{MERAA} Mathematics in Education, Research and Applications

ISSN 2453-6881

Math Educ Res Appl, 2017(3), 2

Received: 2017-10-23
Accepted: 2017-12-05
Online published: 2017-12-31
DOI: http://dx.doi.org/10.15414/meraa.2017.03.02.49-56

Original Paper

Application of the logistic regression analysis to assess credibility of the farm

Jozef Palkovič[1*], Martina Šoporová[1]

[1] Slovak University of Agriculture in Nitra, Faculty of Economics and Management, Department of Statistics and Operations Research, Slovak Republic

ABSTRACT

Logistic regression is useful tool of statistical analysis used in various field of research, especially to classify units according their parameters, or to estimate chance of event occurrence. On the economic field this method is usually used to estimate bankruptcy and credit models, or to predict consumers' behavior. Objective of the proposed paper is to present application of the logistic regression analysis to assess credit of the farm. This paper can be used also as guide through the process of modelling, model verification and interpretation of its results. Data used to estimate logistic regression were individual farm data cover large farms from the database of the Ministry of Agriculture and Rural Development in Slovakia for the period 2009 to 2013. 4000 observations were used to estimate final model, and 427 observations were used as the sample for the model verification. Then, logistic regression model was estimated and verified. From the initial set of 13 variables were selected 7 significant variables to final model. Factor which increased probability of getting loan the most significantly was proportion of loans, on the other hand, factor which decreased this probability the most was the proportion of crop production. Quality and prediction ability of the final model according to standard indicators was fair, however there could be suggested including additional variables to improve model prediction ability, and its further testing by its application on more testing samples. Paper offers better insight into process of logistic regression application, and suggests ways of current topic further developing.

KEYWORDS: logistic regression, credit model, chance model, logit

JEL CLASSIFICATION: C01, C10, C18, B23, B26

INTRODUCTION

In the 19th century Sir Francis Galton introduced to the world very popular statistical method called regression. Method is preferred in various areas of investigation, but in the first, it was used in genetics. Galton estimated regression model for the prediction of height of the child based on the data of parents. He found, that the difference between high of child and average

[*] Corresponding author: Jozef Palkovič, Ph.D., Slovak University of Agriculture in Nitra, Faculty of Economics and Management, Department of Statistics and Operations Research, Trieda Andreja Hlinku 2, 949 76 Nitra, Slovakia, e-mail: jozef.palkovic@uniag.sk

high in the population is proportional to his parent's deviation from typical people in the population.

Classical linear econometric model was not appropriate in case, when dependent variable had binary or categorical character. The main reason was, that probability does not have linear nature, and that predicted values should fall in the interval between 0 and 1. These assumptions were not met in case of linear regression model. The need for a new method, which will satisfy these two conditions led to development of the logistic regression model.

This method has been already applied in various fields of research. For example, it has been applied in healthcare research, social, geographic, ecological, physical research and in the field of economics. Presented paper is focused on application of logistic regression in the field of economics and finance, especially in assessing the credibility or bankruptcy with the use of logistic regression models. Paper shows application of this method to assess farm solvency, resp. to assess a chance, that farm will get bank loan. This method has already been used in this area of research which is described in the examples below.

In France, logistics regression was used for prediction of individual bankruptcy of enterprises. Because of lack of traditional prognostic models, Jabeur [1] developed a model that includes financial ratios as the explanatory variables, and deals with correlation and use penalizing weights for the wrong data in the matrix. This model was applied to predict business failure, which was appreciated not only by bankers but also by investors in France. Suggested model allows them to predict bankruptcy in advance and helps them to avoid bad investment.

When analyzing the financial statements of corporate entities, Nikolic et.al [2] used a model of logistic regression in the prediction of the credit score. Researchers proposed corporation credit scoring model, thanks to which they can predict probability of bankruptcy in 1 year period advance. They used the test sample to verify prediction ability of estimated models. Logistic regression was evaluated as the best predictive credit scoring model from the set of suggested solutions. Their model includes eight explanatory variables that showed the best predictive performance. Analysis was conducted in the Serbian region, so the final model could be implemented in a bank that operates in the same area, or in the region of South Eastern Europe. In other regions, it is possible to build an analogic model based on a similar technique.

Serener [3] analyzed the use of internet banking. They suggested model to estimate the probability of using Internet banking by customers. Factors considered as the explanatory variables were age, gender, income, marital status, education, occupation, experience with online shopping. Results of their analysis suggests, that clients aged 56 - 65 are less likely to use these services than respondents aged 18 - 25. Persons in marriage are less likely to use Internet banking than single respondents. With the increase in the individual's income, the likelihood of using internet banking also increased. Similarly, university graduates have a higher chance to use internet banking than lower educated people. Respondents who already had experience with online shopping showed tendency to use the internet banking. Among the professions considered in research, internet banking is most likely used by bank staff. Further application of the logistic regression can be expanded to marketing support. Appropriate sales support could be focused on the specific group of people suggested by logistic regression results.

There could be mentioned more examples of logistic regression applications. For its valuable properties and availability of software solutions it may be applied in many field of research. Presented paper is focused on the application of logistics regression in the field of finance to assess farm credibility and to determine factors which can influence it. Proposed paper describes the whole procedure of model specification, estimation, verification and interpretation of the results. Therefore, presented paper can be also used as the application instructions to logistic regression.

MATERIAL AND METHODS

Data used to present the logistic regression model were farm data over the period 2009 to 2013 divided in two groups based on the criterion, whether the farm did receive a bank loan or not. Individual farm data cover large farms from the database of the Ministry of Agriculture and Rural Development in Slovakia (Information letters of farms with double entry accounting). Dataset was divided into two parts. First part was used to estimate logistic regression model and included 4000 observations. Second part of dataset was used for verification of model prediction ability and included 427 observations. Model was estimated and verified using R Cran software package.

Model

If the Y is a binary response variable equal to 1 if attribute is present and 0 if it is not present in observation. If x = (x_1, x_2, $x_{3,}$ …, x_k) is a set of explanatory variables which can be discrete, continuous or a combination. First, 13 variables were considered as exogenous factors in the model. After backward elimination and model selection process were left following 7 variables in the final model:
Debt - firm debt, ploan - proportion of loans, Ebitda, size - size of firm, revpha - revenue per ha, own - number of owners and pprv - proportion of crop production.

Logistic regression model presents conditional probabilities (log odds) through a linear function of the predictors expressed as:

$$\ln\left(\frac{P(y_i = 1)}{P(y_i = 0)}\right) = \beta_0 + x_i^T\beta = I_i \tag{1}$$

Where $\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$ is the estimated vector of k predictor coefficients. Vector of parameters β is estimated using maximum likelihood method. Following likelihood function is maximized:

$$\ln[L(\beta)] = \sum_{i=1}^{n}\left\{y_i \ln\left[\frac{exp(I_i)}{1 + exp(I_i)}\right] + (1 - y_i)\ln\left[\frac{1}{1 + \exp(I_i)}\right]\right\} = \sum_{i=1}^{n}(y_i I_i - \ln[1 + \exp(I_i)]) \tag{2}$$

Then predicted probability can be expressed as follows:

$$F_i(I_i) = P(y_i = 1) = \frac{\exp(I_i)}{1 + \exp(I_i)} \tag{3}$$

In case of logistic regression is no more necessary to hold the assumptions of classical linear econometric model based on ordinary least square. Linear relationship between dependent and independent variables, explained variables and error term does not need to be normally distributed. Logistic regression also does not need variances to be homoscedastic and can handle also nominal or ordinal data as explanatory variables.

Model evaluation and diagnostics

Likelihood ratio test

This method compares the likelihood of the data under the full model against the likelihood of the data under a model with fewer predictors.

Let L1 is the maximum value of the likelihood without the additional assumption (unrestricted model) and L2 the maximum value of the likelihood when the parameters are restricted (reduced model). Calculate the ratio:

$$\lambda = \frac{L_2}{L_1} \tag{4}$$

Result is always between 0 and 1. Then test statistics can be calculated:

$$\chi^2 = -2 \ln \lambda \tag{5}$$

And it follows Chi-square distribution with k degrees of freedom (k-number of restriction in the second model). H_0 holds that the reduced model is appropriate, and p-value for the overall model fit statistic less than 0.05 would suggest rejecting the null hypothesis. It provides evidence in favor of current model.

Pseudo R2

Usual R2 cannot be applied in case of logistic regression, due to binary nature of dependent variable. Estimated logistic function does not fit the real observations which can take values of 0 or 1. Therefore, it was necessary to introduce indicator, which better reflect the nature of binary data. McFadden Pseudo R2 can be calculated using following equation:

$$R^2_{McFadden} = 1 - \frac{Ln(L_c)}{Ln(L_{intercept})} \tag{6}$$

Lc – refers to the maximized likelihood value from the current fitted model and Lintercept refers to likelihood value from the model with only the intercept and no covariates. If comparing two models on the same data, McFadden's would be higher for the model with the higher likelihood.

Classification Rate

Classification rate is calculated comparing predicted probabilities with real results on the control group of data. If P(Y=1|X) > 0.5 then predicted Y=1 if P(Y=1|X) < 0.5 then predicted Y is 0. In some other application could be considered different boundaries to assess the model. The higher classification rate means better model.

ROC curve

ROC curve and Area under the curve (AUC) present typical performance indicator for binary classifier. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system: 0.90 - 1 = excellent, 0.80 - 0.90 = good, 0.70 - 0.80 = fair, 0.60 - 0.70 = poor, 0.50 - 0.60 = failed.

RESULTS AND DISCUSSION

First, the dataset was divided into two parts, first part including 4000 observations was used to modelling process (variables selection, model estimation and verification) and second part including 427 observations was used to test prediction ability of the final model. Then, process of variables selection was applied; from total number of 12 variables were selected 7, as factors which significantly influence the probability of getting loan. These variables were used to estimate final logistic regression model. Estimated coefficients and their statistics are shown in Table 1.

Table 1 Estimated logistic regression model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
Intercept	-7.958e-01	1.163e-01	-6.845	7.62e-12	***
Debt	1.356e+00	1.443e-01	9.398	< 2e-16	***
Ploan	5.740e+00	2.611e-01	21.983	< 2e-16	***
EBITDA	-8.289e-07	1.120e-07	-7.399	1.37e-13	***
Size	4.582e-04	5.480e-05	8.362	< 2e-16	***
Revpha	-7.086e-06	2.683e-06	-2.641	0.008261	**
Own	1.057e-03	3.009e-04	3.512	0.000445	***
Pprv	-5.175e-01	1.121e-01	-4.615	3.92e-06	***

Sig. Codes *** 0.001 ** 0.01 *0.05

Source: Authors

Estimated coefficients in this case does not mean direct influence of the explanatory variables to probability of getting loan, but their influence on the log odds ratio of getting loan. Therefore, it is difficult to evaluate influence of each variable on dependent variable by coefficient value; on the other side, it is possible to evaluate significance of each variable according to their p-values in the final model. Significance is indicated in last column of table 1. Most of the factors are significant at alfa = 0.001, only variable revenues per ha is significant at the alfa = 0.01. Overall significance of the model evaluated by the probability of likelihood ratio test is close to 0, which suggest significant model. The same test was applied also to evaluate significance of individual variables. In this case was compared likelihood of model with only intercept (without explanatory variables) and model including additional explanatory variable. If the adding of explanatory variable improved prediction ability of the model, variable is denoted as significant. This procedure differs from the method used in coefficients table above.

Table 2 Evaluation of variables significance

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	Significance
NULL			3999	4989.0		
Debt	1	104.05	3998	4884.9	< 2.2e-16	***
Ploan	1	873.32	3997	4011.6	< 2.2e-16	***
EBITDA	1	09.6	3996	4002.6	0.002616	**
Size	1	118.93	3995	3883.6	< 2.2e-16	***
Revpha	1	6.49	3994	3877.1	0.010839	*
Own	1	18.32	3993	3858.8	1.865e-05	***
Pprv	1	21.50	3992	3837.3	3.536e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Authors

Overall model quality can be evaluated also by using ROC curve and its area under the curve. The curve shows relationship between true positive rate (correctly predicted loans) and false positive rate (when model predicted loan in case when it did not occur). ROC curve is shown on the Figure 1. In this case it can be noticed, that model is slightly in favor of true positive prediction. From ROC curve is derived another important indicator AUC, which denotes Area under curve. AUC indicator of presented model is equal to 0.69. According common rules this means that model is close to fair quality.

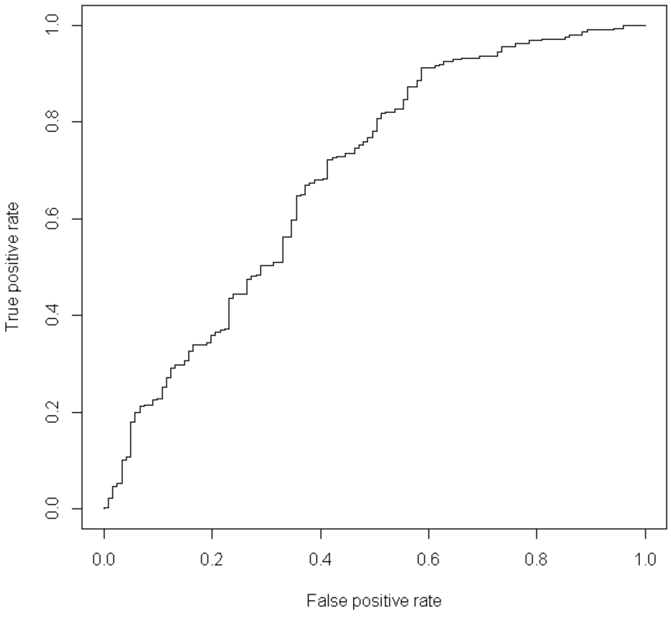

Figure 1 ROC curve

Quality of model can be assessed also using McFadden R square, which value is equal to 0.23. In case of logistic regression R square values are not as important as in case of classical linear regression model. Due to character of dependent variable, their values are usually low.

Real values of dependent variable are equal to 0 or 1; on the other side predicted values are decimal numbers varying between 0 and 1. Better indicator of model quality, which compares predicted and real values, in this case is number of correct predictions. If the predicted value is higher than 0.5, it means predicted 1 (firm will get loan), otherwise 0 (firm will not get loan). Presented model successfully predicted 60% of cases. It suggests fair quality of the model, on the other side the model should be reconsidered to increase its prediction ability. Number of correct predictions and ROC curve were estimated by application of the model on the training set of data (427 observations).

As it was already mentioned above, coefficients in estimated models does not indicate direct influence of explanatory variables on the probability of getting loan, but their influence on the log odds ratio of getting loan. Estimated function is not linear; influence of each explanatory variable on the final probability value will depend on the value of X. It will be different between low and high value of X. To describe the real influence of independent variable to probability it would be necessary to describe influence of change in low, medium and high values of this variable. Much easier is to describe constant effect on the odds ratio. Therefore, to derive influence of each variable on the probability of getting loan, it is necessary to exponentiate each coefficient value to get odds ratio for each variable. Odds ratios of explanatory variables with confidence intervals are shown in the table.

Table 3 Odds ratios with confidence intervals

Variable	OR	2.5 %	97.5 %
Debt	3.88	2.93	5.16
Ploan	310.92	188.09	523.59
EBITDA	0.999999	0.999999	0.999999
Size	1.000458	1.000352	1.000567
Revpha	0.999993	0.999987	0.999998
Own	1.001057	1.000490	1.001673
Pprv	0.60	0.48	0.74

Source: Authors

Odds ratios higher than one mean positive effect of explanatory variable on chance of getting loan, odds ratio value lower than one mean negative effect on chance of getting loan. Significantly positive or negative influence on chance of getting loan means only odds ratio without 1 in its confidence interval (1 means indifferent to positive or negative influence). Confidence intervals present range of values where the odds ratio should be with 95% probability. Highest positive influence on chance of getting loan has proportion of loans, then debt, and only slightly positive influence farm size and number of owners. For example, if the number of owners increases by 1, the chance of getting loan increases by 0.1%. The rest of the odds ratios could be interpreted analogically. On the other side, highest negative influence on chance of getting loan has proportion of crop production and slightly negative influence EBITDA and revenue per ha. In case, when it is necessary to assess credibility of farm, parameters of the individual farm can be filled into model and probability of getting loan can be estimated to assess its solvency.

CONCLUSIONS

The main objective of presented paper was to demonstrate application of logistic regression in case, when it is necessary to classify units according their attributes. In the economic field, this method is usually used for insurance and bank purpose, bonity models, bankruptcy models or in targeted marketing. Application in this paper was evaluation of farms solvency to get loan. First, from the set of 13 explanatory variables it was selected the set of 7 variables as the exogenous variables to final model. All the variables and overall model were considered significant according usual statistical procedures. Model quality was assessed by its application on test set of data and calculating McFadden R square, ROC curve and area under curve, and number of correct predictions. These indicators considered model as fair (60% of correct predictions). However, if the model should be used in practice, another explanatory variable should be considered to improve its prediction ability. It should be also applied on more testing data groups to get cross validated verification and get better insight into its prediction accuracy. When interpreting results of the model it should be noted, that estimated coefficients does not mean direct influence of explanatory variables on final probability of getting loan, but on the log odds ratio of getting loan. Probability in this kind of model is non-linear, therefore there is no constant influence of the variable, but the change in the probability depends on the specific value of explanatory variable and is different between low, medium and high values. This is the reason, why in logistic models is usually estimated and interpreted odds ratio (exponentiated value of coefficients) to assess variables influence. In presented example the highest positive influence on chance of getting farm loan had variable debt and proportion of loans, which suggest, that if the farm have already successfully got loan in the past, it probably would get also new loan. Highest negative influence on the chance of getting loan had proportion of crop production. The output of the model is probability of getting bank loan. In conclusion, producing final model would require further modification and cross validation, but it was presented that logistic regression offers efficient tool for data classification. This method can be applied in various fields of research and help to get better insight into process where outcome is categorical variable.

REFERENCES

[1] Jabeur, S. B. (2017). Bankruptcy prediction using partial least squares logistic regression. *Journal of Retailing and Consumer Services*, 36, 197-202.

[2] Nikolic, N., Zarkic-Joksimovic, N., Stojanovski, D. & Joksimovic, I. (2013). The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements. *Expert Systems with Applications*, 40(15), 5932-5944.

[3] Serener, B. (2016). Statistical Analysis of Internet Banking Usage with Logistic Regression. *Procedia Computer Science*, 102, 648-653.