IOP Conf. Series: Journal of Physics: Conf. Series **995** (2018) 012011 doi:10.1088/1742-6596/995/1/012011

# An Application of Robust Method in Multiple Linear Regression Model toward Credit Card Debt

**Nur Amira Azmi, Mohd Saifullah Rusiman, Kamil Khalid, Rozaini Roslan, Suliadi Sufahani, Mahathir Mohamad, Rohayu Mohd Salleh and Nur Shamsidah Amir Hamzah**

Fakulti Sains, Teknologi dan Pembangunan Insan, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

E-mail: nuramiraazmi39@gmail.com, saifulah@uthm.edu.my, kamil@uthm.edu.my

**Abstract.** Credit card is a convenient alternative replaced cash or cheque, and it is essential component for electronic and internet commerce. In this study, the researchers attempt to determine the relationship and significance variables between credit card debt and demographic variables such as age, household income, education level, years with current employer, years at current address, debt to income ratio and other debt. The provided data covers 850 customers information. There are three methods that applied to the credit card debt data which are multiple linear regression (MLR) models, MLR models with least quartile difference (LQD) method and MLR models with mean absolute deviation method. After comparing among three methods, it is found that MLR model with LQD method became the best model with the lowest value of mean square error (MSE). According to the final model, it shows that the years with current employer, years at current address, household income in thousands and debt to income ratio are positively associated with the amount of credit debt. Meanwhile variables for age, level of education and other debt are negatively associated with amount of credit debt. This study may serve as a reference for the bank company by using robust methods, so that they could better understand their options and choice that is best aligned with their goals for inference regarding to the credit card debt.

## 1. Introduction

Credit card has been known in many countries and become one of the competitive financial industries in the world. Many consumer use credit card as it is convenient to be used as a payment tool and can be accepted everywhere in order to obtain goods and services more easily. The biggest problem is it might lead the consumer to bankruptcy if there were impractical used of the card. Credit card debt is a part of most people lives as credit card ownership and usage have increased substantially in recent decades [1].

Credit defined as "money lent, which an individual intends to repay (often in a set period of time and at a set rate)" [1]. Generally, the choice of credit card as a payment mechanism is often accidental andriven by simpler considerations like convenience (a charge card is always in one's wallet), acceptability (certain retailers might not accept checks), accessibility (there is no convenient automated teller machine to withdraw cash), and habit (rent is typically always paid by checks). Debt, on the other hand, was defined as "financial liabilities, regardless of how these are incurred" [2].

Bakar and Tahir used multiple linear regression and neural network on bank performance. Seven variables including liquidity, credit risk, cost to income ratio, size, concentration ratio, inflation and GDP were used as independent variables. This study concluded that artificial neural network is the more powerful tool in predicting bank performance [3]. Emir et al. studied machine-learning techniques to construct nonlinear nonparametric forecasting models of consumer credit risk. Using conservative assumptions for the costs and benefits of cutting credit lines based on machine-learning forecasts, the cost savings range from 6% to 25% of total losses. This study suggested that aggregated consumer-credit risk

analytics may have important applications in forecasting systemic risk [4]. Regis and Artes analyzed the application of multi-state Markov models and logistic regression model to evaluate credit card risk. As a conclusion, multi-state Markov models performed better in predicting default risk, and logistic regression models performed better in predicting cancellation risk [5]. There are other quite considerable studies were carried out to merge MLR model with other method nowadays such as fuzzy theory [6, 7, 8]. Whereas [9, 10, 11] studied the statistical application used in Malaysia.

The purpose of this study is to investigate the factors that affect the credit card debt. Through this study the investigation of how variables involved and then giving impact toward credit card debt can be discovered. In this study, two methods will be used which are multiple linear regression model and multiple linear regression model with robust method toward credit card debt based on data bank loan. Lastly, the model from both methods will be compared by using mean square error in order to find the best model and factors for the credit card debt.

## 2. Methodology

The data is obtained from United States Census Bureau. In this study, the data includes 7 independent variables. The predictor variables are $x_1$ (age in years), $x_2$ (level of education), $x_3$ (years with current employer), $x_4$ (years at current address), $x_5$ (household income in thousands), $x_6$ (debt to income ration) and $x_7$ (other debt). While for dependent variable is $y$ (credit card debt in thousands). The model of this study should be able to provide an ideal relationship between all independent variables with predictor variables [12]. Besides that, this study should be able to identify the factors that give more impact in having credit card debt.

In this study, there are two analysis will be used which are multiple linear regression analysis and robust method. Regression analysis, also termed regression modelling, is an increasingly common statistical method used to describe and quantify the relation between a clinical outcome of interest and one or more other variables [13]. Worster *et al*. described regression analysis, also termed regression modelling, is an increasingly common statistical method used to describe and quantify the relation between a clinical outcome of interest and one or more other variables. The model is estimated by least squares, which yields parameter estimates such that the sum of squares of errors is minimized. The resulting prediction equation is shown as in (1) [14].

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_k x_{i,k} \quad (1)$$

where the "^" denotes as estimated values.

There are two robust methods that will be used which are least quartile difference (LQD) method and median absolute deviation (MAD). The first step is test the normality of the data. If the data do not fulfil the requirement to be a normal data, then transformation of the data will be done to overcome it. Next, test the multicollinearity between variables. Multicollinearity does not exist if the value of VIF is less than 10. The analysis of multiple linear regression will be done if the data fulfil the assumptions. However, in real life situations because of large data, usually the population is skewed. Lastly, the model from both multiple linear regression analysis and robust method analysis will be compared by using mean square error in order to find the best model and factors for the credit card debt.

Next, the model for multiple linear regression (MLR) is run in Statistical Packages for Social Sciences (SPSS) 20. Then the MSE of the model is obtained. The LS estimator is not robust, as its breakdown value is $1/n$, i.e. a single outlier can have arbitrarily large effects on the estimation [13]. The method of Least Quartile Difference (LQD) and Mean Absolute Deviation (MAD) run in Microsoft Excel. Then the model of MLR with applied LQD and model of MLR with applied MAD are obtained as in (2) and (3) [15, 16].

$$\text{LQD model:} \quad r_i(L) = y_i - \beta x_i - \alpha \quad (2)$$

$$\text{MAD model:} \quad \text{median}\left(|x_i - \text{median}(x)| \;\; i=1, 2, \ldots, n\right) \quad (3)$$

Finally, there are comparison between all the MSE value of the model in order to find the best model with several factors that influence credit card debt.

## 3. Result and discussions

In this study, the normality test was performed to see if the residuals are normally distributed by using Shapiro-Wilk normality test and Kolmogorov Smirnov test. Based on Table 1, both of the test show that the residual are normally distributed since the *p*-value is bigger than 0.05. It can be concluded that the data are normally distributed.

**Table 1.** Test or normality of original data

| |
|---|
| Shapiro-Wilk normality test |
| data: mydata |
| $W$ = 0.978, $p$-value = 0.9493 |
| Lilliefors (Kolmogorov-Smirnov) normality test |
| data: mydata |
| $D$ = 0.1084, $p$-value = 0.9989 |

The Variance Inflation Factor (VIF) which is the reciprocal of the tolerance statistics is considered. A VIF of greater than 10 indicate the existence of multicollinearity. However, based on the analysis, none of the variables have the value of VIF exceeding 10. This shows that there is no exist of multicollinearity. Since this two assumptions are fulfilled, it shows that the data are normally distributed.

The analysis of multiple linear regression model (Table 2), the credit card debt is directly proportional with years with current employer, years at current address, household income and debt to income ratio, whereas the credit card debt is inversely proportional with age, level of education and other debt.

**Table 2.** The MLR model with coefficients

| Model | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|
| | B | Std. Error | | |
| (Constant) | -1.532 | .270 | -5.667 | .000 |
| Age in years, $x_1$ | -.006 | .008 | -.751 | .453 |
| Level of education, $x_2$ | -.023 | .056 | -.415 | .678 |
| Years with current employer, $x_3$ | .024 | .010 | 2.362 | .018 |
| Years at current address, $x_4$ | .001 | .008 | .102 | .919 |
| Household income in thousands, $x_5$ | .031 | .002 | 14.188 | .000 |
| Debt to income ratio (x100), $x_6$ | .175 | .010 | 16.754 | .000 |
| Other debt in thousands, $x_7$ | -.019 | .026 | -.712 | .477 |

From Table 2, the MLR model is shown as in (4),

$$\hat{y} = -1.532 - 0.006 \text{ Age in years } (x_1) - 0.023 \text{ Level of education } (x_2)$$
$$+ 0.024 \text{ Years with current employer } (x_3) + 0.001 \text{ Years at current address } (x_4)$$

+ 0.031 Household income in thousands ($x_5$) + 0.175 Debt to income ratio ($x_6$)

– 0.019 Other debt in thousands ($x_7$) (4)

Next, the data will use MLR model with least quartile difference method where the analysis is shown in Table 3. The credit card debt is directly proportional with years with current employer, years at current address, household income and debt to income ratio, whereas the credit card debt is inversely proportional with age, level of education and other debt.

**Table 3.** The MLR model with LQD method

| Model | Unstandardized Coefficients | | T | Sig. |
|---|---|---|---|---|
| | B | Std. Error | | |
| (Constant) | -1.824 | .426 | -4.282 | .000 |
| Age in years, $x_1$ | -.007 | .013 | -.523 | .602 |
| Level of education, $x_2$ | -.046 | .087 | -.526 | .599 |
| Years with current employer, $x_3$ | .043 | .016 | 2.668 | .008 |
| Years at current address, $x_4$ | .008 | .013 | .620 | .535 |
| Household income in thousands, $x_5$ | .032 | .003 | 9.712 | .000 |
| Debt to income ratio (x100), $x_6$ | .194 | .016 | 12.358 | .000 |
| Other debt in thousands, $x_7$ | -.044 | .039 | -1.118 | .264 |

From Table 3, the MLR model is shown as in (5),

$\hat{y}$ = -1.824 -0.007 Age in years ($x_1$) -0.046 Level of education ($x_2$)

+ 0.043 Years with current employer ($x_3$) + 0.008 Years at current address ($x_4$)

+ 0.032 Household income in thousands ($x_5$) + 0.194 Debt to income ratio ($x_6$)

– 0.044 Other debt in thousands ($x_7$) (5)

Then, the analysis of MLR model with mean absolute deviation method is shown as in Table 4. The credit card debt is directly proportional with years with current employer, years at current address, household income and debt to income ratio, whereas the credit card debt is inversely proportional with age, level of education and other debt.

**Table 4.** The MLR model with MAD method

| Model | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|
| | B | Std. Error | | |
| (Constant) | -1.167 | .144 | -8.109 | .000 |
| Age in years, $x_1$ | -.009 | .004 | -2.040 | .042 |
| Level of education, $x_2$ | -.036 | .030 | -1.219 | .223 |
| Years with current employer, $x_3$ | .009 | .006 | 1.543 | .123 |
| Years at current address, $x_4$ | .010 | .005 | 2.133 | .033 |
| Household income in thousands, $x_5$ | .034 | .002 | 20.009 | .000 |
| Debt to income ratio (x100), $x_6$ | .154 | .006 | 23.832 | .000 |
| Other debt in thousands, $x_7$ | -.091 | .019 | -4.823 | .000 |

From Table 4, the MLR model is shown as in (6),

$$\hat{y} = -1.167 - 0.009 \text{ Age in years } (x_1) - 0.036 \text{ Level of education } (x_2) + 0.009 \text{ Years with current employer } (x_3) + 0.010 \text{ Years at current address } (x_4) + 0.034 \text{ Household income in thousands } (x_5) + 0.154 \text{ Debt to income ratio } (x_6) - 0.091 \text{ Other debt in thousands } (x_7) \quad (6)$$

Finally, to select the best model, the comparison of mean square error within all models has been made by using MSE values as in Table 5.

**Table 5.** MSE Value for 3 models

| **Model** | **MSE Value** |
|---|---|
| Multiple linear regression model | 2.3520 |
| Multiple linear regression model with LQD method | 1.8789 |
| Multiple linear regression model with MAD method | 1.9689 |

Based on the value of MSE in Table 5, the MLR model with LQD method has the lowest value of MSE. This can be interpreted that the MLR model with LQD method is the best model among all model. Therefore, the best chosen model is indicated as in (5).

From the model, it can be concluded that the credit card debt is directly proportional to years with current employer ($x_3$), years at current address ($x_4$), household income in thousands ($x_5$), and debt to

income ratio ($x_6$). Whereas the credit debt is inversely proportional to age in years ($x_1$), level of education ($x_2$) and other debt in thousands ($x_7$).

## 4. Conclusions

In conclusion, according to the MLR and both robust method; LQD and MAD, it is demonstrated that the credit card debt is directly proportional to years with current employer ($x_3$), years at current address ($x_4$), household income in thousands ($x_5$) and debt to income ratio ($x_6$). Whereas the credit debt is inversely proportional to age in years ($x_1$), level of education ($x_2$) and other debt in thousands ($x_7$). Based on the value of MSE for this data, the MLR model with applied LQD method have the lowest value of MSE. This can be interpreted that the MLR model with applied LQD method is the best model among all model. Therefore, this study might serves as a reference for the bank organisation using multiple linear regression and application of robust method in multiple linear regression, so that they can better understand their options and able to make an informed choice that is best aligned with their goals for inference regarding the credit card debt.

## Acknowledgement

This research work is supported by FRGS (Fundamental Research Grant Scheme) grant (Vot 1498), Ministry of Higher Education, Malaysia.

### Reference

[1] Betti G, Dourmashkin N, Rossi M and Verma V and Yin Y 2001 *Study of the Problem of Consumer over indebtedness* : Statistical Aspects, Draft Final Report (London Macro International Social Research)
[2] Soman D 2001 Effects of payment mechanism on spending behaviour .the role of rehearsal and immediacy of payment *The Journal of Consumer Research* 27 (**4**) 460-474
[3] Bakar N M A and Tahir, I M 2009 Applying Multiple Linear Regression and Neural Network to Predict Bank Performance *International Business Research* 2 (**4**) 176-183
[4] Amir E K, Adlar J K and Andrew W L 2010 *Consumer Credit Risk Models via Machine-Learning Algorithms* (MIT Sloan School of Management)
[5] Regis D E and Artes R 2015 Using multi-state markov models to identify credit card risk *Production* 25 (**2**)
[6] Bin Shafi M A, Bin Rusiman M S and Che Yusof N S H 2014 Determinants Status of Patient After Receiving Treatment at Intensive Care Unit: A Case Study in Johor Bahru. *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings 30 September 2014*, 6914150 80 – 82
[7] Rusiman M S, Nasibov E and Adnan R 2011 The Optimal Fuzzy C-regression Models (OFCRM) in Miles per Gallon of Cars Prediction, *Proceedings – 2011 IEEE Student Conference on Research and Development, SCOReD 2011*, 6148760 333-338
[8] Shafi M A and Rusiman M S 2015 The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malaysia *Applied Mathematical Sciences* 9 (**56**) 2749-2759
[9] Rusiman M S, Hau O C, Abdullah A W, Sufahani S F, Azmi N A 2017 An Analysis of Time Series for the Prediction of Barramundi (Ikan Siakap) Price in Malaysia *Far East Journal of Mathematical Sciences* 102(**9**) 2081-2093
[10] Nor M E, Rusiman M S, Mohamad N A I and Lee M H 2017 Directional Change Error Evaluation in Time Series Forecasting *AIP Conference Proceedings* 1830 (**1**) 080013
[11] Sufahani S, Che-Him N, Khamis, A, Rusiman M S, Arbin N, Yee C K, Ramli I N, Suhaimi N A, Jing S S and Azmi Z A 2017 Descriptive Statistics with Box-Jenkins and Marketing Research

for Jewellery Company in Malaysia *Far East Journal of Mathematical Sciences* 101(**10**) 2151-2161

[12] Kevin R 2010 *A Culture of Debt : A Study of Credit Card Spending in America* (University of Harvard)

[13] Baker W L, Michael W C, Cappelleri J C, Kluger J and Coleman C I 2009 Understanding heterogeneity in meta-analysis: the role of meta-regression *International Journal of Clinical Practice* 63 (**10**) 1426-1434

[14] Worster A, Fan J and Ismaila A 2007 Understanding Linear and Logistic Regression Analyses *Canadian Journal of Emergency Medicine* 9 (**2**) 111-113

[15] Bernholt T, Nunkesser R and Schettlinger K 2007 Computing the Least Quartile Difference Estimator in the Plane *Computational Statistics & Data Analysis* 52 763 – 772

[16] Bernholt T 2005 Computing the least median of squares estimator in time o(nd). *Proceedings of ICCSA 2005*, 3480 697–706