**Statistical Modeling Methods in Automobile Insurance**

Elena BURLACU PhD Student  
The Academy of Economic Studies of Moldova  
Republic of Moldova,  
elenaburlacu@gmail.com

**Abstract**  
This article describes the notions of the statistical indicators of automobile insurance and also presents methods that describe interdependence between them. Statistical calculations - the collection, the pre-processing and the further data analysis is based on certain system of statistical indicators. This system is composed of the indicators, which are supposed to be used in all firms in a given sphere of activity, in our case – these are the insurance companies which operating on the auto insurance market. The interdependence between insurance indicators was analyzed by correlation analysis, multiple regressions and testing the hypothesis of normality and model validation and its coefficients.

**Key words:** insurance compensations, the amount of insurance, correlation, linear regression.

**JEL Classification:** G22, C1

**Introduction**  
There is an obvious need of using the calculations and rating in auto insurance. Taking into account the market conditions for the insurance company, the most important criterion for attracting clients is certainly the rate policy, namely to ensure customer in analogous conditions to other companies, providing the same amount but with a lower price for the insured. Determining this value, the insurer collate the need of calculus and likelihood estimate of its business, with need to build mathematical models, to analyze them and to choose the optimal model that best describes certain phenomena on auto insurance market.

1. The statistical modeling methods  
The insurance amount of damages (Y) is, probably, the main indicator of the financial stability of the insurance company, therefore it represents interest and the research of dependent factors has an influence on the formation of this indicator.

1.1 The interdependence of statistical indicators  
Analyzing which is the influence that the size of insurance compensation has, on whether it occur, can lead to the following conclusions:

- The overview numbers of the insurance payments depends on the total number of contracts that form the whole package of insurance company.
- The total size of insurance compensations, for the same reasons, depends on the number of insurance requests to the entire insurance package.

- The type of contract. For example, statistics show that the number of requests into covering damages caused by means of transport is bigger than the number of shares on theft, but it is understood the fact, that the size of the payment for the caused damage, on average is less than the size payment for the theft.

- The operating region of means of transportation, driving experience of the driver, car brand, the number of insured events that occurred previously with the participation of one or another driver etc, may also become influential factors

- Reinsurance. The amount of reinsurance compensations can be treated as a factor that reduces the total compensations.

It has to be mentioned that the influence of these factors on the whole or the degree of their influence on the formation of random value Y, described above, is only at logical reasoning level. Therefore, it presents mathematical research interest of the influence of indicators on the value Y, and the research of interdependence of these indicators.

1.2 Practical approach

From the string of factors that may affect the amount of damages in an insurance company, it is proposed to study the interdependence of the three of them: the amount of insurance payments (X1), the number of insurance contracts (policies) (X2), the number of cases provided (X3). As a basis for calculating time series are going to be used, for 24 months, for each of the indicators selected from the reports out of an insurance company.

- the hypothesis of normality (criterion Pearson)

At the first phase was tested the hypothesis of normality, for each of the three rows data of indicators for which has been used Pearson criterion.

The result of calculating the theoretical frequency data were grouped into six intervals, $\chi^2_{\text{obs.}} = 0.18523$, $\chi^2_{\text{cr.}}$ (for degrees of variance $v=3$ and the importance level 0.05) = 7.815.

Since the value of Pearson's statistical analysis is less than the critical one, according to the criterion of Pearson, the normal hypothesis distributional law is accepted the confidence level 0.05. ($X^2$ calculation was based on Microsoft Excel 7.0 with built CHIDIST).

The assumption of normality for the three indicators was accepted at the confidence level of 0.05.

- the linear regression

The application of linear regression, as a tool for the analysis of the data presented, derives from a compromise between the veracity of assumptions and simplicity the interpretation method. To determine the relations between indicators we will use the correlation analysis and the correlation matrix:
Table 1. The results of correlation analysis

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>0.485578</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>0.500817</td>
<td>0.334214</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>0.65210</td>
<td>0.521037</td>
<td>0.857567</td>
<td>1</td>
</tr>
</tbody>
</table>

The correlation matrix talks about a strong connection between the independent variables and the dependent variable Y and also signals that the independent variables do not correlate with each other.

The correlation analysis shows that the time series have a linear tendency in order to exclude the autocorrelation between the variables, in the regression model it is introduced an additional factor time t. The final multiple regression pattern will have the following form:

$$\hat{Y} = -9244.5 + 0.180035 \times X1 + 751.6681 \times X3 + 818.457 \times t$$  \hspace{1cm} (1.1)

- **Fisher test**
In order to test the significance of the model, the Fisher test is used.
For the given model the tabular value $F_{crit} = 3.34$ at level of significance $\alpha = 0.05$, $\gamma_1 = 3$, $\gamma_2 = 14$, a value which is much lower than $F_{obs} = 32.85537$. Thus the model describes the real structure of influence of the analyzed factors on the amount of compensations.

- **Student test**
It also showed an interest, checking the significance of regression coefficients. The significance of regression coefficients was checked by Student criterion (t-distribution) in a comparison of observation $t_{obs}$ and $t_{cr}$. $t_1 = 2.444$, $t_2 = 2.6182$, $t_4 = 2.6045$ values that are bigger than the tabular value $t_{cr} = 2.145$ the number of thresholds of freedom $d_f = 14$ and confidence level $\alpha = 0.05$.
As observed, only two of the coefficients for the variables X1 (amount of insurance) and X3 (number of claims) can be considered significant from the statistically point of view.

- **Durbin Watson test**
To assess the existence of autocorrelation in regression balances, which creates difficulties into applying the method of the smallest squares regression equation, we used the criterion of Durbin - Watson. Statistics criticism of Durbin - Watson has the form:

$$d = \frac{\sum_{i=2}^{n} (\hat{\epsilon}_i - \hat{\epsilon}_{i-1})}{\sum_{i=1}^{n} \hat{\epsilon}_i^2}$$  \hspace{1cm} (1.2)

where:
\[ \hat{e}_i = Y_i - (m_0 + m_1 \times X_i + \ldots + mK \times X_k) \] - the discrepancy method of the smallest squares.

Frankly speaking, in the absence of autocorrelation balances regression, the significance of critical statistics \( d \), according to the criterion of Durbin - Watson, must not divert more than 2. The significance statistics \( d = 1.964 \) obtained, allows us to make a conclusion regarding the lack of autocorrelation in the examined sequences.

Therefore the regression coefficient for the first factor shows that the increase of the amount of insurance payments by 1000 u.c., the compensation amount will increase on average by 180 u.c. The coefficient \( X_3 \) shows that the increasing number of cases provided by one unit amount of compensation will increase on average by 751 u.c. may represent the average value of compensation for insured event. Regression coefficient of the time factor shows that each month amount of compensation payments will increase by 818 u.c.

Accordingly, to improve the financial stability of the insurance company it is necessary when forming insurance payment, to forecast for each client, the probability of occurrence of accident.

### 2. Models with qualitative factors

This can be done through analysis of qualitative factors. Upon a possible risk of an accident there are a number of qualitative factors that might influence: sex, age of driver, driving experience, type of car (the technical capabilities of the car) and operating region. These features can be described by the vector:

\[ X = (x_1, x_2, \ldots, x_k)^T \] (2.1)

If \( Y \) is a quantitative variable, then
\[ y_i = 1 \quad \text{if the individual is involved in a car accident} \]
\[ y_i = 0 \quad \text{otherwise} \]

Then the vector:
\[ Y = (y_1, y_2, \ldots, y_n)^T \]

will contain dichotomous variables. The analysis of dependence of the characteristics of \((X_1, X_2, \ldots, X_k)\) will be described by the model:

\[ y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \] (2.2)

Where \( i \) - number of observations, \( \beta = (\beta_1, \beta_2, \ldots, \beta_k)^T \) - a set of unknown parameters, \( \epsilon_i \) - random error. As \( y_i \) takes the value of 0 or 1 and \( E(\epsilon_i) = 0 \), then \( E(y_i) = 1 \cdot P(y_i = 1) + 0 \cdot P(y_i = 0) = P(y_i = 1) = x_i^T \beta \). Accordingly the model (2.2) can be expressed by a linear probability model:
The purpose of this complex analysis consists to form some groups from the data taken from a survey on drivers, and then to using the regression to detect the interdependence between the probability of occurrence of accidents according to age, gender, driver experience, technical car specifications.

\[ P(y_i = 1) = x_i^T \beta \]  \hspace{1cm} (2.3)

**Conclusion**

In conclusion we can iterate that any statistical calculations should be based on the system of statistical indicators and use statistical and mathematical computing techniques. To verify the degree of influence of various factors on the value of insurance claims there were used correlational and regression analysis: there were constructed matrices of correlation coefficients and regression equations and factor analysis was applied. It was examined the significance of regression equation obtained and there were tested its coefficients.

The researches allow to make a conclusion about the fact that, the amount of compensation is an amount dependent on the following factors: the amount of insurance and the number of insured cases.

Accordingly, to improve the financial stability of the insurance company it is necessary when forming insurance payment, to forecast for each client, the probability of occurrence of accident, for this purpose the analysis of influence of quantitative and qualitative factors on the occurrence of this risk, is important.

**Bibliography**

Legea nr.414-XVI din 22.12.2006 „Cu privire la asigurarea obligatorie de răspundere civilă pentru pagube produse de autovehicule” publicată în MO nr.32-35/112 din 09.03.2007


Iacob A. I., Tanasoiu O. „Modele econometrice” vol.I, Editura ASE, Bucureşti, 2005

Turturean C. I. “ Metode statistice de analiză a seriilor de timp” Editura Sedcom, Iaşi, 2006

Лемер. Ж., Автомобильное страхование. Актуарные модели,. М.,Янус-Ко, 1998