Applying Regression Models to Predict Business Results

Jelena Rusov
PhD Student
Dunav osiguranje a.d.o, Belgrade

Mirjana Misita
Assistant Professor
University of Belgrade
Faculty of Mechanical Engineering

Dragan D. Milanovic
Full Professor
University of Belgrade
Faculty of Mechanical Engineering

Dragan Lj. Milanovic
Assistant Professor
University of Belgrade
Faculty of Mechanical Engineering

In terms of modern business practice, business prediction results are crucially important for evaluation of future financial performance of a company. Planning and prediction procedures are especially important for companies operating under uncertainty. This paper shows an example of planning and prediction of business results in insurance when calculating premium trend by use of linear and nonlinear regression. Due to the uncertainty associated with the moment of claim occurrence and claim amount, it is necessary to secure enough assets to cover the risks. Asset-liability matching requires the prediction of future premium movement per insurance lines which represents the basic concept of development and operation of insurance companies.

Keywords: *linear regression, nonlinear models, prediction, number of policies, premium.*

1. INTRODUCTION

Definition of business objectives through relation between the value of income, expense and gain is one of the basic goals when predicting the cost-effectiveness of business operations in a specific time frame. Analysis of business predictions aims at determining potentially critical periods that call for additional funds in order to provide the continuing operation of a company.

This paper deals with predictions and profit management on the example of an insurance company. It predicts financial results of an insurance company while describing the possible movements of premium amount, when applying linear regression and nonlinear models. Monitoring and predicting of policy sales flow (premium flow) is very important in order to provide enough funds to cover the risk and expenses and to generate profit. Predicting the financial results provides the basic concept of development and characteristics of insurance company's operations. It stabilizes the company operations, provides growth, development and improvement of insurance market and provides full protection of insureds' interests.

2. REGRESION MODELS

Regression analysis of a phenomenon aims at defining the regularity of a phenomenon development. Based on that regularity, regression analysis enables prediction of the future progress of a phenomenon. Regression model implies the class of stochastic models represented by an equation where a dependent variable is expressed as a linear or nonlinear function of independent variables. Much research has shown that the complexity of method does not have to be in correlation with data accuracy [1-5,8,13]. Complex methods might get too close to less significant data, where as simple methods include only the most important and basic trends in data, thus predicting the future more accurately.

It is possible to determine the dependence between variables by their graphic illustration, i.e. by drawing the scatter diagram based on the empirical data. Scatter diagram represents the set of values of two observed numerical variables, i.e. it shows the movement of function when its argument takes all the values from the domain of definition. Based on the appearance of initial values in scatter diagram, the character and intensity of interdependence of variables may be determined [6,7,12].

The dependence of experimental points cannot always be presented with linear model. As a result, this paper uses different forms of functions to show the development of regression model. Linear model and exponential, polynomial and logarithmic nonlinear models are used to predict the number of sold policies as well as the collected premium amount. Applied curvilinear models can be converted into a model of linear regression by transformation process. Linear regression is the simplest regression model and it turned out to be very significant in modelling of a wide variety of phenomena, primarily in marketing and economic research. In addition, it is characterized by a prompt projection and low complexity level.

For linear model, sample regression curve has the form $y = ax_i + b$, provided that for each pair of sample data (x_i, y_i) $i = 1,...,n$, implies that $y_i = ax_i + b$. The goal of linear regression analysis is to evaluate coefficients aandb, that describe linear dependence in the whole population, based on the initial data (x_i, y_i) $i = 1,...,n$. Parameters of linear trend, a and b, are the variables calculated for each specific case and in this paper, they are evaluated by the least squares method:

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\ ;\ b = \overline{y} - a\overline{x}\ ;\ \overline{x} = \sum_{i=1}^{n} x_i\ ;\ \overline{y} = \sum_{i=1}^{n} y_i \quad (1)$$

Received: November 2014, Accepted: June 2015
Correspondence to: Jelena Rusov
Dunav osiguranje a.d.o,
11000 Belgrade, Serbia
E-mail: jrusov@gmail.com
doi:10.5937/fmet1701198R

Parameter b represents the initial trend value, while parameter a, being the slope of a line, represents the constant variable of incline or decline of trend from one period to the other. In the mentioned example, those parameters are calculated separately for each line of insurance, number of sold policies and amount of collected premium.

Applied exponential model is $y = c \cdot e^{d \cdot x}$. After app–lying the logarithm, the linear model $\ln y = d \cdot x + \ln c$ is obtained. Analysis of the transformed model is the same as the analysis of the linear model. Though, it should be pointed out that when interpreting the results, it is important to pay attention to what variables are the transformed ones. In the mentioned model, values of parameters that are estimated with empirical values, are obtained in logarithmic values, since the transformation of dependent variable was performed. After applying antilogarithm, the value of variable of initial model is obtained.

The general form of logarithmic regression used for prediction is $y = p + q \ln x$ which is by transformation of independent variable ($t = \ln x$)) brought down to linear form. Analysis of the transformed part is the same as the analysis of linear model. Nonetheless, special attention should be paid to transformation of independent variable.

When choosing the curve that best fits the points in diagram, the starting point may be the polynomial regression model. General form of polynomial regression is $y = b_0 + b_1x + ... + b_nx^n$. Polynomial coefficients are the parameters of regression model that should be evaluated. Evaluation is performed with n empirical pairs (x_i, y_i). Prediction of parameters bi, i = 1,...,n is determined by least squares method, just like in case of simple linear regression model, having in mind that number of equation should equal to the number of unknown parameters. Second order polynomial regression is used in this paper.

Coefficient of determination R^2, being the relative measure of regression line adjustment to empirical data, indicates the representative quality of a model. It is the relation between the sum of squares deviation explained by regression model and sum of squares of total deviations. Coefficient of determination value in any regression is $0 \le R^2 \le 1$. It is better if the coefficient is closer to one, which means that the value of residual sum is lower, i.e. the scatter of value around the regression is low. Theoretically, the limit of model's representative quality is set at 0,9. In practice, it is very hard to find the model that properly describes that phenomenon, therefore, the limit is reduced to 0,6. Low value of coefficient of determination does not always mean that graded regression is incorrect. Nonetheless, in practical economy it is always better to consider the significance and the meaning of determination coefficient and to choose the model accordingly.

3. EXPERIMENTAL RESEARCH

Premium is the biggest source of unemployed financial assets of insurance companies. It makes a significant source of investments and ensures financial stability of a country [9,10,11]. Therefore, this paper explores the possibility to predict the number of effected policies and premium amount of the insurance lines with the biggest share in one of the insurance companies operating in the Republic of Serbia (01 – Accident Insurance, 02 – Voluntary Health Insurance, 03 – Motor Insurance, 07 – Goods in Transit Insurance, 08 – Property Insurance against Fire and Allied perils, 09 – Other Property Insurance Lines, 10 – Motor Vehicle Liability Insurance, 13 – General Liability Insurance, 14 – Credit Insurance, 18 – Travel Assistance Insurance, 20 –Life Insurance and 22 – Supplementary Insurance along with Life Insurance (according to the codes of the National Bank of Serbia)).

In order to determine the future flow of collected premium, it is necessary to graphically show the movement of the number of effected insurances and related premiums in the past. Scatter diagrams provide clearer picture of phenomenon movement and indicate the possible type of mathematical model that would describe the movement of observed phenomenon value through the timeframe. In total, 24 scatter diagrams for 12 insurance lines are created. Two graphs correspond to each insurance line. One graph shows the functions of experimental data on number of effected policies per policy year, while the other shows the premium amount per effected insurance policies.

The entire research will be shown through the example of only one insurance line – General Liability Insurance (insurance line IL 13). The Figure1a) shows the scatter diagram of dependency of number of effected policies per policy year, while Figure1b) shows the diagram of dependency of number of effected policies (abscissa) and the amount of collected premium in thousands of Dinars (ordinate) of the same policies issued in the period from 2008 to 2013 for IL 13.

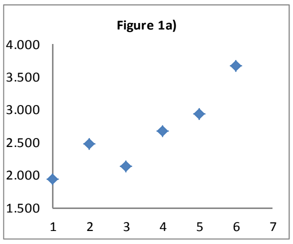

Fig. 1a)

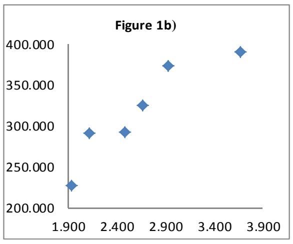

Fig. 1b)

Figure 1a) Scatter diagram for IL 13; Number of effected policies per policy year

Figure 1b) Scatter diagram forIL 13; Number of effected policies and premium collected accordingly

Scatter diagrams created on the basis of initial data on number and amount of effected policies are straight line and curvilinear. For that reason, four regressions have been used (linear, exponential, logarithmic and polynomial) to construct a model with independent variable (years when the policies were effected) and dependent variable (number of effected policies). The aim of this model is to predict future events, i.e. the direction of policy sales for the following year. The second model that was developed is used to predict the total premium amount based on the estimated number of effected policies. Models are created for the above mentioned insurance lines according to the data on total number of effected policies and total premium in the period from 2008 to 2013 in a company that deals with life and nonlife insurance in the territory of the Republic of Serbia.

Application of linear regression is emphasized in this paper since the remaining nonlinear models used in the paper might be simply brought down to linear form. Predicting the total premium amount per insurance lines for the following year requires that the number of effected policies should be estimated based on the historical data. Linear trend applied to predict the number of effected insurance policies is expressed by function $y_i = ax_i + b$, $i = 1,...,n$ where independent variable x_i, represents time, whereas dependent variable y_i, represents the number of effected insurance policies for each time unit and nvariable of the basic set ($n = 6$). The unit for x_i is one year, whereas for y_i, the unit is one sold policy. Total premium per insurance lines is predicted as a function, i.e. regression line, based on the number of policies calculated in this manner.

4. COMPARISON MODEL

The comparison between the number and the amount of policies obtained by linear, logarithmic, exponential and polynomial regression for 12 insurance lines is performed in experimental research. In total, 48 functions were generated to predict the number of policies and the same number of functions was created to predict the amount of collected premium.

It is determined that linear function best describes the movement of empirical data for prediction of number of policies in case of 3 insurance lines, whereas polynomial approximation was the most efficient in case of 5 out of 12 insurance lines that were considered. It should be mentioned that determination coefficients of linear and polynomial regressions have approximately the same values in case of 3 out of 5 insurance lines. Therefore, it might be concluded that for 50% of total number of insurance lines considered in the research, linear trend turned out to be the best option for predicting the number of policies soldin the following year. Regarding 4 insurance lines, this experiment showed that linear regression and nonlinear models were unreliable for predicting the number of policies sold (maximum determination coefficient was $R^2 \leq 0{,}31$). Therefore, in case of 67% of examined insurance lines, at least one of the regression models might be valid to predict the number of policies. If there are many regression models whose trends of movement are good at adjusting to the schedule and movement of initial data (which is the case for 3 insurance lines), the prediction model should be chosen carefully. While applying various regression types to predict number and amount of policies, it should be emphasized that the quality of human prediction might be crucial in method selection, in view of the fact that sometimes, quantitative methods do not provide better estimates than people. The assessment of a manager is significant, provided that it is not the single but one of the prediction methods.

Further research deals with central tendency of movement and development of premium amount based on the predicted number of policies. Analysis showed that premium trend movement corresponded best to linear regression in case of 3, whereas, in case of 4 insurance lines, polynomial regression was the best choice. In terms of the remaining insurance lines that were considered in this paper, the research showed that one insurance line had the greatest determination coefficient when applying exponential regression, whereas in case of 5 insurance lines non of the regression models was found to be suitable. It may be concluded that for 58% of insurance lines that were included in the research, one of the regressions might be used to predict the premium amount based on the previously predicted number of policies. In case of one out of five insurance lines where none of the regression models could be applyed topredict the insurance premium, experimental research showed that three out of four applied models were acceptable for prediction of the number of effected policies. This leads to the conclusion that for particular insurance lines(4 in terms of number of policies, 5 in terms of premium amount) the applied models are unreliable. Having in mind that applied regressions do not correspond to the movement of empirical data, future research may be directed toward the formation of analytical functions. Their movement trend is adjusted to the schedule and movement of original data in the best possible manner and that can be used to show the future sales development for those insurance lines separately. The comparison betweenthe results obtained by application of linear regression and nonlinear functions, showed the deviation of 50%. This deviation indicates that the use of linear regression concept is valid when predicting business results in particular insurance lines.

This paper compares the methods through the example of General Liability Insurance. Figure2 shows the number of sold policies predicted by the aforementioned regressions. The highest determination coefficient of predicted number of effected policies is obtained in polynomial regression ($R^2 = 0.90$), while the lowest is obtained through logarithmic regression ($R^2 = 0.69$ which is in practice considered to be the representative model). Linear regression model shows that 83% of change in dependent variable is explained by the change in independent variable.Based on the data projected in this mannerand used to predict the number of sold policies, it is possible to predict the amount of collected premium, Figure3. In case of number of policies, as well as in case of premium amount, the best approximation is obtained by polynomial function of second degree ($R^2 = 0.92$), whereas the worst (where determination coefficient is 0,81) is obtained by logarithmic function. Relative measure of linear model, where the evaluated regression is adjusted to the value of samples, amounts to 0,86. When the deviations of many different trend functions based on the empirical data (in the stated example, the difference in coefficient determination for polynomial and linear regression for number of policies soldis 0,07, while for the premium amount it is 0,06) are small, the choice of method is achieved by experience and assessment of the human factor. In the Figures 2 and 3, empirical data are represented by rhomb, whereas predictions (number of policies and premium amount) are represented by circle.

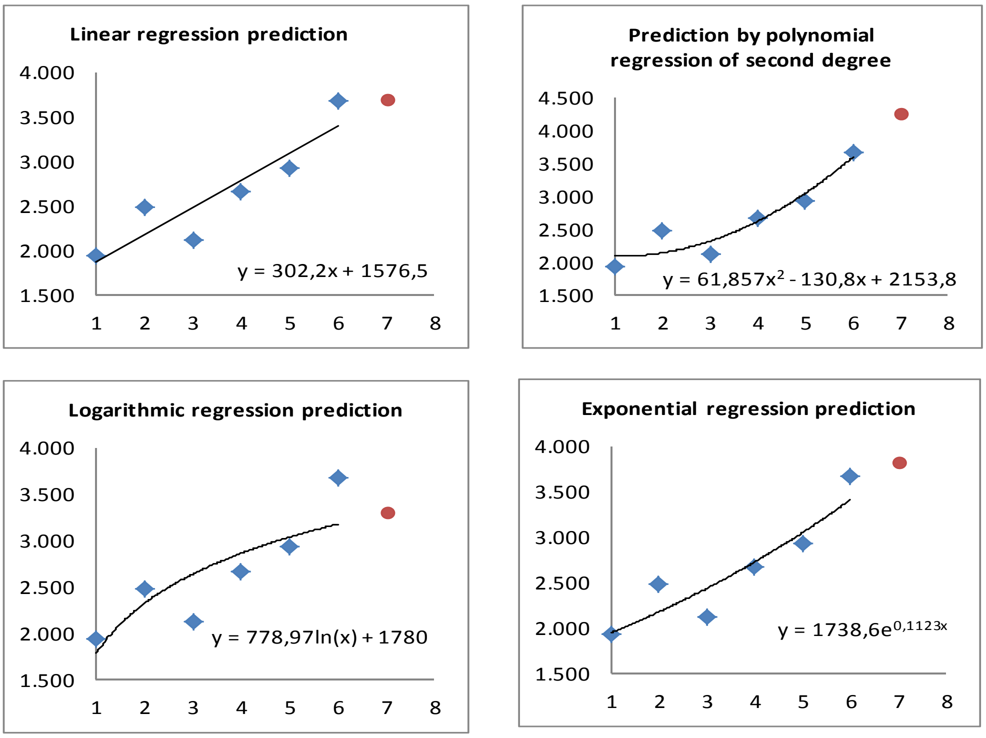

Figure 2. Graphically represented prediction of number of insurance policies effected for IL 13

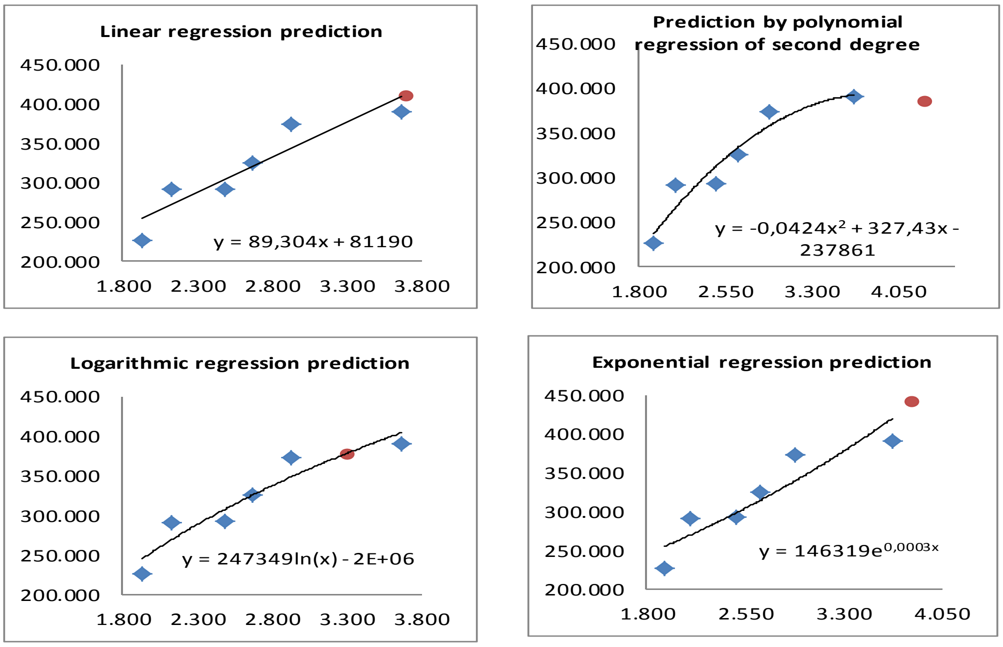

Figure 3. Graphically represented prediction of total premium amount based on the predicted number of sold policies for the following year

5. CONSLUSION

This paper uses the specific example to demonstratethe comparison between the results obtained from the use of central tendency of movement and development in particular insurance lines while applying linear regression and nonlinear dependencies. Conducted experimental research has shown that application of linear regression is valid in certain insurance lines. It is significantly simpler to apply linear regression model when predicting than nonlinear methods. Despite the fact that nonlinear functions of premium trend in insurance more realistically describe the real movement when compared to the linear, the degree of deviation in prediction results as well as the determination coefficient value indicate that approximation by use of linear functions is possible for some insurance lines.

Insurance companies act as the institutional investors in financial system of a country. Risk dispersion is an important segment of their business. Accordingly, this paper analyses prediction mechanisms and demonstrates how its application in insurance companies' operations may decrease the risk of illiquidity.

REFERENCES

[1] Christofides, S., Regression Models Based on Log-incremental Payments, *Claims Reserving Manual*, 1990Vol.2, Institute of Actuaries, London.

[2] Milanovic, LJ. D, Milanovic, D. D., Misita, M., The Evaluation the Risky Investment Porject, *FME Transactions,* 2010, 38(2), pp. 103-106

[3] Gediminaitė I.: On the prediction error in several claims reserves estimation methods, *Master thesis*, 2009, Royal Institute of Technology, School of Engineering Sciences, Stockholm.

[4] Gestel, T.V., Martens, D., Baesens, B., Feremans, D., Huysmans, J. and Vanthienen, J.: Forecasting and analyzing insurance companies' ratings, *International Journal of Forecasting*, 2007, Vol. 23, Issue 3, pp. 513-529.

[5] Mack, T., Venter, G., A Comparison of Stochastic Models that Reproduce Chain Ladder Reserve Estimates (2000), *Insurance: Mathematics and Economics,* vol. 26, 101-107.

[6] Maddala, G.S. and Lahiri, K.: Introduction to Econometrics, 4th Edition, John Wiley and Sons, New York, ISBN : 978-0-470-01512-4 , 2010.

[7] McCullagh, P.: Regression models for ordinal data, *Journal of the Royal Statistical Society, Series B*, Methodological, 1980, Vol. 42 (2) , pp. 109–142.

[8] Pankratz, A., *Forecasting with dynamic regression models*, John Wiley and Sons, New York, ISBN 0-471-61528-5, 1991.

[9] Partachi, I., et al. Statistical Methods Of Estimating Loss *Reserves In General Insurance, Scientific Annals of the "Alexandru Ioan Cuza",* University of Ias;2010 Supplement, p. 357.

[10] Potock, R., and Stehlík, M.: Nonlinear Regression Models with Applications in Insurance, *The Open Statistics & Probability Journal*, 2010, Vol. 2, pp. 9-14.

[11] Sánchez, J.A.: Calculating insurance claim reserves with fuzzy regression, *Fuzzy Sets and Systems*, 2006, Vol. 157, Issue 23, pp. 3091–3108.

[12] Wang G. and Chaman, J., *Regression Analysis, Modeling & Forecasting*, Graceway publishing Company, USA, ISBN 0-932126-50-2, 2003.

[13] Zhang, Y. and Dukic, V.: Predicting Multivariate Insurance Loss Payments under the Bayesian Copula Framework, *Journal of Risk and Insurance*, 2013, Vol. 80, No. 4.

ПРЕДВИЂАЊЕ ПОСЛОВНИХ РЕЗУЛТАТА ПРИМЕНОМ РЕГРЕСИОНИХ МОДЕЛА

Ј. Русов, М. Мисита, Д. Д. Милановић, Д. Љ. Милановић

За савремену пословну праксу резултати предвиђања пословања су од суштинског значаја за евалуацију будуће финансијске ефикасности предузећа. Поступак планирања и предвиђања нарочито је значајан за предузећа која послују у условима неизвесности.

У раду је изложен пример планирања и предвиђања пословних резултата у осигурању приликом прорачуна тренда премије линеарном и нелинеарном регресијом. Због неизвесности која прати тренутак настанка и износа штете неопходно је осигурати довољно средстава за покриће ризика. За усклађивање средстава и обавеза потребно је предвидети будуће кретање премије по врстама осигурања, што чини основни концепт развоја и пословања осигуравајућих друштава.