Auto insurance premium calculation using generalized linear models

Mihaela David*

Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iasi, Iasi 700505, Romania

Abstract

The non-life insurance pricing consists of establishing a premium or a tariff paid by the insured to the insurance company in exchange for the risk transfer. A usual way to obtain the insurance premium is to combine the conditional expectation of the claim frequency with the expected claim amount. The aim of this paper is to present an overview of the Generalized Linear Models techniques in order to calculate the pure premium given the observable characteristics of the policyholders. A numerical illustration based on a French auto insurance portfolio is performed with the statistical software SAS.

Keywords: Non-life insurance pricing; risk factors; tariff class; pure premium; frequency of claims; cost of claims; Generalized Linear Models.

1. Introduction

The fundamental role of insurance is to provide financial protection, offering a method of transferring the risk in exchange of an insurance premium. Considering that not all the risks are equal, it is natural that every insured will pay a premium or tariff corresponding with the gravity of the risk.

The different charging tariffs are emphasized by the insurance portfolio heterogeneity that leads directly to anti-selection phenomenon. This basically presumes charging same tariff for the entire portfolio, meaning that the

* Corresponding author.

E-mail address: mihaela_david88@yahoo.com
unfavorable risks are also assured (at a lower price) and as an adverse effect, it discourages insuring medium risks. The necessity of pricing for non-life insurance comes precisely in an attempt to combat the anti-selection phenomenon by dividing the insurance portfolio in sub-portfolios based on certain influence factors. Therefore, every class will contain policyholders with identical risk profile that will pay the same reasonable premium.

A usual method to calculate the premium is to combine the conditional expectation of the claim frequency with the expected cost of claims, considering the observable risk characteristics. The process of evaluating risks in order to determine the insurance premium is performed by the actuaries, which over time proposed and applied different statistical models. In this context, linear regression, used to evaluate the impact of explanatory variables on the phenomenon of interest (studied risk), has been replaced starting with 1980 by the Generalized Linear Models (GLMs). GLMs allow modeling a non-linear behavior and a non-Gaussian distribution of residuals. This aspect is very useful for the analysis of non-life insurance, where claim frequency and claim cost follow an asymmetric density that is clearly non-Gaussian. GLMs development has contributed to quality improvement of the risk prediction models and to the process of establishing a fair tariff or premium given the nature of the risk.

The main objective of this paper is to apply GLM models in order to assess the premiums applied to each insured, in an equitable and reasonable manner. In this purpose, the paper is structured as follows. Section 2 presents a brief review of the literature regarding the approach of GLMs in non-life insurance pricing. Section 3 describes the methodological framework used in this paper. Each subsection of this part analyzes the estimation methods of frequency and cost of claims, leading to the calculation model of the pure premium. Section 4 is dedicated to a study applied to auto insurance branch in order to highlight how to identify the risk factors that allow dividing the insurance portfolio in tariff classes and how to obtain the corresponding pure premium. Section 5 presents the main conclusions of the study.

2. A brief history of non-life insurance pricing

Historically, actuarial science has been limited to using the standard Gaussian linear regression in order to quantify the exogenous variables impact over the phenomenon of interest. The linear model, proposed by Legendre and Gauss in 19th century, has taken the lead in econometrics, but the applicability of this model in insurance has been found to be difficult. In this context, the linear modeling implies a series of hypothesis that are not compatible with the reality imposed by the frequency and cost of the damages generated by the risks occurrence. Considering this, the most important assumptions are the Gaussian probability density, the linearity of the predictor and homoscedasticity.

While the complexity of the statistical criteria has become more pronounced, the actuaries had to solve the problem of finding some models that explain as realistic as possible the risk occurrence. An important milestone of the non-life insurance pricing development is considered to be the minimum bias procedure implemented by Bailey and Simon (1960). The principle of this method consists of defining randomly the link between the explanatory variables, the risks levels and the distance between the predicted values and the observed ones. Once these elements are established, an iterative algorithm calculates the coefficient associated with each risk level using the minimizing distance criterion. Although it was created outside a recognized statistical framework, this algorithm has been found subsequently to be a particular case of the GLM models.

The implementation merits of these models, both in actuarial science and statistics, goes to British actuaries from City University, John Nelder and Robert Wedderburn (1972). They demonstrate that the generalization of the linear modeling allows the deviation from the assumption of normality, extending the Gaussian model to a particular family of distribution, namely the exponential family. Members belonging to this family include, but not limited to, the Normal, Poisson, Binomial and the Gamma distributions.

Comparing to the minimum bias procedure techniques, the GLM models have the advantage of a theoretical framework that allows the usage of statistical tests in order to evaluate the fitting of models. Also, Nelder and Wedderburn (1972) are suggesting that the estimation of the GLMs parameters to be performed through maximum likelihood method, so that the parameter estimates are obtained through an iterative algorithm. The contribution of Nelder in developing and completing the GLMs theory continues while collaborating with the Irish statistician Peter McCullagh, whose paper (1989) offers detailed information on the iterative algorithm and the asymptotic properties of the parameter estimations.
Since the foundation of GLMs principles, the complexity and abundance of papers is remarkable, many authors and scientists succeeded to highlight, develop or improve the assumptions imposed by the practical application of these models in non-life insurance. Among the precursors of GLMs approach as the main statistical tool in determining the insurance tariffs is noted Jean Lemaire (1985). Resorting to these models, he aims to estimate the probability of risk occurrence in auto insurance, to establish the insurance premium and also to measure the effectiveness of the models used to estimate it. In this area, a significant contribution goes also to Arthur Charpentier and Michel Denuit (2005) who have succeeded to cover, in a modern perspective, all the aspects of insurance mathematics. Some recent studies have pointed out the contribution of Jong and Zeeler (2008), Kaas and al. (2009), Frees (2010), or Ohlsson and Johansen (2010), who have highlighted the GLMs particularities in non-life insurance risk modeling.

3. Research methodology

The methodological section of the paper aims to present some specific issues related to the GLMs and the role of these models within non-life insurance business. The main focus is on the definition, interpretation and presentation of the properties and limits of the insurance premium calculation models.

3.1. Generalized Linear Models (GLMs)

Starting with the actuarial illustration of McCullagh and Nedler (1989), the GLMs have become standard industry practice for non-life insurance pricing. These models are defined as an extension of the Gaussian linear models framework that is derived from the exponential family. The purpose of these models is to estimate an interest variable \( Y \) depending on a certain number of explanatory variables \( X \).

During the actuarial analysis, considering that the variables represent exogenous information about the insured or his assets, the variable \( Y \) can be one the followings:

- a binary variable that can only have the value zero or one, the phenomenon studied in this case being the probability of a risk occurrence, for which it applies the binomial regression models (logit, probit and log-log complementary models);
- a count variable, with values belonging to the set of natural numbers, while following the modeling frequency of claims. In this case the Poisson regression model will be applied;
- a real positive variable, with values belonging to the set of positive real numbers, while following the econometric analysis of the claim cost. In this case the Gamma regression model will be applied.

Conditioned on the explanatory variables \( X \), the random variables \( Y_1, Y_2, ..., Y_n \) are considered to be independently, but not identically distributed, that have the probability density generated by the expression:

\[
f(y_i | \theta_i, \phi) = \exp \left( \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right), \quad y_i \in S
\]

where \( S \) represents a subassembly that belongs to \( \mathbb{N} \) or \( \mathbb{R} \) set, \( \theta_i \) is the natural parameter and \( \phi \) is the scale parameter. In binomial and Poisson distributions, the scale parameter has the value 1, and for the Gamma distribution \( \phi \) is unknown and has to be estimated.

From (1) it can be define the probability density of the variables \( Y_1, Y_2, ..., Y_n \) using the following expression:

\[
f(y | \theta, \phi) = \prod_{i=1}^{n} f(y_i | \theta_i, \phi)
\]
Similar with the Gaussian model approach, the purpose of the econometrical modeling is to obtain the expected values of the dependent variables through conditional means, given independent observations. In this case, there are searched the parameters $\beta_1, \beta_2, \ldots, \beta_p$, through a function ($g$) of the dependent variable mean ($\mu_i$), written as a linear combination of the exogenous variables $X_i$:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = x_i^T \beta = \eta_i$$

the monotonous and differentiable function $g$ is known as a link function because it connects the linear predictor $\eta_i$ with the mean $\mu_i$.

Since the main objective of the paper is to establish the pure premium, further on the focus is on introducing and elaborating the estimation models of the frequency and cost of claims.

### 3.2. Estimation model of claim frequency

Within non-life insurance business, it has been demonstrated that the usage of the GLMs techniques in order to estimate the frequency of claims, has an a priori Poisson structure. Antonio and Valdez (2012) present the Poisson model as the modeling archetype of the “event counts”, also known in actuarial literature as the frequency of claims. According to Dionne and Vanasse (1989, 1992), Denuit and Lang (2004), Gourieroux and Jasiak (2004), the Poisson model represents the main tool for the modeling claim frequency in non-life insurance.

An important milestone in the development of counts regression models is the contribution of Cameron and Trivedi (1998). They have managed to highlight the particularities of Poisson regression approach in modeling the claim frequency as a particular case of GLMs. To confirm this assertion, the authors underlines that the counts analyses within auto insurance branch is not restricted by a limited number of independent trials, but the accent is on the risk exposure such that $n$ increases significantly while $n \times p$, representing the number of “successes”, remains finite. In this case, Poisson distribution represents the correct statistical model to assess the probability of 0, 1, 2... risks occurrence.

Assuming that the discrete random variable $Y$ (claim frequency or observed number of claims), conditioned by the vector of explanatory variables $X_i$ (the observable risk characteristics), is Poisson distributed. Therefore, for the insured $i$, the probability that the random variable $Y_i$ takes the value $y_i$ ($y_i \in \mathbb{N}$), is given by the density:

$$f(Y_i = y_i | x_i) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

The Poisson distribution implies a particular form of heteroskedasticity, leading to the equidispersion hypothesis or the equality of the mean and variance of claim frequency. In other words, the Poisson distribution parameter represents at the same time the mean and the variance of the distribution:

$$E(Y_i | x_i) = V(Y_i | x_i) = \lambda_i = e^{x_i^T \beta}$$

The standard estimator for this model is the maximum likelihood estimator. The likelihood function is defined as follows:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i^T \beta}} (e^{x_i^T \beta})^{y_i}}{y_i!}$$
Using a logarithm in both sides of (7), it is obtained the log-likelihood function:

\[ LL(\beta) = \sum_{i=1}^{n} \left[ y_i \ln \lambda_i - \lambda_i - \ln y_i! \right] = \sum_{i=1}^{n} \left[ y_i x_i' \beta - e^{x_i' \beta} - \ln y_i! \right] \]  \hspace{1cm} (8)

It can be easily verified that the first two partial derivatives of the log-likelihood function exist and are expressed as follows:

\[
\frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \lambda_i)x_{ij} = \sum_{i=1}^{n} \left( y_i - e^{x_i' \beta} \right)x_{ij} \hspace{1cm} (9)
\]

\[
\frac{\partial^2 LL(\beta)}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n} \lambda_i x_{ij} x_{ik} = -\sum_{i=1}^{n} \left( e^{x_i' \beta} x_{ij} x_{ik} \right) \hspace{1cm} (10)
\]

The maximum likelihood estimators \( \hat{\beta}_j \) are the solutions of the previous likelihood equations that are obtained by differentiating the log-likelihood in terms of the regression coefficients and solving them to zero. The equations forming the system are not generating explicit solutions and therefore they have to be solved numerically using an iterative algorithm. The most common iterative methods are considered Newton-Raphson and Fisher information.

3.3. Estimation model of claim cost

Claim amount (or economic compensations to be paid) is more difficult to predict than claim frequency. In this case, the analysis is less clear because there are no distributions for positive real values. The literature argues that the classical method allowing the econometric modeling of claim costs is the Gamma model.

For the modeling of claim cost in auto insurance, Jean Pinquet (1997) describes a simple parametric model, but realistic, based on Gamma distribution, which represents another generalization of the exponential family.

Noting with \( c_{i1}, c_{i2}, \ldots, c_{in_i} \) the costs of the claims caused by insured \( i \) and assuming that they are independently Gamma distributed, the probability density function is given by:

\[ f(c_i) = \frac{1}{\Gamma(v)} \left( \frac{\nu c_i}{\mu_i} \right)^v \exp \left( \frac{-\nu c_i}{\mu_i} \right), \quad c_i > 0 \]  \hspace{1cm} (11)

verifying the mean \( E(c_i) = \mu_i \) and variance \( V(c_i) = \mu_i^2 / \nu \).

The log-likelihood function for the Gamma model is given as follows:

\[ \mathcal{L}(\beta) = \prod_{i \mid y_i > 0} \prod_{k=1}^{y_i} \left( \frac{1}{\Gamma(v)} \left( \frac{\nu c_{ik}}{\mu_i} \right)^v \exp \left( \frac{-\nu c_{ik}}{\mu_i} \right) \frac{1}{c_{ik}} \right) \]  \hspace{1cm} (12)

The equations of the log-likelihood function allowing to obtain the estimators \( \hat{\beta}_j \) are given by:

\[
\frac{\partial LL(\beta|c)}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \sum_{i \mid y_i > 0} \sum_{k=1}^{y_i} \left( -\ln \mu_i - \frac{\nu c_{ik}}{\mu_i} \right) = 0 \hspace{1cm} (13)
\]

or can be simplified as:
\[ \sum_{i\mid y_i > 0} \sum_{k=1}^{y_i} x_{ij} \left( 1 - \frac{\nu c_{ik}}{\mu_i} \right) = 0 \] (14)

Defining \( \hat{c}_i = \hat{\mu}_i = e^{x_i \hat{\beta}} \) the estimated cost of the claims for the insured \( i \), the maximum likelihood estimates \( \hat{\beta} \) is the solution of the following equation:

\[ \sum_{i\mid y_i > 0} \left( y_i - \frac{c_i}{\hat{c}_i} \right) x_i = 0 \] (15)

The previous expression is being interpreted as an orthogonality relationship between explanatory variables and residuals. Actuarial literature argues that the main advantage of applying the Gamma model is due to parameters \( \nu \) and \( \mu \), through which more flexibility is obtained while estimating the cost of claims.

### 3.4. Calculation model of pure premium

In non-life insurance, the pure premium represents the expected cost of all claims declared by policyholders during the insured period. The calculation of the premium is based on statistical methods that incorporate all available information about the accepted risk, thereby aiming at a more accurate assessment of tariffs attributed to each insured.

The basis for calculating the pure premium is the econometric modeling of the frequency and cost of claims depending on the characteristics that define the insurance contracts. The pure premium is the mathematical expectation of the annual cost of claims declared by the policyholders and is obtained by multiplying the two components, the estimated claim frequency and cost:

\[ E \left[ \sum_{i=1}^{N} C_i \right] = E[Y] \times E[C_i] \] (16)

for the claims amount \((C_1, C_2, ...)\) independent of their number \((Y)\).

As shown by Charpentier and Denuit (2005), the separate approach of frequency and cost of claims is particularly relevant because the risk factors, which influence the two components of the pure premium, are usually different. Essentially, the separate analysis of the two phenomena provides a clearer perspective on how the risk factors are influencing the premium.

### 4. Numerical application

The empirical part of this paper includes a brief presentation of the used data, based on which a numerical illustration of the described techniques is performed.

#### 4.1. Data used

The empirical analysis is provided using the version 9.3 of the statistical software SAS. The phenomenon concerns the theft and damage of the vehicles, for which the insurance is covering the losses within the limits of the amount insured. The proposed methods can be also approached in other non-life insurance branches (auto liability insurance, buildings and fire insurance, travel insurance etc.), taking into consideration the features of the corresponding contracts.

In this paper, the data used constitute a French auto insurance portfolio containing 50000 policies registered during the year 2009. The elements included in the policies are the factors considered in this study. Thus, except the explained variables, the frequency and cost of claims, the other ones are considered risk factors, known a priori by
the insurer and are used to customize the profile of each insured. These exogenous variables reflect the insured
characteristics: age, profession (unemployed, employed, housewife, retired, self-employed); the vehicle features: type
(A, B, C, D, E, F), category (small, medium, large), power (<40Kw, 40-70Kw, >70Kw), purpose of vehicle usage
(private, professional); the insurance contracts characteristics: age, bonus-malus coefficient.

Among the explanatory variables introduced in the analysis, bonus-malus coefficient presents a particular interest.
This coefficient assumes the increase or decrease of insurance premium depending on the number of claims
registered by an insured during a reference period. Thus, if the policyholder does not cause any responsible accident,
he receives a bonus, meaning that the insurance premium will be reduced. Contrary, if the insured is responsible for
the accident, he is penalized by applying a malus, which will have the consequence of a premium increase.

These increases and decreases are based on a standard tariff defined by the insurer, depending on which the
premium is multiplied by a coefficient. The basic coefficient is 1 and it corresponds to the reference premium of the
insurance company. If the bonus-malus coefficient is lower to this value, a bonus is applied, and if it is higher, a
malus is considered. More specifically, the French bonus-malus system involves a malus of 25% for a claim
declared and a bonus of 5% for the non-declaration of any claims in the reference period, usually a year. In this way,
the system aims the encouragement of prudent insured drivers and the discouragement of those who, for various
reasons, register sever losses. In the studied portfolio, the calculations corresponding to the bonus-malus coefficient
are already generated, registering negative values, respectively positive, which indicates a decrease, respectively an
increase, of the insurance premium.

4.2. Results

In this subsection, there are presented and interpreted the results obtained through the application of mentioned
models, based on which the insurance pure premium is determined.

Poisson model

Within SAS, the GENMOD procedure is used to fit the regression models, in the framework of GLMs. The Type
3 analysis, generated by using this procedure, enables the contribution evaluation of each variable taking into
consideration all the others exogenous variables. The results of the Type 3 analysis are presented in Table 1. In the
column Chi-Square is calculated, for each variable, two times the difference between the log-likelihood of the model
which includes all the independent variables and the log-likelihood of the model obtained by deleting one of the
specified variables. This statistics test follows the asymptotic Chi-Square distribution with df degrees of freedom,
representing the number of parameters associated to the analyzed variable. The column Pr>ChiSq indicates the
probability associated to the likelihood ratio test which appreciates the impact of each risk factor on the studied
phenomenon.

| Source                      | Chi-Square* | Pr > ChiSq* | Chi-Square** | Pr > ChiSq** |
|-----------------------------|-------------|-------------|--------------|--------------|
| Age                         | 87.75       | <.0001      | 89.87        | <.0001       |
| Occupation                  | 63.76       | <.0001      | 63.71        | <.0001       |
| Type                        | 46.02       | <.0001      | 46.22        | <.0001       |
| Category                    | 4.13        | 0.1268      | -            | -            |
| Power                       | 1.60        | 0.4499      | -            | -            |
| Use                         | 83.61       | <.0001      | 83.58        | <.0001       |
| Bonus-Malus                 | 451.76      | <.0001      | 451.69       | <.0001       |
| Age of insurance contract   | 35.14       | <.0001      | 35.23        | <.0001       |

(*) Poisson regression including all the explanatory variables
(**) Poisson regression including only the significant explanatory variables
It can be observed that the variable denoting the category of vehicle is not statistically significant as it yields a p-value of 0.1902 greater than the risk $\alpha$ of 0.05. In consequence, this variable is excluded from the model and the analysis will continue in the same manner until it is obtained the optimal combination of factors (p-value < 0.05) which can explain the variation of claim frequency. After excluding from the model the non-significant factors (category and power of vehicle), it is noticed that all the other explanatory variables are now statistically relevant, which clearly underlines their influence on the claims frequency.

Analyzing the results from Table 3, a decrease of the claims frequency can be observed along with an increase in the age of the insured and the age of the insurance contracts. On the contrary, when the bonus-malus coefficient increases, the frequency of claims increases as well.

Gamma model
The next stage in establishing the insurance premium resides in estimating the cost of claims based on the risk factors considered by the insurance company. Therefore, for the Gamma model, the obtained results (Table 2) suggest that for the analyzed portfolio, the cost of claims is influenced by the age of insured, the profession of insured and the type of vehicle. The influence factors of the claims cost are different from the factors corresponding to the frequency of claims, fact that confirms the assumption suggested by the actuary literature regarding the isolated analysis of these two phenomena.

| Source         | Chi-Square* | Pr > ChiSq* | Chi-Square** | Pr > ChiSq** |
|----------------|-------------|-------------|--------------|--------------|
| Age            | 46.13       | <.0001      | 47.75        | <.0001       |
| Occupation     | 60.50       | <.0001      | 61.00        | <.0001       |
| Type           | 39.81       | <.0001      | 39.57        | <.0001       |
| Category       | 2.62        | 0.2701      | -            | -            |
| Power          | 2.11        | 0.3491      | -            | -            |
| Use            | 0.87        | 0.3510      | -            | -            |
| Bonus-Malus    | 0.97        | 0.3253      | -            | -            |

(*) Gamma regression including all the explanatory variables
(**) Gamma regression including only the significant explanatory variables

Reviewing the coefficient signs from Table 3, a decrease of the claim cost can be noticed once the age of policyholders increases. Basing on the claims cost, it is not possible to obtain conclusive information regarding the risk occurrence probability and the insurance company cannot properly divide the policyholders. Nevertheless, the amount of cost is a fundamental component considered while establishing the insurance premium.

Pure premium model
The process of establishing the insurance premium resides in using the same procedure GENMOD as noticed in previous cases, the obtained results being summarized in Table 3. In this stage of non-life insurance pricing, the explained variable is the product between the estimated frequency and the estimated cost of claims:

$$E \left[ \sum_{k=1}^{N_i} c_{ik} \right] = E[N_i]E[C_{11}] = \exp \left( (\beta_{freq} + \beta_{cost})^t x_i \right)$$

the calculated value representing the insurance pure premium established for insured $i$, characterized by the variables vector $x_i$.

Considering this relationship within the analyzed insurance portfolio, the pure premium for each category of policyholders is established based on the Gamma regression model, that includes all the statistical relevant tariff
variables that explain the variation of the claim frequency and costs. Therefore, this regression model allows to obtain the pure premium corresponding to each tariff class through the expression: $\exp(\beta^T x_i)$. In summary, the obtained results lead to using tariffs corresponding to the gravity of risks insured by the insurance company.

The default purpose of the non-life insurance pricing is deduced from the idea that the new policies will be mostly established for drivers that fit the profile generated after establishing the insurance tariffs. Thereupon, the pure premium will be used for the new insureds that will be classified in one of the tariff classes already defined.

Table 3. Analysis of Parameter Estimates.

| Source                      | Estimate * | Std Error * | Estimate ** | Std Error ** | Pure premium |
|-----------------------------|------------|-------------|-------------|--------------|--------------|
| Intercept                   | -2.1541    | 0.1193      | 8.4556      | 0.1076       | 6.3029       |
| Age                         | -0.0189    | 0.0020      | -0.0124     | 0.0018       | -0.0314      |
| Occupation (employed)       | -0.2611    | 0.0617      | -0.1674     | 0.0638       | -0.4276      |
| Occupation (housewife)      | -0.4189    | 0.0743      | 0.0238      | 0.0765       | -0.3945      |
| Occupation (retired)        | -0.2348    | 0.1099      | 0.0231      | 0.1113       | -0.2101      |
| Occupation (self-employed)  | 0.0489     | 0.0657      | 0.2972      | 0.0693       | 0.3471       |
| Type (A)                    | -0.3760    | 0.0905      | -0.4205     | 0.0932       | -0.7952      |
| Type (B)                    | -0.4431    | 0.0940      | -0.4457     | 0.0964       | -0.8888      |
| Type (C)                    | -0.3873    | 0.1003      | -0.3000     | 0.1033       | -0.6863      |
| Type (D)                    | -0.2068    | 0.0929      | -0.1718     | 0.0958       | -0.3780      |
| Type (E)                    | -0.0601    | 0.0989      | -0.2026     | 0.1021       | -0.2628      |
| Use (private)               | 0.4282     | 0.0481      | -           | -            | 0.4282       |
| Bonus-Malus                 | 0.0082     | 0.0004      | -           | -            | 0.0082       |
| Age of insurance contract   | -0.0286    | 0.0049      | -           | -            | -0.0286      |

(*) Poisson regression results
(**) Gamma regression results

5. Conclusions

The non-life insurance pricing consists of establishing a premium or a tariff paid by the insured to the insurance company in exchange for the risk transfer. The premium insurance calculation is based on the multiplication of estimated frequency and cost of claims.

This paper considers an analysis of the Generalized Linear Models in order to establish the pure premium given the characteristics of the policyholders. Therefore, as a first stage, the frequency of claims is estimated through Poisson regression model. In the next analysis stage, by using the Gamma model, the estimated average level of the claim cost corresponding to each class of policyholders is determined.

Eventually, the research results have shown that for the new customers, the insurance premium will be established while considering a series of risk factors, like age and profession of the insured, purpose of vehicle usage, bonus-malus coefficient and the age of the insurance contract. Therefore, within the analyzed insurance portfolio, a decrease of the pure premium is observed along with an increase of the insured’s age and the age of the insurance contract, and also an increase along with the bonus-malus coefficient growth.

The conclusions of this study are representative and useful for the insurance company business, but they do not present a generalized character, therefore they cannot be applied to all portfolio or insurance companies. On one side, this aspect is justified by the data used and the risk factors considered during the analysis process, meaning that every insurer can use different information on the insured to their benefit. On the other side, the used data is not obtained through a random selection related to the entire population of policyholders.
References

Antonio, K., Valdez, E.A., 2012. Statistical Concepts of A Priori and A Posteriori Risk Classification in Insurance. Advances in Statistical Analysis 96(2), 187-224.

Bailey, R.A., Simon, L.R.J., 1960. Two Studies in Automobile Insurance Ratemaking. ASTIN Bulletin, 1(4), 192-217.

Cameron, A. C., Trivedi, P. K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Charpentier, A., Denuit, M., 2005. Mathématiques de l’Assurance Non-Vie, Tome II: Tarification et provisionnement. Economica, Paris.

Denuit, M., Lang, S., 2004. Nonlife Ratemaking with Bayesian GAM’s. Insurance: Mathematics and Economics, 35(3), 627-647.

Dionne, G., Vanasse, C., 1989. A Generalization of Automobile Insurance Rating Models: the Negative Binomial Distribution with a Regression Component. ASTIN Bulletin, 19(2), 199-212.

Dionne, G., Vanasse, C., 1992. Automobile Insurance Ratemaking in the Presence of Asymmetrical Information. Journal of Applied Econometrics, 7(2), 149-165.

Frees, E. W., 2010. Regression Modeling with Actuarial and Financial Applications. Cambridge University Press, Cambridge.

Gourieroux, C., Jasiak, J., 2004. Heterogeneous Model with Application to Car Insurance. Insurance: Mathematics and Economics, 34(2), 177-192.

Jong, P., Zeller, G., 2008. Generalized Linear Models for Insurance Data. Cambridge University Press, Cambridge.

Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M., 2009. Modern Actuarial Risk Theory: Using R. 2nd ed. Springer, Berlin.

Lemaire, L., 1985. Automobile Insurance: Actuarial Models. Kluwer Academic, Dordrecht.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. 2nd ed. Chapman and Hall, London.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Interactive Models. Journal of the Royal Statistical Society, A 135(3), 370-384.

Ohlsson, E., Johansson, B., 2010. Non-Life Insurance Pricing with Generalized Linear Models. Springer, Berlin.

Pinquet, J., 1997. Allowance for Cost of Claims in Bonus-Malus Systems. ASTIN Bulletin, 27(1), 33-57.