STATISTICS IN TRANSITION new series, December 2019
Vol. 20, No. 4, pp. 13–31, DOI 10.21307/stattrans-2019-032
Submitted – 02.07.2019; Paper ready for publication – 12.11.2019

GENERAL LINEAR MODEL: AN EFFECTIVE TOOL FOR ANALYSIS OF CLAIM SEVERITY IN MOTOR THIRD PARTY LIABILITY INSURANCE

Erik Šoltés[1], **Silvia Zelinová**[2], **Mária Bilíková**[3]

ABSTRACT

The paper focuses on the analysis of claim severity in motor third party liability insurance under the general linear model. The general linear model combines the analyses of variance and regression and makes it possible to measure the influence of categorical factors as well as the numerical explanatory variables on the target variable. In the paper, simple, main and interaction effects of relevant factors have been quantified using estimated regression coefficients and least squares means. Statistical inferences about least-squares means are essential in creating tariff classes and uncovering the impact of categorical factors, so the authors used the LSMEANS, CONTRAST and ESTIMATE statements in the GLM procedure of the Statistical Analysis Software (SAS). The study was based on a set of anonymised data of an insurance company operating in Slovakia; however, because each insurance company has its own portfolio subject to changes over time, the results of this research will not apply to all insurance companies. In this context, the authors feel that what is most valuable in their work, is the demonstration of practical applications that could be used by actuaries to estimate both the claim severity and the claim frequency, and, consequently, to determine net premiums for motor insurance (regardless of whether for motor third party liability insurance or casco insurance

Key words: general linear model, claim severity, motor third party liability insurance, least squares means.

1. Introduction

In general, two approaches are used to determine net premiums in non-life insurance. Either the target variable is equal to the net premium (euros of loss per exposure) or it is separately modelled the claims frequency (number of claims per

[1] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Slovakia. E-mail: erik.soltes@euba.sk. ORCID ID: https://orcid.org/0000-0001-8570-6536.
[2] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Mathematics and Actuarial Science, Slovakia. E-mail: silvia.zelinova@euba.sk. ORCID ID: https://orcid.org/0000-0002-9932-6857.
[3] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Mathematics and Actuarial Science, Slovakia. E-mail: maria.bilikova@euba.sk.

exposure) and the claim severity (average loss per claim). Goldburd et al. (2016) mention that special modelling of frequency and severity is more stable and leads to a lower variance of the error term compared to when the net premium is directly modelled. In addition, in the case of a separate analysis of frequency and severity we can detect effects in the data that we otherwise would not. On the other hand, the standard techniques of net premium determination based on specific modelling of frequency and severity assume independence between the number and the size of claims. Methods that are appropriate in the case of correlation between frequency and severity components are dealt with by, e.g. Shi et al. (2015). The above facts motivated us to consider a separate modelling, so the paper focuses only on the claim severity in motor third party liability (MTPL) insurance. Since severity refers to the cost of a claim, through this metric we can identify those tariff classes in MTPL insurance which are more expensive and those which are cheaper for an insurance company.

For the calculation of auto insurance premiums, many actuaries use techniques based on regression analysis and analysis of variance in their scientific work. Very popular models include generalized linear models, which are used by, e.g. (De Azevedo et al., 2016), (Kafková and Křivánková, 2014), (Jong and Heller, 2008) and (Frees et al., 2016). The Poisson regression model is frequently used to model claim frequency and the Gamma regression model is used to model claim costs (see, e.g. (David, 2015) and (Duan et al., 2018)). As David (2015) indicates, generalized linear models allow for the modelling of a non-linear behaviour and a non-Gaussian distribution of residuals, which is very useful for the analysis of non-life insurance, where claim frequency and claim cost follow an asymmetric density, which is clearly non-Gaussian. A special case of generalized linear model (GzLM) is the general linear model (GLM), which we use in the article to assess the impact of relevant factors on claim severity. GLM and GzLM are two commonly used families of statistical methods to relate some number of continuous and/or categorical predictors to a single outcome variable. The main difference between the two approaches is that GLM strictly assumes that the residuals will follow a conditionally normal distribution, while GzLM loosens this assumption and allows for a variety of other distributions from the exponential family for the residuals (see, e.g. (Agresti, 2015), (Fox, J., 2015), (Kim and Timm, 2006) and (Littell, et al., 2010)).

GLM includes the t-test, analysis of variance (ANOVA), multiple regression, descriptive discriminant analysis (DDA), multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA) and structural equation modelling (SEM). Therefore, Graham (2008) indicates that the vast majority of parametric statistical procedures in common use are part of the general linear model. Thompson (2015) discusses GLM as a unifying conceptual framework that helps students and researchers understand common features of analyses included in GLM.

The aim of the article is to provide a presentation of the possibility of using general linear models for claim severity analysis in motor third party liability insurance for the purpose of tariffication. The article does not limit itself to an illustration of general linear models by means of a demonstrational example but provides the analysis of an actual data set from an unnamed insurance company operating in Slovakia.

In the past, actuaries often relied on a one-way analysis of pricing. However, one-way analyses do not consider interdependencies between factors in the way they affect claim experience, which is why multivariate methods are more effective (Anderson et al., 2007). For this reason, in this paper we use multivariate methods included in general linear models, which correct the correlation between factors and allow for the investigation of interaction effects.

2. Research methods

The general linear model, which will be the subject of interest in our paper, can be simplified as follows:

$$y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}} + \varepsilon_{ijk} \tag{1}$$

where y_{ijk} is *k*-th observation of the target (explained) variable *Y* in cell ij, i.e. at the *i*-th level of factor *A* and at the same time the *j*-th level of factor *B*. We assume that the random errors ε_{ijk} are independent of each other and identically distributed with the normal distribution $N(0, \sigma^2)$.

Let us denote by μ_{ij} the mean of the target variable for the *i*-th variation of factor *A* and the *j*-th variation of factor *B*. This mean is called the *cell mean* for cell ij and is defined as the sum of the constant μ (intercept), α_i - factor *A* effect, β_j - factor *B* effect and $(\alpha\beta)_{ij}$ – the interaction effect between factors *A* and *B*. Note that more than two factors will be taken into account in the application part of this paper, some will be in the form of quantitative variables and others in the form of categorical variables.

The general linear model can be used to examine several types of effects, such as:

- *simple effects*, which indicate that one factor level affects the target variable, while other factors remain constant at that level;
- *interactions,* which characterize how levels of one factor affect the target variable across levels of another factor. If, at all levels of 2nd factor, 1st factor affects the target variable equally, it is a non-interaction model. If, at different levels of 2nd factor, 1st factor affects the target variable differently, it is an interaction model;
- *main effects,* which reflect the overall differences between the levels of each factor averaged across all levels of another factor.

The focus should be on interaction and then on simple or main effects. If a significant interaction is confirmed, it is appropriate to compare simple effects. One way to compare the means of the target variable at different levels of one factor specifically for different levels of the second factor is to carry out the analysis of variance or general linear model separately for different levels of the second factor. However, by this separate analysis, we discard some of the

information from other levels of the second factor, and this unused information manifests itself in a low number of degrees of freedom for SS (ERROR), which is central to statistical tests associated with the analysis of variance (Littell et al., 2010). This inefficient solution would waste a lot of data, which will severely reduce the strength of the tests. With the tools in the GLM procedure (PROC GLM) of the SAS statistical software, which we use in the paper, it is possible to avoid such a problem.

PROC GLM has options within the LSMEANS statement that allow you to test each factor at a particular level of another factor. The LSMEANS statement calculates the estimate of the so-called least squares mean (LS mean), also referred to as the marginal mean. In unbalanced, multi-way designs, the LS means estimation is often assumed to be closer to reality. LS means correct the design's imbalance. In balanced designs, or in unbalanced one-way ANOVA designs, observed means and least squares means are the same ((Lenth, 2016) and (Cai, 2014)).

The general linear model can be written in the form of a multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \tag{2}$$

PROC GLM for estimating the parameters of such a model, therefore, uses the least squares method, which results in the formula

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{3}$$

In view of the fact that in the GLM procedure generally considered with the classification explanatory variables, which are converted to dummy variables, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not of full rank and therefore has no unique inverse. For such a situation, PROC GLM computes a general inverse $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$ and the parameters of the regression model (2) are estimated according the formula

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{4}$$

where the estimated parameter vector $\hat{\boldsymbol{\beta}}$ has zero values at the locations that correspond to the zero rows in the matrix $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$. The estimate $\hat{\boldsymbol{\beta}}$ thus obtained is not unique. However, there is a set of linear functions $\mathbf{L}\hat{\boldsymbol{\beta}}$ where $\mathbf{L}$ is a linear combination of rows of the matrix $\mathbf{X}$, which are called estimable functions (more detail in (Agresti, 2015, pp. 14–15) and (Littell et al., 2010, pp. 194–203)) and have these features:

- $\mathbf{L}\hat{\boldsymbol{\beta}}$ and its covariance matrix $Var\left(\mathbf{L}\hat{\boldsymbol{\beta}}\right)$ are unique,
- $\mathbf{L}\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\mathbf{L}\boldsymbol{\beta}$.

As with the full rank, covariance matrix $\mathbf{L}\hat{\boldsymbol{\beta}}$ is given by the formula

$$Var\left(\mathbf{L}\hat{\boldsymbol{\beta}}\right) = \sigma_{\varepsilon}^{2}\left[\mathbf{L}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{L}^{\mathrm{T}}\right] \tag{5}$$

wherein the estimate of the variance of the random error σ_ε^2 is the residual variance MSE, which is calculated similarly to the multiple regression analysis, while the sum of squared error SSE (also known as the sum of squared residuals – SSR) is no longer dependent on the general inverse $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$.

3. Preparation of input variables, selection of regressors and verification of assumptions about the error term

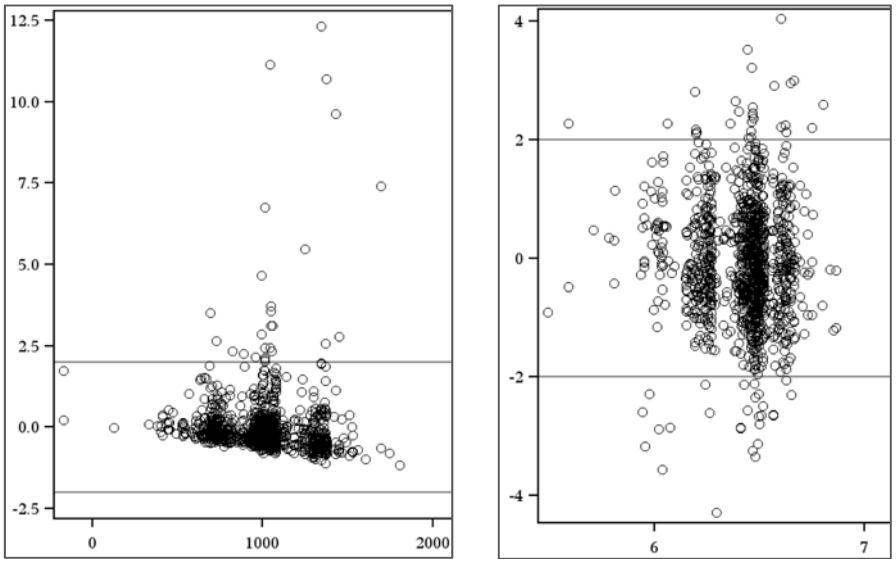
Our analysis focuses on the target variable – the claim severity (average costs per claim) of passenger cars in MTPL insurance. We modelled this variable depending on the following factors:

- relating to the insured vehicle such as Engine Power (kW, abbr. EP), Engine Volume (cm^3), Weight (kg), Age (years) and Car Make,
- relating to the vehicle owner such as Age (years) and Residence.

We categorized the vehicle owner's age and created the Age_group variable, which has six groups: the vehicle owners aged up to 30, aged 30–40, 40–50, 50–60, 60–70 and over 70 (upper limits of the indicated intervals are closed).

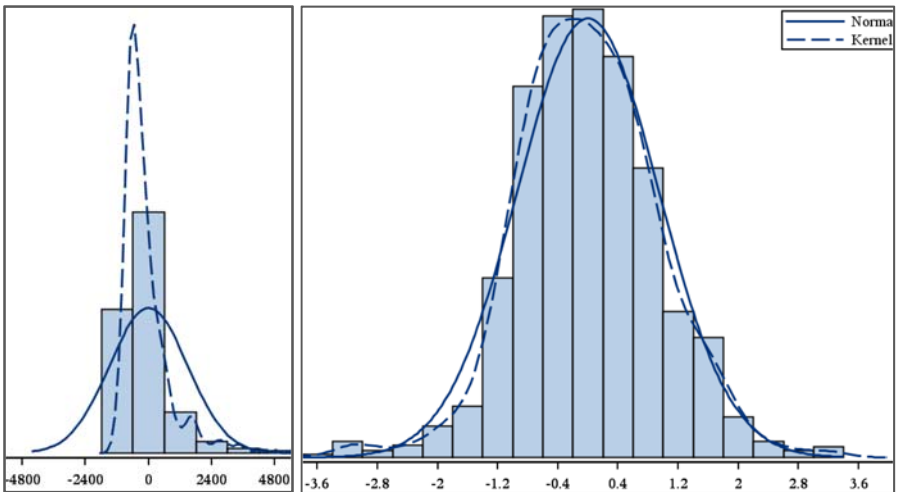
Since the residuals showed heteroscedasticity (Figure 1, on the left) and were markedly right-skewed (Figure 2, on the left) while modelling claim severity, we decided to use the logarithmic transformation of the explained variable. In the log-linear model, which modelled the dependence on the factors considered, the residuals had approximately a normal distribution with zero mean (see (Figure 1, on the right) and (Figure 2, on the right)).

Figure 1. Studentized residuals vs predicted for claim severity (on the left) and vs predicted for logarithm of claim severity (on the right)



Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

Figure 2. Distribution of residuals for claim severity (on the left) and for logarithm of claim severity (on the right)



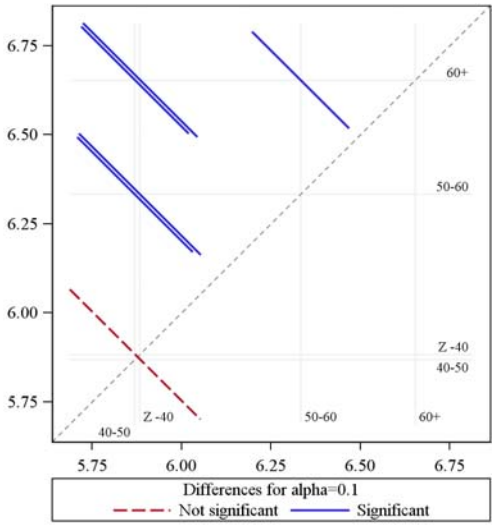
Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

We verified the homoscedasticity of the error term by using the White test. This test uses a model of the second squares of residuals depending on the predicted values and their squares, while the original test is based on the model of squared residuals depending on the original explanatory variables, the squares and cross products of independent variables (see (Wooldridge, 2013, pp. 279–280)). Based on calculated test statistics $\chi^2 = 3.3376$, which had 2 degrees of freedom, we quantified $p - value = 0.1885$. Since $p - value$ is greater than any commonly used confidence level, we do not reject the null hypothesis of homoscedasticity.

Based on the average amount of claims incurred in fixing the other considered factors, we transformed some of the original explanatory variables during the modelling process. We created three groups of vehicle makes and we call the resulting variable in other analyses Vehicle_group. This categorical variable has values 1, 2 and 3, with category 1 being the makes of vehicles with the highest average costs per claim, and category 3 including vehicle makes, where we quantified the lowest average costs per claim (in eliminating the influence of other factors). Similarly, we developed new categories of the Residence variable, using the LS means tests below. This process created a classification variable with 3 values (A, B and C), with category A (including the regional cities of Košice and Trenčín), where we detected the highest average costs per claim, category B (including villages, small towns, all district towns as well as the regional cities of Bratislava and Nitra) and category C (including the regional cities of Banská Bystrica, Prešov, Trnava and Žilina), where we quantified the lowest average costs per claim. We have to emphasize that statistically significant differences in average costs per claim were not confirmed among the owners' residence that fall within the same category.

We included the variables of Engine Power (EP), Engine Volume, Weight, Age, Vehicle_group, Age_group and Residence, as well as the polynomials of numerical explanatory variables, but also the interaction between the considered variables. By the method of backward elimination (Agresti, 2015), factors that did not have a significant effect on the explained variable at the confidence level 0.1 were excluded from the model. At the same time, the equality of the marginal means (LS-means) were tested using the Tukey-Kramer test (adjusted Tukey's test appropriate for unbalanced data, see (Wilcox, 2003)). In the case of insignificant differences between the marginal means of the target variable on two levels of one particular factor and after taking into account the logical context we finally merged the original categories of that factor.

Figure 3. Comparison of LS means of logarithm of claim severity for factor Age_group



Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

In short, we will explain this procedure with the factor of Age_group. This factor originally contained 6 categories, but because of the insignificant differences in average severity between insured aged 70+ and 60–70, we merged these categories to form a category 60+. Similarly, we proceeded in the same way in the case of age categories 30-40 years and up to 30 years. However, we must remark that in the 70+ and under 30 age groups, the insurance company had a low number of claims and, therefore, based on the input database, we cannot persuasively claim that young or old vehicle owners (over 70) do not have higher or lower average severity compared to the other age categories. In the case of the Age_group factor, the next analysis found that insured persons aged 40 to 50 and under 40 report the smallest average severity, with no significant differences between these two age categories, as shown in Figure 3.

We merged these two categories and present the category of those aged up to 50 in the following results.

4. Empirical results

In this section of the paper, we will provide the results of the analysis obtained from the PROC GLM of the SAS statistical software. In Section 4.1 we will focus on assessing the differences between the individual levels of the competent relevant factors and quantifying the impact of these factors on the average claim severity of vehicles in MTPL insurance. Section 4.2 offers examples of the application of the CONTRAST and ESTIMATE statements that actuaries and statisticians can use for further analyses of the impact of the factors on the target variable.

4.1. Estimating the model and quantifying the impact of relevant factors

As we mentioned in the previous section, the method of backward elimination was used to select regressors, in which the statistical significance of a particular factor was assessed by the F-test, which uses the partial sum of squares, called Type II SS in regression analysis, but Type III SS in the GLM procedure (see more in (Kuznetsova et al., 2017), (LaMotte, 2019) and (Littell et al., 2010)). This sum of squares for the particular variable represents an increase in SSM due to the addition of this variable to the model. This type of sum of squares does not depend on the sequence in which the independent variable is loaded into the model and is useful to verify the statistical significance of the effect of the analysed explanatory variable on the target variable *Y*. Table 1 confirms the significance of the influence of the factors left in the resulting model.

Table 1. Verifying the impact of considered factors on claim severity

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EP	1	0.86485345	0.86485345	6.72	0.0096
EP * EP	1	1.03197187	1.03197187	8.02	0.0047
EP * EP * EP	1	1.28920833	1.28920833	10.01	0.0016
Age_group	2	2.87135126	1.43567563	11.15	<.0001
Vehicle group	2	1.82409264	0.91204632	7.08	0.0009
Residence	2	3.26944802	1.63472401	12.70	<.0001
Age_group*Residence	4	2.43942634	0.60985659	4.74	0.0008

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

The regression coefficients (Table 2) of the dummy variables, which encode the categories of Age_group, Vehicle group and Residence, are statistically significant at the 0.1 confidence level. In the above categories, the average severity of insurance claim is significantly different from the reference category of the relevant factor (at the level of confidence of 0.1). Figures 4 and 5 confirm that not only in comparison with the reference category, but among all pairs of particular factor categories there are significantly different LS means of the target

variable at the 0.1 confidence level. The highest average severity when fixing other factors was found for the oldest vehicle owners (in our case over the age of 60), then in the owners of vehicles from the regional cities of Trenčín and Košice and in the makes of vehicles belonging to group 1. On the contrary, we found the lowest average severity under ceteris paribus conditions in the group of vehicle owners aged under 50, as well as in vehicles of the group 3 and for vehicles from the regional cities of Banská Bystrica, Prešov, Trnava and Žilina.

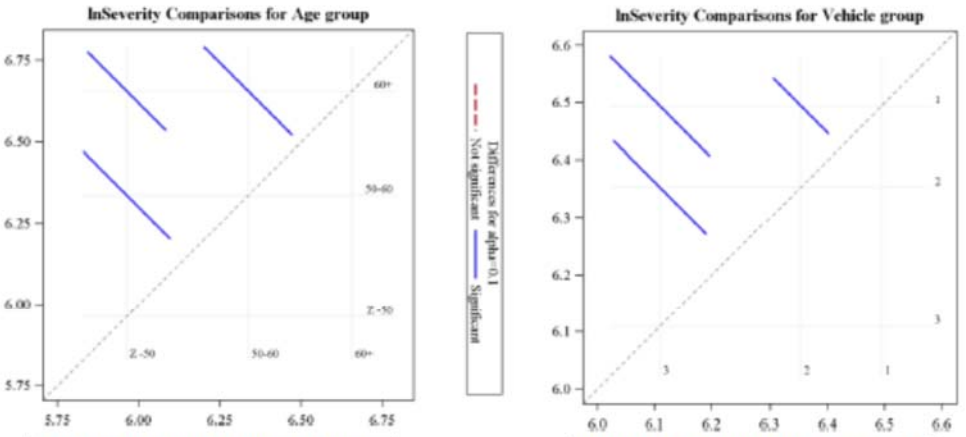
Table 2. Estimate of the parameters of general model for natural logarithm of claim severity

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		3.1170	B	0.8205	3.80	0.0002
EP		0.0907		0.0350	2.59	0.0096
EP * EP		-0.0014		0.0005	-2.83	0.0047
EP * EP * EP		6.7E-6		0.0000	3.16	0.0016
Age_group	**60+**	1.0729	B	0.2958	3.63	0.0003
Age_group	**50-60**	0.8390	B	0.3001	2.80	0.0052
Age_group	**-50**	0.0000	B	.	.	.
Vehicle group	**1**	0.3854	B	0.1065	3.62	0.0003
Vehicle group	**2**	0.2439	B	0.0993	2.46	0.0141
Vehicle group	**3**	0.0000	B	.	.	.
Residence	**A**	1.0997	B	0.3044	3.61	0.0003
Residence	**B**	1.1402	B	0.2380	4.79	<.0001
Residence	**C**	0.0000	B	.	.	.
Age_group*Residence	**60+ A**	-0.1156	B	0.4361	-0.27	0.7910
Age_group*Residence	**60+ B**	-1.0213	B	0.3029	-3.37	0.0008
Age_group*Residence	**60+ C**	0.0000	B	.	.	.
Age_group*Residence	**50-60 A**	-0.6337	B	0.4865	-1.30	0.1929
Age_group*Residence	**50-60 B**	-0.7596	B	0.3060	-2.48	0.0132
Age_group*Residence	**50-60 C**	0.0000	B	.	.	.
Age_group*Residence	**-50 A**	0.0000	B	.	.	.
Age_group*Residence	**-50 B**	0.0000	B	.	.	.
Age_group*Residence	**-50 C**	0.0000	B	.	.	.

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

The interaction between the factors Age_group and Residence showed to be statistically significant. Based on the LS means tests (Figure 5, on the right) for pairs of vehicle owner groups that arose from breaking down by the two mentioned factors, we found that not all pairs report different average severities. It is clear that because of the interaction of Age_group and Residence factors it is significantly the highest claim severity in the case of vehicle owners who live in the villages falling into category A and at the same time are aged 60+. On the other hand, the general linear model quantified that the lowest average severity is among the group of vehicle owners under the age of 50 who live in the regional cities of Banská Bystrica, Prešov, Trnava and Žilina.

Figure 4. Comparison of LS means for factor Age_group (on the left) and for factor Vehicle group (on the right)



Least Squares Means for effect Age_group
Pr > |t| for H0: LSMean(i)=LSMean(j)

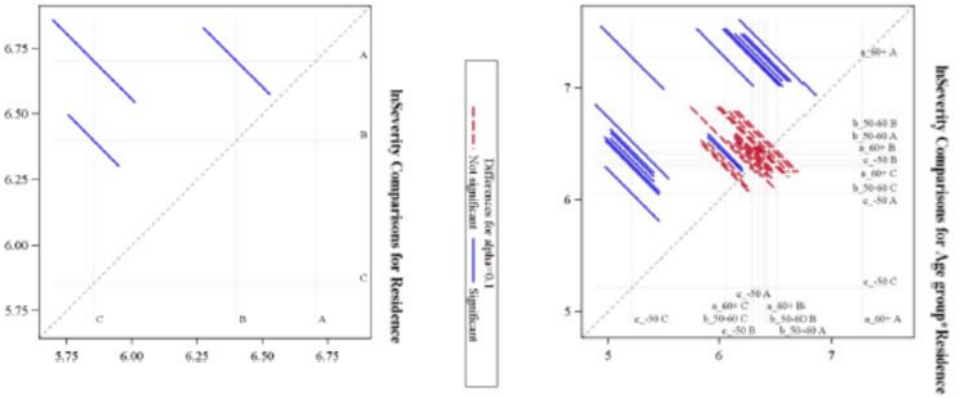
i/j	60+	50-60	-50
60+		0.0542	<.0001
50-60	0.0542		0.0221
-50	<.0001	0.0221	

Least Squares Means for effect Vehicle group
Pr > |t| for H0: LSMean(i)=LSMean(j)

i/j	1	2	3
1		0.0174	0.0003
2	0.0174		0.0141
3	0.0003	0.0141	

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide

Figure 5. Comparison of LS means for factor Residence (on the left) and for interaction Age_group×Residence (on the right)



Least Squares Means for effect Residence Pr > |t| for H0: LSMean(i)=LSMean(j)			
i/j	A	B	C
A		0.0530	<.0001
B	0.0530		<.0001
C	<.0001	<.0001	

Effect Age_group*Residence Pr > |t| for H0: LSMean(i)=LSMean(j)									
i/j	60+ A	60+ B	60+ C	50-60 A	50-60 B	50-60 C	-50 A	-50 B	-50 C
60+ A		0.0010	0.0017	0.0721	0.0013	0.0001	0.0029	0.0004	<.0001
60+ B	0.0010		0.5250	0.7355	0.7008	0.0706	0.6472	0.4325	<.0001
60+ C	0.0017	0.5250		0.5376	0.4287	0.3673	0.9191	0.7125	0.0003
50-60 A	0.0721	0.7355	0.5376		0.7985	0.2200	0.5920	0.6210	0.0013
50-60 B	0.0013	0.7008	0.4287	0.7985		0.0491	0.5480	0.1915	<.0001
50-60 C	0.0001	0.0706	0.3673	0.2200	0.0491		0.3323	0.1151	0.0052
-50 A	0.0029	0.6472	0.9191	0.5920	0.5480	0.3323		0.8374	0.0003
-50 B	0.0004	0.4325	0.7125	0.6210	0.1915	0.1151	0.8374		<.0001
-50 C	<.0001	<.0001	0.0003	0.0013	<.0001	0.0052	0.0003	<.0001	

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

In order to quantify the impact of various factors on the average severity it is necessary to convert the estimate of the model $\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$ shown in Table 2 into the form $y_i = e^{\hat{\beta}_0} \cdot \left(e^{\hat{\beta}_1}\right)^{x_{i1}} \cdot \left(e^{\hat{\beta}_2}\right)^{x_{i2}} \cdot \ldots \cdot \left(e^{\hat{\beta}_k}\right)^{x_{ik}}$. Naturally, in the additive model, the influence of reference categories is at the "0" level, which is transformed into a value $e^0 = 1$ in the multiplicative model. Based on the above transformation, using the parameter estimates from Table 2, we get

$$\hat{y}_i = 22.579 \cdot (1.095)^{EP} \cdot 0.9986^{EP^2} \cdot 1.000067^{EP^3} \cdot 2.924^{Age\ Group=60+} \cdot 2.314^{Age\ Group=50-60} \cdot$$
$$\cdot 1.470^{Vehicle\ Group=1} \cdot 1.276^{Vehicle\ Group=2} \cdot 3.003^{Residence=A} \cdot 3.127^{Residence=B} \cdot$$
$$\cdot 0.891^{Age\ Group=60+ \wedge Residence=A} \cdot 0.360^{Age\ Group=60+ \wedge Residence=B} \cdot$$
$$\cdot 0.531^{Age\ Group=50-60 \wedge Residence=A} \cdot 0.468^{Age\ Group=50-60 \wedge Residence=B}$$

The shape of the function with EP (Engin Power) as an explanatory variable shows that with a normal engine power of between 50 and 100 kW, the claim severity is approximately constant while fixing other factors and starts to rise more quickly for vehicles with engine power over 100 kW. In the case of vehicle makes falling under category 1, we estimate an almost 1.5 times higher average severity than for the 3rd category of vehicles and about 15% higher (1.152 = 1.470 / 1.276) than for the 2nd category of vehicles.

Since there is an interaction between the factors Age_group and Residence, the influence of these factors can be quantified from the exponential bases of the dummy variables belonging to the variables Age_group, Residence and their interactions Age_group × Residence.

Table 3. Multiplier estimates for vehicle owners broken down by-Age_group and Residence factors

Age_group	Residence		
	A	**B**	**C**
60+	7.822	3.293	2.924
50-60	3.688	3.386	2.314
-50	3.003	3.127	1.000

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

It is clear from Table 3 that the highest average loss per claim is for vehicle owners over the age of 60 who live in the municipalities of group A (regional towns of Trenčín and Košice). The fact that it is the riskiest group from the point of view of claim severity was already confirmed in Figure 5 (on the right). Now, we have found that their average severity is up to 7.8 times higher than in the case of the least risky group, which is vehicle owners under the age of 50 living in villages in category C (regional towns Banská Bystrica, Prešov, Trnava and Žilina). Similarly, the other multipliers estimated in Table 3 could also be interpreted as compared to the "-50 C" reference category.

4.1. Use of the CONTRAST and ESTIMATE statements for a deeper analysis of the impact factors

According to Table 3 and the estimated LS means, it appears that in the age group of vehicle owners aged 50 to 60, residence has little impact on average severity. In the age group 50-60 years, between the residence categories A and B, based on the LS means test (Figure 5, $p-value=0.7895$), it did not confirm the significant difference and thus we can assume equality $H_0:\mu_{2A}=\mu_{2B}$. Note that for ease of writing, we will use index 2 to denote the second variation of the Age_group variable (50-60 years). In order to verify the equality of the corresponding 3 means $H_0:\mu_{2A}=\mu_{2B}=\mu_{2C}$, we will test the hypothesis

$$H_0:\left(\mu_{2A}+\mu_{2B}\right)/2=\mu_{2C} \quad \text{or equivalently} \quad H_0:0.5\mu_{2A}+0.5\mu_{2B}-\mu_{2C}=0$$

We will verify this hypothesis in the SAS software with the CONTRAST statement, using Table 4 to determine the coefficients in this statement.

Table 4. Coefficients to the CONTRAST statement to verify the hypothesis

$$H_0 : 0.5\mu_{2A} + 0.5\mu_{2B} - \mu_{2C} = 0$$

Age_group	Residence			Σ
	A	B	C	
1=60+	0	0	0	**0**
2=50-60	0.5	0.5	-1	**0**
3=-50	0	0	0	**0**
Σ	**0.5**	**0.5**	**-1**	**0**

Source: Self-processed.

Then the statement has a syntax

```
contrast 'Age_group*Residence 2A 2B vs 2C' Residence 0.5 0.5 -1
Age_group*Residence 0 0 0 0.5 0.5 -1;
```

The result of the test is given in the first row in the body of Table 5. Depending on the level of confidence, we reject or do not reject the null hypothesis. If we take into account the level of confidence of 0.05, we do not reject the null hypothesis that the average severity of vehicle owners aged 50-60 in categories A and B is the same as that of the residents aged 50-60 in category C. However, at a confidence level of 0.1, we reject this null hypothesis.

A more correct way to verify the hypothesis $H_0 : \mu_{2A} = \mu_{2B} = \mu_{2C}$ is by simultaneously testing hypotheses

$$H_0 : \mu_{2A} = \mu_{2B} \quad \text{and} \quad H_0 : \left(\mu_{2A} + \mu_{2B}\right)/2 = \mu_{2C}$$

To verify these two hypotheses, we use the CONTRAST statement in the form

```
contrast 'Age_group*Residence 2A=2B=2C'
Residence 1 -1  Age_group*Residence 0 0 0 1 -1,
Residence 0.5 0.5 -1  Age_group*Residence 0 0 0 0.5 0.5 -1;
```

The result of the simultaneous testing of the two mentioned null hypotheses is an F-test statistic with degrees of freedom 2 for the nominator, which is also shown in 2nd row of the body of Table 5. Remember that degrees of freedom for the denominator correspond to the degrees of freedom SSE.

Table 5. Results of the CONTRAST statement

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Age_group*Residence 2A 2B vs 2C	1	0.36714636	0.36714636	2.85	0.0915
Age_group *Residence 2A=2B=2C	2	0.51349173	0.25674587	1.99	0.1365

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

Based on simultaneous testing, we find that even at the confidence level of 0.1 residence has no significant impact on the average severity in the age category of people aged 50 to 60. Given the insignificant differences, an insurance company may be interested in the degree of impact on the average severity when the insured person is aged 50-60 (on average over all residences). We can estimate this by using the ESTIMATE statement using Table 6.

Table 6. The coefficients for the ESTIMATE statement to estimate the mean $E(\mu_{2A}, \mu_{2B}, \mu_{2C})$

Age_group	Residence			Σ
	A	B	C	
1=60+	0	0	0	**0**
2=50-60	1	1	1	**3**
3=-50	0	0	0	**0**
Σ	**1**	**1**	**1**	**3**

Source: Self-processed.

The values in the body of Table 6 correspond to the coefficients of the means μ_{2A}, μ_{2B} and μ_{2C} in the required formula $(\mu_{2A} + \mu_{2B} + \mu_{2C})/3$. These coefficients are then taken as coefficients for interaction. The values in the sum column and the sum row are used as coefficients for the effects of factors A and B, and the sum value in the lower right corner represents the coefficient for the intercept. In order to obtain the required average of the three means, we must use the option DIVISOR = 3 to divide by the value of 3. The required statement then has the form

```
estimate 'Age_group*Residence mean 2A 2B 2C'
intercept 3  Age_group 0 3 0 Residence 1 1 1
Age_group*Residence 0 0 0 1 1 1 / divisor=3;
```

In addition to the point estimate LS mean for vehicle owners aged 50 to 60 (across all residences), the first row of the body of Table 8 provides also the test

of significance, i.e. the test result $H_0 : (\mu_{2A} + \mu_{2B} + \mu_{2C})/3 = 0$. In our case, this test is not of great importance, but thanks to the standard error estimate (0.7956) we can easily calculate the interval estimate and possibly verify the hypotheses that may be of interest to the insurance company. Based on the estimated value (4.4479) and its transformation $e^{4.4479} = 85.45$, we get a multiplier for those policyholders aged 50 to 60 (including the intercept). Of course, a part of the estimated regression function, which includes the influence of other factors, has to be used to estimate the average severity. In our case it is the factors Engine power and Vehicle_group, whose impact on the average severity we quantified by the expression

$$(1.095)^{EP} \cdot 0.9986^{EP^2} \cdot 1.000067^{EP^3} \cdot 1.470^{Vehicle\ Group=1} \cdot 1.276^{Vehicle\ Group=2}$$

The estimate of the average severity for vehicle owners aged 50 to 60 is calculated so that the value of the above expression is in addition multiplied by the multiplier 85.45. After adjusting for the intercept, i.e. $e^{4.4479}/e^{3.1170} = e^{4.4479-3.1170} = 3.7844$, the value $e^{4.4479}$ indicates that policyholders aged between 50 and 60 have an average severity, which is 3.7844 times higher than the reference category, which in our case consists of policyholders under the age of 50 from the regional cities of Banská Bystrica, Prešov, Trnava and Žilina.

If the insurance portfolio of policyholders aged 50 to 60 is 20% residence group A, 50% residence group B and residence group C the reminder, then it is necessary to use the weighted average of μ_{2A}, μ_{2B} and μ_{2C} to calculate the overall mean in the group of policyholders aged 50-60. Therefore, the interaction coefficients in the ESTIMATE statement follow the 2:5:3 ratio, which is captured also in Table 7.

Table 7. The coefficients for the ESTIMATE statement to estimate the mean $E(\mu_{2A}, \mu_{2B}, \mu_{2C})$ with weights in the ratio 2:5:3

Age_group	Residence			Σ
	A	B	C	
1=60+	0	0	0	**0**
2=50-60	0.2	0.5	0.3	**1**
3=-50	0	0	0	**0**
Σ	**0.2**	**0.5**	**0.3**	**1**

Source: Self-processed.

The statement ESTIMATE for the required weight mean has the form

```
estimate 'Age_group*Residence w_mean 2A 2B 2C'
intercept 1 Age_group 0 1 0 Residence 0.2 0.5 0.3
Age_group*Residence 0 0 0 0.2 0.5 0.3;
```

and it generates the output shown in row 2 of the body of Table 8.

Table 8. ESTIMATE statements results

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Age_group*Residence mean 2A 2B 2C	4.4479	0.7956	5.59	<.0001
Age_group*Residence w_mean 2A 2B 2C	4.4492	0.7896	5.63	<.0001

Source: Unnamed insurance company, self-processed in SAS Enterprise Guide.

Given the fact that between the means μ_{2A}, μ_{2B} and μ_{2C} significant differences were not confirmed (Table 5), the selected weights have a minimal impact on the overall mean μ_2 as indicated by the negligible differences in the point estimates of the LS means, as shown in Table 8.

5. Conclusions

The paper points to the possibilities of using the general linear model to analyse claim severity in motor third party liability insurance. In order to make adequate use of GLM, it was necessary to apply the logarithmic transformation of the explained variable, thereby eliminating the problem of heavy-tailed distribution and heteroscedasticity of error terms. Thus, the analyses presented in the paper are based on a log-linear model, in which the individual components are in an additive formula, which, however, is converted to a multiplicative formula after the backward exponential transformation. This fact needs to be taken into account when interpreting the results.

Our analyses confirmed that engine power and engine volume are strongly correlated, and their impact on claim severity overlaps significantly. By using the backward elimination method, only the engine power was retained from the two variables in the model, which avoided strong multicollinearity that could lead to problems with the interpretation of the results. Including this variable, categorical variables such as the age group and the owner's residence, as well as their interaction and the Vehicle group factor, were left in the model from the set of explanatory variables (listed in Section 3). Due to the fact that our base is an unbalanced multi-factor model, we could not use arithmetic means to compare the differences in claim severity at different levels of the relevant factors, so we used least squares means. By the gradual merging of categories in which comparable LS means of claim severity were estimated, among which there were no statistically significant differences, we created 3 groups of vehicle makes, 3 age categories of vehicle owners and 3 groups of residential cites. The results of our research reveal the impact of the relevant factors on claim severity, which is quantified by multipliers for each category of relevant factor through the exponential transformation of the respective regression coefficients. Since a significant interaction between the Age group and the Residence factors was confirmed, the paper also quantifies the multipliers for the categories that were created by combining the categories of the mentioned two factors.

Our empirical study shows that the claim severity does not change significantly in vehicles with 50 to 100 kW engine power, and a substantial increase occurs only in vehicles with higher power. The highest average severity was found in owners aged over 60 and in the owners from the regional cities Trenčín and Košice. Vehicle owners who were aged over 60 and had permanent residence in Trenčín and Košice showed 7.8-fold higher average severity, with other variables fixed, as compared to owners under the age of 50 living in the regional towns of Banska Bystrica, Prešov, Trnava and Žilina. That age category (up to 50 years) and the category of residence mentioned (Banská Bystrica, Prešov, Trnava and Žilina) are the least risky in terms of claim severity and their combination reduces the risk.

The benefit of the paper is not only empirical results, but the paper also points to the application of the general linear model to create tariff classes, to estimate average severities for these tariff classes and to detect simple and interaction effects of relevant factors. The general linear model provides such findings through model parameter estimates and least squares means, which are directly available in SAS software or which can be quantified using the CONTRAST and ESTIMATE statements.

The paper shows that the general linear model is an effective tool for the modelling of claim severity because it allows us to use quantitative and categorical regressors and their interactions as well. Unlike other methods, GLM provides estimation of the least square means (besides the arithmetic means) of the target variable. Moreover, PROC GLM in software SAS offers the CONTRAST statement, which is very useful to confirm significant differences between tariff classes in motor insurance. The values of these differences can be estimated using the ESTIMATE statement. Due to the possibility of testing several individual statistical hypotheses for LS means and the possibility of simultaneous testing, the GLM procedure is very flexible and proper for the purpose of tariffication in motor insurance. One disadvantage of the general linear model is the assumptions put on the random error. The error term often does not fulfil the assumption about homoscedasticity. In such a case, a researcher can try to use a logarithmic transformation as we did in our analysis presented in the article. If it does not work, we suggest applying generalized linear models, which are more flexible in this aspect.

Tools of the general linear model applied in the paper can be used by actuaries not only in claims severity, but also in claims frequency, and then for the determination of net premiums in motor insurance.

Acknowledgements

The publication was financed from the fund VEGA number 1/0618/17 *Modern tools for modeling and managing life insurance risks* granted by the Ministry of Education, Science, Research and Sport of the Slovak Republic.

REFERENCES

AGRESTI, A., (2015). Foundations of linear and generalized linear models, John Wiley & Sons.

ANDERSON, D., FELDBLUM, S., MODLIN, C., SCHIRMACHER, D., SCHIRMACHER, E., THANDI, N., (2007). A Practitioner's Guide to Generalized Linear Models: A foundation for theory, interpretation and application (3rd ed.), Towers Watson.

CAI, W., (2014). Making Comparisons Fair: How LS-Means Unify the Analysis of Linear Models. SAS Institute Inc. Paper. SA, S060-2014. [online] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.644.7680&rep=rep1&type=pdf [Accessed on 30 April 2019].

DAVID, M., (2015). Auto insurance premium calculation using generalized linear models, Procedia Economics and Finance, 20, pp. 147–156, DOI: https://doi.org/10.1016/S2212-5671(15)00059-3.

DE AZEVEDO, F. C., OLIVEIRA, T. A., OLIVEIRA, A., (2016). Modeling non-life insurance price for risk without historical information, REVSTAT–Statistical Journal, 14(2), pp. 171–192.
Available at: https://www.ine.pt/revstat/pdf/rs160205.pdf [Accessed on 30 April 2019].

DUAN, Z., CHANG, Y., WANG, Q., CHEN, T., ZHAO, Q., (2018). A Logistic Regression Based Auto Insurance Rate-Making Model Designed for the Insurance Rate Reform. International Journal of Financial Studies, 6(1), 18, DOI: https://doi.org/10.3390/ijfs6010018.

FOX, J., (2015. Applied regression analysis and generalized linear models. Sage Publications.

FREES, E., LEE, G., YANG, L., (2016). Multivariate frequency-severity regression models in insurance, Risks, 4(1), 4, DOI: https://doi.org/10.3390/risks4010004.

GOLDBURD, M., KHARE, A., TEVET, D., (2016). Generalized linear models for insurance rating, Casualty Actuarial Society, CAS Monographs Series, (5).

GRAHAM, J. M., 2008). The general linear model as structural equation modeling, Journal of Educational and Behavioral Statistics, 33(4), pp. 485–506, DOI: https://doi.org/10.3102/1076998607306151.

JONG, DE P., HELLER, G. Z., (2008). Generalized linear models for insurance data, Cambridge University Press.

KAFKOVÁ, S., KŘIVÁNKOVÁ, L., (2014). Generalized Linear Models in Vehicle Insurance. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62(2), pp. 383–388, DOI: https://doi.org/10.11118/actaun201462020383.

KIM, K., TIMM, N., (2006). Univariate and multivariate general linear models: theory and applications with SAS. Chapman and Hall/CRC.

KUZNETSOVA, A., BROCKHOFF, P. B., CHRISTENSEN. R. H. B., (2017). lmerTest package: tests in linear mixed effects models. Journal of Statistical Software, 82(13), DOI: https://doi.org/10.18637/jss.v082.i13.

LAMOTTE, L. R., (2019). A formula for Type III sums of squares. Communications in Statistics-Theory and Methods, pp. 1–11, DOI: https://doi.org/10.1080/03610926.2019.1586933.

LENTH, R. V., (2016). Least-squares means: the R package lsmeans. Journal of statistical software. 69(1), pp. 1–33, DOI: https://doi.org/10.18637/jss.v069.i01.

LITTELL, C. L., STROUP, W. W., FREUND, R. J., (2010). SAS for Linear Models (4th Revised ed.). North Carolina, USA: SAS Institute.

SHI, P., FENG, X., IVANTSOVA, A., (2015). Dependent frequency–severity modeling of insurance claims. Insurance: Mathematics and Economics, 64, pp. 417–428. DOI: https://doi.org/10.1016/j.insmatheco.2015.07.006.

THOMPSON, B., (2015). The Case for Using the General Linear Model as a Unifying Conceptual Framework for Teaching Statistics and Psychometric Theory. Journal of Methods and Measurement in the Social Sciences, 6(2), pp. 30–41, DOI: https://doi.org/10.2458/azu_jmmss_v6i2_thompson.

WILCOX, R. R., (2003). Applying contemporary statistical techniques, Elsevier.

WOOLDRIDGE, J. M., (2013). Introductory econometrics: A modern approach (5^{th} ed.), Mason: South-Western.