Revista Brasileira de
Risco e Seguro

# A Comparative Sudy of Pricing Methods of Automobile Insurance in Brazil

**Yuri Rosembaum Silva**
Master in Actuarial Science – Cass Business School
yuri.rosembaum@gmail.com

**Luís Eduardo Afonso**
Associate Professor – Department of Accounting and Actuarial Science – FEA/USP
lafonso@usp.br

## *Summary*

This work aims to present and compare some pricing techniques in the automobile insurance portfolio. The methods used in the pricing of products offered by an insurer are essential to their competitiveness. The four methods analysed are Modelling of the Pure Premium by historical aggregated claims, Modelling of the Pure Premium by expected aggregated claims, Classical Linear Models and Generalized Linear Models (GLM). The data used came from SUSEP's Automobile Statistics System database (AUTOSEG), taken from the second half of 2007 to the first half of 2011. The data from the last half of 2011 was not used for the pricing in a form that could verify the adherence of the techniques to the database and perform the comparison between the expected claims, discrepancies in relation to actual prices and dispersion of the expected and actual risk premiums. To this end, a variable termed *Efficiency* was created, by means of which the increase in price is calculated so that there is a reduction of one percentage point compared to the actual price in the first half of 2011. The modelling techniques of historical pure premium and GLM had the best pricing efficiency (1.60), but the former showed a lower variation than the latter (standard error 5.26% against 17.52%). The comparative advantage of the historical pure premium can be explained by the low variation of the basic indicators over time and the fact that it contains the data for all insured vehicles in Brazil, enabling high adherence to the technique. However, GLM appears to be an interesting pricing alternative for medium portfolios that are seeing growth, or that explore possible niche markets.

## *Key Words*

Automobile Insurance; Pricing; Generalized Linear Models.

## *Contents*

## *Sinopse*

**Um estudo comparativo sobre métodos de precificação de seguro de automóveis no Brasil**

**Yuri Rosembaum Silva**
Mestre em Ciências Atuariais – Cass Business School
yuri.rosembaum@gmail.com

**Luís Eduardo Afonso**
Professor Associado – Departamento de Contabilidade e Atuária – FEA/USP
lafonso@usp.br

## *Resumo*

Este estudo tem o objetivo de apresentar e comparar algumas técnicas de precificação do portfólio de seguro de automóveis. Os métodos utilizados na precificação dos produtos oferecidos por uma seguradora são essenciais para sua competitividade. Os quatro métodos analizados são a Modelagem do Prêmio Puro por sinistros historicamente agregados, Modelagem do Prêmio Puro por sinistros previstos agregados, Modelos Clássicos Lineares e Modelos Lineares Gerais (GLM). Os dados usados vieram do Sistema de Estatísticas de Automóveis da SUSEP (AUTOSEG), tirados do segundo semestre de 2007 até o primeiro semestre de 2011. Os dados do último semestre de 2011 não foram utilizados para a precificação de forma que pudesse ser verificada a aderência das técnicas no banco de dados e comparar as afirmações previstas, discrepâncias em relação aos preços reais e dispersão do prêmio de risco previsto e real. Para este efeito, a variável *Efficiency* foi criada, pela qual o aumento no preço é calculado para que haja uma redução de um ponto percentual comparado ao preço real no primeiro semestre de 2011. As técnicas de modelagem de prêmios puros históricos e GLM tiveram a melhor eficiência de precificação (1,60), mas a primeira mostrou uma variação mais baixa que a última (erro padrão 5,26% contra 17,52%). A vantagem competitiva do prêmio puro histórico pode ser explicada pela baixa variação dos indicadores básicos ao longo do tempo e o fato de que contém os dados para todos os veículos segurados do Brasil, gerando alta aderência à técnica. No entanto, GLM parece ser uma alternativa de precificação interessante para portfolios médios que estão vendo crescimento, ou que explorem possíveis nichos de mercado.

## *Palavras-chave*

Seguro de Automóveis; Precificação; Modelos Lineares Generalizados.

## *Sumário*

## *Sinopsis*

**Un estudio comparativo sobre métodos de cálculo de precios de seguro de automóviles en Brasil**

**Yuri Rosembaum Silva**
Máster en Ciencias Actuariales – Cass Business School
yuri.rosembaum@gmail.com

**Luís Eduardo Afonso**
Profesor Asociado – Departamiento de Contabilidad y Actuaría – FEA/USP
lafonso@usp.br

## *Resumen*

El objetivo de este estudio es presentar y comparar algunas técnicas de cálculo de precios de la cartera de seguro de automóviles. Los métodos utilizados en el cálculo de los productos ofrecidos por una aseguradora son fundamentales para su competitividad. Los cuatro métodos analisados son el Modelo de la Prima Pura por reclamaciones históricamente agregadas, Modelo de la Prima pura por reclamaciones previstas agregadas, Modelo Clásico Lineal y Modelo Lineal General (GLM). Los datos utilizados vinieron del Sistema de Estadísticas de Automóviles de SUSEP (AUTOSEG), sacados del segundo semestre de 2007 hasta el primer semestre de 2011. Los datos del último semestre de 2011 no fueran utilizados para el cálculo de precios de manera que pudiera ser verificada a adherencia de las técnicas en el base de datos y comparar las afirmaciones previstas, discrepancias en relación con precios reales y dispersión de la prima de riesgo previsto y real. Para este efecto, la variable *Efficiency* fue creada, por la cual el aumento en el precio es calculado para que haya una reducción de un punto porcentual comparado al precio real en el primer semestre de 2011. Las técnicas de Modelo de primas puras históricas y GLM tuvieran la mejor eficiencia de cálculo de precios (1,60), pero la primera demostró una variación más baja que la última (error estándar 5,26% contra 17,52%). La ventaja competitiva de la prima pura histórica puede ser explicada por la baja variación de los indicadores básicos a lo largo del tempo y el facto de que contiene los datos para todos los vehículos asegurados de Brasil, generando alta adhrencia a la técnica. Sin embargo, GLM parece ser una alternativa de cálculo de precio interesante para carteras medianas que están vendo crecimiento, o que exploren possibles nichos de mercado.

## *Palabras-clave*

Seguro de Automóviles; Cálculo de precios; Modelo Lineal Generalizado.

## *Sumario*

## 1. Introduction

The automobile branch has great relevance in the Brazilian insurance market, because the premiums represent about 30% of the total and 50% of the branches of non-life insurance. The sector generated approximately R$ 19.9 billion in retained premiums in 2011 and R$ 23.3 billion in 2012, with nominal growth approximating 17% for branches 0520 (Personal Accident – APP), 0531 (Damage to Vehicle – Body) and 0553 (Damage to Third Parties), according to data from the Superintendence of Private Insurance (SUSEP). However, in recent years the loss ratio (ratio of retained claims and earned premiums) has shown a worrying upward trend, from 64.01% in 2010 to 66.09% in 2011 and 66.26 % in 2012.

The portfolio management process is extremely important for insurers. Items such as management costs and expenses, estimation of technical provisions, claims settlement processes, underwriting rules, pricing methods and other aspects of the policies are fundamental in their activity. And among these, the pricing methods are the focus of this article.

Not all insurers have a statistical modelling process developed to price their portfolios. Many estimate their rate by historical losses in its portfolio, along with a comparison with the prices charged by competitors. To that end, this article aims to present some pricing techniques for an auto insurance portfolio. Some of these strategies involve simple statistical concepts and others employ more elaborate statistical techniques. This way, it seeks by comparison to verify the accuracy of each technique for the determination of future premiums.

The data used is from the database of SUSEP's Automobile Statistics System (AUTOSEG), which contains the information that is sent compulsorily and semi-annually, by all insurance companies operating in the domestic market. With the information of all insured vehicles in the country, the intention is to obtain the value of the actuarially fair rate, with a view to minimizing the effects of dispersion of vehicles amongst insurers, a fact which may cause the under or overestimation of the premium charged to policyholders. This is also performed by risk group, with the aim of distinguishing the largest of the lesser risks and thus enable greater gains through proper pricing.

Studies involving statistical methods for setting automobile insurance tariffs began with Bailey & Simon (1960) who analysed a Canadian automobile portfolio. The authors covered topics such as risk classification and gains in the use of relativities (risk factors/coefficients) for each risk profile in the multiplicative form. In the work of Weisberg, Tomberlin, & Chatterjee (1984), the aim was to make a comparison of the various methods of pricing an automobile portfolio, using ordinary least squares, maximum likelihood and Bayesian estimation. In the text, there is the use of pure premium calculation as the basis for the comparative analysis of the models, as well as the calculation of the premium with frequencies and severities estimated separately; techniques that are to be used in this study.

Later, Santos (2008) addressed the methodological differences between the methods of Linear Models and Generalized Linear Models (GLM). There are gains in the use of the latter, mainly due to the fact that the variable response has a distribution different from the Normal (which can be any exponential family distribution and without the need for constant variance). In addition, the response variable does not need to be written on the basis of a linear combination and, yes, by means of the linear predictor function. These two models will also be used in the current study.

Based on this framework, the four techniques to be analysed are:

- Modelling of the Pure Premium by the aggregated claims history;
- Pure Premium Modelling estimating the distributions of frequencies and severities for the formulation of the expected aggregated claims;
- Classic Linear Models;
- General Linear Models.

The methods of comparison between the techniques are:

- Model explanation capacity, through the standard errors in risk premiums;
- Linear adjustment between historical prices and those obtained with the techniques;
- Analysis of the adequacy of future premiums with the use of the technique, since the last half of data is not used in the pricing to check the result that the portfolio would get with each strategy.

This study is divided into four sections, including this introduction: the theoretical basis is presented in section 2. The methodology and the data are reported in section 3. In section 4 the results are shown. Finally, section 5 draws the conclusions.

## 2. Theoretical Basis

The study of pricing was pioneered with the work of Bailey & Simon (1960) which focused on two main points: (1) the use of risk classification strategy amongst tariff classes to achieve the separation between good and bad risks, using the experience of Canada's automobile portfolio; (2) the presentation of methods for obtaining the relativities of a tariff, achieving a comparison between the multiplicative and additive methods. The authors suggest using the least squares method for the estimation of the parameters of the model and presentation of the minimum bias method. The latter consists of verifying that subclasses created for the tariff are not skewed, in so far as the total model does not present bias, in order to ensure the credibility of the classes used. Bailey (1963) widened this study, continuing the comparison between multiplicative and additive models using two techniques: least squares and the average absolute difference. The concept of zero bias was also presented, in which average difference between the observed rates and adjusted should be equal to zero. As shown by Zehnwirth (1994), the works of Bailey and Simon provided the basis for the development of Generalized Linear Models (GLM) and encouraged texts such as Brown (1988), which carried

out an application of Bailey's methodologies to the GLM. Weisberg *et al*. (1984) presents several models for estimation of the premiums. These models will be used in this article for predicting future premiums, based on the mean squared error of each model and its accuracy. This is one of the objectives of the present work. Following is the presentation of the models employed.

### 2.1 Pure Premium by aggregated claims history

According to McClenahan (2001), the expected average loss for the insured persons belonging to a certain class of risk is known as Pure Premium (or risk premium). In this technique, the expected value of claims is based on the portfolio history. Thus, one can define the pure premium *PP* by expression 1, in which $L_x$ is the historical value of each claim occurred for a certain class of risk and *e* represents the number of vehicles exposed to that risk class.

$$PP = \frac{\sum_{0}^{k} L_x}{e} \tag{1}$$

Because they do not estimate frequencies and severities, assumptions are not made about the change of the parameters over time. In addition, similar to Weisberg *et al*. (1984), this method will serve as a basis for comparison to the other methods that will be discussed later.

### 2.2 Pure Premium by expected aggregated claims

As shown in Santos (2008), because the occurrence of an accident is random, the total amount of the indemnities arising from a particular risk group can be defined by a composite process called *Aggregated Claims S(t)*, formed by two principal components:

*N(t)*: amount of claims incurred in the range *0* to *t*. This is a discrete random variable, used in the process of counting events. Known distributions of this type are, for example, Binomial, Poisson and Negative Binomial.

$X_i$: value of the $i^{th}$ claim indemnity occurring in the range *0* a *t*. This continuous variable has values distributed independently of the behaviour of *N(t)*. Known distributions with this profile are Normal, Gamma and Inverse Gaussian.

The value of *S(t)* is obtained by the use of equation 2:

$$S(t) = \sum_{i=0}^{N(t)} X_i \tag{2}$$

According to Boland (2007), in order to obtain *E(S)* and *Var(S)* the following expressions are used:

$$E(S) = E(X)E(N) \tag{3}$$

$$Var(S) = E^2(X)Var(N) + Var(X)E(N) \tag{4}$$

This way, one can define the *Pure Premium* of each class of risk through expression 5:

$$PP = E(S) + \theta \sqrt{Var(S)} \quad (5)$$

In which θ is the confidence level.

For both Pure Premium models it is possible to cite the McClenahan (2001) procedure in order to obtain the relativities for use in calculating the tariffs. For each class of risk it should be decided which group will serve as a basis for category comparison (e.g. in the sex category, male will be the base group). Therefore, simply divide the pure premium of each class by the Pure Premium base, generating the tariff relativity. The calculated relativities will make the composition of tariff for both models.

## 2.3 Classic Linear Models

The traditional multiple linear regression model is defined by equation 6, in which *Y* is the phenomenon to be explained by the explanatory variables, *Xp* are the explanatory variables, *βp* are the coefficients to be estimated and ε is the random error term:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \quad (6)$$

So that the linear model can be estimated by ordinary least squares, some assumptions must be valid (Gujarati, 2009):

- *Error with Zero Mean*. The expected value of the error is null for each observation, E(*ε*)= 0.
- *Homoscedasticity*. The variance of the errors does not change with the increase of $X_j$; in other words, the error is a random variable with constant variance. One can define *Var* (*ε*)= $\sigma^2$ Thus, we can affirm that the errors and the variable *Y* have the same variance, being that this fact implies that *Y* has the same form regardless of the values of *X*.
- *Normality*. The distribution of *ε* must be normal with constant variance $\sigma^2$. Thus, $\varepsilon \sim N(0, \sigma^2)$
- *Absence of autocorrelation of the residuals*. The error associated with each observation is independent of the errors associated with the other observations. In this way, $Cov(\varepsilon_i, \varepsilon_j) = 0$.
- *Linearity*. It is assumed that the relationship between the dependent variable *Y* and the explanatory variables $X_j$ is linear. It is worth mentioning that this relationship can be the result of a transformation applied to a non-linear relationship. Thus, the adjustment of several non-linear functional forms to the classic model is possible.

## 2.4 Generalised Linear Models (*GLM*)

Generalised linear models are an extension of the classical linear model. According to Cordeiro (1986), GLM are defined by the probability distribution for the response variable, a set of explanatory variables that describe the linear structure of the model and a link function between the response variable and the linear structure. According to de Jong & Heller (2008) the two main differences between the GLM and classical linear models are:

- It is not necessary the distribution to be normal, and can be any from the family of exponential distributions;
- Even if there is a linearity structure for the model, the function that performs the relationship between the expected value and the vector of variables can be any differentiable function.

The GLM can be expressed with the use of three components:

i) **Random Component**
The response variable $Y$ is represented by a set of independent random variables with a distribution belonging to the exponential family, with probability density function *PDF* described by equation 7. In this, $f$ is the dispersion parameter, $\theta$ is the canonical parameter and $b(\theta)$ and $c(y,\phi)$ are real known functions. The choice of these functions determines the probability function $f(Y)$ that can be a Binomial, Negative Binomial, Poisson, Normal, Inverse Normal or Gamma.

$$f\left(Y|\theta,\phi\right)=\exp\left(\frac{y\theta-b\left(\theta\right)}{\phi}+c\left(y,\phi\right)\right) \quad (7)$$

ii) **Structural Component**
The structural component (linear predictor) $\eta = (\eta_1, \eta_2, .....\eta_n)^T$ is a linear function of the unknown parameters $\beta = (\beta_1, \beta_2, ..... \beta_n)^T$, described by equation 8, in which $X$ is a matrix $n$ x $k$ and $k < n$:

$$\eta = X\beta \quad (8)$$

iii) **Link Function**
The expected value $\mu$ is related to a linear predictor through the link function $g(\mu)$.

$$\eta = X\beta = g\left(\mu\right) \quad (9)$$

Once the main definitions are presented, four steps should be followed for estimation of the model:

- Choose the distribution for the response variable $f(Y)$;
- Obtain the link function $g(\mu)$;
- Set the predictive explanatory variables $X_i$;
- Estimating the model parameters $\beta_i$.

## 3. Methodology and Data

The data used is from SUSEP'S Automobile Statistics System (AUTOSEG), which contains the information that is submitted compulsorily and semi-annually by insurers operating in the domestic market. It is worth mentioning that the database in question has only information about the insured vehicle (Own damage coverage). The qualitative variables available are tariff category, model/year of the vehicle, the driver's age, sex and SUSEP regions. Quantitative variables are average amount insured, amount of exposure, frequencies and indemnities for collision, fire, theft and other losses. The data employed in this article includes the period from the second half of 2007 to the first half of 2011.

For each half-year period there is a file named *base_auto.mdb*. This work used the main table *arq_casco_comp* and auxiliary tables *auto_cau* (causes of the claim), *auto_cob* (types of cover), *auto_idade* (age group), *auto_reg* (description of the SUSEP region), *auto_sexo* (sex of insured) and *auto_vei2* (vehicle table). The aggregate of seven half-year periods resulted in a database with 13.450.949 records. The principle files of the database *arq_casco_comp* are:

- *COD_TARIF* (category code of the vehicle);
- *REGIAO* (Group codes for SUSEP region);
- *COD_MODELO* (FIPE code of the vehicle);
- *ANO_MODELO* (Reference year of vehicle model);
- *SEXO* (sex of the insured, or indication of legal person);
- *IDADE* (code for the age of the insured);
- *EXPOSICAO1* (value of vehicles exposed during the analysis period);
- *PREMIO1* (Premium exposed – or gained – based on the amount paid by the insured)
- *IS_MEDIA* (average value of the vehicle at the time of contracting – or amount insured– IS);
- *FREQ_SIN1* e *INDENIZ1* (Quantity of occurrences and indemnities for theft or robbery);
- *FREQ_SIN2* e *INDENIZ2* (Quantity of occurrences and indemnities for partial collisions);
- *FREQ_SIN3* e *INDENIZ3* (Quantity of occurrences and indemnities for total collisions);
- *FREQ_SIN4* e *INDENIZ4* (Quantity of occurrences and indemnities for fire);
- *FREQ_SIN9* e *INDENIZ9* (Quantity of occurrences and indemnities for other losses); and
- *ENVIO* (half-yearly submission to the SUSEP database).

The *COD_TARIF* field of the database was not used as, for several records, the same FIPE code (unique code for each vehicle) was found for more than one category of the database (e.g. the same FIPE code classified passenger car and Pick-up), which skew the analysis for this variable. So, the *auto_vei2* table containing the FIPE code and the SUSEP vehicle groups was used in order that unique categories were created for each vehicle. This new categorization will be used in this work.

All nominal values were deflated with the use of the National Consumer Price Index (IPCA) for each half-year period, based on the first half of 2011.

In relation to the sex of the insured, there are four categories: male, female, legal and others. A discrepancy was found in the legal category because it contained almost entirely premiums data (only). By contrast, the other categories had only data for indemnities and very little data on premiums. Thus, the two categories were grouped into a new category – *others*.

Tables 1 to 5 summarizes the characteristics of the main variables. The *Freq* fields are the percentages of claims for each type of loss (*PC* – partial collision; *TC* – total collision; e *RT* – robbery and theft) in relation to the exposure. The fields with the initials *Ind* are the average values of indemnification for each loss type, in R$. The *Avg SA* field represents the average sum assured, in constant currency, for the first half of 2011.

To avoid any inaccuracies or having only a small quantity of records by category, records with the following characteristics were excluded: Year Model prior to 1990; Indeterminate tariff region; Age group not reported; sex equal to *other*; and the categories of *motorbikes, minibuses, vans*, *trucks and busses* and *undetermined.* In this way, these exclusions reduced the database by 18% of exposure, 26% of premiums and 29% of indemnities.

**Table 1** – Basic Descriptive Statistics (by semester)

| Semester | Exposure | Freq PC | Freq TC | Freq RT | Ind PC | Ind TC | Ind RT | Avg SA |
|---|---|---|---|---|---|---|---|---|
| 2007/02 | 4.813.083 | 6,52 | 0,60 | 1,21 | 4.112 | 36.191 | 26.974 | 31.489 |
| 2008/01 | 5.022.046 | 6,55 | 0,56 | 1,20 | 4.101 | 37.280 | 26.979 | 32.438 |
| 2008/02 | 5.031.608 | 7,15 | 0,59 | 1,19 | 3.955 | 37.267 | 27.591 | 35.440 |
| 2009/01 | 5.479.060 | 6,34 | 0,52 | 1,19 | 4.141 | 36.384 | 26.874 | 34.964 |
| 2009/02 | 5.016.378 | 7,16 | 0,55 | 1,09 | 4.139 | 35.498 | 25.789 | 34.092 |
| 2010/01 | 5.943.202 | 6,46 | 0,51 | 0,96 | 3.945 | 34.068 | 25.977 | 33.938 |
| 2010/02 | 6.269.831 | 7,02 | 0,58 | 0,95 | 3.897 | 34.676 | 25.780 | 34.172 |
| 2011/01 | 6.060.089 | 7,60 | 0,60 | 0,94 | 3.790 | 33.932 | 25.813 | 35.511 |

Source: Autoseg database, authors' tabulation

**Table 2** – Basic Descriptive Statistics (by age group)

| Age Group | Exposure | Freq PC | Freq TC | Freq RT | Ind PC | Ind TC | Ind RT | Avg SA |
|---|---|---|---|---|---|---|---|---|
| 18-25 | 1.582.968 | 8,68 | 1,01 | 1,26 | 3.751 | 29.834 | 22.534 | 27.645 |
| 26-35 | 8.305.881 | 6,29 | 0,59 | 1,20 | 3.742 | 33.530 | 25.893 | 31.217 |
| 36-45 | 11.270.204 | 6,84 | 0,41 | 1,07 | 3.818 | 34.821 | 24.719 | 33.660 |
| 46-55 | 8.958.735 | 5,79 | 0,51 | 0,91 | 3.880 | 33.701 | 26.481 | 34.133 |
| 55 + | 8.882.220 | 5,22 | 0,44 | 0,77 | 3.703 | 32.429 | 25.501 | 32.261 |
| NI | 4.635.290 | 12,53 | 1,04 | 1,76 | 4.857 | 44.605 | 31.650 | 45.689 |

Source: Autoseg database, authors' tabulation

**Table 3** – Basic Descriptive Statistics (by sex)

| Sex | Exposure | Freq PC | Freq TC | Freq RT | Ind PC | Ind TC | Ind RT | Avg SA |
|---|---|---|---|---|---|---|---|---|
| Male | 20.882.603 | 6,54 | 0,61 | 1,13 | 4.192 | 34.858 | 26.255 | 33.607 |
| Female | 17.076.766 | 6,40 | 0,44 | 0,96 | 3.415 | 31.678 | 24.363 | 30.570 |
| Outros | 5.675.929 | 9,43 | 0,75 | 1,27 | 4.683 | 44.581 | 32.056 | 46.280 |

Source: Autoseg database, authors' tabulation

**Table 4** – Basic Descriptive Statistics (by region)

| Region | Exposure | Freq PC | Freq TC | Freq RT | Ind PC | Ind TC | Ind RT | Avg SA |
|---|---|---|---|---|---|---|---|---|
| SE | 26.188.464 | 6,40 | 0,53 | 1,34 | 3.855 | 33.541 | 25.996 | 32.761 |
| S | 8.460.938 | 6,81 | 0,60 | 0,72 | 4.482 | 36.431 | 29.447 | 35.582 |
| NE | 4.886.662 | 7,34 | 0,62 | 0,74 | 3.807 | 38.550 | 25.564 | 35.016 |
| CW | 3.277.018 | 9,32 | 0,64 | 0,56 | 3.983 | 39.911 | 29.309 | 37.643 |
| NW | 816.746 | 9,39 | 0,65 | 0,54 | 4.346 | 46.093 | 21.140 | 40.189 |
| NI. | 5.468 | 8,67 | 1,32 | 0,79 | 5.468 | 34.935 | 39.428 | 37.627 |

Source: Autoseg database, authors' tabulation

**Table 5** – Basic Descriptive Statistics (by vehicle model year)

| Model Year | Exposure | Freq PC | Freq TC | Freq RT | Ind PC | Ind TC | Ind RT | Avg SA |
|---|---|---|---|---|---|---|---|---|
| < 1990 | 258.538 | 2,73 | 0,48 | 0,93 | 4.803 | 26.273 | 27.776 | 17.059 |
| 1991-1997 | 2.227.788 | 3,31 | 0,84 | 1,53 | 3.555 | 17.855 | 16.093 | 15.965 |
| 1998-2002 | 6.883.691 | 4,94 | 0,73 | 1,33 | 3.946 | 26.095 | 21.603 | 22.389 |
| 2003-2007 | 17.970.722 | 6,73 | 0,53 | 1,04 | 4.148 | 37.749 | 28.326 | 33.828 |
| 2008-2012 | 16.294.558 | 8,37 | 0,48 | 0,96 | 3.893 | 43.295 | 29.412 | 42.009 |

Source: Autoseg database, authors' tabulation

For the pricing, four explanatory variables are used: category, region, age group and sex. So that the values can be calculated and compared, it is necessary to define an expected claim target for the first half of 2011. We made the assumption that this target is equal to 63.37%, a value equal to the average claims ratio index from the second half of 2007 until the second half of 2010. Therefore, a 36.63% loading will be aggregated to the Pure Premium obtained by the four techniques in order to obtain the Commercial Premium. In this way, the expected loss for the first half of 2011 for each technique can be verified.

## 4. Results

### 4.1 Pure Premium by historical aggregated claims

This method considers the average loss for each insured individual as the total retained claims of risk classes divided by the exposure of vehicles. As can be seen in the example of Table 6, there were 2.4 million exposures during the second semester of 2007 and 2010 for passenger cars in the Southeast, for people aged between 36 and 45 years old, male. For this group the value of R$ 1.7 billion in damages incurred were registered. With this data it is possible to obtain the average indemnity payment for each participant of this group (Pure Premium), which is the ratio between the total indemnity and exposure, i.e. R$ 706,35 (1.759.411.580 / 2.490.851). Thus, in order to obtain an expected loss ratio of 63.37%, i.e. the rate that maintains the loss ratio in the first half of 2011 similar to the base between the second half of 2007 and 2010 – there needs to be a loading of 36,63% on the *Pure Premium* to obtain the *Commercial Premium* of R$ 1.114,64 (706,35/(1 – 36,63%).

**Table 6** – Example of pricing by Historical Pure Premium

| Category | Region | Age Range | Sex | Exposure | Indemnity | Pure. P | Coml. P. |
|---|---|---|---|---|---|---|---|
| National Passenger Car | SE | 36-45 | M | 2.490.851 | 1.759.411.580 | 706,35 | 1.114,64 |
| National Passenger Car | SE | 36-45 | F | 2.394.609 | 1.397.133.514 | 583,45 | 920,70 |

Source: Authors' calculations

After performing this procedure for all classes of risks, the Commercial Premium that would be charged in the first half of 2011 was calculated, making it possible to compare the tariff increase the study would propose, while maintaining same number of exposures. In addition, it is possible to compare the differences in terms of claims and of the deviation in relation to the risk premium effectively obtained in the period. These results are presented in tables 7 and 8 below.

The tariff based on historical aggregated pure premium presented an increase compared to the effective tariff applied in the first half of 2011 in the amount of 7.56%. Assuming that during this time the indemnities are equal to the claims actually made and in the proposed tariff, the increase above would cause the expected loss ratio (ratio between indemnity *S* and premium *P*) to remain at 62.36%, i.e. a reduction

of 4.72 p.p. in relation to the actual claims of the period in question. There is an implicit assumption that demand is price inelastic, i.e. an increase by 7.56 % in the value of the premiums collected does not affect the amount of exposures during the period. In this way, the total premium of the proposed rate is the effective exposure in the first half of 2011 multiplied by the estimated average premium of each class of risk, based on the rate by the method of aggregated pure premium.

**Table 7** – Claims Comparison: Actual Tariff and Proposed Tariff

| | Claims 2011_1 (S) | Premium 2011_1 (P) | S/P |
|---|---|---|---|
| **Actual** | 3.362.130.275 | 5.012.837.940 | 67,07% |
| **Proposed Tariff** | 3.362.130.275 | 5.391.908.290 | 62,36% |

Source: Authors' calculations

To check the variability of the risk premium of the proposed tariff and of that which was effectively applied, two metrics can be used (table 8). The first, which measures the standard deviation, was R$ 109,54. Similarly, the standard error, which is the measure of percentage of deviation weighted by the size of the sample, presented the value of 5.26% of variation in the actual and proposed risk premiums.

**Table 8** – Increase in relation to Original Pricing and Dispersion of Pure Premiums

| | |
|---|---|
| **Increase** | 7,56% |
| **Standard Dev.** | 109,54 |
| **Standard Error** | 5,26% |

Source: Authors' calculations

**4.2 Pure Premium by expected aggregated claims**

In order to estimate the expected aggregated claims of a period, the first step requires estimating the expected exposure for the first half of 2011. For this, the average variation in exposure values among the semesters was used in the value of 4.80% (according to the values of the Table 1).

To estimate the amount of occurrences in the period the Poisson distribution was used and for the loss amount, the gamma distribution, both presenting a better adjustment in the graphical comparison of the theoretical distribution and that obtained with the database. In relation to the estimation of the quantities, the average frequency of each risk group within the analysis period was obtained and it was multiplied by the projected exposure to obtain the number of expected occurrences $E(N) = \lambda$ (Poisson average). The standard deviation for the Poisson distribution is defined by $\sigma(N) = \sqrt{\lambda}$ .

For the calculation of indemnities, it was necessary to obtain the mean and standard deviation of the indemnities of each group to make it possible to compute the parameters $\alpha$ and $\beta$ necessary for the gamma distribution. So, after performing the calculation of such parameters for each one of risk classes, $E(X) = \alpha\beta$ e $\sigma(X) = \sqrt{\alpha\beta^2}$ was obtained. Once these values were calculated for each class, it was possible to calculate the expected claim and the aggregated standard deviation, in addition to pure premium. The calculations were made by adopting the confidence level for $\theta$ in 90%, 95% and 99%. The results are presented in Tables 9 and 10.

The rate obtained by the aggregated claims resulted in increases of 10.49%, 11.32% and 12.88% to 90%, 95% and 99% of confidence, respectively. The highest increase in relation to pricing by the historical premium can be related the use of much larger confidence levels than necessary for an automobile portfolio, and mainly for the stability of this product over time, as can be observed in Table 1. Therefore, it is believed that the rate based on the 90% confidence level has presented sufficient increase and expected loss ratio. For this reason, it will be used as the basis of comparison with other models. The variability in relation to the actual and projected risk premium presented a standard deviation of R$147,84% and standard error of R$ 6.73%.

**Table 9** – Claims Comparison: Actual Tariff and Proposed Tariff

| | Claims 2011_1 (S) | Premium 2011_1 (P) | S/P | Increase |
|---|---|---|---|---|
| **Actual** | 3.362.130.275 | 5.012.837.940 | 67,07% | |
| **Tariff 90%** | 3.362.130.275 | 5.538.838.845 | 60,70% | 10,49% |
| **Tariff 95%** | 3.362.130.275 | 5.580.491.616 | 60,25% | 11,32% |
| **Tariff 99%** | 3.362.130.275 | 5.658.625.287 | 59,42% | 12,88% |

Source: Authors' calculations

**Table 10** – Measures of dispersion

| | |
|---|---|
| Standard Dev. | 147,84 |
| Std. Error | 6,73% |

Source: Authors' calculations

### 4.3 Classic Linear Model

To estimate the model for each class of risk, it was necessary to create a series of dummy variables. To obtain the risk premium, it was necessary to carry out separately the estimation of frequency and severity. For the linear model of frequency of occurrence, all submissions were used, i.e. those in which the claim occurred, but also those in which there had been no occurrence. However, for the average severity model, observations were used only where there had been claims, so it would be possible to obtain the expected value of average compensation in cases where there was occurrence and so prevent the zeroed values of the other observations underestimating the average value. Therefore, multiplying the frequency by the estimated average severity generated a risk premium for each group. The results of this technique are reported in Tables 11 and 12.

The tariff based on the classical linear model generated an increase of 9.27% compared to what has been effectively deployed by market, a value that would generate a reduction in the claims rate of 5.69 percentage points, reaching the value of 61.38% for the first half of 2011. Compared to previous models, the variability of this technique has a higher value, with a standard deviation of R$ 502.66 and a standard error of 49.98%.

**Table 11** – Claims Comparison: Actual Tariff and Proposed Tariff

| | Claims 2011_1 (S) | Premium 2011_1 (P) | S/P |
|---|---|---|---|
| **Actual** | 3.362.130.275 | 5.012.837.940 | 67,07% |
| **Proposed Tariff** | 3.362.130.275 | 5.477.430.329 | 61,38% |

Source: Authors' calculations

**Table 12** – Increase in relation to Original Pricing and Dispersion of Pure Premiums

| | |
|---|---|
| Increase | 9,27% |
| Standard. Dev. | 502,66 |
| Standard Error | 49,98% |

Source: Authors' calculations

### 4.4 Generalized Linear Models (*GLM*)

For the GLM, it was also necessary to carry out the estimation of frequencies and severities separately. For the frequency model, the Poisson distribution was used as the response variable, due to its better adherence to the data. Therefore, the canonical link function used for the model was logarithmic, which implies $X\beta = ln(\mu)$. For the severities model, the distribution with the highest adequacy to the data was the Gamma. Thus, the link function that provides the relationship between the linear predictor and the distribution function for this distribution is the inverse, given by $X\beta = -\mu^{-1}$. Also, it is worth mentioning that in the estimation of the relativities, the variable of gender and age was used in combined form by virtue of the fact that there is high correlation between them.

The rate estimated by GLM generated expected claims of 62.58%, i.e. a reduction of 4.49 p. p in relation to actual claims for the first half of 2011. The standard deviation of the risk premium tariff in relation to those occurred was R$ 384,66, and the variability of the sample averages was 17,52%.

**Table 13** – Claims Comparison: Actual Tariff and Proposed Tariff

| | Claims 2011_1 (S) | Premium 2011_1 (P) | S/P |
|---|---|---|---|
| Actual | 3.362.130.275 | 5.012.837.940 | 67,07% |
| Proposed | 3.362.130.275 | 5.372.440.390 | 62,58% |

Source: Authors' calculations

**Table 14** – Increase in relation to Original Pricing and Dispersion of Pure Premiums

| | |
|---|---|
| **Increase** | 7,17% |
| **Stand. Dev.** | 384,66 |
| **Standard Error** | 17,52% |

Source: Authors' calculations

### 4.5 Comparison between the methods

Table 15 presents a comparative summary of the results performed in the study. With the use of equation 10 the *Efficiency* variable was created, which measures the necessary increase in price for there to be a reduction of one percentage point in relations to the actual price in the first half of 2011. In this expression, *Increase* is the percentage increase in estimated tariffs in relation to the tariff effectively applied in the first half of 2011. The *Reduction* p.p. corresponds to the reduction in percentage points in relation to claims estimated by the calculated tariff and that occurred effectively in the first half of 2011.

$$Efficiency = \frac{\text{Increase}}{\text{Reduction p. p.}} \quad (10)$$

Thus, it is possible to measure the efficiency of each model in a standardized manner. The smaller the value found for this variable, the more efficient is the technique. For example, for the historical pure premium there was a 7.56% increase and reduction of claims at 4.72 p.p., which implies an efficiency value of 1.60%. Thus, in order to be able to reduce by a percentage point the claims using the historical pure premium, there needs to be an increase of 1.60% in the rate effectively applied in the first half of 2011.

It can be observed that pricing by expected aggregated claims (2) showed the worst efficiency index, although showing the second less variability. Pricing by the classical linear model (3) showed the third best efficiency. However, this generated the greatest variability of all the models presented. A possible explanation is related to the need to transform the frequency and severity distributions into a normal distribution, which ultimately may not be adequate to explain the phenomena studied.

Standing in the first place in terms of efficiency is pricing by historical pure premium (1) and by GLM (4). However, this first method presented less variability than the second did. It is believed that the GLM shows greater variability mainly because of performing the adjustment to a hypothetical distribution, which is not suitable for all actual data. Pricing by historical pure premium presented a comparative advantage, because the automobile insurance portfolio is large and quite stable over time, allowing proper pricing considering only the average expected claims.

**Table 15** – Pricing – Results Summary

| Technique | S/P (%) | Reduction p.p | Increase (%) | Standard-Deviation | Standard-Error (%) | Efficiency (%) |
|---|---|---|---|---|---|---|
| **Historical Pure Prem. (1)** | 62,36 | 4,72 | 7,56 | 109,54 | 5,26 | 1,60 |
| **Agg. Pure Prem. (2)** | 60,70 | 6,37 | 10,49 | 147,84 | 6,73 | 1,65 |
| **Class. Lin. Model (3)** | 61,38 | 5,69 | 9,27 | 502,66 | 49,98 | 1,63 |
| **MLG (4)** | 62,58 | 4,49 | 7,17 | 384,66 | 17,52 | 1,60 |

Source: Authors' calculations

## 5. Final Comments

The results showed that a simpler technique was more efficient than the others that were more technically refined. This result, at first counterintuitive, can be explained by several factors. First, the study considered the total database of insured vehicles in Brazil over the years studied. In addition, in an automobile portfolio, the risks are not exposed to disasters or high severity claims – such as a property risk portfolio, for example. In this way, historical data of the portfolio is a fairly accurate estimator for future periods. Taking all of this in context, the historical pure premium pricing structure proved to be advantageous in comparison with other methods. Another aspect to be considered is the low historical variability of the automobile database in the studied period, as can be confirmed by the data presented in Table 1.

However, this situation is not common with the majority of insurers, mainly because none of them have the information of all market risks in their portfolio. Thus, if only this technique is used, the result could be a biased analysis for certain classes of risk, because there is no historical basis with sufficient experience. For insurers with the largest portfolios in the market, that have considerable participation in relation to the total, the method of historical Pure Premium, if aligned with any other metric of positioning in relation to the market (e.g. ratio of number of accepted proposals and the number of quotes performed) can be a simple and effective pricing technique. However, for the portfolios with a smaller market share and a low number of insured vehicles, this technique is less efficient, which decreases the credibility of the sample data and may distort the results, resulting in misguided pricing strategies, reduced competitiveness and financial losses to the company.

In this context, the use of GLM becomes an interesting alternative for smaller portfolios, by presenting a proper pricing technique, even in those in the process of growth or that present greater variability. This conclusion is based primarily on the GLM having a theoretical distribution for estimation of frequencies and severities, achieving adequacy of risk premiums, even without having a large exposure in the portfolio. However, there are costs that need to be taken into account when implementing this pricing method. First, it is necessary to have a team with expertise in statistics, actuarial science or a related field, to accomplish the formulation, interpretation and appropriate updating of the models. In addition, there are the high costs involved in installation and maintenance of computer programs in this process.

Thus, the decision as to which pricing technique is more adequate is best left to the strategy of each insurer and should be analysed in relation to the cost of each and how much this will impact positively on the result of the portfolio.

## 6. Bibliographical References

BAILEY, R. A.; SIMON, L. J. **Two Studies in automobile insurance ratemaking.** Casualty Actuarial Society Proceedings. Volume XLVII, Numbers 87, 1960.

BAILEY, R. **Insurance rates with minimum bias.** Casualty Actuarial Society Proceedings. 1963.

BOLAND, P. J. **Statistical and Probabilistic Methods in Actuarial Science**. Chapman & Hall/CRC. 2007.

BRASIL. Superintendência de Seguros Privados. **Sistema de Estatísticas da SUSEP.** Available at: < http://www2.susep.gov.br/menuestatistica/SES/principal.aspx>. Accessed on 15 Nov. 2012.

BRASIL. Superintendência de Seguros Privados. **Sistema de Estatísticas de Automóveis da SUSEP.** Available at: <http://www2.susep.gov.br/menuestatistica/ Autoseg/principal.aspx>. Accessed on 15 Nov. 2012.

BRASIL. Instituto Brasileiro de Geografia e Estatística. **Índice Nacional de Preços ao Consumidor Amplo – IPCA**. Available at: <http://www.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca/defaultinpc.shtm>. Accessed on 30 May. 2013.

BROWN, R. L., **Minimum Bias with Generalized Linear Models.** Proceedings of Casualty Actuarial Society, Vol. LXXV, Casualty Actuarial Society. 1988.

CORDEIRO, G. M. **Modelos Lineares Generalizados**. VII Simpósio Nacional de Probabilidades e Estatística, Campinas, São Paulo. 1986.

GUJARATI, D. **Essentials of Econometrics**. New York: McGraw-Hill. 1992.

JUNG, J. **On Automobile Insurance Ratemaking**. Casualty Actuarial Society. Astin Bulletin, Volume V, Part I. 1968.

JEWELL, W. S. **The credible distribution**. Astin Bulletin, volume VII, n.3, 1974.

JONG, P.; HELLER, G. Z. **Generalized Linear Models for Insurance Data**. Cambridge University Press, 2008.

KLUGMAN, S. A., PANJER, H. H., WILLMOT, G. E. **Loss Models from Data to Decisions**. John Wiley & Sons, Inc. 2004.

MCCLENAHAN, C.L., **Ratemaking**. Foundations of Casualty Actuarial Science, Casualty Actuarial Society. 2001.

MORGADO, W. L. **Método de Classificação de Risco Aplicado ao Mercado de Seguros de Automóveis.** Pontifícia Universidade Católica, Rio De Janeiro. 2004.

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear Models**. Journal of the Royal Statistical Society. Series A Vol. 135, No. 3, pp. 370-384, 1972.

SANTOS, S. T. **Construção De Uma Tarifa de Responsabilidade Civil Automóvel**. Universidade Nova de Lisboa, Lisboa. 2008.

WEISBERG, H. I.; TOMBERLIN, J. T.; CHATTERJEE, S. **Predicting Insurance Losses under Cross-Classification: A Comparison of Alternative Approaches**. Journal of Business & Economic Statistics. Vol. 2, No. 2, pp. 170-178, 1984.

ZEHNWIRTH, B. **Ratemaking: from Bailey and Simon (1960) to Generalized Linear Regression Models**. Casualty Actuarial Society Winter Forum, 615-659. 1994.