Equilibrium. Quarterly Journal of Economics and Economic Policy
Volume 17 Issue 3 September 2022
p-ISSN 1689-765X, e-ISSN 2353-3293
www.economic-policy.pl

ORIGINAL ARTICLE

Citation: Reiff, M., Šoltés, E., Komara, S., Šoltésová, T., & Zelinová, S. (2022). Segmentation and estimation of claim severity in motor third-party liability insurance through contrast analysis. *Equilibrium. Quarterly Journal of Economics and Economic Policy, 17*(3), 803–842. doi: 10.24136/eq.2022.028

Contact to corresponding author: Marian Reiff, marian.reiff@euba.sk

Article history: Received: 19.05.2022; Accepted: 1.09.2022; Published online: 30.09.2022

Marian Reiff
University of Economics, Slovakia
orcid.org/0000-0002-4064-704X

Erik Šoltés
University of Economics, Slovakia
orcid.org/0000-0001-8570-6536

Silvia Komara
University of Economics, Slovakia
orcid.org/0000-0001-6641-7456

Tatiana Šoltésová
University of Economics, Slovakia
orcid.org/0000-0002-0953-2519

Silvia Zelinová
University of Economics, Slovakia
orcid.org/0000-0002-9932-6857

Segmentation and estimation of claim severity in motor third-party liability insurance through contrast analysis

JEL Classification: *C12; C21; G22*

Keywords: *general linear model; claim severity; motor third party liability insurance; least squares means; contrast analysis*

Abstract

Research background: Using the marginal means and contrast analysis of the target variable, e.g., claim severity (CS), the actuary can perform an in-depth analysis of the portfolio and fully use the general linear models potential. These analyses are mainly used in natural sciences, medicine, and psychology, but so far, it has not been given adequate attention in the actuarial field.

Purpose of the article: The article's primary purpose is to point out the possibilities of contrast analysis for the segmentation of policyholders and estimation of CS in motor third-party liability insurance. The article focuses on using contrast analysis to redefine individual relevant factors to ensure the segmentation of policyholders in terms of actuarial fairness and statistical correctness. The aim of the article is also to reveal the possibilities of using contrast analysis for adequate segmentation in case of interaction of factors and the subsequent estimation of CS.
Methods: The article uses the general linear model and associated least squares means. Contrast analysis is being implemented through testing and estimating linear combinations of model parameters. Equations of estimable functions reveal how to interpret the results correctly.
Findings & value added: The article shows that contrast analysis is a valuable tool for segmenting policyholders in motor insurance. The segmentation's validity is statistically verifiable and is well applicable to the main effects. Suppose the significance of cross effects is proved during segmentation. In that case, the actuary must take into account the risk that even if the partial segmentation factors are set adequately, statistically proven, this may not apply to the interaction of these factors. The article also provides a procedure for segmentation in case of interaction of factors and the procedure for estimation of the segment's CS. Empirical research has shown that CS is significantly influenced by weight, engine power, age and brand of the car, policyholder's age, and district. The pattern of age's influence on CS differs in different categories of car brands. The significantly highest CS was revealed in the youngest age category and the category of luxury car brands.

Introduction

One of the primary tasks of the actuary is to calculate the pure premium so that the insurer (insurance company) covers, on average, the paid insurance benefits. In order to establish a tariff structure that reflects the various risk profiles in a portfolio, actuaries usually use econometric models. Those models include various classifying variables to create risk classes corresponding to each risk profile. Standard industry practice for pricing risks in non-life insurance became general linear models (GLM) and generalized linear models (GzLM). They are now commonly used for estimating the pure premium through the frequency–severity approach, based on a priori characteristics of the insurance policy (Zahi, 2021).

However, most non-life GLM and GzLM applications do not use contrast analysis, which allows for a deeper analysis of the impact of risk factors through testing and estimating linear combinations of model parameters. The article aims to point out the use of contrast analysis for the segmentation of policyholders and to estimate claim severity (CS).

In the interest of actuarial fairness and statistical correctness, segmentation of policyholders will be understood as the creation of such segments which include categories of policyholders between which there are no significant differences from the point of view of CS, and at the same time, there are demonstrable differences in CS between the segments. The article aims to answer the following research questions:

RQ1: *Can the relevant categorical factor be redefined through contrast analysis in such a way as to ensure the segmentation of policyholders in the sense of the above definition?*

RQ2: *Does segmentation, based on partial factors, which are appropriately set for the case of main effects, guarantee adequate segmentation even in the case of interaction of these factors? If not, how can contrast analysis be further used for segmentation in case of interaction of factors?*

RQ3: *How to estimate CS for segments that were created based on contrast analysis?*

The software SAS is used for the contrast analysis associated with GLM, specifically the SAS EG, SAS JMP, and PROC GLM in the SAS programming language. Within PROC GLM, the research presented in the article is based on the analysis of marginal means (least-squares means) and contrast analysis using the CONTRAST and ESTIMATE statements, which are used for the above testing and estimation.

The article shows that the analysis of marginal means and contrast analysis can be instrumental in the actuarial field, especially in motor insurance. The article presents a case study on a portfolio of 176,000 insurance contracts from motor third-party liability (MTPL) insurance. The dataset was provided by an unnamed insurance company operating in Slovakia.

In the next part of the article, we will provide an overview of scientific works that motivated us to research and use the methods listed in the Research methodology section. The results themselves are presented in the Results section, divided into four subsections. The first part describes the input variables and verifies the model's assumptions. The second part deals with segmentation based on marginal means analysis and contrast analysis. The final model is constructed in the third, and its parameters are interpreted. In the fourth part, testing and estimating linear combinations of GLM parameters are used to estimate claim severity in the case of risk factors interaction.

Literature review

In non-life insurance, two approaches are used to determine net premiums in general. The target variable is directly the loss per exposure (loss cost), or the number of claims per exposure (claim frequency — CF), and the average loss per claim (claim severity — CS) is modeled separately. Goldburd *et al.* (2016) state that separate claim frequency and severity modeling leads to lower variance of the error term compared to directly modeled loss cost. In addition, in the case of separate analyses, we can reveal effects that could go unnoticed when modeling loss costs. The standard techniques of net premium determination with claim frequency and severity separate modeling assume independence between the frequency-severity components. In reality, these components are largely dependent, especially in motor insurance (Su & Bai, 2020). However, some procedures can eliminate the problem of correlation between the two components, and these were dealt with by Shi *et al.* (2015). The above facts motivated us to consider separate modeling, while the article only focuses on the claim severity in MTPL insurance.

Many actuaries use techniques based on regression and the analysis of variance in their scientific works to calculate motor insurance premiums. Popular models include generalized linear models, which have been used by De Azevedo *et al.* (2016), Frees *et al.* (2016), de Jong and Heller (2008) and Kafková and Křivánková (2014). Claim frequency is modeled frequently by the Poisson regression model, and claim severity by the Gamma regression model (David, 2015; Duan *et al.*, 2018). As David (2015) indicates, generalized linear models allow for modeling a non-linear behavior and cases where residuals do not follow Gaussian (normal) distributions. This approach is very useful in non-life insurance, where claim frequency and claim severity follow asymmetric distributions, significantly deviating from the non-Gaussian distribution. The article uses the general linear model (GLM), a special case of the generalized linear model (GzLM). GLM and GzLM include statistical methods used to assess the effect of numerical continuous regressors and categorical factors on the target variable. The major difference is that GLM assumes that the error term is normally distributed, while GzLM does not require this assumption and allows for various other distributions that belong to exponential family distributions (Agresti, 2015; Fox, 2015). Although this fact favors GzLM models, in many cases a simple transformation of the target variable will solve the problem of violating homoscedasticity and normality, which allows the correct application of general linear models.

For modeling frequency-severity components in vehicle insurance, traditional risk factors such as customer age, vehicle age, vehicle engine power (Kafková, 2015), and others or telematics factors such as distances driven during a given period, the drivers' habits and behavior are used. Telematics and traditional rate-making factors give better outcomes in insurance pricing (Ayuso *et al.*, 2019). Unfortunately, we did not have such factors in our research. In addition to the above traditional factors, we also used car brand and geographical location, whose significant impact on claim severity was demonstrated Fung *et al.* (2021).

The main tools used in the presented research were marginal means and subsequent contrast analysis. For unbalanced data with a larger number of effects, either in the form of categorical factors or numerical covariates, group arithmetic means do not provide an adequate picture of the response of the target variable for the particular factor. The reason is that they do not consider other effects, which may lead to the Simpson paradox (Wang *et al.*, 2018). Cai (2014) states that if the data are unbalanced, arithmetic means are not appropriate, because they do not consider that not all factors have the same chance of influencing the target variable. In such cases, it is appropriate to estimate the marginal means based on the model, in our case, on the GLM. Marginal means actually correct the imbalance. Marginal means are group means, assuming that the influence of other explanatory variables is fixed. The marginal mean is also referred to as the LS-mean (*Least Squares mean*, see (Goodnight & Harvey, 1997)) or the EM-mean (*Estimated Marginal mean*; see (Searle *et al.*, 1980)). LS-means are predicted means calculated from the fitted model and adjusted appropriately for any other variable (Suzuki *et al.*, 2019).

The SAS software used in our analysis has the LSMEANS tool as part of the PROC GLM procedure. For example, IBM SPSS statistical software uses the EMMEANS tool, and the R software environment has created a special package for calculating marginal means (Lenth, 2016). It was originally called *lsmeans*, but its newer versions are called *emmeans* (Lenth *et al.*, 2022). For comparing other software (see Tabachnick & Fidell, 2013).

Marginal means and contrast analysis are mainly used in the natural sciences, e.g., in ecology and environmental science (Rivers *et al.*, 2017; Quigley *et al.*, 2018), in plant science (Byrne *et al.*, 2017; Huzar-Novakowiski & Dorrance, 2018), in biological science (Colin *et al.*, 2018; Singh *et al.*, 2015; Zhao *et al.*, 2019), in the human sciences, e.g., in medicine and sports medicine (Bae *et al.*, 2017; Bergelt *et al.*, 2020; Ennour-Idrissi *et al.*, 2016), and in psychology (Olivera-La Rosa *et al.*, 2020), but their application in actuarial science is rare. While some scientific articles use marginal means analysis, the application of contrast analysis is much less widespread, even though it is a relatively simple and effective statistical method for testing the differences between groups of means (Šoltés *et al.*, 2019). As Haans (2018) states, the reason for the occasional use of this method (despite its advantages) is that it is not implemented in a comfortable way in many statistical software packages.

As mentioned, we use SAS software, but the contrast analysis procedures are not specifically designed for SAS software but are also universally applicable in other software such as SPSS (George & Mallery, 2019; Haans, 2018), STATA (Haans, 2018), STATISTICA (de Sá, 2007), Statgraphics centurion (Statgraphics Technologies Inc., 2017), in the S-Plus system (Ugarte *et al.*, 2008; Wicklin, 2018) or in the R environment (Schad *et al.*, 2020; Tattar *et al.*, 2016). In addition, the contrast analysis procedures are not only used in GLM, but also in GzLM. According to Thompson (2006), the CONTRAST statement allows many more sophisticated questions to be addressed by procedures such as PROC GLM, PROC MIXED, PROC GENMOD, etc.

Research methods

The research interest in our article, the general linear model, 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 *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 indicate μ_{ij} the mean of the target variable for *i*-th category of factor *A* and *j*-th category of factor *B*. 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 $(\alpha\beta)_{ij}$ which denotes the *A* and *B* interaction effect. In the application part of the article, more than two factors will be taken into account. The factors will be in the form of quantitative variables and others in the form of categorical variables.

In terms of interpreting the results, it is important to note that in our research, we used factors with fixed effects (Searle & Gruber, 2017), and for categorical factors, we used indicator (dummy) coding (Darlington & Hayes, 2016). The interaction was based on the crossed classification structure (Littell *et al.*, 2010).

In the case of indicator coding, the general linear model can be written in the form of a multiple regression model

$$y_{ij} = \mu + \tau_1 x_{1j} + \tau_2 x_{2j} + \ldots + \tau_k x_{kj} + \varepsilon_{ij} \tag{2}$$

where $X_1, X_2, \ldots, X_k$ are dummy variables, while the variable X_j takes the value 1 for the observations from category *j*, otherwise, it takes the value 0. The parameter τ_i represents the difference between the mean in the *i*-th category and the mean in the reference category (in our case the *k*-th) form

$$\tau_i = \mu_i - \mu_k \tag{3}$$

while for the last parameter $\tau_k = 0$. Model (2) can be expressed by matrix notation as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

Whether in indicator coding or effect coding, the parameter τ_k is a linear combination of other parameters, τ_i. In this case, the matrix $\mathbf{X}$ is of non-full-rank, and a generalized inverse method is used to estimate the vector of parameters of the model (4), the result of which is an estimate

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-} \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{5}$$

where the matrix $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}$ is a generalized inverse matrix that must satisfy at least the first of the Penrose conditions (Searle & Gruber, 2017). In the application presented in this article, we use PROC GLM within SAS software, where g$_2$-inverse is used (Wicklin, 2018).

Although the estimation of the vector of parameters $\boldsymbol{\beta}$ obtained by the generalized inverse method is not unique, there is a group of linear functions of the mentioned parameters, which we refer to as estimable functions, for which there is a single solution (Agresti, 2015; Littell *et al.*, 2010). Estimable functions $\mathbf{L}\boldsymbol{\beta}$ have several properties (Searle & Gruber, 2017), and the following property will be important for our purposes

$$Var(\mathbf{Lb}) = \sigma_{\varepsilon}^{2}\left[\mathbf{L}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{L}^{\mathrm{T}}\right] \tag{6}$$

The reader can find its proof in (Elswick *et al.*, 1991; O'Brien, 2014).

In non-full rank models, we can test general linear hypotheses $H_0 : \mathbf{L\beta} = \mathbf{m}$, while $\mathbf{L\beta}$ must be an estimable function. General linear hypotheses and testable hypotheses are discussed in more detail in (McFarquhar, 2016; Poline *et al.*, 2007; Searle & Gruber, 2017). A special case of general linear hypotheses is the case when $\mathbf{m} = \mathbf{0}$. In such situation, to test the null hypothesis

$$H_0 : \mathbf{L\beta} = \mathbf{0} \tag{7}$$

uses an F-test or a t-test. For the F-test numerator, the sum of squares is calculated (SAS Institute Inc., 2017)

$$SS\left(H_0 : \mathbf{L\beta} = \mathbf{0}\right) = \left(\mathbf{Lb}\right)^{\mathrm{T}} \cdot \left[\mathbf{L}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{L}^{\mathrm{T}}\right]^{-1} \cdot \left(\mathbf{Lb}\right) \tag{8}$$

which has degrees of freedom l expressing the number of independent rows of the matrix $\mathbf{L}$. The test statistic is then given by formulae

$$F = \frac{\dfrac{SS\left(H_0 : \mathbf{L\beta} = \mathbf{0}\right)}{l}}{\dfrac{SSE}{n-p}} \tag{9}$$

whereas the sum of squared errors can be calculated according to the relation

$$SSE = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{Xb} \tag{10}$$

and $\dfrac{SSE}{n-p}$ is an unbiased estimate of the residual variance, and thus the following applies

$$\widehat{\sigma_{\varepsilon}^{2}} = MSE = \frac{SSE}{n-p} \tag{11}$$

We reject the null hypothesis if the value of the test statistic satisfies the inequation

$$F > F_{1-\alpha}\left(l; n-p\right) \tag{12}$$

The test mentioned above is used to verify simple hypotheses (if $l=1$), and also to simultaneously test multiple hypotheses (if $l \geq 2$). To verify simple hypotheses, of course, a *t*-test is also used, or alternatively, an interval estimate is constructed as well (Kuznetsova *et al.*, 2017; Westfall & Tobias, 2007).

If we use a multi-categorical factor in the general linear model and want to verify whether there is a significant difference between the different pairs of categories of the relevant factor in terms of target variable mean, then we use *Multiple Comparison Methods* (Lee & Lee, 2018; Rafter *et al.*, 2002; Rahardja, 2020). Various multiple comparison methods are known, and for our purposes the suitable ones are those that perform pairwise comparisons of the target variable means $\mu_i = \mu_{i'}$ for all pairs of factor categories. In these tests, all pairwise comparisons form the so-called *family*. If we perform all paired tests at the same level of significance α (i.e., at the same type I error), then it is desirable that for such a family the probability of the incorrect rejection of at least one of the null hypotheses is also at level α. This probability is called *familywise error rate* (FWER). Some tests have FWER under control, which means that if individual tests are performed at the significance level α, then also $FWER = \alpha$. Other tests are conservative ($FWER < \alpha$) or liberal ($FWER > \alpha$). In the application, we use the Tukey-Kramer test (hereinafter referred to as the "T-K test"). It is a modification of Tukey's test, also known as Tukey's HSD (Honestly Significant Difference) or Tukey's WSD (Wholly Significant Difference) test.

Although Tukey's test has the FWER under control, it is only suitable for balanced data. The T-K test allows proper testing even for unbalanced data. However, according to Rafter *et al.* (2002) the T-K test is slightly conservative. In addition, some configurations of imbalance and heteroscedasticity may cause greater conservatism in the Tukey-Kramer test (Herberich *et al.*, 2010). The T-K test is based on standardized pairwise differences (SAS Institute Inc., 2018)

$$t_{ii'} = \frac{\left(\overline{y}_i - \overline{y}_{i'}\right)}{\hat{\sigma}_{ii'}} \tag{13}$$

where $\overline{y}_i$ and $\overline{y}_{i'}$ are the means or LS-means for category i and category i', $\hat{\sigma}^2_{ii'}$ is the estimated variance of the difference $\left(\overline{y}_i - \overline{y}_{i'}\right)$, calculated for LS-means according to equation (6), whilst the variance σ^2_ε is substituted by its estimate given by equation (11).

The critical area for the T-K test is defined by the inequality

$$\left|t_{ii'}\right| \geq \frac{q_\alpha\left(k;\nu\right)}{\sqrt{2}} \tag{14}$$

where $q_\alpha\left(k;\nu\right)$ is a critical value for the significance level α, which is commonly tabulated and depends on k — the number of compared factor categories and ν — the degrees of freedom for *SSE*.

Results

Data description, a verification of assumptions and the transformation of the target variable

The presented analyses are based on a database provided by an unnamed insurance company for MTPL insurance for a period of approximately 4.5 years, more precisely from 1 January 2016 to 15 June 2020. The database contained almost 176,000 insurance contracts relating to passenger cars for everyday use, of which, only those on which there were claims and had records of regressors, which we used in the general linear model, were entered into the analysis. 7,776 insurance contracts were included in the analysis, which represented approximately 4.4% of the original set.

The target variable in our analysis is claim severity (CS). Due to the fact that the durations of the insurance contracts were different, the number of insurance benefits for each insurance contract on which an insured event was recorded in the observed period was proportionally converted to a calendar year (more precisely, 365 days). Thus, such a standardized explanatory variable was included in all analyses. As we expected, the distribution of the CS was strongly skewed to the right. Residuals of the preliminary model for the target variable CS, therefore, showed a significant deviation from the normal distribution and, in addition, a high degree of heteroscedasticity. We solved this problem by the logarithmic transformation of the explained variable CS (Figure 1 and Figure 2).

From a regression analysis point of view, we will further analyze the logarithmic-linear model

$$\ln CS = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

in which we considered variables related to the insured car as regressors (variables marked *name*_C), namely 4 quantitative variables:

- *Weight_C* – car weight (in kg),
- *Volume_C* – engine volume (in cm^3),
- *Age_C* – age of the car (in years),
- *Engine_C* – engine power of the car (in kW)
- and one categorical factor
- *Brand_C* – the car brand.

In addition, two explanatory variables characterizing the policyholder were included in the analysis, namely his or her age and the district in which the policyholder has permanent residence.

Our analysis revealed that the engine volume alone had a significant effect ($p = 0.0047$) on the target variable, but its correlation with engine power ($r = 0.6657$; $p < 0.0001$), led to the fact that the inclusion of the *Volume_C* variable was not confirmed ($p = 0.9822$) in the full model. In addition, engine volume has been shown to have contributed most significantly to multicollinearity ($\mathit{Variance\ Inflation\ Factor} = 3.0085$). According to Kim (2019) multicollinearity is present when the VIF is higher than 5 to 10 or the condition indices are higher than 10 to 30. After excluding the *Volume_C* factor from the model, the VIF did not exceed 2, and the condition number (maximum condition index) decreased from value $\eta_j = 23.873$ to the value $\eta_j = 19.117$. Although the condition number was still above level 10, two or more variance decomposition proportions corresponding to condition indices higher than 10 to 30 were not identified, which would exceed 80%, which would identify that multicollinearity is present between the explanatory variables corresponding to the exceeding variance decomposition proportions (Kim, 2019). This shows that after excluding the variable *Volume_C*, there is no evidence of collinearity among the variables.

We will add a note to the question of engine volume vs. engine power. It is clear that the engine power compared to the engine volume includes more relevant information on the dynamic characteristics of the car, which have an impact on the claim amount. Today, car dynamics cannot be adequately assessed by engine volume, e.g., also due to the so-called downsizing (reducing engine volume while maintaining performance) or due to an increase in the penetration of electric cars and hybrids into the automotive market.

Segmentation based on the analysis of marginal means

Apart from marketing strategies, the segmentation of policyholders should be statistically demonstrable. Although modeling is common in this area, less attention is paid to the use of marginal means analysis, which is based on the relevant model, whereas this kind of analysis provides an effective and correct tool for segmentation. In our analysis, we used the SAS programming language and the LSMEANS and CONTRAST statements within the PROC GLM procedure (Dean *et al.*, 2017; Kim & Timm, 2006; Littell *et al.*, 2010; Schad *et al.*, 2020). We divided the original continuous variable *Age* of the policyholder into six categories: up to 25 years (inclusive), 25–35, 35–45, 45–55, 55–65, and more than 65 years old. Based on the tests for the difference between the marginal means of the target variable for all pairs of age categories (Table 1), we find that at the significance level of 0.05 there are no statistically significant differences between the age categories 55–65 and 65$^+$ ($p = 0.5747$) and between individual pairs of 3 age categories: 25–35, 35–45, and 45–55 ($p = 0.0613$, $p = 0.1944$, $p = 0.4303$).

This means that we do not have enough evidence to be able to assume a different claim severity in the age categories 55-65 and 65$^+$, and therefore we will merge these two age categories. However, regarding the age groups 25-35, 35-45, and 45-55, it is necessary to verify the hypothesis

$$H_0 : \mu_2 = \mu_3 = \mu_4$$

We emphasize that the failure to reject the equality of individual pairs of means does not yet entitle us to assume the equality of the above 3 means. To verify the null hypothesis, we will use the simultaneous testing of 2 hypotheses, e.g., these two hypotheses

$$H_0 : \mu_3 = \mu_4 \qquad \wedge \qquad H_0 : \mu_2 = \frac{1}{2}(\mu_3 + \mu_4)$$

which we will rewrite into linear combinations

$$H_0 : \mu_3 - \mu_4 = 0 \qquad \wedge \qquad H_0 : \mu_2 - 0,5\mu_3 - 0,5\mu_4 = 0$$

The coefficients at means can be used directly in the CONTRAST statement. When testing multiple hypotheses simultaneously, linear combinations are separated by a comma within a single CONTRAST statement. The statement to simultaneously test the above hypotheses then has the notation:

CONTRAST 'Age 2=3=4' Age_cat **0 0 1 -1**, Age_cat **0 1 -0.5 -0.5**;

and its result is in Table 2.

Since we do not reject the null hypothesis ($p = 0.1707$), we will continue to consider the age category 25–55 (merging categories 25–35, 35–45, and 45–55). Next, we will thus work with the *Age_cat* factor, which has three categories: A — up to 25 years (inclusive), B — from 25 to 55 years (inclusive), and C — over 55 years.

We also applied the procedure of reducing factor categories to other multi-categorical factors for which a significant effect on CS was confirmed. There were two factors, namely the district and the car brand. From 79 districts of the Slovak Republic, we created four categories of districts (A to D), and from the 45 brands of passenger cars, that were included in the database, we obtained four categories of vehicles (A to D). In the case of all the categorical factors, we arranged the new categories in descending order in terms of CS. Using the above procedure, we created categories for each of the considered factors (*Age_cat*, *Brand_C*, and *District_cat*), among which there were significant differences in terms of claim severity, while for marginal means the following relation applies

$$\mu_A > \mu_B > \mu_C \qquad \text{or alternatively} \qquad \mu_A > \mu_B > \mu_C > \mu_D$$

which is confirmed by Figures 3–5.

General linear model with interaction

In our analysis, the interactions between the explanatory variables were also assessed, and we found that at a significance level of 0.05, the interaction between the *Age_cat* and *Brand_C* factors is statistically significant. It means that for different categories of car brands, the pattern of age’s influence on CS may vary. However, in the model with interaction, the influence of the *Age_cat* factor itself has not been confirmed ($p = 0.7109$), and therefore the *Age_cat* factor is further incorporated in the considered model only through the interaction. The statistical significance of the influence of individual regressors is verified in Table 3. Note that in Table 3, a Type IV SS is used to verify the significance of the effect of the particular factor because the table for sorting the set according to the considered categorical factors listed in Table 3 contained empty cells and in this case the use of the commonly used Type III SS would not be correct (Kuznetsova *et al.*, 2017; LaMotte, 2020).

The estimation of the parameters of the general linear model with regressors (including interaction), which are given in Table 3, provides the output in Table 4. The parameters of the model were estimated by the generalized inverse method, therefore the estimates of regression coefficients for the categories of factors are not unique, which is indicated by the letter B at the relevant parameters. The parameters, of course, depend on the choice of the reference category of the factor, but when changing the reference categories, we obtain equivalent results, which lead to the same estimates of the marginal means.

To quantify the impact of individual factors on the claim severity, it is necessary to convert the estimate of the model $\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$ into the form $\hat{y}_i = e^{\hat{\beta}_0} \cdot \left(e^{\hat{\beta}_1}\right)^{x_{i1}} \cdot \left(e^{\hat{\beta}_2}\right)^{x_{i2}} \cdot \ldots \cdot \left(e^{\hat{\beta}_k}\right)^{x_{ik}}$. In the additive model, the influence of reference categories is at the level "0" (see Table 4), which is transformed into the value $e^0 = 1$ in the multiplicative model. Based on the above transformation, using the parameter estimates from Table 4, we get

$$\begin{aligned}
\widehat{CS} = {} & 133.7147 \cdot (1.001785)^{Engine} \cdot 0.999895^{Weight} \cdot 1.010698^{Age} \cdot \\
& \cdot 4.7829^{Brand=A} \cdot 1.6220^{Brand=B} \cdot 1.1918^{Brand=C} \cdot \\
& \cdot 1.8451^{District=A} \cdot 1.4712^{District=B} \cdot 1.2787^{District=C} \cdot \\
& \cdot 0.2943^{Age\_cat=A \wedge Brand=B} \cdot 2.8938^{Age\_cat=A \wedge Brand=C} \\
& \qquad \cdot 3.4337^{Age\_cat=A \wedge Brand=D} \cdot \\
& \cdot 0.6343^{Age\_cat=B \wedge Brand=A} \cdot 1.0323^{Age\_cat=B \wedge Brand=B} \\
& \qquad \cdot 1.1933^{Age\_cat=B \wedge Brand=C} \cdot \\
& \cdot 1.2667^{Age\_cat=B \wedge Brand=D}
\end{aligned}$$

After the exponential transformation of the estimated intercept, we obtain a value of € 133.71, which can be understood as the basic claim severity, which, however, has no logical interpretation because it applies to an insured vehicle with zero power, zero weight and zero for driver’s age, further, for a vehicle from the group of brands D, from the group of districts D, and for a policyholder in the age category C (over 55 years old).

Subsequently, we will interpret the regression coefficients, more precisely $e^{\hat{\beta}_j}$, under the ceteris paribus assumption, i.e., under the condition that the other factors considered in the model remain unchanged. If the engine power increases by 1 kW, the claim severity increases by 0.1785%, and an increase in engine power by 10 kW results in an average increase of 1.80% [$(1.001785)^{10} = 1.0180$]. With an increase in vehicle weight of 100 kg, the claim severity will be reduced by 1.04% on average [$(0.999895)^{100} = 0.9896$]. An increase in the car's age by one year will cause an average increase in claim severity of 1.07%.

Let us now look at the influence of the *District_cat* factor, which is not in interaction with other factors. As we have already mentioned, the riskiest in terms of claim severity are districts in category A, followed by districts in categories B and C, and we have quantified the smallest claim severity in the category of districts D. Compared to the districts in category D, districts in category A, have on average, 84.51% higher claim severity, and districts B and C have, compared to the reference category (category D), claim severity higher by 47.12%, and 27.87%, respectively.

Since there is an interaction between the *Age_cat* and *Brand_C* factors, the influence of these factors can be calculated from the exponential bases for the *Brand_C* factor and for their interaction. In the 4.5-year period thus far, the insurance company has not had any insurance contract with a loss in group AA, and therefore this cell is empty. There was only one observation in group AB, so we also have insufficient information in this group. If we look away from these two groups, in Table 5 it is clear that we have estimated the highest claim severity for policyholders under the age of 25 and for vehicles in group A. In the corresponding groups (AC, AD, BA, and CA), the claim severity is more than three times higher (3.034 times to 4.783 times higher) than in the CD group, which includes insurance contracts in which the policyholder is over 55 years old, and the insured car belongs to the category of brands D. The CD group is the least risky in terms of claim severity. However, when interpreting these results, the insurance company must be careful and take into account the size of the sample.

Since the up to 25 years old age category and the brands of cars belonging to category A, where luxury brands were included by the above procedure, have a low count (only 0.5% of the whole set; Table 6), it is important to look at the relevance of the results for these categories.

The estimate of claim severity for groups determined by the Age_cat × Brand_C interaction

Tests of pairwise comparisons of LS-means for individual pairs of groups, which have arisen based on the interaction *Age_cat* × *Brand_C* (Table 7), show that within age category B (25-55 years old, blue matrix of *p*-values) and within age category C (55^+, green matrix of *p*-values) at a significance level of 0.05 there are significant differences between different categories of car brands. However, for age category A (up to 25 years old; yellow matrix of *p*-values), differences in claim severity for the cars' categories B, C and D ($p = 0.1027$; $p = 0.1159$; $p = 0.9937$) are not confirmed and group AA is even empty.

Until the insurance company has a sufficient number of observations to show significant differences in LS-means of claim severity for different categories of car brands within the age group of policyholders under 25 years old, it is reasonable to assume the same claim severity. We also confirmed the insignificant difference in LS-means of claim severity across the age category of policyholders up to 25 years old by a test of equality of marginal means in these three groups (AB, AC, AD; $p = 0.2563$). In Table 7, we see that the AC and AD groups, in particular, are very similar to the BA and CA groups, while there are no significant differences at significance level of 0.05 between any pair formed from these four groups. This also applies to the AB group, however, it is not so convincing there. However, since there was only one observation in this group, we will also assess it together with the AC and AD groups, even on the basis of not rejecting the hypothesis $H_0 : \mu_{AB} = \mu_{AC} = \mu_{AD}$. To test the null hypothesis

$$H_0 : \mu_{AB} = \mu_{AC} = \mu_{AD} = \mu_{BA} = \mu_{CA}$$

the simultaneous testing of four null hypotheses is required, e.g., these:

$$H_0 : \mu_{AB} = \mu_{AD} \quad \wedge \quad H_0 : \mu\left(\mu_{AB}, \mu_{AD}\right) = \mu_{AC} \quad \wedge$$
$$\wedge \quad H_0 : \mu_{BA} = \mu\left(\mu_{AB}, \mu_{AC}, \mu_{AD}\right) \quad \wedge \quad H_0 : \mu_{CA} = \mu\left(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA}\right)$$

For each of the null hypotheses, it is necessary to determine the coefficients that will enter the CONTRAST statement. We proceed by rewriting the null hypotheses in the form of a linear combination. We will show this only for the example of the last (fourth) hypothesis:

$$H_0 : \mu_{CA} = \mu\left(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA}\right)$$

which we will rewrite as follows $\mu_{CA} - \frac{1}{4}\left(\mu_{AB} + \mu_{AC} + \mu_{AD} + \mu_{BA}\right) = 0$, and further adjust it to a linear combination

$$-0.25 \cdot \mu_{AB} - 0.25 \cdot \mu_{AC} - 0.25 \cdot \mu_{AD} - 0.25 \cdot \mu_{BA} + \mu_{CA} = 0$$

Here it is important to emphasize that into the CONTRAST statement we do not enter coefficients at means but coefficients at effects (Littell *et al.*, 2010). While in the model without interaction, the coefficients at the effects and the coefficients at the means are identical, in the case of the model with the interaction, this does not apply, and the tested means should be overridden through the effects (using relation (1)). An easier way is to create a contingency table (Table 8). In its field, we write the coefficients from the above linear combination. However, we must realize that the AA cell is empty in our case.

In the sum row we get the coefficients for the factor *Brand_C*, in the sum column, we calculate the coefficients for the factor *Age_cat*, and the sum of all the coefficients (the cell in the lower right corner) represents the coefficient for the intercept. Because the *Age_cat* factor itself is not included in our model, the corresponding coefficients (in the sum column) will not be used in the CONTRAST statement. In Table 8 are determined the coefficients for the fourth partial hypothesis, and similarly, we would determine the coefficients for the first three hypotheses ($H_0 : \mu_{AB} = \mu_{AD}$; $H_0 : \mu\left(\mu_{AB}, \mu_{AD}\right) = \mu_{AC}$; $H_0 : \mu_{BA} = \mu\left(\mu_{AB}, \mu_{AC}, \mu_{AD}\right)$)

The resulting statement has the following syntax:

```
CONTRAST 'AB=AC=AD=BA=CA'
    Brand_C 0 1 0 -1 Age_cat*Brand_C 1 0 -1,
    Brand_C 0 0.5 -1 0.5 Age_cat*Brand_C 0.5 -1 0.5,
    Brand_C 1 -0.333 -0.333 -0.333 Age_cat*Brand_C -0.333 -0.333
    -0.333 1,
    Brand_C 1 -0.25 -0.25 -0.25 Age_cat*Brand_C -0.25 -0.25 -0.25
    -0.25 0 0 0 1;
```

Statements for partial hypotheses are separated by a comma. Figure 6 is the output from the SAS JMP software, which shows a matrix $\mathbf{L}$ for verifying the general linear hypothesis (7), the result of the partial *t*-tests, and the overall *F*-test (9).

Based on the stated p-values ($p = 0.1159$; $p = 0.1600$; $p = 0.2908$; $p = 0.1214$) at a significance level of 0.05, we do not reject the above-mentioned partial null hypotheses. However, the overall result is a *p*-value of 0.4661, which leads us to conclude that we cannot reject the equality of the means $\mu_{AB} = \mu_{AC} = \mu_{AD} = \mu_{BA} = \mu_{CA}$ at any commonly used level of significance.

Let us assume that the current representation of groups AB, AC, AD, BA, and CA, which can be expressed by the ratio 2:31:12:40:15, will not change in the future. Then, to estimate the LS-mean of claim severity for the cluster of groups AB, AC, AD, BA, and CA, we use the ESTIMATE statement (Dean *et al.*, 2017; Littell *et al.*, 2010; SAS Institute Inc., 2018), in which we consider the above weights. These weights are the coefficients for the *Age_cat* × *Brand_C* interaction and are listed in Table 9.

As with the CONTRAST statement, the sum row and the sum column also now contain the coefficients for the *Brand_C* and *Age_cat* factors, respectively. However, unlike the CONTRAST statement, the weight for the intercept is non-zero and has a value of 100. Since the intercept (grand mean) only needs to be counted in once and other effects need to be counted in proportionally, the ESTIMATE statement uses the Divisor option with a constant of 100.

```
ESTIMATE 'mean_w(AB, AC, AD, BA, CA)'
        intercept 100 Brand_C 55 2 31 12
        Age_cat*Brand_C 2 31 12 40 0 0 0 15/divisor=100;
```

This statement generates the output in Table 10. After taking into account the above weights, the point estimate of the claim severity mean for the cluster of groups AB, AC, AD, BA, and CA has the value

$$e^{6.4025} = 603.35 \text{ €}$$

We can also calculate an interval estimate using the standard error from the Table 10. 95% confidence interval is

$$\left(e^{6.4025-1.96\cdot 0.19825}; e^{6.4025+1.96\cdot 0.19825}\right)$$

$$\left(409.08; 889.89\right)$$

To interpret this result correctly, let us look at an estimable function. In the general form of an estimable function, the coefficients for a reference category of a categorical factor are a linear combination of the coefficient at the intercept and the coefficients at the other categories of the factor. Since the *Age_cat* factor has a reference category of C, all categories that include the C category of the *Age_cat* factor are referenced for the *Age_cat* × *Brand_C* interaction. We can see this fact in Table 4, where there are zero regression coefficients for categories CA, CB, CC, and CD. Thus, for the *Age_cat* × *Brand_C* interaction, not 12 coefficients (3 categories of the *Age_cat* factor × 4 categories of the *Brand_C* factor) are determined for the estimable function, but only 7, because the AA category was empty and the coefficients for the 4 categories (CA, CB, CC, and CD) are a linear combination of other coefficients. In Table 4, there is a model with 23 parameters, but 6 (category D of *Brand_C* factor, category D of *District_cat* factor, and 4 mentioned categories of *Age_cat* × *Brand_C* interaction) are zero because they are a linear combination of other parameters. These linear combinations are written in the general form of an estimable function (see Table 11):

$$\begin{aligned}\mathbf{Lb}^{\circ} = {} & L_1\mu + L_2 Engine + L_3 Weight + L_4 Age + \\ & + L_5\alpha_A + L_6\alpha_B + L_7\alpha_C + \left(L_1 - L_5 - L_6 - L_7\right)\alpha_D + \\ & + L_9\gamma_A + L_{10}\gamma_B + L_{11}\gamma_C + \left(L_1 - L_9 - L_{10} - L_{11}\right)\gamma_D + \\ & + L_{13}\left(\beta\gamma\right)_{AB} + L_{14}\left(\beta\gamma\right)_{AC} + L_{15}\left(\beta\gamma\right)_{AD} + \\ & + L_{16}\left(\beta\gamma\right)_{BA} + L_{17}\left(\beta\gamma\right)_{BB} + L_{18}\left(\beta\gamma\right)_{BC} + L_{19}\left(\beta\gamma\right)_{BD} + \\ & + \left(L_9 - L_{16}\right)\left(\beta\gamma\right)_{CA} + \left(L_{10} - L_{13} - L_{17}\right)\left(\beta\gamma\right)_{CB} + \left(L_{11} - L_{14} - L_{18}\right)\left(\beta\gamma\right)_{CC} + \\ & + \left(L_1 - L_9 - L_{10} - L_{11} - L_{15} - L_{19}\right)\left(\beta\gamma\right)_{CD}\end{aligned}$$

whereas we used the symbols α_i, β_j a γ_l for the *District_cat, Age_cat* a *Brand_C* factors to shorten the notation. Using the CONTRAST and ESTIMATE statements, we can test or estimate any linear combination that satisfies the above relation, where the values of the 17 coefficients can be any real numbers. In the case of the ESTIMATE statement, which we used to estimate the weighted marginal mean $\mu\left(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA}, \mu_{CA}\right)$, the estimable function has the form:

$$\begin{aligned}\mathbf{Lb}^{\circ} = {} & 1\cdot\mu + 0\cdot Engine + 0\cdot Weight + 0\cdot Age + \\ & + 0.25\cdot\alpha_A + 0.25\cdot\alpha_B + 0.25\cdot\alpha_C + 0.25\cdot\alpha_D + \\ & + 0.55\cdot\gamma_A + 0.02\cdot\gamma_B + 0.31\cdot\gamma_C + 0.12\cdot\gamma_D + \\ & + 0.02\cdot(\beta\gamma)_{AB} + 0.31\cdot(\beta\gamma)_{AC} + 0.12\cdot(\beta\gamma)_{AD} + \\ & + 0.4\cdot(\beta\gamma)_{BA} + 0\cdot(\beta\gamma)_{BB} + 0\cdot(\beta\gamma)_{BC} + 0\cdot(\beta\gamma)_{BD} + \\ & + 0.15\cdot(\beta\gamma)_{CA} + 0\cdot(\beta\gamma)_{CB} + 0\cdot(\beta\gamma)_{CC} + 0\cdot(\beta\gamma)_{CD}\end{aligned}$$

The coefficients of this function are generated using option E within the ESTIMATE statement. While in the ESTIMATE statement, we used integer coefficients, and we divided them by the value 100 using the option DIVISOR = 100, the coefficients of the estimable function are already in the form of decimal numbers (after dividing by the value 100). Let us note that the coefficients for the continuous variables *Engine*, *Weight*, and *Age* are zero, and the coefficients for factor variations α_i (*District_cat*) are all the same (with a value of 0.25), thereby eliminating the influence of this factor and considering the average value across all the categories of the *District_cat* factor. This is important for the correct understanding of the point and interval estimate we obtained above. These are, therefore, estimates of the claim severity, adjusted for the effect of other continuous numerical variables and averaged across all the categories of categorical factors included in the GLM. Estimates of claim severity for all groups determined by the *Age_cat* × *Brand_C* interaction (adjusted for the effect of the continuous variables *Engine*, *Weight*, and *Age* and averaged across all categories of the *District_cat* factor) are given in Table 12. The youngest policyholders have the largest CS mean (category A of *Age_cat* factor; up to 25 years (inclusive)) and luxury car brands (category A of *Brand_C* factor). At the same time, these categories have the largest standard error of the estimate, which was reflected in the widest interval estimate. (409.1-889.1; see Table 12). On the contrary, we found the smallest CS mean and the smallest standard error of estimate in the CD group of *Age_cat* × *Brand_C* interaction, which includes insurance contracts in which the policyholder is over 55 years old, and the insured car belongs to the category of brands D.

Let us note that the ESTIMATE statement can also be used to estimate the claim severity for any contract, the profile of which is characterized by specific values of relevant factors contained in the model, estimated in Table 4.

Discussion

A recent topic in non-life insurance is being investigated on improving the original tariffs based on GLMs. Several researchers and actuaries are reaching for machine learning methods such as neural networks (Burka *et al*., 2021; Staudt & Wagner, 2021), tree-based methods (Burka *et al*., 2021; Staudt & Wagner, 2021; Henckaerts *et al*., 2021) and gradient boosting machines (Henckaerts & Antonio, 2022; Henckaerts *et al*., 2021). Burka *et al.* (2021) found in their empirical research that all the separate models (GLM, GAM, Random forest, Neural network) showed good figures; however, the best model was a mixture of them. Staudt and Wagner (2021) came to a similar conclusion when in their analysis, no model (GLM, GAM, Random forest) was outperforming the other ones throughout all the criteria. Henckaerts *et al*. (2021) state that the gradient boosting machine can be used to discover the important variables and interactions between those variables, which can then be included in a GLM for deployment. Although machine learning methods achieve good results, several researchers recommend these methods to improve GLM. Thus, GLM continues to have an irreplaceable position in non-life insurance, especially in motor insurance. The article shows that GLM provides much more information than is presented in scientific works in the actuarial field. This valuable information can be obtained by contrast analysis and subsequently used for segmentation and prediction.

The current article points out the possibilities of using contrast analysis in the segmentation of policyholders and the estimation of claim severity based on actual data from 7,776 insurance contracts from MTPL insurance of passenger cars for common use. In addition, it is about the insurance contracts with a loss within the entire portfolio of approximately 176,000 insurance contracts. An unnamed insurance company provided these data for a period of approximately 4.5 years. The presented results are based on our application of a GLM with fixed effects with a crossed classification structure. We must emphasize that indicator coding of multi-categorical factors was used, which we considered when interpreting the results. The specifics of other coding methods and corresponding interpretations are provided by Darlington and Hayes (2016).

The GLM parameters were estimated by the least-squares method. Several multi-categorical factors entered the model as the regressors, so the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ had a non-full rank, and we had to use the generalized inverse method. Since the distribution of residuals was significantly skewed to the right, and the variance of residuals was heteroscedastic, we used a logarithmic transformation of the target variable. Note that in the insurance claim severity modeling, the log-normal distribution is traditionally applied in GLM (Frees *et al.* (Eds.), 2014). We obtained a model in which homoscedasticity was not violated, and the deviation of the empirical residuals from the normal distribution was minimal. Since the sample was relatively large, we could rely on a central limit theorem, which means least squares estimates have an asymptotically normal distribution (Wooldridge, 2013).

Another practical problem we had to deal with was a large number of categories of some of the considered factors. We verified between which categories of the particular factor there are no significant differences based on the LS-means (marginal means) using the Tukey-Kramer test. The analyses presented in the article are based mainly on contrast analysis.

Let us look at the main findings based on our statistical analyses. We found that claim severity is significantly affected by engine power, weight, age, the brand of the car, the policyholder's age, and the district in which the policyholder resides. Besides these regressors, we also considered the engine volume, which was correlated with the engine power and significantly contributed to the multicollinearity. Therefore, it was excluded from the model. Of these factors, it had the most significant impact on the car’s age and brand, the policyholder's age, and the district. The influence of these factors on the claim of severity was also confirmed by Fung *et al.* (2021). Henckaerts and Antonio (2022), like we, found that a car's brand and geographical location are among the factors that largely determine CS. At the same time, however, they found that the most fundamental influence is the vehicle's weight. This factor turned out to be significant, but less substantial, in our analysis, which may be due to the fact that we only considered passenger cars, while Henckaerts and Antonio (2022) analyzed all vehicle categories.

We used the analysis of marginal means and simultaneous testing to reduce the number of categories of categorical factors. Based on these analyses, the *Age_cat* factor (the age category of the policyholder) was created with three categories, while the significantly highest claim severity (under the ceteris paribus assumption) was quantified in the youngest age category (up to 25 years old) and the lowest in the oldest policyholders’ category (over 55 years old). The higher claim severity in the young policyholders’ segment was also detected Fung *et al.* (2021) and Staudt and Wagner (2021). Our finding is consistent Henckaerts *et al.* (2018) findings, who, based on the MTPL insurance portfolio from a Belgian insurer in 1997, concluded that very young drivers are involved in more severe car accidents. At the same time, they found that the average claim cost starts to increase for policyholders older than sixty rapidly. Which explained the assumption that older policyholders drive more expensive cars and repairing costs increase.

In contrast to Henckaerts *et al.* (2018), we had data on car brands and confirmed that luxury cars have significantly higher average claim costs than common cars. Henckaerts and Antonio (2022) came to a similar conclusion, arguing that some more expensive brands lead to higher severities. The *Brand_C* factor included 45 car brands, and their analysis led us to create four statistically significantly different (in terms of claim severity) categories. The riskiest category includes luxury car brands. From the 79 districts of the Slovak Republic, we created 4 clusters of districts using the above analyses. In the case of the *Age_cat* factor and the *Brand_C* factor, the little numerous riskiest group was created, which in both cases included only 0.5% of insurance contracts with a loss. Our analysis showed that even though the insurance company has a large insurance portfolio (in our case, there were approximately 176,000 insurance contracts, of which 7,776 had a loss), some tariff classes may have a low count. This risk increases with the number of categories of tariff factors. Despite reducing the categories of the above three multi-categorical factors, we did not avoid this problem either.

In the analysis, we also assessed the interactions between the explanatory variables. We revealed that the influence of the car brand on claim severity is different in different age categories of policyholders. Among other factors, no significant interaction was confirmed. Thus, in addition to the regressors mentioned above, the final model also includes the interaction between the *Age_cat* and *Brand_C* factors. After the backward transformation of the logarithmic-linear model, we obtained an estimate of the model in exponential form:

$$\begin{aligned}\widehat{CS} = {} & 133.7147 \cdot (1.001785)^{Engine} \cdot 0.999895^{Weight} \cdot 1.010698^{Age} \cdot \\ & \cdot 4.7829^{Brand=A} \cdot 1.6220^{Brand=B} \cdot 1.1918^{Brand=C} \cdot \\ & \cdot 1.8451^{District=A} \cdot 1.4712^{District=B} \cdot 1.2787^{District=C} \cdot \\ & \cdot 0.2943^{Age\_cat=A \wedge Brand=B} \cdot 2.8938^{Age\_cat=A \wedge Brand=C} \\ & \qquad \cdot 3.4337^{Age\_cat=A \wedge Brand=D} \cdot \\ & \cdot 0.6343^{Age\_cat=B \wedge Brand=A} \cdot 1.0323^{Age\_cat=B \wedge Brand=B} \\ & \qquad \cdot 1.1933^{Age\_cat=B \wedge Brand=C} \cdot \\ & \cdot 1.2667^{Age\_cat=B \wedge Brand=D}\end{aligned}$$

In the riskiest group of the policyholders in terms of policyholder's age and the vehicle's brand, i.e. in the group of policyholders up to 25 years old with a car from category A (group -25A), we did not have any information, so we abstracted from this group. With simultaneous testing, we found out that the insurance company did not have enough evidence to be able to assume a different claim severity among groups -25B, -25C, -25D, 25-55A, and 55$^+$ A, i.e., that in the riskiest age category — up to 25 years old (across categories of car brands B, C, and D) as well as in the riskiest category of car brands — the category of brands A (across age categories 25-55 years old and 55$^+$), we can assume the same claim severity.

We revealed this for five groups (-25B, -25C, -25D, 25-55A, and 55$^+$ A) using the ESTIMATE statement at a level of 603.35 Euros, while using weights based on the proportional representation of individual groups in the empirical set, that is, assuming that this ratio replicates the ratio in the hypothetical population. The coefficients of the estimable function showed that it is the claim severity adjusted for the impact of the continuous variables *Engine*, *Weight*, and *Age* and averaged across all the categories of the *District* factor. We must emphasize that an actuary with additional information may approach a contrast analysis differently. There were significant differences between the other groups as determined by the *Age_cat* × *Brand_C* interaction. We quantified the point estimates and 95% confidence intervals of the adjusted means of claim severities for these groups, which are shown in Table 12.

The smallest mean of claim severities is in group 55$^+$ D, i.e., for policyholders aged 55$^+$ with a vehicle from the least risky category of brands (category D). Based on point estimates, we can say that for drivers under the age of 25 and for the riskiest vehicles (category A including luxury brands), the mean of claim severities is approximately 231% higher. Opposite the group 55$^+$ D, in the age category of policyholders from 25–55 years old, in individual categories of vehicle brands B, C, and D, the means of claim severities are higher by 67.4%, 42.2%, and 26.7%, respectively. Finally, in groups 55$^+$ B and 55$^+$ C, the means of claim severities are 62.2% and 19.2% higher than in group 55$^+$ D.

The article shows that the analysis of marginal means and contrast analysis, which we performed using the LSMEANS, CONTRAST, and ESTIMATE statements in the SAS EG and SAS JMP, are effective tools for reducing factor categories and assessing differences in the means of the target variable for different factor categories and for different groups created by the interaction of factors, but also for the prediction of the target variable. Haans (2018) states that contrast analysis is an efficient and effective means for conducting post-hoc analyses but is used relatively little because the method is not implemented, at least not in a convenient point-and-click manner, in most statistical software packages. In the article, however, we have shown that the application of contrast analysis in statistical software requires intervention in the programming code, but it is relatively simple, and we agree with Haans (2018) that contrast analysis is even understandable to researchers with a minimal background in statistics. We also agree with the statement by Schad *et al.* (2020), who say that contrast coding makes it possible to implement comparisons in a very flexible and general way.

Since contrast analysis is one of the modern quantitative procedures used in modeling, e.g., in GLM and GzLM models with fixed as well as random effects or in modeling categorical data, the procedures presented in the article are universal. Contrast analysis is not only universal in terms of the form of modeling, but also in terms of software support. By using contrast analysis of the marginal means of the target variable (whether claim severity, claim frequency or loss cost), the actuary can perform an in-depth analysis of the portfolio and make full use of the GLM's and GzLM's potential.

Conclusions

The article focuses on applying contrast analysis associated with the general linear model (GLM) to analyze claim severity in motor third-party liability insurance. Contrast analysis makes it relatively easy to test and estimate different linear combinations of general linear model parameters, thus answering most research or practical questions that have led researchers and analysts to use GLM. The article points out that the analysis of marginal means and contrast analysis is mainly used in natural sciences, medicine, and psychology. However, it has not been given adequate attention in actuarial or economics research.

Analyzes presented in the paper confirmed that contrast analysis could be a useful tool for the segmentation and estimation of claim severity in actuarial, especially in motor vehicle insurance, for the reasons listed below. With the help of contrast analysis, it is possible to pre-define the relevant categorical factor in such a way as to ensure a statistically correct segmentation of the policyholders (RQ1). Segmentation based on partial factors, suitable for the case of main effects, does not guarantee adequateness in the case of interaction of these factors (RQ2). However, this problem can be solved through contrast analysis. The article presents the contrast analysis approach and the procedure for point and interval estimates of the claim severity for segments created based on contrast analysis (RQ3).

The advantage of segmentation is provability, which can be well communicated to all stakeholders (e.g., managers and clients), the key requirement for implementing the segmentation into practice. The practical use of contrast analysis also lies in the fact that it can be applied within a wide range of models that fall into the GLM or GzLM category, through professional analysis software such as SAS, presented in the article, or through open-source systems, e.g., R.

We want to emphasize that the article provides an empirical analysis based on the portfolio of one insurance company operating in Slovakia. Although we believe that many conclusions apply to a portfolio of other insurance companies that also operate in other countries, at least in the CEE countries, it needs to be verified by further research. The analysis results have their limitations, which relate to the factors that were used in GLM. In our analysis, telematic factors such as distances driven during a given period and the drivers' habits and behavior were not used. The insurance company did not provide these factors, and we did not have information about the driving experience of policyholders. However, it should be noted that although several scientific studies have shown that driving experience significantly reduces the claim (see Ayuso *et al.*, 2019; Ordaz *et al.*, 2011), this may not apply in the case of claim severity (Alemany *et al.*, 2020).

Obtained empirical results also have a time limit, which is especially true nowadays, at the time of the COVID-19 pandemic, the energy crisis, and high inflation. Spilbergs *et al.* (2022) showed that the COVID-19 period also significantly impacted MTPL claims due to the change in traffic intensity. However, macroeconomic indicators also have a demonstrable influence on MTPL claims, as confirmed by Spilbergs *et al.* (2021). Therefore, the war in Ukraine and the energy crisis associated with inflation will impact claims in MTPL. For this reason, risk factors and classification models need to be regularly validated based on the most up-to-date information available.

Since the latest research has confirmed that machine learning methods also achieve good results. We will examine how contrast analysis linked to GLM or GzLM combined with machine learning methods can improve actuarial modeling, setting an adequate tariff structure and pricing in motor insurance or, more generally, in non-life insurance. It is a challenge for further research to improve the mentioned processes from the actuary's and the stakeholder's points of view when introducing them into insurance practice.

References

Agresti, A. (2015). *Foundations of linear and generalized linear models*. New York: John Wiley & Sons.

Alemany, R., Bolancé, C., Rodrigo, R., & Vernic, R. (2020). Bivariate mixed Poisson and Normal Generalised Linear models with Sarmanov dependence—an application to model claim frequency and optimal transformed average severity. *Mathematics*, *9*(1), 73. doi: 10.3390/math9010073.

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, *46*(3), 735–752. doi: 10.1007/s11116-018-9890-7.

Bae, J., Kim, Y. Y., & Lee, J. S. (2017). Factors associated with subjective life expectancy: comparison with actuarial life expectancy. *Journal of Preventive Medicine and Public Health*, *50*(4), 240. doi: 10.3961/jpmph.17.036.

Bergelt, M., Fung Yuan, V., O'Brien, R., Middleton, L. E., & Martins dos Santos, W. (2020). Moderate aerobic exercise, but not anticipation of exercise, improves cognitive control. *PloS One*, *15*(11), e0242270. doi: 10.1371/journal .pone.0242270.

Burka, D., Kovács, L., & Szepesváry, L. (2021). Modelling MTPL insurance claim events: can machine learning methods overperform the traditional GLM approach? *Hungarian Statistical Review*, *4*(2), 34–69. doi: 10.35618/hsr2021. 02.en034.

Byrne, K. M., Adler, P. B., & Lauenroth, W. K. (2017). Contrasting effects of precipitation manipulations in two Great Plains plant communities. *Journal of Vegetation Science*, *28*(2), 238–249. doi: 10.1111/jvs.12486.

Cai, W. (2014). Making comparisons fair: how LS-means unify the analysis of linear models. *SAS Institute Inc. Paper SA, S060-2014*.

Colin, T., Bruce, J., Meikle, W. G., & Barron, A. B. (2018). The development of honey bee colonies assessed using a new semi-automated brood counting method: CombCount. *PLoS One*, *13*(10), e0205816. doi: 10.1371/journal.pone. 0205816.

Darlington, R. B., & Hayes, A. F. (2016). *Regression analysis and linear models: concepts, applications, and implementation*. Guilford Publications.

David, M. (2015). Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, *20*, 147–156. doi: 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), 171–192. doi: 10.57805/revstat.v14i2.185.

de Jong, P., & Heller, G. Z. (2008). Generalized linear models for insurance data. *Cambridge Books*.

de Sá, J. P. M. (2007). *Applied statistics using SPSS, Statistica, MatLab and R*. Springer Science & Business Media.

Dean, A., Voss, D., & Draguljić, D. (2017). *Design and analysis of experiments* Springer, Cham.

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: 10.3390/ijfs6010018.

Elswick Jr, R. K., Gennings, C., Chinchilli, V. M., & Dawson, K. S. (1991). A simple approach for finding estimable functions in linear models. *American Statistician*, *45*(1), 51–53. doi: 10.1080/00031305.1991.10475766.

Ennour-Idrissi, K., Têtu, B., Maunsell, E., Poirier, B., Montoni, A., Rochette, P. J., & Diorio, C. (2016). Association of telomere length with breast cancer prognostic factors. *PLoS One*, *11*(8), e0161903. doi: 10.1371/journal.pone.016 1903.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.) (2014). *Predictive modeling applications in actuarial science (Vol. 1).* Cambridge University Press.

Frees, E. W., Lee, G., & Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, *4*(1), 4. doi: 10.3390/risks4010004.

Fung, T. C., Badescu, A. L., & Lin, X. S. (2021). A new class of severity regression models with an application to IBNR prediction. *North American Actuarial Journal*, *25*(2), 206–231. doi: 10.1080/10920277.2020.1729813.

George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: a simple guide and reference*. Routledge.

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2016). Generalized linear models for insurance rating. *Casualty Actuarial Society, CAS Monographs Series*, *5*.

Goodnight, J. H, & Harvey, W. R (1997). *SAS technical report R-103. Least Squares Means in the Fixed Effects General Model*. Cary, NC: SAS Institute Inc.

Haans, A. (2018). Contrast analysis: a tutorial. *Practical Assessment, Research, and Evaluation*, *23*(1), 9. doi: 10.7275/7dey-zd62.

Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *8*, 681–705. doi: 10.1080/03461238.2018.1429300.

Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, *25*(2), 255–285. doi: 10.1080/10920277.2020.174 5656.

Henckaerts, R., & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance: Mathematics and Economics*, *105*, 79–95. doi: 10.1016/j.insmath eco.2022.03.011.

Herberich, E., Sikorski, J., & Hothorn, T. (2010). A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PloS one*, *5*(3), e9788. doi: 10.1371/journal.pone.0009788.

Huzar-Novakowiski, J., & Dorrance, A. E. (2018). Genetic diversity and population structure of Pythium irregulare from soybean and corn production fields in Ohio. *Plant Disease*, *102*(10), 1989–2000. doi: 10.1094/PDIS-11-17-1725-RE.

Kafková, S., & Křivánková, L. (2014). Generalized linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *62*(2), 383–388. doi: 10.11118/actaun201462020383.

Kafková, S. (2015). Bonus-malus systems in vehicle insurance. *Procedia Economics and Finance*, *23*, 216–222. doi: 10.1016/S2212-5671(15)00354-8.

Kim, K., & Timm, N. (2006). *Univariate and multivariate general linear models: theory and applications with SAS*. Chapman and Hall/CRC.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558. doi: 10.4097/kja.19087.

Kuznetsova. A., Brockhoff. P. B., & Christensen. R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*. *82*(13), 1–26. doi: 10.18637/jss.v082.i13.

LaMotte, L. R. (2020). A formula for Type III sums of squares. *Communications in Statistics-Theory and Methods*, *49*(13), 3126–3136. doi: 10.1080/03610926.2019.1586933.

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, *71*(5), 353. doi: 10.4097/kja.d.18.00242.

Lenth, R., V. (2016). Least-squares means: the R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. doi: 10.18637/jss.v069.i01.

Lenth, R., Buerkner, P., Herve, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). Estimated marginal means, aka least-squares means. R package ‘emmeans’, version 1.7.2. Retrieved from https://cran.r-project.org/web/packages/emmeans/emmeans.pdf (15.03.2022).

Littell, R. C., Stroup, W. W., & Freund, R. J. (2010). *SAS for linear models.* Cary, NC: SAS Institute Inc.

McFarquhar, M. (2016). Testable hypotheses for unbalanced neuroimaging data. *Frontiers in Neuroscience*, *10*, 270. doi: 10.3389/fnins.2016.00270.

O’Brien, R. M. (2014). Estimable functions in age-period-cohort models: a unified approach. *Quality & Quantity*, *48*(1), 457–474. doi: 10.1007/s11135-012-9780-6.

Olivera-La Rosa, A., Chuquichambi, E. G., & Ingram, G. P. (2020). Keep your (social) distance: pathogen concerns and social perception in the time of COVID-19. *Personality and Individual Differences*, *166*, 110200. doi: 10.1016/j.paid.2020.110200.

Ordaz, J. A., del Carmen Melgar, M., & Khan, M. K. (2011). An analysis of Spanish accidents in automobile insurance: the use of the Probit model and the theoretical potential of other econometric tools. *Equilibrium. Equilibrium. Quarterly Journal of Economics and Economic Policy*, *6*(3), 117–134. doi: 10.12775/EQUIL2011.024.

Poline, J. B., Kherif, F., Pallier, C., & Penny, W. (2007). Contrasts and classical inference. In W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel & T. E. Nichols (Eds.) (2011). *Statistical parametric mapping: the analysis of functional brain images* (126–139). Elsevier.

Rafter, J. A., Abell, M. L., & Braselton, J. P. (2002). Multiple comparison methods for means. *Siam Review*, *44*(2), 259–278. doi: 10.1137/S0036144501357233.

Rivers, J. W., Newberry, G. N., Schwarz, C. J., & Ardia, D. R. (2017). Success despite the stress: violet-green swallows increase glucocorticoids and maintain reproductive output despite experimental increases in flight costs. *Functional Ecology*, *31*(1), 235–244. doi: 10.1111/1365-2435.12719.

Rahardja, D. (2020). Multiple comparison procedures for the differences of proportion parameters in over-reported multiple-sample binomial data. *Stats*, *3*(1), 56–67. doi: 10.3390/stats3010006.

Quigley, M. Y., Rivers, M. L., & Kravchenko, A. N. (2018). Patterns and sources of spatial heterogeneity in soil matrix from contrasting long term management practices. *Frontiers in Environmental Science*, *6*, 28. doi: 10.3390/stats3010006

SAS Institute Inc. (2017). *The four types of estimable functions*. In SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2018). *SAS/STAT® 15.1 User's Guide. The GLM Procedure*. Cary, NC: SAS Institute Inc.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *Journal of Memory and Language*, *110*, 104038. doi: 10.1016/j.jml.2019.104038.

Searle, S. R., & Gruber, M. H. J. (2017). *Linear models*. John Wiley & Sons.

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: an alternative to least squares means. *American Statistician*, *34*(4), 216–221. doi: 10.1080/00031305.1980.10483031.

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance Mathematics and Economics*, *64*, 417–428. doi: 10.1016/j.insmatheco.2015.07.006.

Singh, N., Wang, C., & Cooper, R. (2015). Role of vision and mechanoreception in bed bug, Cimex lectularius L. behavior. *PLoS one*, *10*(3), e0118855. doi: 10.1371/journal.pone.0118855.

Spilbergs, A., Fomins, A., Krastins, M. (2021). Impact of Covid-19 on the dynamics of MTPL insurance premiums and claims paid in Latvia. *WSEAS Transactions on Computer Research*, *9*, 33–42. doi: 10.37394/232018.2021.9.5

Spilbergs, A., Fomins, A., & Krastins, M. (2022). Road traffic accidents risk drivers' analysis – multivariate modelling based on Latvian motor third party liability insurance data. In D. Tipuric, A. Krajnovic & N. Recker (Eds.). *Economic and social development: book of proceedings* (pp. 246–264). Varazdin, Croatia: Varazdin Development and Entrepreneurship Agency.

Statgraphics Technologies Inc. (2017). *General linear models*. Statgraphics centurion 18.

Staudt, Y., & Wagner, J. (2021). Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks*, *9*(3), 53. doi: 10.339 0/risks9030053.

Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PloS one*, *15*(8), e0238000. doi: 10.1371/journal.pone.0238 000.

Suzuki, M., Taniguchi, T., Furihata, R., Yoshita, K., Arai, Y., Yoshiike, N., & Uchiyama, M. (2019). Seasonal changes in sleep duration and sleep problems: a prospective study in Japanese community residents. *PLoS One*, *14*(4), e0215345. doi: 10.1371/journal.pone.0215345.

Šoltés, E., Zelinová, S., & Bilíková, M. (2019). General linear model: an effective tool for analysis of claim severity in motor third party liability insurance. *Statistics in Transition New Series*, *20*(4), 13–31, doi: 10.21307/stattrans-2019-032.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics.* Boston, MA: Pearson.

Tattar, P. N., Ramaiah, S., & Manjunath, B. G. (2016). *A course in statistics with R*. John Wiley & Sons.

Thompson, P. A. (2006). *The "handy-dandy, quick-n-dirty" automated contrast generator-A SAS/IML R c macro to support the GLM, MIXED, and GENMOD procedures*. SUGI 31 *Statistics and data Analysis.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.176.736&rep=rep1&type=pdf (11.12.2021).

Ugarte, M. D., Militino, A. F., & Arnholt, A. T. (2008). *Probability and statistics with R*. CRC press.

Wang, B., Wu, P., Kwan, B., Tu, M. X., & Feng, Ch. (2018). Simpson's paradox: examples. *Shanghai Archives of Psychiatry*, *30*(2), 139. doi: 10.11919/j.issn.10 02-0829.218026.

Westfall, P. H., & Tobias, R. D. (2007). Multiple testing of general contrasts: Truncated closure and the extended Shaffer–Royen method. *Journal of the American Statistical Association*, *102*(478), 487–494. doi: 10.1198/0162 14506000001338.

Wicklin R. (2018). *Generalized inverses for matrices*. Retrieved from https://blogs.sas.com/content/iml/2018/11/21/generalized-inverses-for-matrices.html (23.02. 2022).

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. Elsevier.

Wooldridge, J. M. (2013). *Introductory econometrics: a modern approach.* Mason: South-Western.

Zahi, J. (2021). Non-life insurance ratemaking techniques. *International Journal of Accounting, Finance, Auditing, Management and Economics*, *2*(1), 344–361. doi: 10.5281/zenodo.4474479.

Zhao, J., Wang, C., Totton, S. C., Cullen, J. N., & O'Connor, A. M. (2019). Reporting and analysis of repeated measurements in preclinical animals experiments. *PloS one*, *14*(8), e0220879. doi: 10.1371/journal.pone.0220879.

Acknowledgments

This work was supported by the VEGA projects *Analysis of insurance risks in relation to the economy of the life insurance company* (No. 1/0410/22) and *Determination of the Capital Requirement to Cover Selected Catastrophic Risks in Life and Non-life Insurance* (No. 1/0166/20).

Annex

Table 1. Tests for differences between LS-means for the *Age_cat* factor with 6 categories (the matrix of *p*-values)

Least Squares Means for effect Age_cat
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: lnCS

i/j	-25	25-35	35-45	45-55	55-65	65^+
-25		<.0001	<.0001	<.0001	<.0001	<.0001
25-35	<.0001		**0.0613**	**0.1944**	0.0001	0.0001
35-45	<.0001	**0.0613**		**0.4303**	0.0049	0.0046
45-55	<.0001	**0.1944**	**0.4303**		0.0006	0.0008
55-65	<.0001	0.0001	0.0049	0.0006		**0.5747**
65^+	<.0001	0.0001	0.0046	0.0008	**0.5747**	

Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Table 2. The simultaneous hypothesis test $H_0 : \mu_3 = \mu_4$ and $H_0 : \mu_2 = \mu(\mu_3, \mu_4)$ for the *Age_cat* factor in GLM for $\ln CS$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Age 2=3=4	2	5.6764	2.8382	1.77	0.1707

Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Table 3. Verification of the statistical significance of the influence of regressors (including the *Age_cat* × *Brand_C* interaction) to the target variable $\ln CS$ in GLM

Source	DF	Type IV SS	Mean Square	F Value	Pr > F
Engine_C	1	7.7743	7.7743	5.71	0.0169
Weight_C	1	9.2094	9.2094	6.76	0.0093
Age_C	1	21.5204	21.5204	15.79	<.0001
Brand_C	3	14.1900	4.7300	3.47	0.0154
District_cat	3	167.1081	55.7027	40.88	<.0001
Age_cat *Brand_C	7	71.8179	10.2597	7.53	<.0001

Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Table 4. The basic analysis of the general linear model for $\ln CS$

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		4.89571	B	0.10632	46.05	<.0001
Engine_C		0.00178		0.00075	2.39	0.0169
Weight_C		-0.00010		0.00004	-2.60	0.0093
Age_C		0.01064		0.00268	3.97	<.0001
Brand_C	**A**	1.56505	B	0.44574	3.51	0.0004
Brand_C	**B**	0.48366	B	0.10042	4.82	<.0001
Brand_C	**C**	0.17546	B	0.05962	2.94	0.0033
Brand_C	**D**	0.00000	B	.	.	.
District_cat	**A**	0.61252	B	0.06010	10.19	<.0001
District_cat	**B**	0.38609	B	0.04894	7.89	<.0001
District_cat	**C**	0.24586	B	0.05179	4.75	<.0001
District_cat	**D**	0.00000	B	.	.	.
Age_cat*Brand_C	**A B**	-1.22319	B	1.17120	-1.04	0.2963
Age_cat*Brand_C	**A C**	1.06257	B	0.32452	3.27	0.0011
Age_cat*Brand_C	**A D**	1.23363	B	0.46112	2.68	0.0075
Age_cat*Brand_C	**B A**	-0.45525	B	0.51108	-0.89	0.3731
Age_cat*Brand_C	**B B**	0.03179	B	0.10441	0.30	0.7608
Age_cat*Brand_C	**B C**	0.17673	B	0.03866	4.57	<.0001
Age_cat*Brand_C	**B D**	0.23641	B	0.05850	4.04	<.0001
Age_cat*Brand_C	**C A**	0.00000	B	.	.	.
Age_cat*Brand_C	**C B**	0.00000	B	.	.	.
Age_cat*Brand_C	**C C**	0.00000	B	.	.	.
Age_cat*Brand_C	**C D**	0.00000	B	.	.	.

Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Table 5. Estimation of multipliers for insurance contracts broken down by the *Age_cat* and *Brand_C* factors

Age_cat	***Brand_C***			
	A	**B**	**C**	**D**
A (-25)	•	0.477	3.449	3.434
B (25-55)	3.034	1.674	1.422	1.267
C (55⁺)	4.783	1.622	1.192	1

Source: own processing based on data provided by an unnamed insurance company.

Table 6. The representation of individual categories of factors in a set of insurance contracts with a loss

Category	Factor *Age_cat*	Factor *District_cat*	Factor *Brand_C*
A	0.5%	12%	0.5%
B	71.5%	50%	8.0%
C	28.0%	28%	64.0%
D	–	10%	27.5%
Sum	**100%**	**100%**	**100%**

Source: own processing based on data provided by an unnamed insurance company.

Table 7. Tests for differences between LS-means for the *Age_cat* × *Brand_C* interaction (the matrix of p-values)

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

Dependent Variable: lnCS

i/j	**AB**	**AC**	**AD**	**BA**	**BB**	**BC**	**BD**	**CA**	**CB**	**CC**	**CD**
AB		**0.1027**	**0.1159**	**0.1223**	0.2832	0.3500	0.4035	**0.0651**	0.2963	0.4336	0.5270
AC	**0.1027**		**0.9937**	**0.7565**	0.0275	0.0062	0.0020	**0.5499**	0.0239	0.0011	0.0002
AD	**0.1159**	**0.9937**		**0.8143**	0.1206	0.0549	0.0300	**0.6032**	0.1082	0.0214	0.0075
BA	**0.1223**	**0.7565**	**0.8143**		0.0250	0.0036	0.0009	**0.3731**	0.0218	0.0004	<.0001
BB	0.2832	0.0275	0.1206	0.0250		0.0090	<.0001	0.0185	0.7608	<.0001	<.0001
BC	0.3500	0.0062	0.0549	0.0036	0.0090		0.0021	0.0062	0.1399	<.0001	<.0001
BD	0.4035	0.0020	0.0300	0.0009	<.0001	0.0021		0.0028	0.0076	0.1796	<.0001
CA	**0.0651**	**0.5499**	**0.6032**	0.3731	0.0185	0.0062	0.0028		0.0164	0.0017	0.0004
CB	0.2963	0.0239	0.1082	0.0218	0.7608	0.1399	0.0076	0.0164		0.0009	<.0001
CC	0.4336	0.0011	0.0214	0.0004	<.0001	<.0001	0.1796	0.0017	0.0009		0.0033
CD	0.5270	0.0002	0.0075	<.0001	<.0001	<.0001	<.0001	0.0004	<.0001	0.0033	

Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Table 8. Coefficients for the CONTRAST statement to test the null hypothesis $H_0 : \mu_{CA} = \mu(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA})$ for the *Age_cat* × *Brand_C* interaction

Age_cat	***Brand_C***				**Sum**
	A	**B**	**C**	**D**	
A	•	-0.25	-0.25	-0.25	**-0.75**
B	-0.25	0	0	0	**-0.25**
C	1	0	0	0	**1**
Sum	**0.75**	**-0.25**	**-0.25**	**-0.25**	**0**

Table 9. Coefficients for the ESTIMATE statement to estimate the weighted marginal mean $\mu(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA}, \mu_{CA})$ for the *Age_cat* × *Brand_C* interaction

Age_cat	***Brand_C***				**Sum**
	A	**B**	**C**	**D**	
A	•	2	31	12	**45**
B	40	0	0	0	**40**
C	15	0	0	0	**15**
Sum	**55**	**2**	**31**	**12**	**100**

Table 10. The estimate of the weighted marginal mean $\mu(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA}, \mu_{CA})$ for the *Age_cat* × *Brand_C* interaction

Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Mean_w(AB, AC, AD, BA, CA)	6.4025	0.19825	32.29	<.0001

Source: own processing in the SAS programming language based on data provided by an unnamed insurance company.

Table 11. General form of estimable function for GLM estimated in Table 4

General Form of Estimable Functions		
Effect		**Coefficients**
Intercept		L1
Engine_C		L2
Weight_C		L3
Age_C		L4
District_cat	**A**	L5
District_cat	**B**	L6
District_cat	**C**	L7
District_cat	**D**	**L1-L5-L6-L7**
Brand_C	**A**	L9
Brand_C	**B**	L10
Brand_C	**C**	L11
Brand_C	**D**	**L1-L9-L10-L11**

Table 11. Continued

General Form of Estimable Functions		
Effect		**Coefficients**
Age_cat*Brand_C	**A B**	L13
Age_cat*Brand_C	**A C**	L14
Age_cat*Brand_C	**A D**	L15
Age_cat*Brand_C	**B A**	L16
Age_cat*Brand_C	**B B**	L17
Age_cat*Brand_C	**B C**	L18
Age_cat*Brand_C	**B D**	L19
Age_cat*Brand_C	**C A**	**L9-L16**
Age_cat*Brand_C	**C B**	**L10-L13-L17**
Age_cat*Brand_C	**C C**	**L11-L14-L18**
Age_cat*Brand_C	**C D**	**L1-L9-L10-L11-L15-L19**

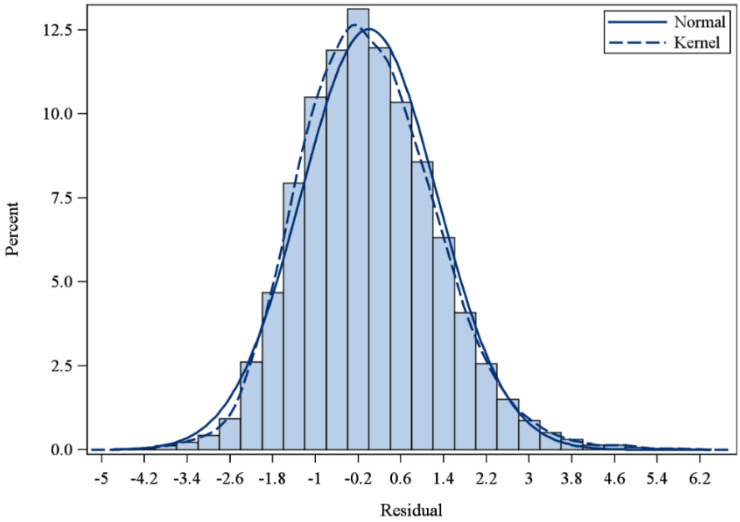
Source: own processing in the SAS programming language based on data provided by an unnamed insurance company.

Table 12. The point and 95% interval estimates of the adjusted means of claim severities (in Euros) for groups determined by the *Age_cat* × *Brand_C* interaction

Age_cat	*Brand_C*			
	A	B	C	D
A	•		603.4 (409.1-889.9)	
B	603.4 (409.1-889.9)	305.6 (244.9-381.3)	259.6 (217.3-310.0)	231.2 (193.9-275.7)
C		296.0 (230.2-380.8)	217.5 (181.2-261.2)	182.5 (150.9-220.8)

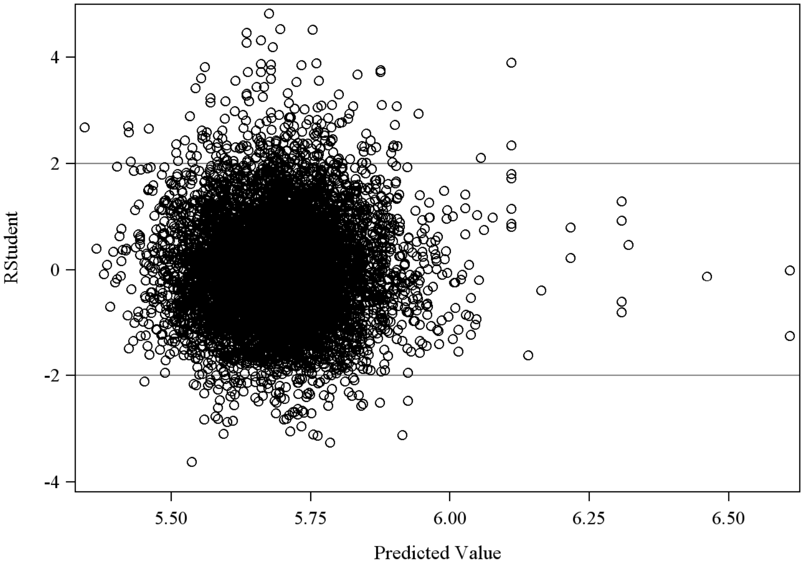
Source: own processing based on data provided by an unnamed insurance company.

Figure 1. Distribution of residuals for the model of logarithmic transformation of claim severity



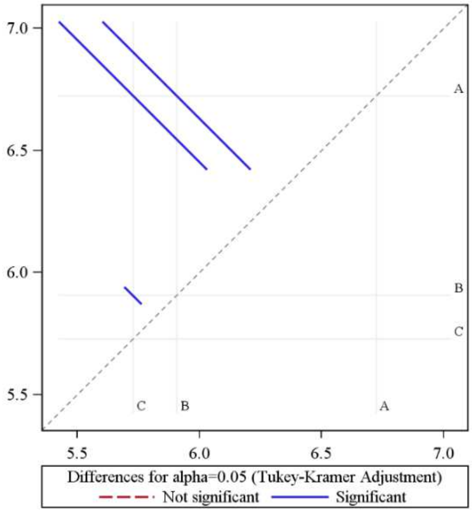
Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Figure 2. Residuals for the model of logarithmic transformation of claim severity



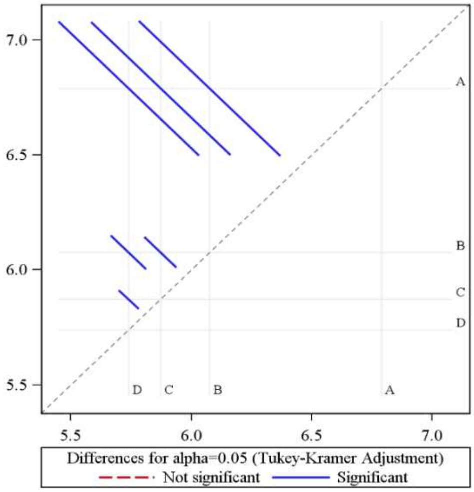
Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Figure 3. Diffogram for pairwise comparisons (and associated 95% Tukey-Kramer-adjusted confidence intervals) of LS-means of $\ln CS$ for the *Age_cat* factor



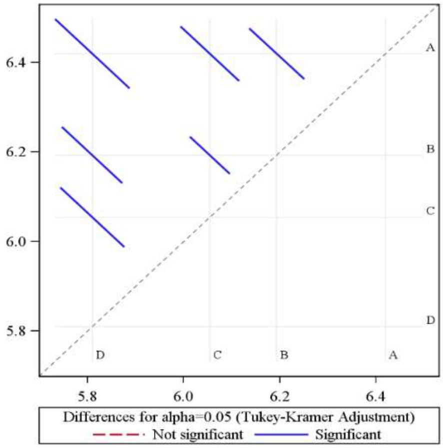
Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Figure 4. Diffogram for pairwise comparisons (and associated 95% Tukey-Kramer-adjusted confidence intervals) of LS-means of $\ln CS$ for the *Brand_C* factor



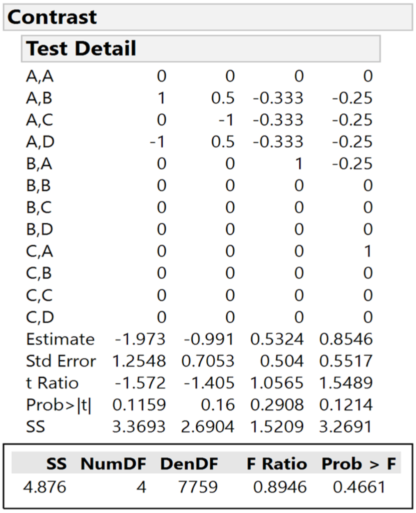
Source: own processing in the SAS EG based on data provided by an unnamed insurance company

Figure 5. Diffogram for pairwise comparisons (and associated 95% Tukey-Kramer-adjusted confidence intervals) of LS-means of $\ln CS$ for the *District_cat* factor



Source: own processing in the SAS EG based on data provided by an unnamed insurance company.

Figure 6. Verifying the null hypothesis $H_0 : \mu_{CA} = \mu(\mu_{AB}, \mu_{AC}, \mu_{AD}, \mu_{BA})$ for the $\mathit{Age\_cat} \times \mathit{Brand\_C}$ interaction



Contrast

Test Detail

A,A	0	0	0	0
A,B	1	0.5	-0.333	-0.25
A,C	0	-1	-0.333	-0.25
A,D	-1	0.5	-0.333	-0.25
B,A	0	0	1	-0.25
B,B	0	0	0	0
B,C	0	0	0	0
B,D	0	0	0	0
C,A	0	0	0	1
C,B	0	0	0	0
C,C	0	0	0	0
C,D	0	0	0	0
Estimate	-1.973	-0.991	0.5324	0.8546
Std Error	1.2548	0.7053	0.504	0.5517
t Ratio	-1.572	-1.405	1.0565	1.5489
Prob>|t|	0.1159	0.16	0.2908	0.1214
SS	3.3693	2.6904	1.5209	3.2691

SS	NumDF	DenDF	F Ratio	Prob > F
4.876	4	7759	0.8946	0.4661

Source: own processing in the SAS JMP based on data provided by an unnamed insurance company.