Estimation of a credit scoring model for lenders company

Felipe Alonso Arias-Arbeláez[*] Juan Sebastián Bravo-Valbuena[†]
Francisco Iván Zuluaga-Díaz[‡]

November 22, 2015

Abstract

Historically it has seen that banks developed their own system of risk for loans to their customers; because such information is privileged, it is very difficult to know how to measure credit risk. That is why it was decided to propose a model of credit scoring to answer basic questions like: should we lend to this client? What is the loan limit? How to reduce the risk of default? Among others. Linkvest Capital is a private equity firm that identifies, analyzes, structure, leads and supervises the businesses in which it invests. Linkvest Capital have different branches (business lines), one of them is being a mortgage loan originator and seller in Florida, United States.

Basically what is sought with the model is to decrease the probability of default using variables of individuals as their income, how long they have been a customer, among others. The results of this research includes a methodology and the steps needed to define the model we are going to estimate. As a conclusion, we defined an econometric model based on binary logistic regression that fits the data of the people that already paid in Linkvest Capital LLC.

[*] Departament of Mathematical Sciences, School of Science, EAFIT University, Medellín, Colombia. email: fariasa@eafit.edu.co

[†] Operations coordinator, Linkvest Capital, Miami, United States; email: operaciones2@linkvestcapital.com

[‡] Departament of Mathematical Sciences, School of Science, EAFIT University, Medellín, Colombia. email: fzuluag2@eafit.edu.co

1 Introduction

Historically it has seen that banks developed their own system of risk for loans to their customers, because such information is privileged, it is very difficult to know how to measure credit risk. That is why it was decided to propose a model of credit scoring to answer basic questions like: Should we lend to this client? What is the loan limit? How to reduce the risk of default? Among others.

Linkvest Capital is a private equity firm that identifies, analyzes, structure, leads and supervises the businesses in which it invests. Linkvest Capital have different branches (business lines), one of them is being a mortgage loan originator and seller in Florida, United States. Basically what is sought with the model is to decrease the probability of default using variables of individuals as their income, how long they have been a customer, among others.
According to Hand and Henley (1997): "Credit scoring is the term used to describe formal statistical methods used for classifying applicants for credit into 'good' and 'bad' risk classes."

The credit scoring is assessed in terms of predictive models of payment or reimbursement by a score that measures the risk of a borrower or the operation itself. The analysis is done in terms of score is seeking to explain the financial behavior in terms of the services requested, the relationship between risk and return and the cost of the operation (Cantón, Rubio, & Blasco, 2010).

Clearly, many models have been used to solve the credit score problem in lenders companies, we will mention some of the most representative and choose one to apply. The discriminant analysis is a good role model when discriminating customers between good and bad payers, the problem that has primarily is impossible to calculate the probability of default and fails to satisfy the basic assumptions of econometrics (homoscedasticity, linearity, normality and independence) Cantón et al. (2010).

In Altman (1968), we can see an application of this method, using the explanatory variables as ratios, he found the probability of default ratios using net income/sales, retained earnings/assets, among others.

There is also the linear probability models, in these models, least squares regression where the dependent variable takes the value of one if the client is failed and zero if the customer meets its payment obligation is used, the equation is a linear function of the explanatory variables (Hand & Henley, 1997). Perhaps the precursor of this model is Orgler (1970) who proposed models for commercial loans, but can be successfully applied to personal loans, as in our case.

The logit model calculates the probability to belong to a group of payer or not payer. Wiginton (1980) in his study compares the logit model with discriminant analysis in the estimation of a model of default and concluded that the logit model offered a better fit in terms of probability that the discriminant analysis.

Finally, the neural networks try to imitate the nervous system, so that generate a certain degree of intelligence. Nodes that respond to certain input signals are interconnected. Rosenberg and Gleit (1994) show a summary of quantitative methods for handling credit modeling, neural networks are mentioned as a possible model but not proposed a solution on credit score but in credit fraud.

The paper proceeds as follows. Section 2 provides the model specification, Section 3 shows the logistic regression strategy to estimate our model and Section 4 exhibits the outcomes of our simulation exercises. Finally, Section 5 contains some concluding remarks.

2 Specification

The methodology adopted will be based on the realization of a model by linear regression and logit model using a variation of the model called model of binary logistic regression (Cantón et al., 2010).

This model in most cases does not produce inefficient estimators, easily supports categorical variables, estimates the probability of loan default to the values of the independent variables and determine the influence of each independent variable on the dependent variable depending on the odd ratio, if this value is close to one then there is an increased probability of default and if it is close to zero indicates a better chance of fulfillment (Hand & Henley, 1997).

The logistic regression model can be formulated as:

$$Z = \Sigma\Psi + \mu \tag{1}$$

where

$$\Sigma = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_{28}] \qquad \Psi = \begin{bmatrix} 1 \\ V_1 \\ V_2 \\ \vdots \\ V_{27} \end{bmatrix} \tag{2}$$

μ is the disturbance and p is going to be the probability of default and can be estimated as follows:

$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \tag{3}$$

The variables defined to analyze the probability of default are taken from the questionnaire the lenders have to answer to get the loan, some questions are mandatory from the US Government:

1. Amount of the loan

2. Interest rate of the loan

3. Months of the loan

4. Age of the lender

5. Years of school of the lender

6. If the lender is married or not

7. Number of dependents of the lender

8. If the lender have an own house or a rent house

9. If the lender is self employed or employed

10. Monthly income of the lender

11. Monthly expense of the lender

12. Average cash of the lender

13. Total fixed assets of the lender

14. Are there any outstanding judgments against you?

15. Have you been declared bankrupt within the past 7 years?

16. Have you had property foreclosed upon or given title or deed in lieu thereof in the last seven years?

17. Are you a party to a lawsuit?

18. Have you directly or indirectly been obligated on any loan which resulted in foreclosure, transfer of title in lieu of foreclosure, or judgment?

19. Are you presently delinquent or in default on any Federal debt or any other loan, mortgage, financial obligation, bond or loan guarantee?

20. Are you obligated to pay alimony, child support, or separate maintenance?

21. Is any part of the down payment borrowed?

22. Are you co-maker or endorser on a note?

23. Are you a U.S. citizen?

24. Are you a permanent resident alien?

25. Do you have intend to occupy the property as your primary residence?

Our independent variable is going to be if the lender paid or not the debt; the information we are using is based on the information Linkvest have about people they already paid the amounts.

Based on Equation (2), we define the variables we are going to use from the questionnaire based on what Cantón et al. (2010) defined:

$V_1 :$ Amount

$V_2 :$ Interest

$V_3 :$ Months

$V_4 :$ Age

$V_5 :$ Years school

$V_6 :$ Married

$V_7 :$ Dependents

$V_8 :$ Home

$V_9 :$ Employment

$V_{10} :$ Income

$V_{11} :$ Expense

$V_{12} :$ Cash

$V_{13} :$ Assets

$V_{14} :$ Questions $(V_{14}, V_{15}, \ldots, V_{25})$

$V_{26} :$ Ethnicity

$V_{27} :$ Gender

3 Estimation

After estimation the model with binary logistic regression is:

$$Z = \widehat{\Sigma}\widehat{\Psi} \tag{4}$$

where

$$\widehat{\Sigma}' = \begin{bmatrix} -18.42007 \\ 0.02773 \\ 0.02606 \\ 0.00793 \\ -0.16936 \\ -0.14049 \\ 0.75352 \\ -0.73294 \\ -0.00029 \\ 0.00012 \\ 0.39758 \\ -0.136413 \end{bmatrix} \qquad \widehat{\Psi} = \begin{bmatrix} 1 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \\ V_7 \\ V_8 \\ V_9 \\ V_{10} \\ V_{11} \\ V_{26} \\ V_{27} \end{bmatrix}$$

Only those 12 variables were statistically significant, so the other ones does not represent the model because at a significant value of 5% they are not explaining the model. It is important to say that the first value is going to be the constant, remember the linear regression is a curve fitting in the data which has an intercept with Y, in this case, the constant is the intercept.

4 Simulation Exercises

Using the model estimated on Section 3 we can determine the probability of default of each individual. In Table 1 we can see the estimated probability of default and the real response (if the individual paid (1) or not (0)) of 15 individuals.

Estimated probability	Real response
0.8754	0
0.9347	0
0.2365	1
0.7912	0
0.7890	0
0.0012	1
0.0233	1
0.3134	0
0.9899	0
0.4679	1
0.0126	1
0.1543	1
0.1456	0
0.9744	0
0.9543	0

Table 1: Calculated probability of default and if the individual paid or not.

Remember that if the probability of default is near to 1 it means the individual is not going to pay the debt, and if it's near 0, the individual is going to pay the debt. Because this is an estimation of the probability of default, not all the results are going to fit, but we can see that almost every probability makes sense with the real response.

A basic but important assumption is that the residuals should have a normal distribution with mean zero and constant variance, which was carried out by a histogram adjustment to a normal to demonstrate this. Figure 1 shows the histogram. Clearly it sees that in general the residuals lead a normal distribution on which meet the assumption of normality in the residuals.

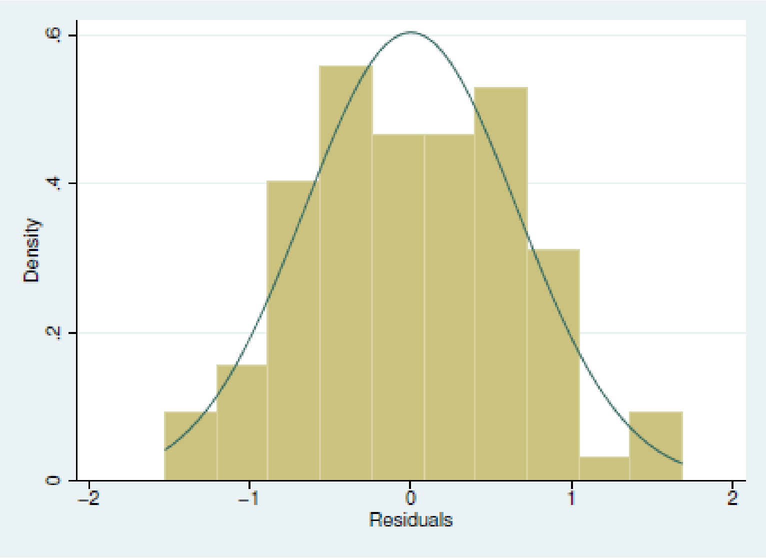

Figure 1: Residuals with normal distribution

Now the aim is to check that the variance of the residuals is homogeneous, or that they present homoskedasticity. One way to check this assumption making a graph between the residuals and the estimated values as shown in Figure 2, where residuals are tending to zero, so it is concluded that there are homoskedasticity.

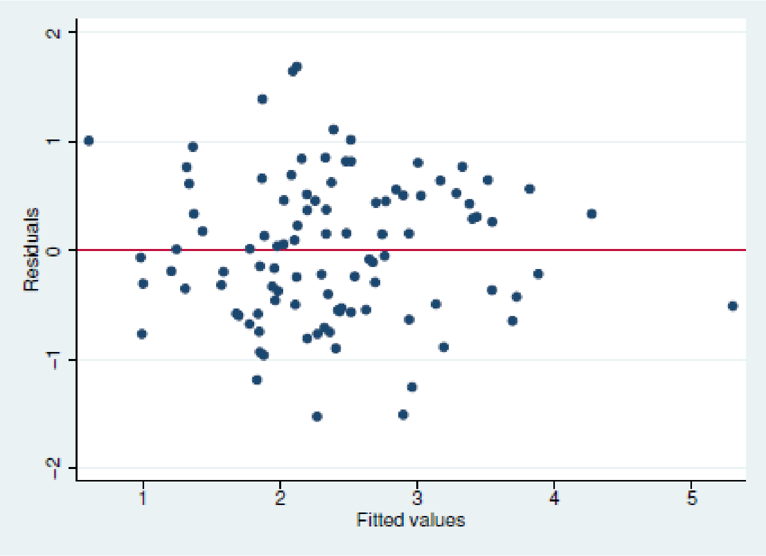

Figure 2: Residuals vs estimated values

Another way to search homoskedasticity is to use a test that proves heteroskedasticity, such a White test, literature make the assumption where the null hypothesis says the variance of the residuals is homoskedasticity (Wiginton, 1980). In Figure 2 we can see the test results for this regression. The p-value for heteroskedasticity is higher than the level of significance of 5%, so the literature make the null hypothesis accepted, rejecting the presence of heteroskedasticity in the model.

Source	**chi2**	**df**	**p**
Heteroskedasticity	15.77	17	0.5400
Skewness	10.30	5	0.0671
Kurtosis	1.15	1	0.2827
Total	27.23	23	0.2463

Table 2: White Test

Finally it is very important to define whether the model correctly specified, an error in the model specification can occur when a relevant variable is omitted or otherwise, one or more irrelevant variables are included in the model. To test this, we use the Ramsey- RESET test. Next we show Ramsey - RESET test for model. The model is correctly specified, because the null hypothesis is accepted.

Ramsey RESET test using powers of the fitted values

Ho: model has no omitted variables

$$F(3,91) = 0.69$$

$$Prob > F = 0.5617$$

5 Conclusions

We have estimated a credit scoring model that efficiently calculates probabilities of default in the case of mortgage loans, to calculate this probability is necessary to have the complete information requested in the relevant variables in the model. You can see that this model meets the basic econometric assumptions of residual normality, correct specification and homoscedasticity.

It is clear that initially we developed a basic model to evaluate the possibility of building a credit scoring model in a brief period of time; as the results were favorable, it is proposed to work in a more complete model of neural networks that can take into account the financial performance of the person requesting the loan. This model has only evaluated variables into account cross-section and does not take into account time series or occurrences over time, this worth to be valued.

It is proposed as future work to assess a more complete model taking into account social variables such as the current socioeconomic stratum, the ratio of loans and payments in other banking institutions, among others.

References

Altman, E. I. (1968, September). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bakruptcy. *The Journal of Finance*, *XXIII*(4), 589-609.

Cantón, S. R., Rubio, J. L., & Blasco, D. C. (2010, June). A Credit Scoring Model for Institutions of Microfinance under the Basel II Normative. *Journal of Economics, Finance and Administrative Science*, *15*(28).

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Costumer Credit Scoring: A review. *Journal of the Royal Statistical Association*, *160*(Part 3), 523-541.

Orgler, Y. E. (1970, November). A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit and Banking*, *2*(4), 435-445.

Rosenberg, E., & Gleit, A. (1994, August). Quantitative Methods in Credit Management: A Survey. *Journal of Operations Research*, *42*(4), 589-613.

Wiginton, J. C. (1980, September). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitaty Analysis*, *15*(3), 757-770.