Multinomial Logistic Regression and Spline Regression for Credit Risk Modelling

M R Adha\textsuperscript{1,*}, S Nurrohmah\textsuperscript{1} and S Abdullah\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, Universitas Indonesia, Kampus Baru UI, Depok, 16424, Indonesia

*Corresponding Author: muhammad.rizky41@sci.ui.ac.id

Abstract. Regression modelling has been adapted in retail banking because of its capability to analyze the continuous and discrete data. It is an important tool for credit risk scoring, stress testing and credit asset evaluation. In this paper, the approach used is multinomial logistic regression model to gain the information regarding the factors that affect the occurrence of default and attrition events on credits. In addition, this paper will also introduce spline regression approach using truncated power basis to model the hazard functions of default and attrition events. The flexibility of spline function allows us to model the nonlinear and irregular shapes of the hazard functions. Then, by using spline regression and multinomial logistic regression model, there will be a better result and interpretation. There are several advantages by using those both models. First, by using the flexible spline regression function, it can model nonlinear and irregular shapes of the hazard functions. Second, it is easy to understand and implement, and its simple parametric form from multinomial logistic regression model can make it easy in model interpretation. Third, the multinomial logistic regression model has the ability to do prediction. Furthermore, by using a credit card dataset, we will demonstrate how to build these models, and we also provide statistical explanatory and the prediction accuracy of multinomial logistic regression model in classifying customers based on the prediction of default and attrition is 95.3%.

1. Introduction

In the last 50 years, credit users have been the driving force of the economy in almost all developed countries \cite{1}. By 2018, Bank Indonesia expects credit growth to increase to 10-12%. In line with these figures, Indonesia's economic growth is in the range of 5.1-5.5%. Therefore, the magnitude of credit growth will affect the magnitude of economic growth in a country.

The rapid growth of consumer credit brings the demand for better credit scoring system to managing credit lending business. The credit scoring has been successfully applied in credit risk management \cite{2}. However, with the fierce competition, the credit scoring has also been expanded from only controlling the credit default risk to maximizing the profit \cite{1}. The profitability has become the most significant business consideration in credit decision making. To better estimate customer’s profitability, the lenders need more sophisticated models which can estimate the dynamics of risk and revenue over time.

Based on the consideration of profitability, required a model that can be used as the basis of decision making related to the provision of credit, interest rates and credit line. For this purpose, in this paper will be used approach by using multinomial logistic regression model to know the factors that influence the default and attrition events on a credit. We use multinomial logistic regression model because this model also can handle time-dependent and time-independent covariates and deal with
competing risks. In this paper, the object of competing risk is credit because it can risk default, attrition (i.e., customers pay off their balance and close their accounts) or not both (open status).

In addition, we also introduce spline regression approach using truncated power basis to model hazard functions of default and attrition events. In this paper, the dependent variable is hazard (the number of individuals who experience the target event in certain time period by the number of individuals at risk during that time period) and the independent variable is time (in months). The flexibility of spline function allows us to model the nonlinear and irregular shape of the hazard functions.

In the end of the discussion, by using a credit card dataset, we will demonstrate how to build these models, and we also provide statistical explanatory and prediction accuracy.

2. Multinomial Logistic Regression Model

In this paper, there are 3 types of event in credit dataset. Define \( j = 1 \) as the default event, \( j = 2 \) as the attrition event, and \( j = 0 \) as the open status. Thus, in other words \( J \) states competing risk in the data.

In the multinomial logistic regression model with the dependent variable having three categories, model formation requires two logit functions. Assume the categories of the dependent variable, \( y_j \), are coded with 0, 1, and 2. In this paper, \( y_j = 0 \) will be used as a reference or basis. To build the model, assume the loan portfolio has \( p \) covariates, denote by vector \( x_i, i = 1, 2, ..., n \). Thus, obtained a logit model as follows.

\[
g_j(x_i) = \ln \frac{\Pr (y_i = j | x_i)}{\Pr (y_i = 0 | x_i)} = \ln \frac{\pi_j(x_i)}{\pi_0(x_i)} = \ln \frac{\pi_j(x_i)}{1 - \sum_{j=1}^{p} \pi_j(x_i)} = \beta_{0j} + \beta_{1j} x_{i0} + \cdots + \beta_{pj} x_{ip} \quad j = 1, 2
\]

Then, the probability of each event is as follows.

\[
\pi_0(x_i) = \frac{1}{1 + \exp(g_1(x_i)) + \exp(g_2(x_i))} \quad i = 1, 2, ..., n
\]
\[
\pi_1(x_i) = \frac{\exp(g_1(x_i))}{1 + \exp(g_1(x_i)) + \exp(g_2(x_i))} \quad i = 1, 2, ..., n
\]
\[
\pi_2(x_i) = \frac{\exp(g_2(x_i))}{1 + \exp(g_1(x_i)) + \exp(g_2(x_i))} \quad i = 1, 2, ..., n
\]

2.1 Parameter Estimation

The method used to estimate the parameters in multinomial logistic regression is the maximum likelihood method. In this method, we will find the parameter value of \( \beta = (\beta_0, \beta_1) \) which maximizes the likelihood function. To simplify the notation, consider \( \pi_{ji} = \pi_j(x_i) \). The \( \beta \) value that maximizes \( \ln L(\beta) \) is obtained by differentiating \( \ln L(\beta) \) against parameters \( \beta_{0j}, \beta_{1j}, ..., \beta_{pj} \), which is then equated with 0 as follows.

\[
\frac{\partial \ln L(\beta)}{\partial \beta_{kj}} = \sum_{i=1}^{n} x_ki (y_{ij} - \pi_{ji}) = 0
\]

for \( j = 1, 2 \) and \( k = 0, 1, 2, ..., p \), with \( x_{0i} = 1 \) for each observation.

2.2 Parameter Testing

Partial test is used to test the effect of each independent variable individually. The test statistic used is Wald test.
3. Spline Regression Model
Spline regression is a regression analysis method that is piecewise polynomial, i.e., a polynomial that has a segmented property on \( r \) intervals formed by knot points [3]. Knot point is a common fusion point that occurs because there are changes in pattern behavior at different intervals. Spline has an advantage in overcoming the pattern of data that shows a sharp rise or fall with the assist of knot points, and the resulting curve is relatively smooth [4].

The general form of spline regression using the truncated power basis is as follows.

\[
y = \sum_{j=0}^{d} \alpha_{j} t^{j} + \sum_{p=1}^{r} \alpha_{p+d} (t - K_{p})^{d} + \varepsilon
\]

with

\[
(t - K_{p})^{d} = \begin{cases} 
(t - K_{p})^{d} , & t > K_{p} \\
0, & t \leq K_{p}
\end{cases}
\]

in which \( a = (\alpha_{0}, \alpha_{1}, \ldots, \alpha_{p}, \alpha_{p+d}, \ldots, \alpha_{r+d})^{T} \) is the parameter or regression coefficient for the truncated power basis function, \( d \) is degree of polynomial on truncated power basis, \( r \) is number of knots used, and \( K_{1}, K_{2}, \ldots, K_{r} \) are location of knots.

3.1 Parameter Estimation
Eubank [5] states that parameter estimation on spline regression using OLS (Ordinary Least Square). The parameter estimator for spline regression using truncated power basis is

\[
a = (T^{T}T)^{-1} T^{T} y
\]

with

\[
a = \begin{bmatrix} \alpha_{0} \\ \vdots \\ \alpha_{d} \\ \alpha_{d+p} \\ \vdots \\ \alpha_{d+r} \end{bmatrix},
T = \begin{bmatrix} 1 & x_{1} & x_{1}^{2} & \ldots & (x_{1} - K_{1})^{d} & \ldots & (x_{r} - K_{s})^{d} \\ 1 & x_{2} & x_{2}^{2} & \ldots & (x_{2} - K_{1})^{d} & \ldots & (x_{r} - K_{s})^{d} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & x_{r} & x_{r}^{2} & \ldots & (x_{r} - K_{1})^{d} & \ldots & (x_{r} - K_{s})^{d} \end{bmatrix},
y = \begin{bmatrix} y_{1} \\ y_{2} \\ \vdots \\ y_{r} \end{bmatrix}
\]

3.2 Location and Number of Knots
Ngo & Wand [6] uses a rule of determining the optimal number of knots on spline regression as follows.

\[
r = \max \left( 5, \min \left( \frac{1}{4} N, 35 \right) \right)
\]

with \( N \) is number of unique \( x_{i} \)’s, and rule for the knot locations is \( K_{s} = \frac{s+1}{r+2} \)th sample quantile of unique \( x_{i} \)’s for \( s = 1, 2, \ldots, r \).

3.3 Error Variance Estimation
To measure the quality of regression parameter estimates by using OLS (Ordinary Least Square), it is necessary to assess the unknown variance of errors. The estimator for error variance is MSE (Mean Square Error). MSE obtained from Sum Square Error divided by error’s degrees of freedom.

3.4 Coefficient of Determination
The coefficient of determination is a measure that states how considerable the contribution of independent variables in predicting the value of the dependent variable.
4. Numerical Example

The dataset used is credit card data of 400 customers with 24 months observation period, in which the data contains information about the occurrence time of the event, customer status (default, attrition, or open status), credit line, monthly income, credit score, average monthly debt, number of credit cards, gender, type of credit card (additional or major), relationship status (married or not).

4.1 Multinomial Logistic Regression Model

In the model, the event that will be used as a reference/basis is the open status event \( (j = 0) \), where \( x_{1i} \) represents credit line, \( x_{2i} \) represents income, \( x_{3i} \) represents credit score, \( x_{4i} \) represents average debt, \( x_{5i} \) represents the number of credit card, \( x_{6i} \) represents gender, \( x_{7i} \) represents the type of credit card, and \( x_{8i} \) represents the relationship status.

The parameters in multinomial logistic regression are estimated using the maximum likelihood method. The results of parameter estimation are shown in Table 1 as follows.

| Variable              | Default Event | Attrition Event |
|-----------------------|---------------|-----------------|
|                       | Parameter     | Coefficient     | p-value | Parameter | Coefficient | p-value |
| Intercept             | \( \beta_{01} \) | 15.420          | 0.000   | \( \beta_{02} \) | 26.057      | 0.000   |
| Credit line           | \( \beta_{11} \) | 0.135           | 0.002   | \( \beta_{12} \) | -0.015      | 0.789   |
| Income                | \( \beta_{21} \) | -0.516          | 0.000   | \( \beta_{22} \) | -1.092      | 0.000   |
| Credit score          | \( \beta_{31} \) | -0.029          | 0.000   | \( \beta_{32} \) | -0.022      | 0.000   |
| Average debt          | \( \beta_{41} \) | 0.122           | 0.397   | \( \beta_{42} \) | 0.076       | 0.790   |
| Number of credit cards| \( \beta_{51} \) | 0.791           | 0.057   | \( \beta_{52} \) | 1.677       | 0.000   |
| Gender                | \( \beta_{61} \) | 0.440           | 0.502   | \( \beta_{62} \) | -0.392      | 0.672   |
| Type of credit card   | \( \beta_{71} \) | 1.964           | 0.121   | \( \beta_{72} \) | 0.257       | 0.813   |
| Relationship status   | \( \beta_{81} \) | 0.661           | 0.406   | \( \beta_{82} \) | 0.157       | 0.845   |

For partial test can be seen in Table 1. at \( \alpha = 0.05 \), it can be seen that the independent variable has significant effect, that is, \( p\text{-value} < \alpha = 0.05 \). The independent variables that significantly influence the default event are credit line, income, and credit score. Then, the independent variables that significantly influence the attrition event are income, credit score, and number of credit cards.

The logistic regression model obtained is used to classify customers based on predictions of default events and attrition events. The degree of accuracy of the logistic regression model in the classification is calculated using the classification table shown in Table 2.

| Observed       | Predicted | Percent Correct |
|----------------|-----------|-----------------|
|                | Open Status | Default | Attrition |                |
| Open Status    | 290        | 2       | 6         | 97.3%          |
| Default        | 2          | 60      | 0         | 96.8%          |
| Attrition      | 7          | 2       | 31        | 77.5%          |
| Overall Percentage | 74.8%   | 16.0%   | 9.3%      | 95.3%          |

The percentage of classification accuracy is calculated as the ratio between the number of observations classified precisely by the model to the total number of observations. Thus, the
percentage accuracy of multinomial logistic regression model obtained in classifying customers based on the prediction of default and attrition is 95.3%.

4.2 Spline Regression Model
In this paper, spline regression is used to model the hazard functions for default and attrition events. In this numerical example, $y$ is the dependent variable which represents hazard and $t$ is the independent variable which represents time (in months). The degree of the polynomial on the truncated power basis used is $d = 3$ (cubic spline), and we will use the rules of location and number of knots that introduced by [7], then the required number of knots is $r = 6$.

Next, we will find the model of spline regression using the cubic truncated power basis with $r = 6$. For comparison, we also use the spline regression with $r = 5$ and $r = 7$. In order to know which regression model is better, it can be known through MSE and $R^2$ of each regression model as follows.

| Type of Event  | Number of Knots | MSE            | $R^2$       |
|---------------|-----------------|----------------|-------------|
| Default       | $r = 5$         | 0.0000128675   | 88.99%      |
|               | $r = 6$         | 0.0000112607   | 91.01%      |
|               | $r = 7$         | 0.0000143916   | 89.33%      |

From the three models above, the best model will be chosen based on the criteria of MSE and $R^2$, i.e., the smallest MSE and the largest $R^2$. Thus, the best model chosen is a spline regression model using cubic truncated power basis with 6 knots, with estimated models as follows.

$$
\hat{y} = 0.0145062 - 0.0129881 t + 0.0041186 t^2 - 0.0003220 t^3 + 0.0008372 (t - K_1)_+^3 \\
-0.0008378 (t - K_2)_+^3 + 0.0004749 (t - K_3)_+^3 - 0.0003707 (t - K_4)_+^3 \\
+0.0006080 (t - K_5)_+^3 - 0.0009193 (t - K_6)_+^3 
$$

The plots for spline regression using the cubic truncated power basis with $r = 6$ is shown on Figure 1 as follows.

![Cubic spline with $r = 6$ for default event](image)

The interpretation of the best model chosen or in accordance with Figure 1, namely that the risk of default increased significantly until the 7th month, after which, there was a significant decrease as well until the 10th month. The gradual increase in risk occurs after the 10th month, followed by a decrease in the 16th month until the 19th month, and after that the risk increases back until the end of observation period.

For the attrition event, we get the models with results as follows.

| Type of Event  | Number of Knots | MSE            | $R^2$       |
|---------------|-----------------|----------------|-------------|
| Attrition     | $r = 5$         | 0.00000695429  | 86.72%      |
|               | $r = 6$         | 0.00000673517  | 87.99%      |
|               | $r = 7$         | 0.00000716070  | 88.15%      |
From the three models above, the best model will be chosen based on the criteria of MSE and $R^2$, i.e., the smallest MSE and the largest $R^2$. Thus, the best model chosen is a spline regression model using a cubic truncated power basis with 6 knots, with estimated models as follows.

$$
\hat{y} = 0.01278 - 0.008524t + 0.0002081r^2 - 0.0001407t^3 + 0.0002279(t - K_1)^3
$$

$$+ 0.00008648(t - K_2)^3 - 0.000601(t - K_3)^3 + 0.0008499(t - K_4)^3
$$

$$- 0.0006539(t - K_5)^3 + 0.0004447(t - K_6)^3
$$

The plots for spline regression using the cubic truncated power basis with $r = 6$ is shown on Figure 2 as follows.

![Figure 2. Cubic spline with $r = 6$ for attrition event](image)

Interpretation of the best model chosen or in accordance with Figure 4.5, namely that the risk of attrition begins to increase after the 3rd month until the 8th month. After that, the risk of attrition always decreases until the 12th month, then experienced a significant increase risk until the 15th month. After the 15th month, the risk decreases again until the 19th month, and increases slowly until the end of observation period.

5. Conclusion
In the formation of a multinomial logistic model with categories which mutually exclusive and exhaustive, the parameters are represented in the logit function of category $j$ compared to one of the $J$ categories that serve as the reference/basis, so that in the formation of the model $J - 1$ is required for the logit function. Based on numerical example in this paper, factors that significantly affect the default event are the credit line, income, and credit score. Then, the factors that significantly affect the attrition event are income, credit score, and number of credit cards. Presentation of accuracy of multinomial logistic regression model obtained in classifying customers based on the prediction of default and attrition is 95.3%.

The relationship between the dependent variable and the independent variable is represented by a curve called the regression curve. If the form of the regression curve is unknown, then the regression curve form can be estimated through the spline regression approach using the truncated power basis. Based on numerical example in this paper, best model chosen for default event and attrition event is spline regression model using cubic truncated power basis with 6 knots.

6. References
[1] Sirong L, Xiao K, and Tingting N 2016 Spline Based Survival Model for Credit Risk Modelling (Shanghai: Shanghai University of Finance and Economics)
[2] Baklouti I and Bouri A 2014 Can Credit-Scoring Models Effectively Predict Microloans Default? Statistical Evidence from the Tunisian Microfinance Bank pp. 57-69
[3] Wang J and Yang L 2009 Polynomial Spline Confidence Bands for Regression Curves Statistica Sinica 19, pp. 325-342
[4] Härđle W 1990 Applied Nonparametric Regression. (New York: Cambridge University Press)
[5] Eubank R L 1999 Nonparametric Regression and Spline Smoothing Second Edition (New York: Marcel Dekker)
[6] Ngo L and Wand M P 2004 Smoothing with Mixed Model Software (Massachusetts: Harvard University)
[7] Nielsen D Jason 2007 Splines Models for The Analysis of Recurrent Event Panel Data (Columbia: Simon Fraser University)
Acknowledgements
This work is supported by Hibah PITTA 2018 funded by Directorate of Research and Community Service of Universitas Indonesia (DRPM UI) No.2308/UN2.R3.1/HKP/05.00/2018.