International Journal of Economics and Finance; Vol. 6, No. 10; 2014
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Censored Regression Techniques for Credit Scoring: A Case Study for the Commercial Bank of Zimbabwe (Bulawayo)

Thandekile Hlongwane[1], Precious Mdlongwa[1], Hausitoe Nare[1] & Isabel L. Moyo[1]

[1] National University of Science and Technology, Department of Statistics and Operations Research, Bulawayo, Zimbabwe

Correspondence: Precious Mdlongwa, Statistics and Operations Research Department, National University of Science and Technology, P O Box AC939, Ascot, Bulawayo, Zimbabwe. Tel: 263-9-282-842. E-mail: precious.mdlongwa@nust.ac.zw

Received: July 3, 2014 Accepted: July 23, 2014 Online Published: September 25, 2014
doi:10.5539/ijef.v6n10p227 URL: http://dx.doi.org/10.5539/ijef.v6n10p227

Abstract

Credit creation is the main income generating activity for banks. However this activity involves huge risks to both the lender and the borrower. The risk of a trading partner not fulfilling his or her obligation as per the contract on due date or any time thereafter can greatly jeopardise the smooth functioning of a bank's business. Credit risk therefore is one of the greatest concerns to most banking authorities and banking regulators. This paper is aimed at coming up with a model that can be used by the Commercial Bank of Zimbabwe in calculating the risk associated with credit scoring. The data set used covered personal loans from January 2010 to January 2012. Linear and Buckley James regression tests were employed to find the explanatory variables influencing time to default and repayment. In investigating customer classification, linear discriminant analysis was applied. Age, marital status, loan purpose and time at current job were found to be linearly related to time to default. Time to repayment was found to be linearly related to age, marital status and loan purpose. 67.5% of the original cases were found to be correctly classified. Buckley James regression out performed linear regression hence it was found to be the most suitable method in determining variables affecting risks in loan lending.

Keywords: credit risk, regression, default, repayment, credit scoring

1. Introduction

There is no instrument that can be used to predict the future accurately but when dealing with lending, banks try to predict the outcome of that loan. Banks have to find out the possibility of a customer either defaulting on the loan or not. Like all debt instruments, a loan entails the redistribution of financial assets over time, between the lender and the borrower. The borrower initially receives an amount of money from the lender, which he pays back, but sometimes not always in regular installments, to the lender.

Credit scoring uses quantitative measures of the performance and characteristics of past loans to predict the future performance of loans with similar characteristics (Caire & Kossman, 2003). Credit scoring is a scientific method of assessing the credit risk associated with new credit applications. Statistical models derive predictive relationships between application information and the likelihood of satisfactory repayment. Models are empirically designed; that is, they are developed entirely from information gained through prior experience. Therefore, credit scoring is an objective risk assessment tool, as opposed to subjective methods that rely on a loan officer's opinion. Clearly, credit scoring is a risk management tool. Scoring systems can help a bank ensure more consistent underwriting and can provide management with a more insightful measure of credit risk.

Credit scoring cannot predict individual loan loss; rather it predicts the likelihood or odds of a bad outcome, as defined by each bank; usually this will be some level of average or total days in arrears at which associated costs make the loans unprofitable, nor should a credit scoring system alone approve or reject a loan application; rather the underwriter must decide how he or she will incorporate the credit score into the loan review. Finally, credit scoring is not meant to increase approval rates; rather, it promotes consistency and efficiency while maintaining or reducing historic delinquency rates. It also allows the users to focus their attention and time on applications that are not obvious approvals or obvious declines (Caire & Kossman, 2003). Hence the research aims at coming up with a censored regression model that can be used in calculating the risk associated with credit.

CBZ is a registered commercial bank in Zimbabwe which was established in 1980 offering a wide range of innovative banking and financial services to personal and corporate customers. Banks generally provide a variety of services that include but are not limited to cash and cheque deposits and withdrawals; provision of credit facilities such as loans, overdrafts and credit cards; processing payments; asset financing; mortgages; clearing; foreign exchange; money transfer; advisory services; safe keeping services; and custodial services (Ambira & Kemoni, 2011). CBZ has banking products which include savings accounts, current accounts, foreign currency accounts, fixed deposits, cash manager accounts, personal loans, private home and commercial loans, micro leasing, asset finance, agribusiness finance, micro finance loans, offshore credit, and business loans. The company also offers foreign currency services, trade finance, international banking, investment banking, small to medium enterprises financing, treasury management, wealth management, agribusiness, custodial services, and bancassurance. Credit scoring at CBZ is done using a credit score sheet which is a standard document with specific attributes used when appraising a loan application. The score sheet is user friendly as there are guidelines on every attribute. The credit risk faced by CBZ is that of the customer defaulting on their loan. Hence the research aims at coming up with a censored regression model that can be used in calculating the risk associated with credit.

2. Literature Review

Decisions on whom to grant credit, and of how much credit to grant, originally relies purely on the skill of a loans officer. The loans officer uses his experience and personal judgement, and guided by attributes that affect the credit worthiness of the applicant, he then makes a decision on whether or not to grant credit. The attributes deemed most important are referred to collectively as the five Cs of credit (Thomas et al., 2002). They are:

1). Character - The willingness to pay debt. For example, how long has the applicant been at their current job?

2). Capacity - The borrower’s capacity to pay the debt. Wages and other income are major determinants here.

3). Collateral - Possessions that might be used to secure the debt are classed as collateral. For a mortgage, the home purchased is used as collateral.

4). Capital - A well-resourced individual is more likely to be granted a loan.

5). Conditions - Current and projected economic conditions are also taken into account.

A number of factors led to the introduction of automated credit scoring in the 1940’s and according to Durand (1941), at the end of World War II there was an explosion in the demand for credit and it became clear that the subjective methods did not scale well to large numbers of applicants. The credit explosion, spurred on by the introduction of credit cards a few decades later, motivated lenders to automate the credit granting decision giving birth to objective credit scoring systems. In parallel with the growth of credit demand, increases in computing power made it possible to analyse large quantities of data with (relative) ease. More recently, the development of scoring systems has been driven by the regulatory environment. As a part of the capital adequacy requirements placed upon banks with the introduction of the Second Basel Accord (Basel Committee for Banking Supervision, 2001), institutions are required to closely monitor the risks associated with their loan portfolios. Since the introduction of the first credit scoring systems, a number of statistical and mathematical methods have been used. Most techniques have a statistical background, such as Markov Analysis, Linear Regression, Logistic Regression and the Buckley James method.

The credit score is determined by a complex formula that takes into account many different factors. Credit scoring models compute a person’s score primarily from information contained in his credit report. The models might also take information from credit applications into consideration, including the person’s age, time with bank (months), number of dependants, time at current address (months), time at current job (months), sex, refinancing of other financial institution’s loan flag, self-employed flag, marriage status and purpose of loan. The person’s payment history reflects the various accounts that he has, including credit cards, mortgage loans, and retail accounts. Collections, foreclosures, lawsuits, and other collection items also fall into this factor. Each factor is given a weight (Credit Risk Scoring Analytics, Issue No: 0710511).

Historically, a credit officer uses information relating to the creditworthiness of an applicant to determine whether or not to grant a loan. Current credit scoring systems work in much the same, although objective, way. Assume that the customer population consists of two classes, good and bad. The information that a customer provides when they apply for a loan is used by banks to determine which group the customer is likely to belong to. Rather than being examined in a subjective way, the information is coded to form quantitative variables that can be input into a statistical model. For an individual, if there are k explanatory variables, they are collected as a vector, to form the input to the model. The explanatory variables can then be used to produce a score to estimate the probability, p, of that individual belonging to the good or bad class. The relationship between the explanatory variables and the probability of default is usually found by fitting to a historical set of completed loans, some of which are bad.

$$X^T = (x^1,\ x^2,\ \dots\ x^k) \tag{1}$$

Credit scoring techniques were originally developed to help organisations automate the credit granting decision. As a result, the primary aim of a traditional credit scoring system is to classify potential customers as either being good or bad so the appropriate action can be taken. A bad customer may be deemed as one who fails to repay the loan in full, but this definition can be expanded to cover a range of undesirable behaviour. Surveys by Rosenberg and Gleit (1994), Hand and Henley (1997) outline the different modelling techniques that can build such systems.

The definition of bad can be somewhat arbitrary and is often driven by regulatory demands. While the definition can include early repayment, churn, or fraudulent activity, the most common definition of bad is default. Default could be taken as one missed payment, three consecutive missed payments, or perhaps when the debt becomes unrecoverable. If the definition of bad is too stringent, or not stringent enough, it may have a negative impact on the quality of the final scorecard (Siddiqi, 2005).

3. Methodology and Data

Assume that a body of loans data has been collected with *n* data points, perhaps it was collected to construct a classification score card. Typically, CBZ would use the data to predict the chance of that loan being in default at some cut off date (Siddiqi, 2005). Now suppose that instead of estimating the probability that the loan will go bad, we wish to estimate the time to default, T_d and time to repayment, T_r, where time to default and time to repayment are assumed to be independent events. As the observation period may end while the loan is still underway there will be censoring at *t* months, where *t* is the length of the maximum observation period for the loan. The total observed time of the loan, *T* is then

$$T = min(T_d,\ T_r,\ t) \tag{2}$$

If two events are assumed to be independent, then the overall hazard function for the loan can be expressed as

$$h(y) = h_d(y) + h_r(y) \tag{3}$$

where h_d and h_r are the individual hazards for each mode of failure. One implication of the independence assumption is that when estimating time to default, early repayment is viewed as a censoring mechanism. Time to repayment can be modelled separately and any defaults viewed as censored observations. Hence, for modelling default, the censoring indicator is given as:

$$\delta_d = \begin{cases} 1 & If\, T_d \le t \quad and \quad T_d \le T_r \\ 0 & otherwise \end{cases} \tag{4}$$

and for modelling repayment:

$$\delta_r = \begin{cases} 1 & If\ T_r \le t \quad and \quad T_r \le T_d \\ 0 & otherwise \end{cases} \tag{5}$$

Figure 1 shows different mechanisms acting when considering default and how they affect the coding of the data. Figure 2 shows how the same events are coded differently for repayment. The hollow circle refers to a censored point.

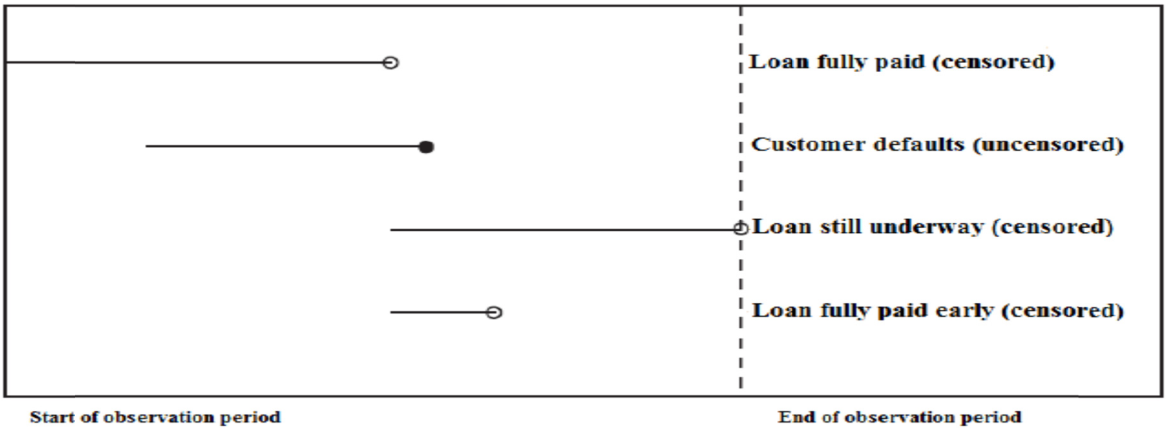

Loan fully paid (censored)
Customer defaults (uncensored)
Loan still underway (censored)
Loan fully paid early (censored)
Start of observation period
End of observation period

Figure 1. Time to default with censored loans data

Source (Glasson, 2007).

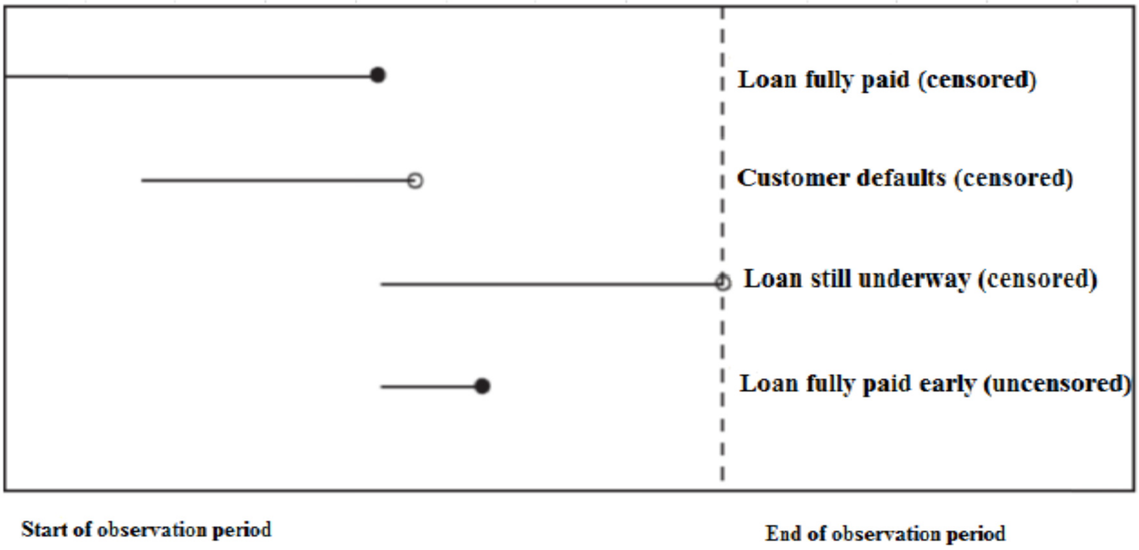

Loan fully paid (censored)
Customer defaults (censored)
Loan still underway (censored)
Loan fully paid early (uncensored)
Start of observation period
End of observation period

Figure 2. Time to repayment with censored loans data

Source (Glasson, 2007).

For the purposes of this study, random censoring is assumed. This is because the observation period is fixed, as in Type I censoring, but as the entry time is random, *t* is scattered rather than fixed. The pattern of censoring seen here is typical of what might be seen in a clinical trial for which survival analysis techniques are often used. The advantage of using censored data is that the models are able to reflect current borrowing patterns rather than relying on historic trends

Suppose again for the applicant, a vector of explanatory variables, $X^T = (x_1, x_2, \ldots, x_k)$ is recorded. As with a classification scorecard, *x* contains application characteristics that will affect the default and repayment of the loan. Therefore the time to default is a function of both the explanatory variables and time. The time to default can then be estimated by any of the various survival regression methods. Depending on the type of method used, either an estimated hazard will be produced in the case of hazard-based methods, or the estimated mean survival time in the case of Buckley-James method. The advantage of the survival analysis over classification methods is that the survival models give an estimate of the timing of the defaults and the repayments rather than just the probability of occurrence. This is one of the main novelties of applying lifetime modelling to scorecards. While time to repayment is the event of interest under this framework, it is equally valid to use time to early repayment. Therefore, the status of some data points will change. In particular, loans paid on time will be now classified as censored. Further work may have to be done to work out whether one method is consistently better than the other for prediction over a number of datasets.

3.1 Linear Discriminant Analysis (LDA)

LDA will be used to classify customers into classes that is either good or bad. A score, Z, will be constructed which is a linear function of the explanatory variables *x*,

$$z = \beta^T x = \beta_1 x_x + \beta_2 x_2 + \ldots + \beta_k x_k \tag{6}$$

$$\text{Let} \quad M = \frac{\beta^T (m_G - m_B)}{\sqrt{\beta^T} \sum^B} \tag{7}$$

Where m_G and m_B are the vector group means for the good and bad classes respectively and Σ is the common covariance matrix.

Discriminant function analysis determines which continuous variables discriminate between two or more naturally occurring groups. In LDA, the explanatory variables are the predictors and the dependent variables are the groups. LDA is usually used to predict membership in naturally occurring groups. It answers the question: can a combination of variables be used to predict group membership? Several variables are included in this study to see which ones contribute to the discrimination between groups.

Discriminant function analysis will be broken into a 2-step process:

1). Testing significance of a set of discriminant functions.

This step is computationally identical to MANOVA. There is a matrix of total variances and covariances; likewise, there is a matrix of pooled within-group variances and covariances. The two matrices are compared via multivariate F tests in order to determine whether or not there are any significant differences (with regard to all variables) between groups. Multivariate test is performed firstly, and, if statistically significant, proceeds to see which of the variables have significantly different means across the groups. Once group means are found to be statistically significant, classification variables is undertaken. LDA automatically determines some optimal combination of variables so that the first function provides the most overall discrimination between groups, the second provides second most, and so on. Moreover, the functions will be independent or orthogonal, that is, their contributions to the discrimination between groups will not overlap. The first function picks up the most variation; the second function picks up the greatest part of the unexplained variation, etc. Computationally, a canonical correlation analysis is performed that will determine the successive functions and canonical roots.

2). Classification.

Classification is then made from the canonical functions. Subjects are classified in the groups in which they had the highest classification scores. The maximum number of discriminant functions will be equal to the degrees of freedom, or the number of variables in the analysis, whichever is smaller. One of the main criticisms of linear discriminant analysis as a credit scoring method involves the assumptions of distributional form (Eisenbeis, 1978): Firstly, the assumptions require that the covariance matrices of the predictor variables are equal for the two groups; furthermore, the predictor variables are required to follow a multivariate normal distribution. In credit scoring applications the predictor variables are often discrete or follow otherwise non-normal distributions. This clearly violates the second assumption. However, even if the normality assumption is violated, linear discriminant analysis is still widely applicable in separating groups and that the violation only affects the validity of significance tests (Hand & Hanley, 1997).

When interpreting multiple discriminant functions, which arise from the analysis of than two groups and more than one continuous variable, the different functions are first tested for statistical significance. If the functions are statistically significant, then the groups can be distinguished based on predictor variables. Standardized *β* coefficients for each variable are determined for each significant function. The larger the standardized *β* coefficient, the larger is the respective variable's unique contribution to the discrimination specified by the respective discriminant function. In order to identify which independent variables help cause the discrimination between dependent variables, one can also examine the factor structure matrix with the correlations between the variables and the discriminant functions. The means for the significant discriminant functions are finally examined in order to determine between which groups the respective functions seem to discriminate.

3.2 Linear Regression

Linear regression models will be used to formulate a credit scoring model, assume a linear model where the probability p that an applicant is bad is related linearly to k explanatory variables,

$$p = \boldsymbol{\beta}^T x = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{8}$$

Where **β** is the vector of parameters ($\beta_1, \beta_2, \ldots, \beta_k$).

3.3 The Buckley James Method

The Buckley James method will be used to correct any bias present in linear regression with censored data by replacing censored points with their expected values, $E(Y_i|Y_i > t_i)$ This is equivalent to creating a new response variable $Y_i^*(b)$ defined as

$$Y_i^*(b) = Y_i\delta_i + E(Y_i|Y_i > t_i)(1-\delta_i) \tag{9}$$

Where b is the arbitrary slope to be estimated by the algorithm, Y is the survival random variable, δ is the censoring indicator, Y_i^x is the response for the i^{th} observation and t_i is the censoring time for the i^{th} observation.

3.4 Simulation

Monte Carlo Simulation is used to compare Linear and Buckley James regression and then select the best model to use to calculate the risk involved in loan lending. Monte Carlo simulation, or probability simulation, is a technique used to understand the impact of risk and uncertainty in financial, project management, cost, and other forecasting models. Thus in our case simulation is used to compare all the regression methods used in this study.

3.5 Data

Key characteristics about debtors and debts includes: residential status, employment status, marital status, time at address, time in occupation, time at the bank, loan purpose, sex and age. Monthly performance data for each loan was recorded from the time each loan was opened until January 2012. The monthly performance data for each loan included whether the loan was still under way, whether the loan was more than 30 days in arrears, or if the loan had been fully repaid. However the data was for 200 customers who had either defaulted or repaid their loans.

The general format of the monthly performance data was supplied as a Structured Query Language, (SQL), dataset; SQL being the data analysis package used by CBZ. Rejected applicants were not included in the data because no reject inference was to be carried out. For each month, a loan could be (G) Good, (B) Bad, or closed (blank value). Good refers to a loan that was not 30 days behind in repayments. Bad refers to a loan that, at any time prior to that month, had been more than 30 days behind in repayments.

Take for example Loan 1 which is in the first data row of Table 1 above, it was opened in January, 2010 as that is the position of the first G, and was closed (repaid in full) in April, 2010. Hence the survival time, z, for this loan was 3 months, and because repayment was the observed mode of failure, $\delta_r = 1$. The loan was not seen to default, accordingly $\delta_d = 0$.

Table 1. Loan performance

Loan No.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	G	G	G	.	.	.	.	.	.	.	.	.
2	.	G	G	G	G	G	G	G	G	G	G	G
3	G	G	G	B	B	B	B	B	B	B	B	B
4	.	.	.	G	G	G	G	G	G	G	G	G

Table 2 shows the results from converting the examples in Table 1 into survival times. This conversion is needed because most survival regression programs require the data to be expressed as a combination of survival times and censoring indicators. A loan is generally classified as bad if it is in default at any stage in the 12 months after opening. If the loan is fully paid off, or is still under way at the 12 month cut-off, the customer is classified as good

Table 2. Survival times of the loans

Loan	Observed Times	δ_d	δ_r
1	3	0	1
2	11	0	1
3	3	1	0
4	9	0	1

4. Results

The package SPSS was used for analysing data using linear regression, Buckley James regression and linear discriminant analysis. In the following subsections are the results obtained and discussions.

4.1 Scatter Plots

Figure 3(a) shows a positive linear relationship between time to default and age. Age consists of two groups the young (< 50 years) and the old (≥ 50 years). For older people default is very high as compared to the young ones. This is because of the fact that the young ones have more years to work, they are ambitious, they want more assets and they are energetic and are more willing to work unlike older people. With the current economic Zimbabwe situation, pension given to retired workers is very low to cover for major expenses thus this puts older people at a default risk. In Figure 3(b) no relationship is seen between time to default and sex. This is because both males and females have an equal opportunity of getting the same income. Even women have their own business because of women empowerment programmes allowing women to work and even run families thus everyone is likely to default despite the fact that they are males or females. Figure 3(c) shows a relationship between time to default and marital status. Those that were single (1) were at most risk of defaulting than the married (2) people. For the married, the spouse could help with the finances unlike the single who face paying back on their own.

Loan purpose was categorised into groups which are:

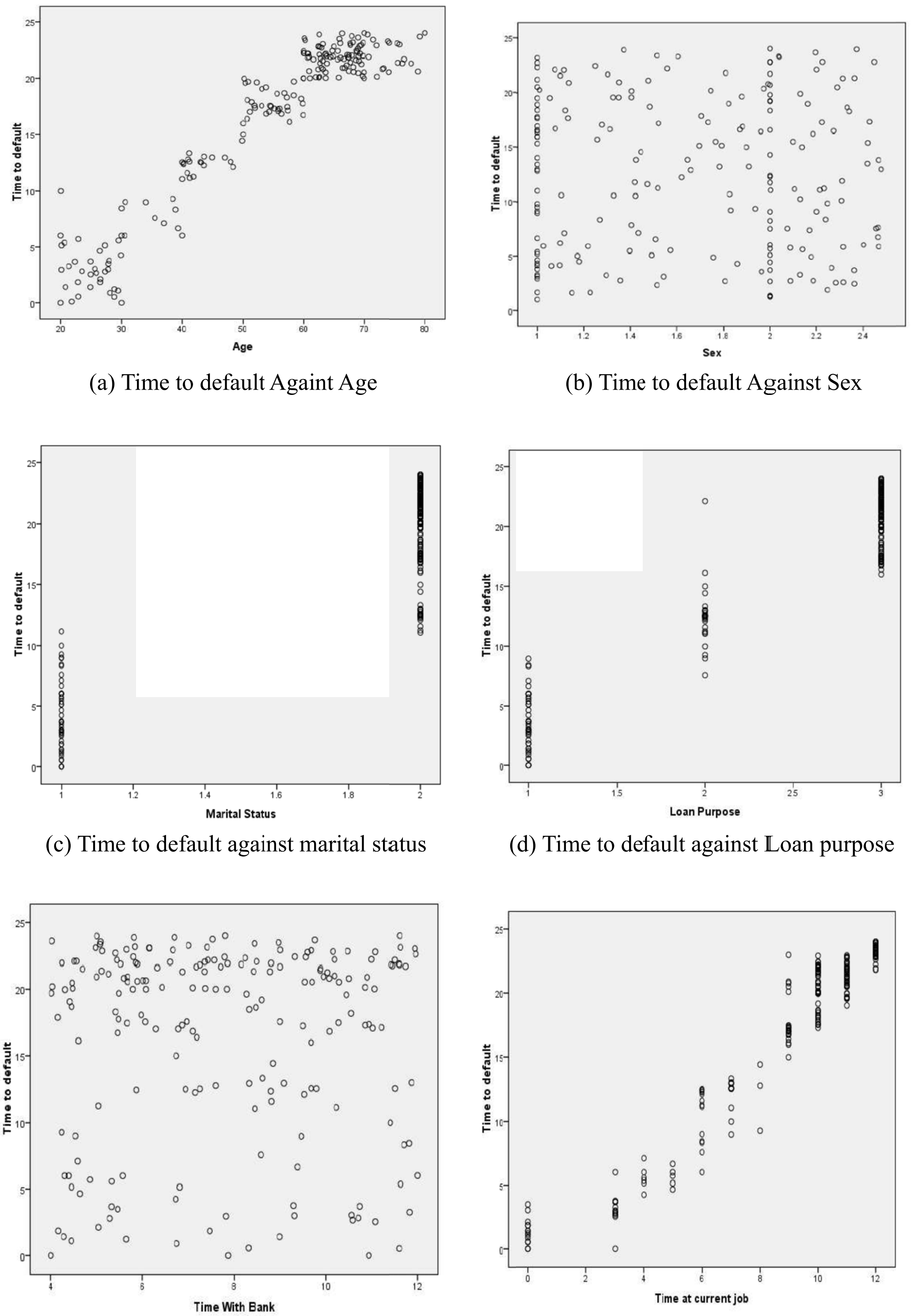
1). High risk purpose loans, that is loans for starting a business.

2). Medium risk purpose loans, that is loans for buying a car.

3). Low risk purpose loans, that is loans for paying school fees.

For high risk loans default occurs within the first months, could be because of the business plans that have not failed as such but taking off is hard and income starts flowing in later and thus the customer has a zero income for that month and the ones to follow up to 9 months. For medium risk loans default starts from 7 months to 16 months. For low risk loans default starts from a year upwards. Low risk loans have most clients because they are easy to pay back and most customers with low income jobs can afford to repay the loan on monthly basis. In Figure 3(d), there is no relationship between time to default versus time with bank meaning that default is not influenced with the time a customer has with the bank. It shows that whether a customer has 2 years or more with the bank it does not mean that they will not default, the same applies for less time with the bank. Default time is not predictable when considering time with bank. In Figure 3(e), there is a positive linear relationship between time to default and time at current job. There was positive correlation. Time at current job shows how stable a customer is, thus, the more stable one is, the less the chances of defaulting. Customers have a tendency to default, through mostly salary divert but if they have proof of where they are working the employer can help clear the credit of the customer by paying the customer’s salary directly into the bank.

Figure 4(a) shows a linear relationship between time to repayment and age.For older people repayment is very low as compared to the young ones, can relate to Figure 3(a). According to Figure 4(b) no relationship is seen between time to repayment and sex. Figure 4(c) shows a relationship between time to repayment and marital status. Those that were married (2) took longer to repay than the single (1) people. Figure 4(d) shows that Repayment was feasible in all the loan purpose categories. Figure 4(e) shows no relationship between time to repayment versus time with bank meaning that repayment is not influenced with the time a customer has with the bank. Figure 4(f) shows a positive correlation between time to repayment and time at current job. Time at current job shows how stable a customer is, thus, the more stable one is the higher the chances of early repayments. Following are models were obtained after removing the insignificant variables that is sex, time with the bank and loan purpose



(a) Time to default Againt Age

(b) Time to default Against Sex

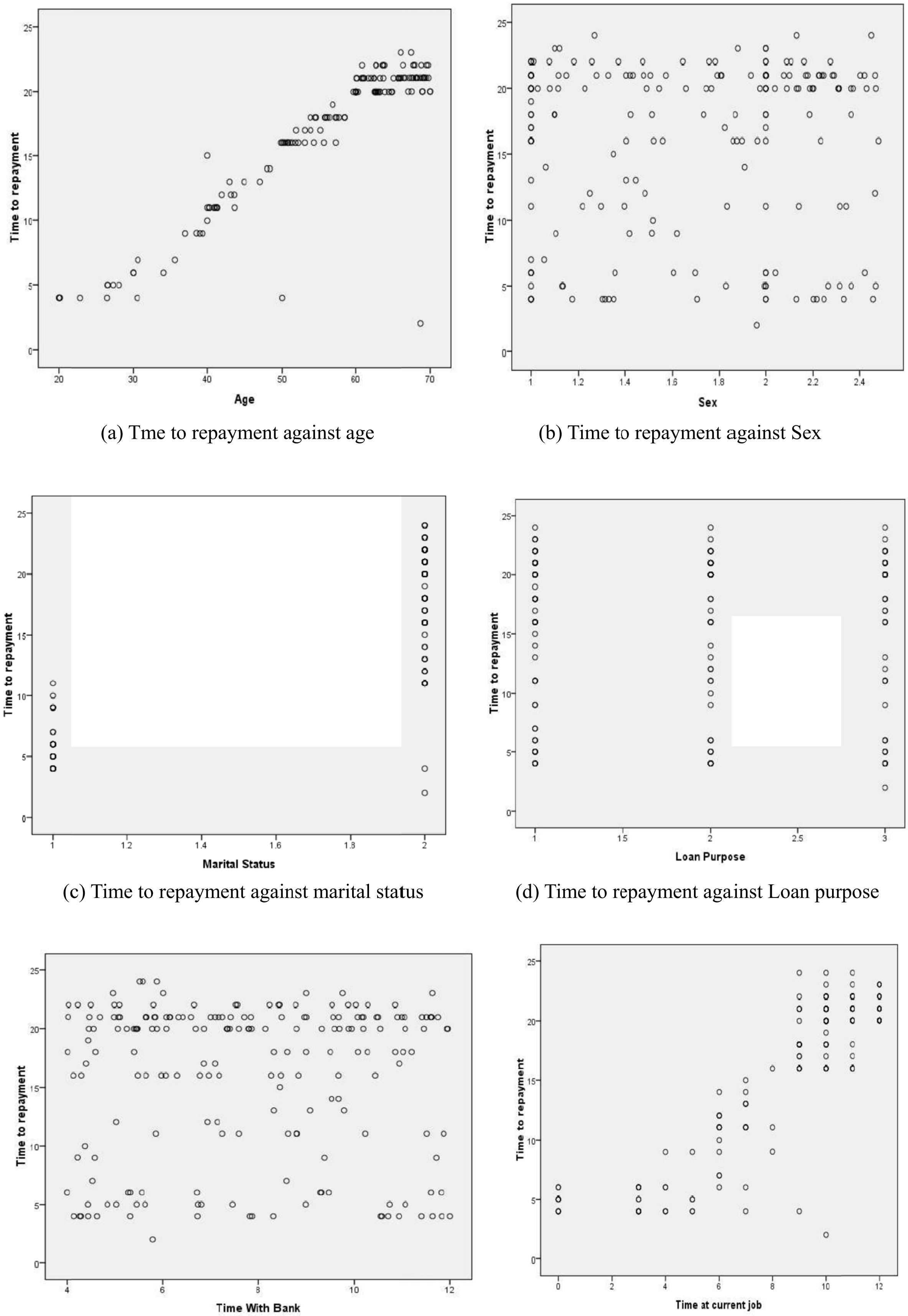
(c) Time to default against marital status

(d) Time to default against Loan purpose

(e) Time to default against time with bank

(f) Time to default against time at current job

Figure 3. Time to default against variable



(a) Tme to repayment against age

(b) Time to repayment against Sex

(c) Time to repayment against marital status

(d) Time to repayment against Loan purpose

(e) Time to repayment against time with bank

(f) Time to repayment against time at current job

Figure 4. Time to repayment against variable

4.2 Defaulting Regression Model

$$Z=-5.514+0.024A+4.043\ MS+1.541TCJ \qquad (10)$$

Table 3. Coefficient summary for defaulting regression model

Model	Coefficient	*t-value*	*sig*
Constant	-5.514	-9.158	0.000
Age(*A*)	0.024	2.947	0.004
Marital Status(*MS*)	4.043	7.621	0.000
Time at Current Job(*TCJ*)	1.5411	23.066	0.000

Model (10) has $R^2 = 0.954$ indicating that the model is good with 95.4% of variability in defaulting explained by age, marital status and time at current Job.

Model Validation.

Normal P-P Plot of Regression Standardized Residual

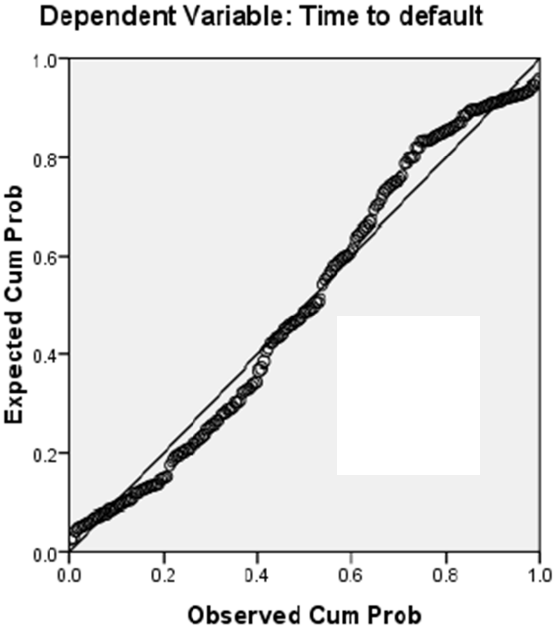


Figure 5. Normal P-P plot for defaulting model

Figure 5 shows that the errors for model (10) are almost normally distributed.

4.3 Repayment Regression Model

$$Z=-4.039+0.049+5.210MS+0.965TCJ \qquad (11)$$

Table 4. Summary coefficients for repayment model

Model	Coefficient	*t-value*	*sig*
Constant	-4.039	-4.487	0
Age(*A*)	0.049	3.635	0
Marital Status(*MS*)	5.21	6.014	0
Time at Current Job(*TCJ*)	0.965	8.883	0

Model (11) has $R^2 = 0.844$ indicating that the model is good with 84.4% of variability in defaulting explained by age, marital status and time at current Job.

Model Validation

Normal P-P Plot of Regression Standardized Residual

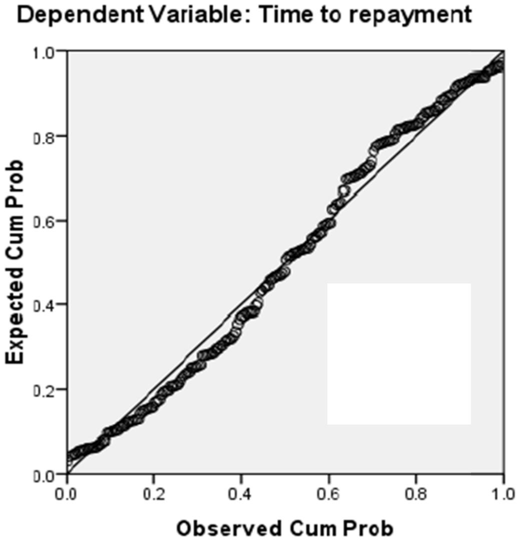


Figure 6. Normal P-P plot for repayment model

Figure 6 shows that the errors for model (11) are almost normally distributed.

4.4 Discriminant Analysis

This analysis is used to classify customers into two groups that is good or bad. Z score was constructed using (6). Using (7) a good result is obtained if Box's **M** value is greater or equal to 20.

Table 5 Univariate ANOVA's results

Variable	**Wilks' Lamba**	**F**	**df1**	**df2**	**sig**
Age(A)	1	0.094	1	198	0.752
Sex (S)	0.998	0.373	1	198	0.542
Loan Purpose (LP)	0.947	11.02	1	198	0.001
Marital Status (MS)	0.991	1.715	1	198	0.192
Time with the bank (TB)	0.9	22.107	1	198	0
Time at current Job (TCJ)	0.995	0.98	1	198	0.323

Table 5 is the representation of the results of Univariate ANOVA's, carried out for each explanatory variable. Here, only time with bank and Loan Purpose differed significantly (Sig α = 0.05) for the two groups, that is, good and bad.

Table 6. Box's test of equality of covariance matrices

	Test Results
Box's **M**	24.355
F Approx	1.122
	21
	1.031
	0.315

In Table 6, for the Box's M tests the null hypothesis is that the observed covariance matrices of the dependent variables are equal across groups, since **M** =24.355 is > 20. The significance value of 0.315 indicates that the data do not differ significantly from multivariate normal. This means that we can proceed with the analysis.

Table 7. Summary of canonical discriminant functions

Function	Eigen Value	% of Variance	Cumulative	Canonical Corr
	1.177	100	100	0.888

An eigen value indicates the proportion of variation explained between-groups sums of squares divided by within-groups sums of squares. The larger the eigen value, a value ≥ 1.1, the stronger the function and the discriminatory power. Therefore from Table 7 we have a stronger function since our eigen value is 1.177. The canonical relation is a correlation between the discriminant explanatory variables and the levels of the dependent variable. A high correlation indicates a function that discriminates well. The present correlation of 0.888 is extremely high between the dependent variables, that is, time to default and time to repayment and the significant explanatory variables.

Table 8. Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	sig
1	0.85		31.738	6

Wilk's Lambda is the ratio of within-groups sums of squares to the total sums of squares. This is the proportion of the total variance in the discriminant variables not explained by the differences among groups. From Table 8 Lambda is 0.85 which is close to 1 this shows that the group means are almost equal to 1 and all the variance is explained by factors other than difference between those means. Here Lambda has a significant value, thus, the group means appear to differ.

Table 9. Classification table for loans

	Group	Good	Bad	Total
Original Count	Good	60	26	86
	Bad	39	75	114
%	Good	69.8	30.2	100
	Bad	34.2	65.8	100

67.5% of the original cases were correctly classified.

In Table 9, a classification result is a simple summary of number and percentage of subjects classified correctly and incorrectly. For our data 67.5% of original grouped cases were correctly classified meaning that 32.5% customers were misclassified. High losses incurred since the cost of misclassification is the same for both groups, this prompts bank failure.

4.5 Buckley James Results

Buckley-James estimation was done in simple linear regression applied to the unsecured personal loans data. The number of limiting values of Buckley-James estimates exhibits chaotic behaviour. Table 10 shows a summary of the coefficient values after the bias had been removed from the linear regression models. In the time to repayment model using Buckley James time at current job was insignificant at 5% level of significance showing that there is poor score assignment for the variable.

From Table 10, Sex and time with the bank have sig. values >0.10 indicating that they do not contribute to the discriminant model otherwise all values are significant. Table 10 also suggests age as best variable followed by time at current Job, marital status and loan purpose.

Table 10. Buckley james regression summary

Model	Variable	Default Coefficients	Repayment Coefficients
	Constant	175.08	20.76
	Age(A)	0.543	0.321
	Marital Status (MS)	0.98	1.89
	Loan Purpose (LP)	-2.5	2.13
	Time with the bank (TB)	0	0
	Time at current Job (TCJ)	0.02	0

The equations after bias correction are as follows:

Defaulting Buckley James Model:

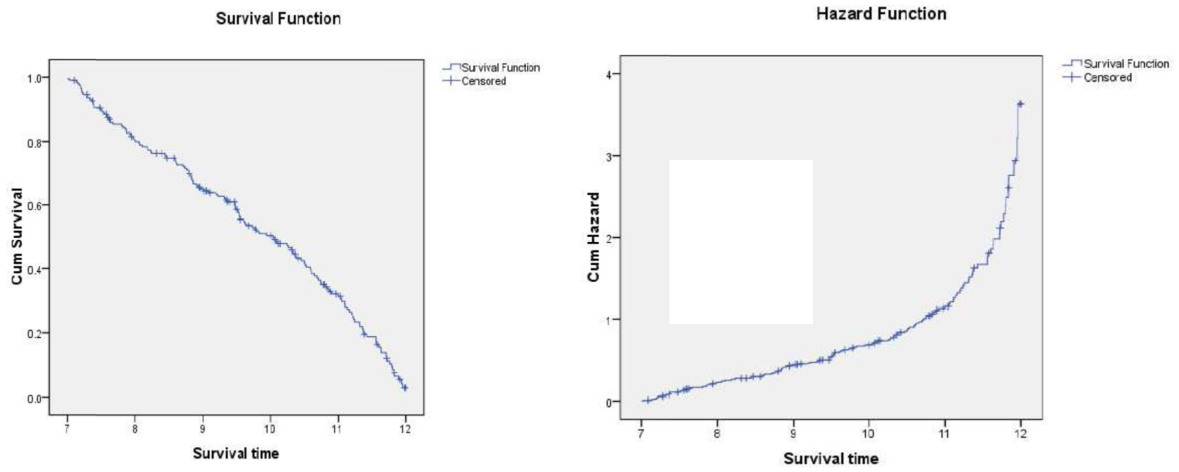
$$Z=175.08+0.543Age+0.98MS-2.5L+0.02TCJ \tag{12}$$

Repayment Buckley James Model:

$$Z=20.76+0.321Age+1.89MS+2.13L.P \tag{13}$$

4.6 Survival Analysis

Survival Function for default

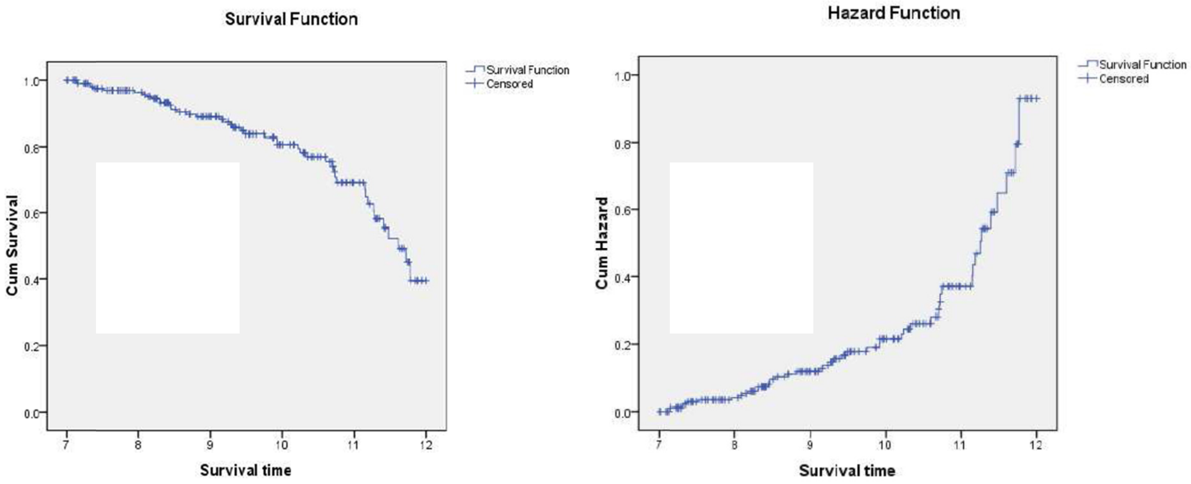


(a) (b)

Figure 7. Survival Analysis and Hazard function for time to default

Figure 7(a) shows the rate of default from 7 months up to the end of the year by the Kaplan - Meier estimator. The rate of default decreases gradually as the time increases. The rate of default decreases gradually as the months go down and it is vice versa for figure 7(b). In 7 months there were 100% customers that had defaulted. As the months increased the defaulting rate decreased by 20% up to the year end. At the end of the year there were no defaults.

Survival Function for repayment



(a) (b)

Figure 8. Survival analysis and hazard function for time to repayment

Figure 8 shows the repayment graphs which were also done using the Kaplan - Meier.Figure About 40% of the customers had repaid their loans. At 7 months there was cumulative survival of 100% then the repayment rate decreased gradually up to the cumulative survival of 0.4. It became constant there up to the end of the year.

4.7 Simulation Results

Table 11. Simulation table for comparing linear regression and buckley james regression.

Variable	Measure	Pearson's R	Std error	t	sig
Age (A)	Linear Regression	0.937	0.025	9.632	0.000
	Buckley James	0.996	0.001	93.123	0.000
Loan Purpose (LP)	Linear Regression	-1	0		
	Buckley James	-0.935	0.028	-9.54	0.000
Marital Status (MS)	Linear Regression	0.849	0.076	5.791	0.000
	Buckley James	0.852	0.072	5.875	0.000
Time at current Job (TCJ)	Linear Regression	0.901	0.041	7.481	0.000
	Buckley James	0.936	0.024	9.624	0.000

The Pearson's R measures the strength of the linear relationship. A positive Pearson's R value means positive correlation, that is, if one variable increases in value, the other variables also increase in value. A negative Pearson's R value means negative correlation, that is, if one variable increases in value, the other variables decrease in value. The higher the value of T, the stronger the function at the same time the function should be significant at 5% level of significance Table 11 shows that Buckley James has high person coefficient and t-values implying it outperforms the simple linear regression model

The study that has been done has revealed some factors that affect time to default and repayment in loan lending done at CBZ. Focusing on the findings from the study certain conclusions and recommendations can be made.

Credit scoring is done in most banks but it is associated with more risks, that is, default risks and repayment risks. Loan lending is the main income for CBZ thus very much precaution should be taken when lending because failure to manage loans leads to bank failure.

This paper involved censored regression techniques (explained in section 3) being done on the unsecured personal loans dataset obtained from the bank. When regression was done, linearity was seen between:

1. Time to default and age, marital status, loan purpose and time at current job.

2. Time to repayment and age, marital status and loan purpose.

It was found that older people and single people have high risk of default. According to Buckley James defaulting model (12) Loan purpose is negatively correlated with credit score that is low risk loans reduces the credit score more than high risk loans. On the other hand for repayment low risk loans increase the score more than high risk loans. Age, Marital Status and Loan Purpose were found to be crucial in both defaulting and repayment model and Time at current job was found to be a significant variable in the defaulting model.

Observing the linearities, this, therefore means that CBZ should reconsider reassigning scores given to the explanatory variables that showed linearity in the credit score sheet as these values are influential to time to default or repayment.

To correct the bias present in linear regression model, the Buckley James method was used and it also performed well in simulation. Survival analysis was applied to the personal loans to estimate the time to default or to early repayment.

6. Recommendations

The commercial bank of Zimbabwe should try and observe the loan performance of each customer and act as soon as the loan goes bad. It is suggested that the bank should establish a credit risk management team that should be responsible for the following actions that will help in minimising credit risk;

- Reconstructing the credit score sheet and reassign scores to all the variables that affect defaulting and repayment.
- Implementing the Buckley James method, as it proved to be better performing.
- Reconsidering the minimum age for a loan applicant, as the study showed that 21 years is not valid for loan application.
- Reviewing the customers that fall under single and married in the credit score sheet as there are widows and widowers.
- Closely monitoring the loan performance of each customer taking survival analysis into consideration as well.

References

Ambira, C. M., & Kemoni, H. (2011). Records management and risk management at Kenya Commercial Bank Limited, Nairobi. *SA Journal of Information Management, 13*(1), 11. http://dx.doi.org/10.4102/sajim.v13i1.475

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society, 50*, 1185−1190. http://dx.doi.org/10.1057/palgrave.jors.2600851

Basel Committee on Banking Supervision. (2001, January). *The New Basel Capital Accord*. Retrieved from http://www.bis.org/press/p010116.htm

Caire, D., & Kossmann, R. (2003). *Credit Scoring: Is It Right for Your Bank?* Bannock Consulting.

Durand, D. (1941). Risk elements in consumer instalment financing. *National Bureau of Economic Research*.

Eisenbeis, R. A. (1978). Problems is applying discriminant analysis in credit scoring models. *Journal of Banking and Finance, 2*, 205−219. http://dx.doi.org/10.1016/0378-4266(78)90012-2

Glasson, S. (2007). *Censored Regression Techniques for Credit Scoring*. PHD Thesis, RMIT University.

Hand, D. J., & Henley, W. E. (1993/4). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry, 5*, 45–55.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, 160*, 523-541. http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x

Rosenberg, E., & Gleit, A. (1994). Quantitative Methods in Credit Management: A survey. *Operations Research, 42*(4), 589−613. http://dx.doi.org/10.1287/opre.42.4.589

Siddiqi, N. (2005). *Credit Risk Scorecards, Developing and Implementing Intelligent Credit Scoring*. New York: Wiley.

Thomas, C. L., Crook, N. J., & Edelman, D. (2002). *Credit Scoring and Credit Control*. Oxford University Press. Retrieved from http://www.decisioncraft./comdmdirect/creditriskscoring.htm

Copyrights