
powerc	>=66 kW	0.09483	109.9%
split	Monthly	0.41265	151.1%
	Thrice	0.52045	168.3%
	Twice	0.22683	125.5%

Source: Authors’ own study.

The reference class (for which $X_1 = X_2 = \ldots = X_1 = 0$) corresponds to a male, whose car does not belong to a fleet, who does not drive a sports car, who has MTPL coverage, uses Petrol, his vehicle is older than 5 years, power of his car is less than 66 kW and he pays premium once a year.

The $\exp(\hat{\beta}_0) = \exp(-2{,}189) = 11{,}2$ % is the annual expected claim frequency for the policy in the reference class.

Since $\hat{\beta}_1 = 0{,}102$ and $\exp(\hat{\beta}_1) = 1{,}107$, this means that the expected claim frequency for female drivers is 10,7% higher compared to males (assumed that other characteristics are not changed).

$\hat{\beta}_3 = 0{,}216$ and $\exp(\hat{\beta}_3) = 1{,}241$, this means that the expected claim frequency for sports cars is 24,1% higher compared to non-sports cars. Let us notice that the $\exp(\hat{\beta}_3) = 1{,}241$ is applied to both males and females or all other modalities.

4. Conclusion

We have applied the GLM model to achieve risk classification of a MTPL portfolio. We used a particular portfolio of policies with variables determined a priori. Different insurance companies could collect different explanatory variable. Portfolio of other insurance company could include different variables such as the age, the period a driver has held a driving licence, marital status, etc. Still they can use the same approach as we propose to find relevant explanatory variables and modalities. There is a number of software programs that insurance industry has developed, for instance SAS GENMOD is used in Denuit et al. [2007]. We decided to use 'R' software, which is a free language.

The result of the GLM is a multiplicative model where the claims frequency of a category is given by the frequency of the reference class * the relativities $\exp(\hat{\beta}_j)$ of the category. The relativities measure the relative difference with respect to the reference class. The Poisson regression model imposes a strong constraint that the mean equals to the dispersion. Equidispersion is often violated in practice, suggesting that Poisson assumption is not appropriate. This can be corrected by applying Mixed Poisson model or negative binomial distribution.

We can see how to split a heterogeneous portfolio into more homogeneous classes with all policyholders belonging to the same class paying the same premium. However, tariff cells are still quite heterogeneous (some risk characteristics are unobservable) despite the use of many *a priori* variables. So, there is a need of the *a posterior* corrections. In *a priori* ratemaking, the actuaries aim to identify the best predictors and to compute the risk premium. In *a posterior* ratemaking, they aim to compute premium corrections according to past claims history. This experience rating is based on a 'crime and punishment' mechanism: claim-free policyholders are rewarded by premium discounts (bonus) and others (who report one or more claims) are penalized by premium surcharges (malus). Past claims experience can reveal the hidden features.

References

Cox D.R., 1972, *Regression Models and Life-tables*, Journal of the Royal Statistical Society. Series B (methodological), no. 34 (2), pp. 187–220.

Chongsuvivatwong V., 2012, *Epicalc: Epidemiological calculator. R package version 2.15.1.0*., http://CRAN.R-project.org/package=epicalc.

Denuit M., Charpentier A., 2005, *Mathématiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement. Collection Economie et Statistique Avancées*, Economica, Paris.

Denuit M., Maréchal X., Pitrebois S., Walhin J.-F., 2007, *Actuarial Modeling of Claim Counts: Risk Classification, Credibility and Bonus Malus Systems*, John Wiley & Sons, New York.

Fox J., Weisberg S., 2011, *An {R} Companion to Applied Regression, Second Edition*, Thousand Oaks CA, Sage, http://socserv.socsci.mcmaster.ca/ jfox/Books/Companion.

Jee B., 1989, *A comparative analysis of alternative pure premium models in the automobile risk classification system*, Journal of Risk and Insurance, no. 56, pp. 434–459.

Lambert D., 1992, *Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing*, Technometrics, no. 34 (1), pp. 1–14.

Nelder J.A., McCullagh P., 1989, *Generalized linear models (Second edition)*, Chapman & Hall, London.

R Core Team, 2015, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, http://www.R-project.org/.

Ter Berg P., 1980, *On the loglinear Poisson and gamma model*, ASTIN Bulletin, no. 11, pp. 35–40.

Warnes G.R., 2013, *Gmodels: Various R programming tools for model fitting. R package version 2.15.4.1.,* includes R source code and/or documentation contributed by Ben Bolker, Thomas Lumley, Randall C Johnson. Contributions from Randall C. Johnson are Copyright SAIC-Frederick, Inc. Funded by the Intramural Research Program, of the NIH, National Cancer Institute and Center for Cancer Research under NCI Contract NO1-CO-12400, http://CRAN.R-project.org/package=gmodels.