Using R In Generalized Linear Models

Mihaela COVRIG (mihaelacovrig@gmail.com)
Iulian MIRCEA (mirceaiulian91@yahoo.com)
Bucharest University of Economic Studies
Gheorghiţă ZBĂGANU (georghitazbaganu@yahoo.com)
University of Bucharest
Alexandru COŞER (alexandru.coser@gmail.com)
Vodafone Romania
Alexandru TINDECHE (tindeche_alex@yahoo.com)
BCR Asigurări de Viață (BCR Life Insurance),

ABSTRACT

This paper aims to approach the estimation of generalized linear models (GLM) on the basis of the `glm` *routine package in R. Particularly, regression models will be analyzed for those cases in which the explained variable follows a Poisson or a Negative Binomial distribution. The paper will briefly present the GLM methodology for count data, while the practical part will revolve around estimating and comparing models in which the response variable shows the number of claims in a portfolio of automobile insurance policies.*

Key words*: GLM, count data, insurance, Poisson regression, Negative Binomial Regression, R*

JEL: *C250*

INTRODUCTION

Generalized linear models (GLM) appeared as a generalization of classical linear regression models, and they were introduced by Wedderburn and Nelder in their article "Generalized linear models", published in 1972 in Journal of the Royal Statistical Society. The differences from the classical linear regression model are the following: the response variable distribution is not necessarily normal, but it belongs to the exponential family of distributions, while the relationship between the dependent variable mean and a linear form built on the explanatory variables is given by a link function. The exponential family of distributions includes discrete distributions such as Bernoulli, Poisson, Binomial, Negative Binomial, and also continuous distributions such as Normal, Inverse-Gaussian, or Gamma.

This paper will focus on generalized linear models that are used in actuarial practices, in decisions related to pricing different types of insurance, and in particular

in the pricing of automobile insurance policies. The response variable shows the number of a relevant event occurrence in a specified interval, and in this particular case it represents the number of damage claims in 1-year interval for one insurance policy. We will presuppose that such a variable follows a Poisson or a Negative Binomial distribution, thus we will analyze the estimation of regression models for count data, using `glm` and `glm.nb` routines in R, respectively.

The numerical illustration will be effectuated with the aid of one of the data sets in De Jong, Heller (2008), the vehicle insurance data set, which contains information about an automobile insurance portfolio, namely characteristics of drivers, and of insured vehicles, as well as the number of claims per policy, their cost, or the exposure of each policy. Our purpose is to estimate the mean claim frequency, considering different explanatory variables and using several models provided by the application of the methodology for generalized linear models, which will be briefly explained in the subsequent section. Based on such methodologies, several models were estimated and compared.

The paper is structured as follows: section 2 contains the theoretical presentation of GLM, with a close focus on Poisson regression models. Section 3 presents the portfolio of the insurance policies, emphasizing the influence of different explanatory variables on the number of claims, with the help of various descriptive techniques in R. Some of the variables were conveniently transformed in order to be used in `glm`. Section 4 is dedicated to fitting several Poisson and Negative Binomial regression models, with a view to identify a good possible model for the considered sample. The last section is dedicated to discussions.

GENERALIZED REGRESSION MODELS FOR COUNT DATA

The specific elements of a generalized linear model are (McCullagh, Nelder, 1989, Denuit, Charpentier, 2005):

1. The model's random component is given by the independent random variables Y_1, Y_2,...,Y_n, which are not identically distributed, their distributions have the same form in the exponential family, depending on the parameters θ_i and on the same scale parameter ϕ; parameter θ_i characterizes statistical unit *i*, thus parameters θ_i may be different. The probability density function or the probability mass function of Y_i is:

$$f(y_i;\theta_i,\phi)=e^{\frac{y_i\theta_i-b(\theta_i)}{\phi}-c(y_i,\phi)},\ y_i\in S, \tag{1}$$

where the support set *S* of the random variable Y_i is a subset of **N** or **R**.

The expectation is $M[Y_i]=b'(\theta_i)=\mu_i$.

2. The model's systematic component or the "linear predictor" is built with $p+1$ parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ and with p explanatory variables:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, i = 1,2,...,n. \quad (2)$$

3. The "link function" between the random and systematic components, namely a monotonic and differentiable function g, so that

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, i = 1,2,...,n. \quad (3)$$

The estimation of the parameters is achieved through the maximum likelihood method.

The most used GLM for count data is the Poisson regression model, which we shall briefly present below. Let us suppose that the random variable Y, which shows the number of claims per one insurance policy in a vehicle insurance portfolio, is modeled through Poisson(λ) distribution, whose parameter represents the average frequency of the claims. The choice of Poisson distribution is justified by the fact that it shows the number of a relevant event's incidence in a specified time interval or in a specified space. Its probability mass function is:

$$Y \sim Poisson(\lambda), \ f(y;\lambda) = \frac{\lambda^y}{y!} \cdot e^{-y}, y \in \mathbf{N}, \lambda > 0, \quad (4)$$

the expectation and variance are equal to parameter λ, $E[Y] = Var[Y] = \lambda$.

Eliminating indices, the link between natural canonical parameter θ in (1) and parameter λ of Poisson distribution is $\theta = \ln\lambda$. If sample variables $Y_i \sim Poisson(\lambda_i)$, then $\mu_i = E[Y_i] = \lambda_i$, and the link function is

$$\eta_i = g(\lambda_i) = \ln\lambda_i, \text{ that is } \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad (5)$$

so the mean frequency will be estimated or predicted by the model

$$\lambda_i = d_i \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right) \text{ or } \lambda_i = d_i \cdot e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot e^{\beta_2 x_{i2}} \cdot ... \cdot e^{\beta_p x_{ip}}, \quad (6)$$

also named the multiplicative model, where a new variable appears, d_i, representing the risk exposure of policy i, namely the time interval, usually expressed in years, from the initial moment when the policy was issued until the moment when the sample is observed and analyzed.

The Poisson regression model for the estimation of claim frequency in a policy portfolio is:

$$\ln \lambda_i = \ln d_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip}, \quad (7)$$

where $\ln d_i$ is the natural logarithm of risk exposure, called the "offset variable".

Explanatory variables are dummy variables, built from multilevel categorical variables or from continuous variables transformed in categorical variables. Continuous variables are transformed in categorical variables through „banding", which is achieved in two steps: first the sample data will be systematized per variation intervals, then each interval will be attributed a level or category, thus obtaining a categorical variable or factor. The introduction and transformation of categorical or continuous variables in sets of dummy variable is recommended for the estimation of GLM, especially in the case of those used for count data (De Jong, Heller, 2008).

In R, the particularity of the estimation of Poisson regression models revolves around the fact that for each explanatory variable introduced in the model, previously transformed in categorical variable, one must specify the reference category or level. The glm function may work with two types of explanatory variables: either numeric variables (continuous), or factor or categorical variables. For the latter, the implemented algorithm of the glm function builds dummy variables corresponding to each level and in the estimation of the Poisson regression model, the first level of the factor variable is considered a reference level, by default. Various authors, such as De Jong, Heller (2008), Frees (2010), Denuit et al. (2007), Cameron, Trivedi (2013), recommend that for each factor variable one should exclude the dummy variable with the biggest exposure in the sample, namely the dummy variable corresponding to the level with the biggest absolute or relative frequency.

In addition to the Poisson regression models, for count data one can also use Negative Binomial regression models, as the assumption that the mean and variance of the number of claims are equal is restrictive and does not hold true in reality, where data exhibits the "overdispersion" phenomenon, namely the variance is much larger than the mean, which is a feature of the Negative Binomial distribution.

Recently, new approaches for basic count data models have been developed. Zero-inflated models characterize those situations in which the number of zero values for the response variable is excessive. Actuaries may face this situation when the insured hesitates to report damages, for fear their multiple damage history may induce a penalty applied by the insurer, resulting in a bigger premium to pay. The zero-inflated count models were introduced by Mullahy (1986), subsequently used in numerous studies, both in insurance, Yip, Yau (2005), Perumean-Chaney et al. (2013), and in studies on medical data, such as Cheung (2002) or Lam, Xue, Cheung (2006). The zero-inflated count distribution for the number of damages is built on the basis of a mixture of distributions: probability for the zero value is expressed with the aid of a logit model, while for strictly positive values the Poisson or the Negative Binomial models will be used.

Hurdle models are similar to zero-inflated models and are also used to describe situations in which an excessive number of zero values occur. Hurdle models

can be justified by the sequential decision-making process of individuals. In the insurance sector, the decision made by an insured person to report a first damage may be radically different from that to report subsequent damages, thus this decision model is a "principal-agent" model type. Although there are a lot of papers dedicated to this issue that actuaries have to deal with, we shall mention here only one of them: Boucher, Denuit, Guillen (2007), who analyzed a portfolio including a big number of automobile insurance policies in Spain through the estimation of the claim frequency with the aid of zero-inflated models, hurdle or heterogeneity models, showing that these represent an improvement as compared to the basic Poisson model.

Among other papers focused on the analysis of different sets of data from the field of automobile insurance it would be worth mentioning Frees, Valdez (2008), who proposed a hierarchical model for three components corresponding to frequency, type and severity of the damages, and who used a set of data from Singapore.

More recent research, such as Antonio, Valdez (2012) or Boucher, Inoussa (2014) concentrate on longitudinal and panel data, with a two-step approach to pricing: a priori ratemaking, and posteriori ratemaking, the second phase being possible when the insurers have long-term information about their policy holders.

Another recent study, Klein et al. (2014), analyzes the situation in which the assumption that the response variable has the distribution in the exponential family is relaxed, this approach permitting the actuary to include risk factors, not only in the mean, but also in certain key parameters that influence the behavior of the one who brings damage claims.

DESCRIPTION OF THE DATA SET

The data set from De Jong and Heller (2008), called vehicle insurance, contains 67,856 1-year automobile insurance policies, in the interval 2004-2005. For each policy, there is information about: policy risk exposure, drivers' characteristics, vehicles' characteristics, number of claims and their cost.

The *exposure* variable shows the length of the time interval, expressed in years, from the moment of issuing the insurance policy to the moment when the observation period ended. This variable follows a uniform distribution on [0,1], which implies that the analyzed company's portfolio consists of insurance policies that are issued on a continuous basis throughout the period of observation, the average exposure is 0.4686, or about half year.

The characteristics of the insured or of the vehicles are given by the following variables: *veh_age*, the age of the vehicle, a variable with four levels: 1 – the newest, 2, 3 and 4 – the oldest category; *agecat*, the age of the driver, with six categories: 1 – the youngest, 2, 3, 4, 5, 6 – the oldest; *clm* with two values: 0, renamed 'No claim', meaning that the policy had no claims, and 1, renamed 'At least one claim', if the policy is associated with at least one claim; *area*, the region or area in which the driver lives and drives, a factor variable with six categories: A, B, C, D, E, and F; *gender*; *veh_body*, the type of vehicle, a factor variable with thirteen categories. The variables

veh_age, *agecat* and *clm* were transformed in categorical variables or factor variables, using the following commands:

```
date$veh_age<-factor(date$veh_age)
date$agecat<-factor(date$agecat)
date$clm<-factor(date$clm, labels=c('No claim','At least one claim'))
```

The continuous numerical variable *veh_value* shows the value of the vehicle, expressed in $ 10,000, most insurance policies, 81.01%, were issued for cars under $ 25,000, in the portfolio there are also 33 insured vehicles worth more than $ 125,000.

The distribution of insurance policies by age, represented in Figure 1, shows that most clients, i.e. 23.86% of them, are to be found in the 4th age group. In the analyzed portfolio, the least represented age group, only 8.46%, is that of the youngest drivers. However, for this last mentioned group, the observed mean claim frequency, calculated taking into account the policies exposure in the portfolio, is 20.10%, the highest annual frequency of the claims in all six age categories, as we can see in Figure 2. The evolution of the mean annual claim frequency decreases as the age increases, the last two groups exhibiting an almost equal annual claim frequency, 12.53%, for the 5th category, and 12.58% for the oldest drivers.

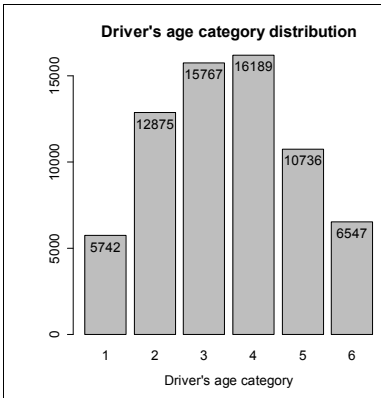


Fig. 1. Drivers' age category distribution.

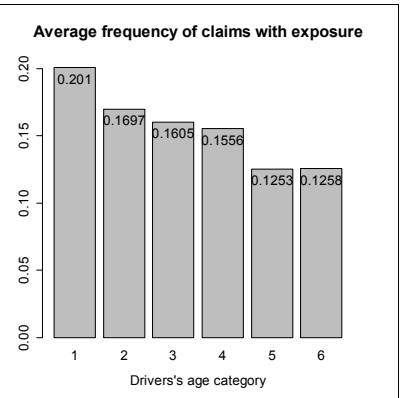


Fig. 2. Observed mean annual claim frequency depending on driver's age.

Similarly, the influence of the drivers' age on the frequency of damage claims can be analyzed on the basis of the policies' bidimensional distribution by variables age and *clm,* which shows if damages were reported for a particular policy. The bidimensional distribution of age (*agecat*) and *clm* is presented in a contingency table, Table 1, in which we considered it useful to calculate proportions for each age category. Thus, from 8.64%, the proportion of the youngest drivers with at least one claim, the figures are decreasing till the oldest drivers, a category in which only 5.58% reported at least claim. This behavior is also represented in Figure 3, a mosaic plot

representation, in which for the category 'At least one claim' one can notice a decrease along with ageing, from the youngest to the oldest. Another particularity of this type of graph is that the width of the bars corresponding to the different age groups is given by their proportion in the sample.

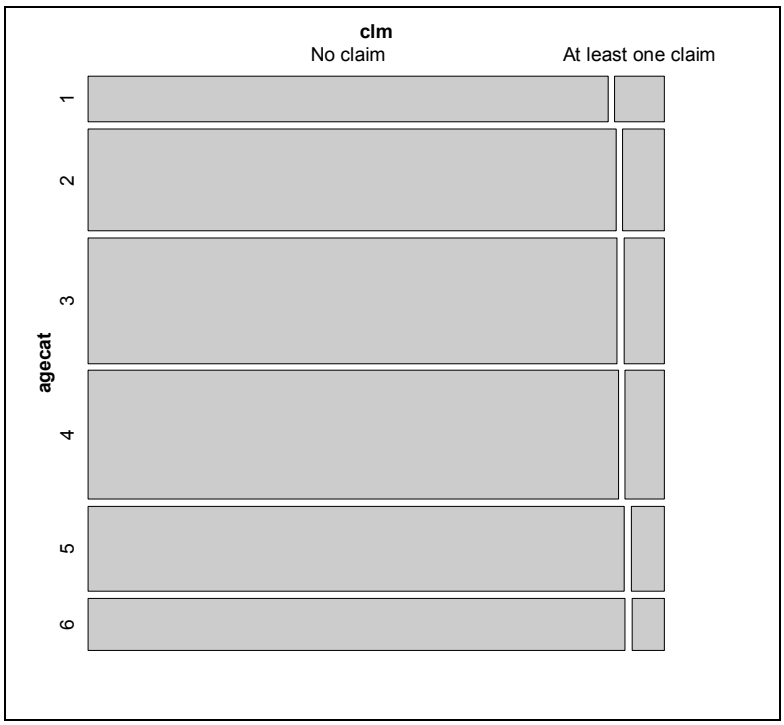
Occurrence of at least one claim for a policy: proportions on age categories

Table 1

		Occurrence of claims	
		No claim	At least one claim
Age category	1	0.9136	0.0864
	2	0.9276	0.0724
	3	0.9294	0.0706
	4	0.9318	0.0682
	5	0.9428	0.0572
	6	0.9442	0.0558

Mosaic plot for the bidimensional distribution of insurance policies by age and the dummy variable showing if at least one claim was reported

Figure 3



Another mosaic plot presents the structure of the portfolio by age, gender and number of claims (Figure 4). This represents the bidimensional distribution of policies by age and number of claims, the conditioning variable being gender. In the contingency table (Table 2) and in Figure 4 we can clearly see that women drivers represent a bigger risk for the insurance company, because the proportion of women drivers who report at least one damage is bigger than that of men. Particularly, the biggest number of damage claims for one policy, i.e. 4, are reported by women drivers in the 3rd and 5th age group. Table 2 and Figure 4 are obtained with the following commands:

```
STD <- structable(~gender + agecat + numclaims, data = date)
mosaic(STD, condvars = "gender", split_horizontal = c(TRUE, FALSE, TRUE))
```

Crosstab of driver's age and number of claims, given the gender

Table 2

		Driver's age category					
Gender	Number of claims	1	2	3	4	5	6
Female	0	2998	7075	8631	8735	5447	3069
	1	258	498	644	600	306	171
	2	18	36	39	41	15	11
	3	0	3	4	2	0	0
	4	0	0	1	0	1	0
Male	0	2248	4868	6023	6350	4675	3113
	1	210	371	400	427	277	171
	2	9	22	24	32	14	10
	3	1	2	1	2	1	2
	4	0	0	0	0	0	0

Mosaic plot for age and number of claims, by gender

Figure 4

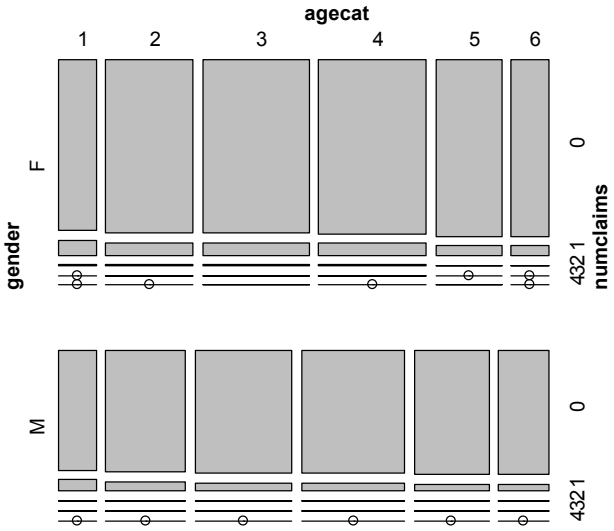


Table 3 presents the most frequent category in the insurance portfolio for each possible predictor variable.

Modal classes for the categorical variables

Table 3

Variable	Modal class or category	Name of modal class
agecat	4	modal_agecat
veh_age	3	modal_veh_age
veh_body	SEDAN	modal_veh_body
gender	F	modal_gender
area	C	modal_area
veh_value_intervals	(-0.01, 2.5]	modal_veh_value_intervals

THE ESTIMATION OF SOME POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

We shall estimate the mean annual claim frequency in the vehicle insurance portfolio through Poisson and Negative Binomial regression models. The explained variable is *numclaims*, the one that shows the number of claims per one policy per one driver.

A first possible approach is to suppose that the response variable follows a Poisson distribution, so that we aim to estimate the mean claim frequency through a Poisson regression model given by (7), using the `glm` function.

In calling the `glm` function, the following arguments need to be specified: formula, namely the response variable and the explanatory variables, and if the latter are categorical variables, one needs to specify the reference level or class which can be the modal class; `family` specifies error distribution and the link function, in our case `poisson(link="log")`; `data`, namely the data frame that contains the variables which will be used in the model; `weights`, an optional vector of weights that can be used in the estimation process; `offset`, that is the variable representing a component known apriori, with the coefficient 1, to be introduced in the linear predictor.

First, let us consider a model where the insured's age is the unique explanatory variable:

```
model_glm_agecat<-glm(numclaims~relevel(agecat,modal_agecat),
family=poisson(link="log"), data=date, offset=log(exposure))
```

The output of the estimation is presented below, where *agecat* variable was found to be a significant predictor.

```
> model_glm_agecat<-glm(numclaims~relevel(agecat,modal_agecat), family=poisson(link="log"),
data=date, weight=NULL, offset=log(exposure))
> summary(model_glm_agecat)

Call:
glm(formula = numclaims ~ relevel(agecat, modal_agecat), family = poisson(link = "log"),
    data = date, weights = NULL, offset = log(exposure))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6338  -0.4544  -0.3482  -0.2227   4.5443

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.86058    0.02905 -64.048  < 2e-16 ***
relevel(agecat, modal_agecat)1  0.25600    0.05243   4.883 1.05e-06 ***
relevel(agecat, modal_agecat)2  0.08701    0.04294   2.026 0.042744 *
relevel(agecat, modal_agecat)3  0.03094    0.04105   0.754 0.451066
relevel(agecat, modal_agecat)5 -0.21635    0.04886  -4.428 9.50e-06 ***
relevel(agecat, modal_agecat)6 -0.21232    0.05838  -3.637 0.000276 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 25507  on 67855  degrees of freedom
Residual deviance: 25415  on 67850  degrees of freedom
AIC: 34862

Number of Fisher Scoring iterations: 6
```

Using this simple model, the estimated annual claim frequencies for each age category are:

- $0.20=e^{-1.8605} \cdot e^{0.2560}$, for the youngest (1st category), which means that their claim frequency is $1.29(=e^{0.2560})$ fold (times) higher than the claim frequency of drivers in the reference age category, namely the 4th category of age,
- $0.17=e^{-1.8605} \cdot e^{0.0870}$, for the 2nd age category, their claim frequency being 1.09 fold higher than the drivers in the 4th age category,
- $0.16= e^{-1.8605} \cdot e^{0.03094}$, for the 3rd age category, and increase of 3% with respect to the baseline (the 4th age category),
- about 0.13 in each of the last age categories, 5th and 6th, which means a decrease of about 19% of the claim frequency in the 4th category.

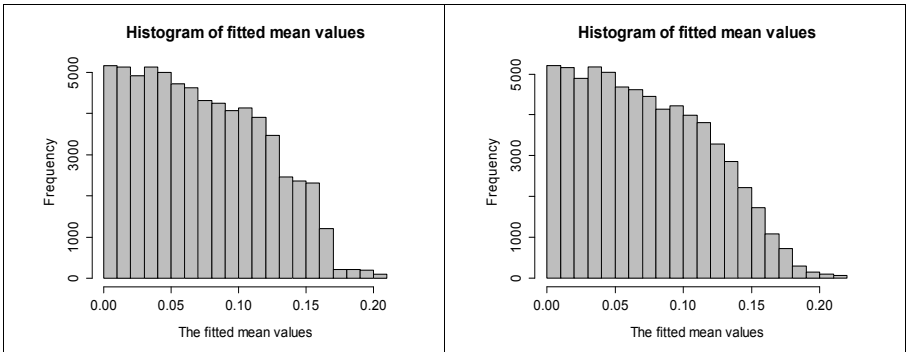
Another model is one in which we introduce a second variable showing the age of the vehicle, *veh_age*:

```
model_glm_agecat_vehage<-glm(numclaims~relevel(agecat,modal_
agecat)+relevel(veh_age,modal_veh_age),
family=poisson(link="log"), data=date, weights=NULL, offset=log(exposure))
```

Figure 5 contrasts the histograms of fitted mean values of the response variable for the two estimated models. If we compare the values of the informational criterion Akaike AIC, we see that for the first model the value is 34,862, and 34,841 for the second, which suggests that the model in which we additionally considered the automobile value is preferable to the first.

The histogram of fitted mean values for model_glm_agecat (left panel) and `model_glm_agecat_vehage` (right panel)

Figure 5



Another approach to the estimation of the mean claim frequency is to assume that the variable that shows the number of claims follows a Negative Binomial distribution. To exemplify, we considered the predictors of such a regression model the variables *agecat* and *veh_age*. Below we present the syntax of calling the `glm.nb` function for this model, as well as that of generating the histogram of the fitted values and the density plot of residual values, the two graphs being presented in Figure 6.

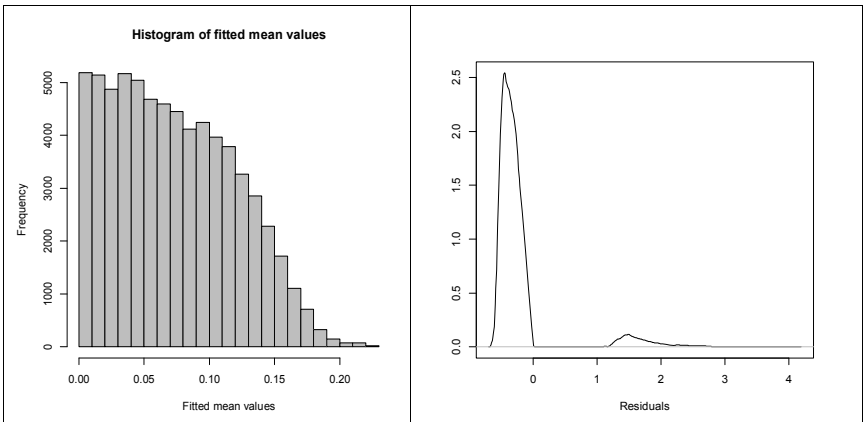
```
library(MASS)
model_age_vehage_nb <- glm.nb(numclaims ~ relevel(agecat,modal_agecat) +
relevel(veh_age,modal_veh_age) + offset(log(exposure)), data=date)
summary(model_age_vehage_nb)
estimat_model_age_vehage_nb<-fitted(model_age_vehage_nb)
summary(estimat_model_age_vehage_nb)
hist(estimat_model_age_vehage_nb, col='grey',main="Histogram of fitted mean
values", xlab="Fitted mean values")
residuals_model_age_vehage_nb<-residuals(model_age_vehage_nb)
summary(residuals_model_age_vehage_nb)
hist(residuals_model_age_vehage_nb, col='grey')
plot(density(residuals_model_age_vehage_nb),  xlab="Residuals",  ylab="",
main="")
```

The predicted mean frequency of claims for insurance policy *i*, from model_age_vehage_nb, is:

$$\hat{\mu}_i = exposure_i \cdot e^{-1.8840} \cdot e^{0.2495 \cdot agecat1_i} \cdot e^{0.0857 \cdot agecat2_i} \cdot e^{0.0330 \cdot agecat3_i} \cdot e^{-0.2185 \cdot agecat5_i} \cdot e^{-0.2120 \cdot agecat6_i} \cdot e^{0.0748 \cdot veh\_age1_i} \cdot e^{0.1214 \cdot veh\_age2_i} \cdot e^{-0.0688 \cdot veh\_age4_i}$$

Histogram of fitted mean values (left panel) and density plot of residuals (right panel)for `model_age_vehage_nb`

Figure 6



DISCUSSIONS

If we compare the results of the three above-presented regression models, we notice that AIC is smaller in the case of the Negative Binomial model, AIC=34802, which suggests that the Negative Binomial distribution models the distribution of the number of claims better, whose variance is bigger than the mean in the sample. In a similar way one can estimate other models that allow the introduction of more explanatory variables available in the data set, as well as considering interactions between them. Comparing these models enables one to choose the most appropriate one by first checking if the model is significant, that is if the ratio between Residual deviance and the corresponding number of degrees of freedom doesn't have a value significantly bigger than 1, and then by retaining the one with a smaller value of the informational criterion Akaike AIC. Another point of view in the selection of a model used to estimate the mean frequency must, at the same time, account for the ultimate objective of the insurance companies, that decide calculating insurance premiums taking into account as many characteristics of the insured drivers and vehicles as possible. There should be investigated other methods in order to capture the random mechanism which generates claims.

ACKNOWLEDGMENT

This paper has been financially supported within the project entitled ***„SOCERT. Knowledge society, dynamism through research",*** contract number POSDRU/159/1.5/S/132406. This project is co-financed by European Social Fund through Sectoral Operational Programme for Human Resources Development 2007-2013. **Investing in people!**

References:

1. Antonio, K., Valdez, E. A. (2012), Statistical concepts of a priori and a posteriori risk classification in insurance, Advances in Statistical Analysis, Volume 96, Issue 2, pp. 187-224.
2. Boucher, J.-P., Denuit, M., Guillen, M. (2007), Risk classification for claim counts: A comparative analysis of various zero-inflated mixed Poisson and hurdle models, North American Actuarial Journal, Volume 11, no. 4, pp. 110-131.
3. Boucher, J.-P., Inoussa, R. (2014), A Posteriori Ratemaking with Panel Data. ASTIN Bulletin, Volume 44, Issue 03, pp. 587-612.
4. Cameron, A. C., Trivedi, P. K. (2013), Regression Analysis of Count Data, Cambridge University Press, Cambridge.
5. Cheung, Y. B. (2002), Zero-inflated models for regression analysis of count data: a study of growth and development, Statistics in Medicine, Volume 21, Issue 10, pp. 1461-1469.
6. Dalgaard, P. (2008), Introductory Statistics with R, Second Edition, Springer, New York.
7. David, M. (2015), Auto insurance premium calculation using generalized linear models, Procedia Economics and Finance, Volume 20, pp. 147-156.
8. De Jong, P., Heller, G. Z. (2008), Generalized Linear Models for Insurance Data, Cambridge University Press, Cambridge.
9. Denuit, M, Charpentier, A. (2005), Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement. Collection Economie et Statistique Avancees, Economica, Paris.
10. Denuit, M., Marechal, X., Pitrebois, S., Walhin, J.-F. (2007), Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems. Wiley, New York.
11. Frees, E. W. (2009), Regression Modelling with Actuarial and Financial Applications, Cambridge University Press, Cambridge.
12. Frees, E. W., Valdez, E. A. (2008), Hierarchical insurance claims modeling, Journal of the American Statistical Association, vol. 103, pp. 1457-1469.
13. Klein, N., Denuit, M., Lang, S., Kneib, T. (2014), Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale and shape, Insurance: Mathematics and Economics, Volume 55, pp. 225-249.
14. Lam, K. F., Xue, H., Cheung, Y. B. (2006), Semiparametric Analysis of Zero-Inflated Count Data, Biometrics, Volume 62, Issue 4, pp. 996-1003.
15. McCullagh, P., Nelder, J. A. (1989), Generalized Linear Models, Chapman & Hall, London.
16. Mullahy J. (1986), Specification and testing of some modified count data models, Journal of Econometrics, Volume 33, Issue 3, pp. 341–365.
17. Nelder, J. A., Wedderburn, R. W. M. (1972), Generalized linear models, Journal of the Royal Statistical Society, Series A, 135, 370-384.
18. Perumean-Chaney, S. E., Morgan, C., McDowall, D., Aban, I. (2013), Zero-inflated and overdispersed: what's one to do?, Journal of Statistical Computation and Simulation, Volume 83, Issue 9, pp. 1671-1683.
19. Yip, K. C. H., Yau, K. K. W. (2005), On modeling claim frequency data in general insurance with extra zeros, Insurance: Mathematics and Economics, Volume 36, Issue 2, pp. 153-163.
20. http://cran.r-project.org/