Fitting standard claims by Generalized Linear Models

Ján Gogola [1]

Abstract

This article deals with the issue of creating homogenous tariff classes of non-life insurance and modelling the cost of standard claims of each tariff class. We use a Generalized Linear Models (GLM) for the purpose of finding significant risk factors and also to determine the cost of standard claims of the individual tariff classes. The theoretical part will be completed by application on a typical heterogeneous portfolio of the Motor Third Party Liability (MTPL). All calculations were performed using the R software. The result of the GLM is a multiplicative model where the cost of standard claim of a particular tariff class is given as a product of the cost of standard claim of the reference class and the relativities of the tariff class.

Key words

Generalized linear models, Gamma distribution, R language, likelihood function, risk classification

JEL Classification: C10, C38, G22

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 heterogeneous. They mix individuals with different risk levels.

There is a need of risk classification (see Denuit and Charpentier (2005)). Nowadays, it has become extremely difficult for insurance companies to maintain cross subsidies between different risks 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 (see 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.).

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 by Jee (1989) relies on a statistical model incorporating all the available information about the risk. The ratemaking aims to evaluate as possible the pure premium for each policyholder.

Usually, the total claim *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

[1] Ján Gogola, RNDr. Ph.D, Fakulta ekonomicko-správní, Univerzita Pardubice, jan.gogola@upce.cz

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. Generalized linear models (GLM)

Generalized Linear Models (GLM), Nelder and 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 *risk factors* or *risk characteristics* in insurance.

GLM describes 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?

2.1 Severity model – Gamma regression for standard claims

Our explanatory variables are assumed to be categorical. A categorical variable with κ levels separates the portfolio into κ 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:

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^T$ is a vector containing the parameters, β_0 called intercept,

$\boldsymbol{X_i} = (1, x_{i1}, x_{i2}, \cdots, x_{ip})^T$ is a vector containing the explanatory variables (or observable characteristics) for the *i*-th policyholder.

Let C_k be the cost of the *k*-th standard claim. We assume C_1, C_2, … that are independent. Let each C_k conform to the Gamma distribution.

The $\text{Gam}(\mu, \alpha)$ probability density function can be cast into

$$f(x \mid \mu, \alpha) = \frac{1}{x \cdot \Gamma(\alpha)} \cdot \left(\frac{\alpha \cdot x}{\mu} \right)^{\alpha} \cdot \exp\left(-\frac{\alpha \cdot x}{\mu} \right). \tag{3}$$

If X~ $\text{Gam}(\mu, \alpha)$, then $E(X) = \mu$, $D(X) = \dfrac{\mu^2}{\alpha}$

and the coefficient of variation $\text{cvar}(X) = \dfrac{\sigma(X)}{E(X)} = \dfrac{1}{\sqrt{\alpha}}$ is constant.

The annual expected cost of standard claim (in a Gamma regression we use *log* link function) for each class is given by:

$$\exp\left(\boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}\right) = \exp\left(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_p \cdot x_p\right), \tag{4}$$

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 cost of standard claim 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 cost of standard claim is then equal to:

$$\mu_i = E(C_i) = \exp\left(\boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}_{\boldsymbol{i}}\right) = \exp(\beta_0) \cdot \prod_{j=1}^{p} \exp(\beta_j \cdot x_{ij}) = \exp(\beta_0) \cdot \prod_{j=1;\, x_{ij}=1}^{p} \exp(\beta_j), \tag{5}$$

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 cost of standard claim. The $\exp(\beta_j)$ is the multiplicative effect on the annual expected cost of standard claim 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 cost of standard claim.

We are estimating the parameters β_j by the maximum likelihood approach.

Let n_i be the number of claims reported by policyholder *i*, and let $c_{i1}, c_{i2}, \ldots, c_{in_i}$ be the corresponding claim costs.

The likelihood function associated with the observations writes

$$L(\boldsymbol{\beta}) = \prod_{i \mid n_i > 0} \prod_{j=1}^{n_i} f(c_{ij} \mid \mu_i, \alpha) = \prod_{i \mid n_i > 0} \prod_{j=1}^{n_i} \left(\frac{1}{c_{ij} \cdot \Gamma(\alpha)} \cdot \left(\frac{\alpha \cdot c_{ij}}{\mu_i} \right)^{\alpha} \cdot \exp\left(-\frac{\alpha \cdot c_{ij}}{\mu_i} \right) \right), \tag{6}$$

where $\mu_i = \exp\left(\boldsymbol{\beta}^{\boldsymbol{T}} \cdot \boldsymbol{X}_{\boldsymbol{i}}\right)$.

The log-likelihood equations are

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = 0 \Leftrightarrow \sum_{i \mid n_i > 0} \sum_{j=1}^{n_i} x_{ik} \left(1 - \frac{c_{ij}}{\mu_i} \right) = 0 \tag{7}$$

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 of one Belgian insurance company . 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)) and “gmodels” (Warnes (2013)).

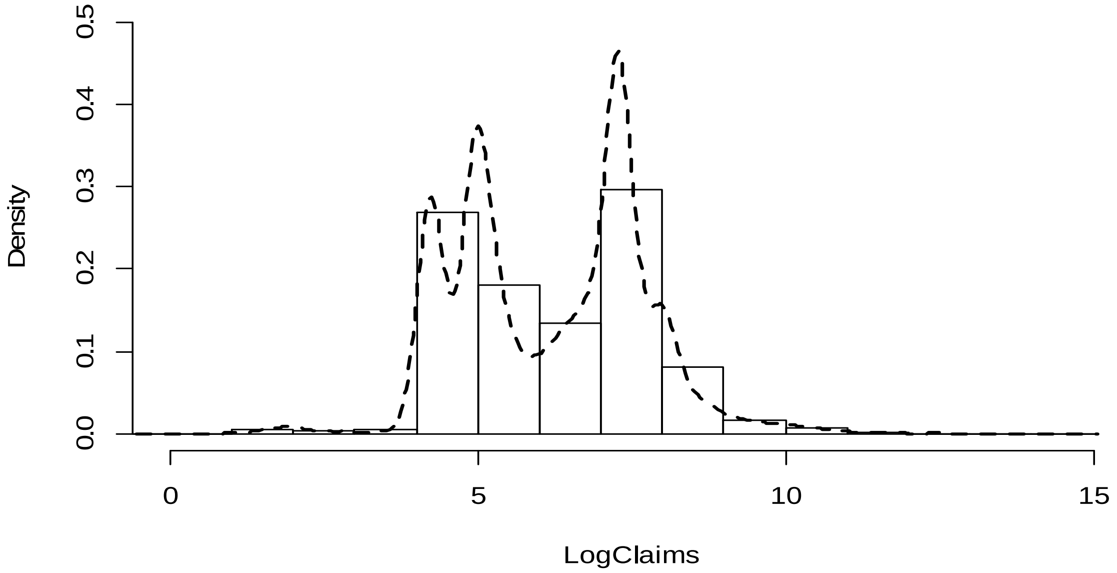
Table 1. Description of variables with their modalities

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: Author’s own study.

First we eliminate atypical extreme losses. By the look at the histogram or pdf of costs (Figure1.) we have chosen a threshold of 50 000 €. Out of the 163 657 data 18 305 represents standard costs (less than 50 000). It means we omit 40 largest costs. The reference class is composed of the modalities of the variables with the largest risk-exposure. In our case it is: Male, agecar 6-10, Petrol, split Once, Private use, No fleet, No sportcar, MTPL cover and power <66 kW. Often only the total claim amount $C_{i\bullet}$ is available and not the individual c_{ij} ´s. In this case, it is convenient to work with the mean claim amount or mean cost $\overline{C}_i = \frac{C_{i\bullet}}{n_i}$, where n_i is the number of claims reported by policyholder *i*.

Figure 1. Pdf of log-costs



In R, the GLM analysis is performed via Gamma regression:

glm(*mean_cost* ~ *agecar* + *sexp* +…+ *powerc*, *family* = *Gamma*(*link*=*log*), *weigths*=*nbrtotc*)

In this case the response variable is the mean cost.

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

With the help of **Anova** analysis, we observe which variables are significant.

```
> Anova(GLM_AnalysisCosts, type='III', test.statistic='F')
  Response: DataCosts[["MeanCharge"]]
                             SS    Df        F    Pr(>F)
  DataCosts[["sexp"]]         1     1  0.1189  0.730271
  DataCosts[["usec"]]         0     1  0.0011  0.973738
  DataCosts[["fleetc"]]       3     1  0.5919  0.441697
  DataCosts[["sportc"]]       0     1  0.0786  0.779140
  DataCosts[["coverp"]]     273     2 24.3595 2.722e-11 ***
  DataCosts[["split"]]        7     3  0.3999  0.753046
  DataCosts[["fuelc"]]        1     1  0.1695  0.680581
  DataCosts[["agecar"]]      72     3  4.2983  0.004879 **
  DataCosts[["powerc"]]       3     2  0.3086  0.734473
  Residuals              102353 18289
  ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We remove ”usec” (*p*-value equals to 0.973738). We repeat this process until obtain only significant variables.

```
> Anova(GLM_AnalysisCosts, type='III', test.statistic='F')
  Response: DataCosts[["MeanCharge"]]
                             SS    Df       F    Pr(>F)
  DataCosts[["coverp"]]    275     2 24.4359 2.522e-11 ***
  DataCosts[["agecar"]]     73     3  4.3501  0.004538 **
  Residuals             102786 18299
  ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At the end of the process the significant variables are: the type of **cover** and **age** of the vehicle.

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_AnalysisCosts,level = 0.95)
                                      2.5 %     97.5 %
  (Intercept)                    7.10008164  7.2133231
  DataCosts[["coverp"]]B/MTPL+   -0.30591194 -0.1461062
  DataCosts[["coverp"]]B/MTPL+++  0.02844816  0.2530795
  DataCosts[["agecar"]]B/0-1      0.05554839  0.3699192
  DataCosts[["agecar"]]B/2-5     -0.13953754  0.0269593
  DataCosts[["agecar"]]C/>10     -0.06202141  0.1105743
```

The process of gathering can be summarized as follows: Variable “agecar”: 2-5 and >10 are overlapping and they overlapping “0” the value of the reference class 6-10 of this variable. This means we try to gather them.

```
> fit.contrast(GLM_AnalysisCosts,DataCosts[["agecar"]],c(-1,1,0,0))
                                        Estimate Std. Error  t value    Pr(>|t|)
DataCosts[["agecar"]] c=( -1 1 0 0 ) 0.2127338  0.0801981 2.652604 0.007994252
> fit.contrast(GLM_AnalysisCosts,DataCosts[["agecar"]],c(-1,0,1,0))
                                          Estimate Std. Error   t value Pr(>|t|)
DataCosts[["agecar"]] c=( -1 0 1 0 ) -0.05628912 0.04247446 -1.325246 0.185106
> fit.contrast(GLM_AnalysisCosts,DataCosts[["agecar"]],c(-1,0,0,1))
                                         Estimate Std. Error   t value  Pr(>|t|)
DataCosts[["agecar"]] c=( -1 0 0 1 ) 0.02427644 0.04403032 0.5513572 0.5813955
> fit.contrast(GLM_AnalysisCosts,DataCosts[["agecar"]],c(0,-1,1,0))
                                         Estimate Std. Error   t value     Pr(>|t|)
DataCosts[["agecar"]] c=( 0 -1 1 0 ) -0.2690229 0.07848303 -3.427784 0.0006098684
> fit.contrast(GLM_AnalysisCosts,DataCosts[["agecar"]],c(0,-1,0,1))
                                         Estimate Std. Error   t value   Pr(>|t|)
DataCosts[["agecar"]] c=( 0 -1 0 1 ) -0.1884573 0.08500623 -2.216983 0.02663652
> fit.contrast(GLM_AnalysisCosts,DataCosts[["agecar"]],c(0,0,-1,1))
                                         Estimate Std. Error  t value  Pr(>|t|)
DataCosts[["agecar"]] c=( 0 0 -1 1 ) 0.08056556 0.05060952 1.591905 0.1114233
```

The modalities ">10" and "6-10" of "agecar" should be gathered (as ">5"), as the *p*-value equals 0.5814. We repeat the previous process with other modalities.
Consequently, we gather modalities "B/2-5" and "C/>5" with the reference modality of "agecar" (as "A/>1").

Let us look at the summary of our fitting (output of *R*):

```
> summary(GLM_AnalysisCosts)
  Call: glm(formula = DataCosts[["MeanCharge"]] ~ DataCosts[["coverp"]] +
        DataCosts[["agecar"]],  family  =  Gamma(link  =  log),  weights  =
        DataCosts[["nbrtotc"]])

  Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
  (Intercept)                    7.15477    0.02150 332.831  < 2e-16 ***
  DataCosts[["coverp"]]B/MTPL+   -0.24636    0.03895  -6.325 2.59e-10 ***
  DataCosts[["coverp"]]B/MTPL+++  0.10340    0.05269   1.962  0.04973 *
  DataCosts[["agecar"]]B/0-1      0.24007    0.07618   3.151  0.00163 **
  ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 2. Estimated parameters and their relativities.

		Costs	
Variable	Modality	Estimate $\vec{\beta}$	Relativities
Intercept		7,15477	**1 280,20** €
sexp	Female	0	100 %
fleetc	Yes	0	100 %
sportc	Yes	0	100 %
coverp	MTPL+	– 0.24636	78,16 %
	MTPL+++	0,10340	110,89 %
fuelc	Gasoil	0	100 %
agecar	0-1	0.24007	127,13 %
powerc	>=66 kW	0	100 %
split	Other	0	100 %

Source: Author’s own calculation.

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 variables. 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 are 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 free available software.

The result of the GLM is a multiplicative model where the cost of standard claim of a category is given by the cost of standard claim 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 Gamma regression model for standard claims can be replaced by Inverse-Gaussian or Lognormal distribution. For extreme claims we would suggest Generalized Pareto distribution which provides good approximation to the excess distribution over large threshold (for instance 50 000 €).

We can see how to partition 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

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

[2] Chongsuvivatwong V. (2012). *Epicalc: Epidemiological calculator. R package version 2.15.1.0*., http://CRAN.R-project.org/package=epicalc (accessed).

[3] Denuit M., Charpentier A. (2005). *Mathématiques de l’Assurance Non-Vie. Tome II: Tarification et Provisionnement. Collection Economie et Statistique Avancées*, Economica, Paris.

[4] 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.

[5] 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 (accessed).

[6] 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.

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

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