International Journal of Statistical Distributions and Applications
2018; 4(1): 29-37
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20180401.14
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Modelling Factors Affecting Probability of Loan Default: A Quantitative Analysis of the Kenyan Students' Loan

Pauline Nyathira Kamau, Lucy Muthoni, Collins Odhiambo*

Institute of Mathematical Sciences, Strathmore University, Nairobi, Kenya

Email address:

nyathirapauline@gmail.com (P. N. Kamau), lmuthoni@strathmore.edu (L. Muthoni), codhiambo@strathmore.edu (C. Odhiambo)

*Corresponding author

To cite this article:

Pauline Nyathira Kamau, Lucy Muthoni, Collins Odhiambo. Modelling Factors Affecting Probability of Loan Default: A Quantitative Analysis of the Kenyan Students' Loan. *International Journal of Statistical Distributions and Applications*. Vol. 4, No. 1, 2018, pp. 29-37. doi: 10.11648/j.ijsd.20180401.14

Received: June 13, 2018; **Accepted**: July 17, 2018; **Published**: August 13, 2018

Abstract: In this study, we perform a quantitative analysis of loan applications by computing the probability of default of applicants using information provided in the Kenya Higher Education Loans application forms. We revisit theoretical distributions used in loan defaulters' analysis particularly, when outliers are significant. Log-logistic, two-parameter Weibull, logistic, log-normal and Burr distribution were compared via simulations. Logistic and log-logistic model performs well under concentrated outliers; a situation that replicates loan defaulters data. We then apply logistic regressions where the binomial nominal variable was defaulter or re-payer, and different factors affecting default probability of a student were treated as independent variables. The resulting models are verified by comparing results of observed data from the Kenyan Higher Education Loans Board.

Keywords: Student Loans, Default Rates, Multiple Logistic Regression

1. Introduction

A student loan is designed to assist students to pay college education and associated expenses such as tuition fees, purchase of books and stationery, hostel/rent expenses among other living costs. Conventionally, student loan defaulting is usually associated with other competing events such as, whether the student is a first time borrower/defaulter, or if the student borrowed several times and defaulted frequently. Like in most cases, Kenya's students loan funds has been created as a self-replenishing pool of money, utilizing interest and principal payments on old loans to issue new ones [1]. Some of the main factors that affect the operation the fund are the interest rate, administrative expenses, and levels of premiums, repayments failure, inflation and liabilities. Whereas analysis of loan defaulters is usually carried out using Cox regression model, this study focuses on the first time the student defaulted given several variables. The understanding of loan repayment distribution is critical to researchers and policy makers as it not only provide better understanding the excessive debt process of but also describing determinants of loan defaulting. Some of the articles that covered models that determine the likelihood of loan defaults and their associated factors include [1-9]. Though exploring association is critical to understand the determinants of loan defaulter, consideration of data structure particularly outliers is important to accurately predict factors that directly influence loan defaulting and solve practical problems that arise. Due to convenient interpretation and implementation, the logistic regression has been routinely used for estimation and prediction of determinants of loan defaulting. More so, applying a nonflexible link function to the data with this special feature may result in link misspecification. We revisit theoretical distributions used in loan defaulters' analysis particularly, when outliers are significant. Specifically, we consider performance of log-logistic, logistic, two-parameter Weibull, log-normal and Burr distribution through simulations study. The main purpose of this paper is to identify the major factors that explain what causes student loan default by using the best model that utilizes structured outlier. The analytic technique of choice is log-logistic regression given its ability to predict

a nominal dependent variable from one or more independent variables.

The next section covers various models for modelling loan defaulters, then simulations, applications of logistic model then discussions.

2. Methods

Here we describe specific models that have been used to model loan defaulters' data i.e. 2-parameter Weibull distribution, Burr Type III distributions, logistic distribution, the log-normal distribution, and the log-logistic distribution.

2.1. Models

2.1.1. Two-Parameter Weibull Distribution

The two-parameter Weibull distribution is defined as

$$f(x) = \frac{\gamma}{\beta^{\gamma}} x^{\gamma-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{\gamma}\right\} \quad (1)$$

Where γ represents shape parameter and β represents the scale parameter. Classic extensions of two-parameter Weibull have been covered in [17-18]

2.1.2. Burr Type III Distributions

The density function of the Burr Type III distribution is described as

$$f(x) = \frac{cb(-a)x^{-(b+1)}}{\left[1+\exp(-a)x^{b}\right]^{c+1}} \quad (2)$$

The values a, b, c are distribution parameters. Estimation and further derivations of Burr Type III distribution have been covered in [19].

2.1.3. The Log-normal Distribution

The probability density function of a log-normal distribution

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right\} \quad (3)$$

Where: μ, σ - distribution parameters (μ - location parameter, σ - shape parameter).

Further discussions regarding parameter estimation together with their properties have been discussed in [20-21]

2.1.4. Log-logistic Distribution

The density function in the log-logistic distribution is described as:

$$f(x) = \frac{\beta \exp(-(\alpha+\beta \ln x))}{x[1+\exp(-(\alpha+\beta \ln x))]^2}. \quad (4)$$

Where α and β are distribution parameters

2.1.5. Logistic Distribution Model

The dependent variable in logistic regression is dichotomous, meaning it can take the value 1 or 0 with a probability of defaulting and repaying respectively. This type of variable is called a binary variable. As mentioned earlier, predictor variables can take any form i.e. multiple logistic regression does not make any assumptions on them. They need not be normally distributed, linearly related or of equal variance within each category. Taking our binary outcome as Y with covariates $X_1, \ldots$ Xp, the logistic regression model assumes that;

$$\ln(P(Y = 1 \mid X_1, \cdots X_p)) = \ln \frac{P(Y = 1 \mid X_1, \cdots X_p)}{1 - P(Y = 1 \mid X_1, \cdots X_p)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (5)$$

In terms of probabilities this is written as;

$$\ln(P(Y = 1 \mid X_1, \cdots X_p)) = \frac{\exp(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}{1+\exp(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)} \quad (6)$$

The unknown model parameters βo through to βp are the coefficients of the predictor variables estimated by maximum likelihood, and X_1 through to X_p are the distinct independent variables. The right hand side of equation (6) above looks similar to a multiple linear regression equation. However, the method used to estimate the regression coefficients in a logistic regression is different from the one use to estimate regression coefficients in a linear regression model.

2.2. Data Description

The data set used in this study was extracted from Kenya's Higher Education Loans Board (HELB). A total of 5,100 clients were included in the analysis with age distribution being <24 years were 738 (14.5%), 24-30 years were 1,341 (26.3%), 30-35 years were 908 (17.8%), 35-40 year were 698 (13.7%), 40-45 years were 468 (9.2%) and > 45 years were 948 (18.6%). Data also consisted of different independent variables and one dependent variable. Dependent variable was defined as to whether there was loan defaulting or not (1 and 0).

Independent Variables considered were

1) loan amount
2) overdue days
3) age
4) interest rate
5) employer
6) gender
7) marital
8) father's education level
9) whether the father is employed or not
10) whether the father is alive or not
11) whether the mother is employed or not
12) whether the mother is alive or not
13) whether bursary was awarded or not
14) whether the client have dependents or not

The entire population included HELB financial statements and data from 1995, when its operation began to 2014. Since the sample population was too large, raw and unpolished, our study only took data from 2009 to 2014 as this is the period when the Board had begun experiencing major improvements in their disbursement and recovery policies. For this study, we focused on individuals who had completed their higher education studies from within the first year of completion up to 50 years since completion. The inclusion criteria therefore include individuals from ages 23 to 75 who both had and had not completed paying for their student loans.

The study sample consists of Kenyan students who studied both in private and public universities and colleges, and had benefited from the loans. The six year period (2009-2014) was chosen because it is more current and it was a time when HELB had made major changes and was experiencing better results from their operations.

2.3. Statistical Analysis

Data was analyzed by first codding in Visual Basic for Applications (VBA). Coding was necessary particularly in the initial stage of polishing the large amount of information in the data in order to gather a sample where only the relevant information was present. The polished data sample was entered into R Studio to build the multiple logistic regression models. This required a number of steps including creating dummy variables for the loan amount and the number of days for which the applicant had delayed their loan payments. This method was applicable in this study, as any categorical variable was made into a dummy variable for ease of functioning of the model. Variable selection used the regressions approach because of the consideration that all possible subsets of the pool of explanatory variables and are fitted according to a given criteria. The criteria used for this study is the Akaike Information Criterion (AIC), which assigns scores to each model and allows us to choose the model with the best score. We used the step function to perform variable selection. All analysis was done using R Studio and SPSS version 20.

3. Simulations

We conducted extensive simulations datasets to compare goodness of fit for the five distributions when fitted to dataset with structured outliers. Our primary aim was to establish the distribution that best fit data and show flexibility in fitting simulated data generated from various models. The true parameters were set such that the proportion of simulated data sets is around 70%, similar to the proportions in the outliers and the HELB defaulters' data set. We perform Akaike Information Criteria (AIC) analysis for a given simulated data set and assess the models using criteria of during the period in question, it repaid and interest on loan. To match data scenario close to the HELB defaulters' data, we simulated 4 covariates with intercept in our model. The types of covariates represent those that occurred in the real data. It includes one intercept (x_1), one continuous covariate generated from normal distribution (x_2) and two discrete covariates. Among the two discrete covariates, one dummy for nominal categorical data with 3 groups (x_3) and the other is binary categorical data (x_4). All covariates are generated for sample sizes n = 50, 100, 200, 500 and 1,000. The results of simulations are displayed using tables at the appendix. Simulation results showed varying performance of characteristics of different theoretical distributions with the empirical distribution shows that for, each of them has certain drawbacks. In the case of log-normal and log-logistic distributions was overstated the value of the mean, while in the distribution of log-logistic and Burr type III inflated value is mode. In turn, the distribution log-normal and log-logistic undercut the value of the median. The evaluation of model fitting based on descriptive parameters indicates that the best model from the proposed ones is logistic distribution. However, it should be noted, that methods of assessing goodness of fit yielded inconclusive results. It all makes the research on modelling the distribution of debt repayments need to continue. It would be appropriate to compare methods for parameter estimation and the inclusion of analysis of other models used for example in the analysis of distribution of income.

4. Results

The main objective of this research was to develop a quantitative model that returns an individual's risk of default. This model can be used by HELB to categorize new loan applicants as highly likely to default or not likely to default. Multiple logistic regressions was developed using the standardized coefficients which are the multiplier of the independent variables and their predictors. Based on the summary of the logistic 28 regression presented in the table below, the most significant variable in the model was the loan amount. Using the predictors and their coefficients, the logistic regression equation is given as below;

$$\begin{aligned} Y = {} & 0.1 + 0.04\text{loan amount} + 0.13\text{employment} - 0.13\text{age} - 0.18\text{gender} + 0.38\text{father alive} - 0.08\text{mother employed} \\ & - 0.07\text{mother alive} - 0.10\text{bursary} + 0.004\text{dependents} - 0.07\text{overdue days}. \end{aligned}$$

The coefficients above indicate the partial contribution of each variable to the regression equation by holding other variables constant.

The model will be given by the equation below;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

Where β_0 = Intercept, β_p = coefficients, X_p = Predictors and ϵ = Error term. We also checked the strength of the model by conducting an Analysis of Variance test. The significance value on the Analysis of Deviance table was tested at 95 percent confidence level and 5 significant levels. The test showed that the model is very strong.

4.1. Odds and Log of Odds

Odds express the likelihood of an event occurring relative to the likelihood of it not occurring. From the analysis, p is the probability of the event of default occurring, and is given by p = 0.44, then the probability of repaying is 1-0.44 = 0.56. The odds of defaulting will be given by;

$$\text{odds} = \frac{\text{P}}{1 - \text{P}} = \frac{0.44}{1 - 0.44} = 0.79.$$

The results imply that the odds of defaulting are 0.79 to 1, and the odds of repaying is 1.27 to 1. Logistic regression uses the log of the odds ratio rather than the odds ratio itself, therefore;

$$\ln \text{odds} = \ln \frac{\text{P}}{1 - \text{P}} = \ln \frac{0.44}{0.56} = -0.1047,$$

including other probabilities. We carried out crude and an adjusted odds ratio in R. The adjusted odds ratio is the crude odds ratio modified or adjusted to take into account data in the model that could be important. The table below shows the results we got.

Table 1. *Crude and adjusted odds ratio for significant covariates (Loan amount and whether the father is alive or not).*

	Crude odds	**Adjusted odds**
Variable in percentages	OR, 2.5 to 97.5	OR, 2.5 to 97.5
loan amount	1.60, 0.02 to 113.94	1.60, 0.02 to 113.76
Father alive	1.12, 0.41 to 3.09	1.12, 0.41 to 3.09

4.2. Deviance

Deviance is specifically useful for model selection. We see two types of deviance in our outcome, namely null and residual deviance. The residual deviance is a measure of lack of fit of the model taken a whole while the null deviance shows how well the dependent variable is predicted by a model that includes only the intercept. In our results, we have a null deviance of 6,360.5 on 5,099 degrees of freedom. The independent variables being included resulted in the decrease of the residual deviance to 6,227.1 on 5,088 degrees of freedom. The residual deviance reduced by 133.4 with a loss of 11 degrees of freedom.

4.3. Fisher Scoring

Fisher scoring iteration is concerned with how the model was estimated. An iterative approach known as Newton-Raphson algorithm is used by default in R for logistic regression. The model is fit based on an approximation about what the estimates might be. The algorithm searches to find out if the fit can be improved by using different estimates instead. If so, it engages in that direction using higher values for the estimates and fits the model again. The algorithm quits when it perceives that searching again would not yield any additional improvement. In our model, we had 4 iterations before the process quit and output the results.

4.4. Hosmer-Lemeshow Test

The strength of the model was tested by use of the Hosmer-Lemeshow goodness of fit test. This test evaluates the goodness of fit by initializing several ordered groups of variables and then comparing the number in each observed group to the number predicted by the logistic regression model. Therefore, the test statistic is a chi-square statistic with a desirable outcome of non-significance, meaning that the model predicted does not differ from the one observed. The ordered groups are created according to their estimated probability where those with the lowest probability are placed in one group and those with higher probability in different groups, up to the highest one read. These groups are further divided into two groups based on the actual observed outcome variable i.e. defaulter or re-payer. The expected frequencies are obtained from the model. If the model is strong, then most of the variables with success are classified in the higher deciles of risk and those with failure in the lower deciles of risk. The Hosmer-Lemeshow goodness of fit test gave us df = 8 and a p-value of less than 2.2e-16, which is very small and definitely less than 0.05, meaning that our model fit the data.

4.5. Multicollinearity

Multicollinearity occurs when you have two or more independent variables that are highly correlated. This result in problems with understanding which variables contribute to the explanation of the dependent variable, which leads to complications in calculating a multiple logistic regression. It reduces the model's legitimacy and predictive power. To ensure the model is well specified and functioning properly, there are tests that can be run. Variance Inflation factor is one such tool used to reduce multicollinearity.

4.6. Variance Inflation Factor (VIF)

This helps to identify the severity of any multicollinearity issues in order for the model to be adjusted accordingly. It measures how much the variance of an independent variable is affected by its interaction with other independent variables. VIFs are usually calculated by the software as part of the regression analysis. VIFs are calculated by taking a predictor variable, X_i and fitting it against every other predictor variables in the model. This gets you the unadjusted R-squared values which can then be injected into the VIF formula. The variance inflation factor ranges from 1 upwards, where the numerical value, in decimal form, informs us the percentage the variance is inflated for each coefficient. For instance, a VIF of 1.065709 tells us that the variance of a particular coefficient is 6.5709 percent larger than what we would expect if there was no correlation with other predictors. Generally, a VIF of 1 indicates zero correlation, if the VIF is between 1 and 5 then there is moderate correlation and anything greater than 5 indicates a high level of correlation. In our sample data, the VIF is as follows; loan amount = 1.001370, employment = 1.008483, age = 1.001269, gender = 1.026480, father alive = 2.981585,

mother employed = 1.064755, mother alive = 3.011166, bursary = 1.065709, dependents = 1.009704, overdue days = 1.152670. The variance between the coefficients used to build the model were only moderately correlated, therefore our model is without extreme multicollinearity.

4.7. Presence of Outliers

Outliers are observations identifiable as distinctly separate from majority of the sample, (Hair et al., 2010). The study developed two box plots of account status against the loan amount given to the student, and as well against the number of overdue days that the 24 individual had delayed their payments. The outliers on both of them were quite extreme, especially small amounts ranging from 700 to 4,200 shillings on the one showing loan amounts. This indicates that the individuals had minimal loan balance left to clear but had not yet done so and this amount remained dormant on their accounts, and is now revealed as outlier variables. The whiskers on the box plots were longer than the size of the box itself. A well-proportioned tail would produce whiskers about the same length as the box, or slightly longer. The box plot for defaulters is slightly bigger than that of non-defaulters indicating the difference between the highest loan amounts to the lowest is larger for the defaulters than it is for their counterparts. The median on the defaulter's box plot is visually equidistant from the upper quartile to the lower quartile, meaning that loan defaulters are well spread whether they took a larger loan amount or a smaller loan amount. However, for the non-defaulters, the number of individuals who took up larger loans are closer together than those who took lower amounts in loans.

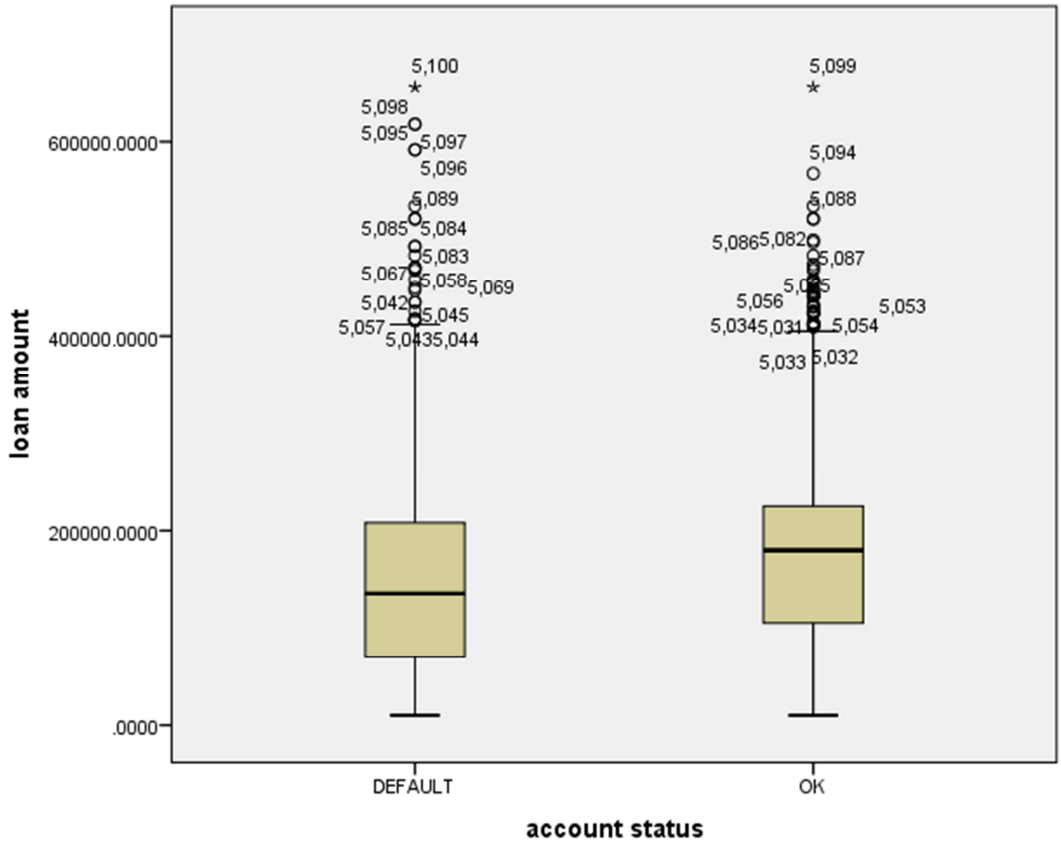

Figure 1. *Box-and-whisker plots for comparing loan amount between defaulters and non-defaulters.*

The box plot on overdue days showed that the majority of beneficiaries delayed their payments by about 50 days. For the non-defaulters, the box plot is very short meaning that there is certain agreement with taking a shorter number of days to pay off the loans as opposed to taking long. This is contrary to the defaulter's box plot which is longer and more evenly spread. The outliers on these two box plots tell the tale of those individuals who completed school a very long time ago and have not yet cleared their student loans. They are the extreme values indicated above the whiskers.

To treat the outliers' setting, we converted the variables in the sample population into probabilities. This allowed for ease of estimation and guaranteed lower errors in the model fit. Converting the variables into probabilities also allowed us to properly gauge the likelihood that an individual had certain characteristics that led them to default.

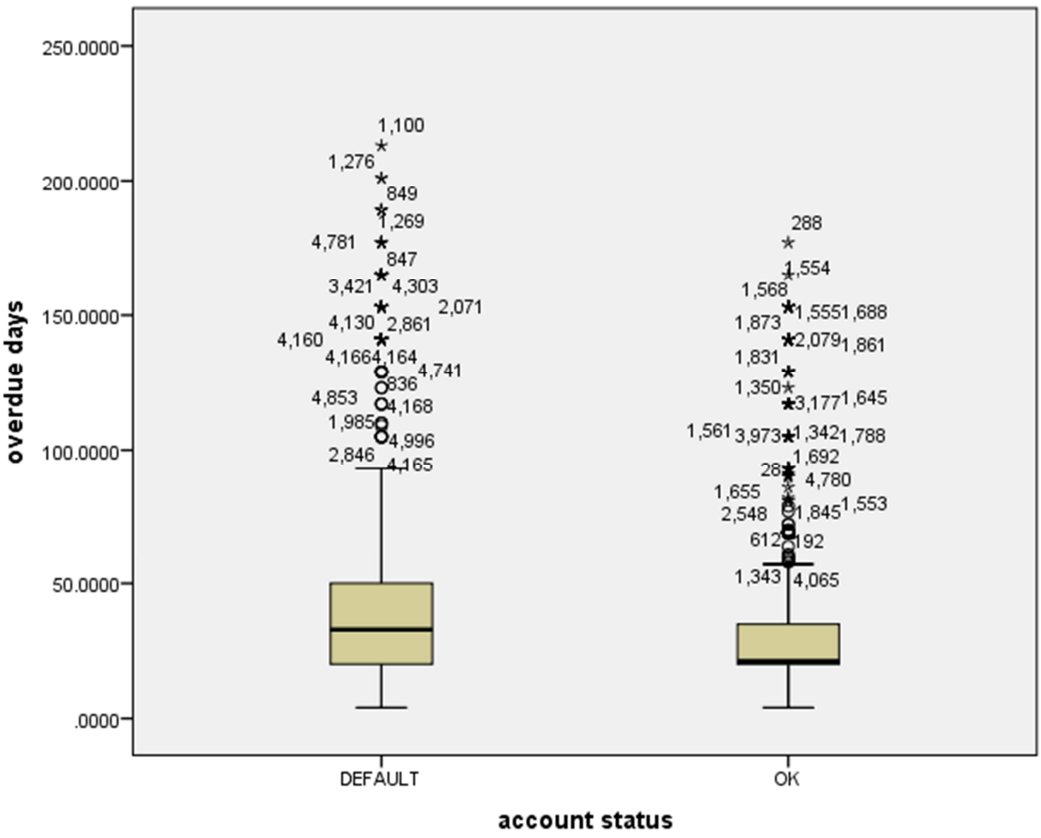

Figure 2. *Box-and-whisker plots for comparing overdue days between defaulters and non-defaulters.*

5. Discussion

This study is designed to find causes of higher education loans payment default among students. Personal characteristics and attributes were found to be key variant with unemployment being the highest by far. Since it is apparent to say that unemployment or lack of lucrative employment is the major cause of student loan default, we placed more focus on the other variants. The findings of the study with regards to cumulative amount of loan given to the student and default indicated a positive relationship indicated by the significance of its p-value. To validate the performance of logistic, we determine factors affecting loan default. Data provided by HELB is qualitative in nature and is provided by loan applicants at the point of application. It contains information about the student's background and parent's employment status among other details. Numerous studies have been done concerning student loan default using different models and methodologies [10-12]. This study explains this matter specifically by use of Multiple Logistic Regression which will have an outcome that will tell us if the individual either defaulted (1) or did not default (0) on their loans. We then confirmed that our model is correctly specified and relevant by use of several tests to ensure unbiasness, consistency, test the variance inflation properties among other tests. Then, we interpreted the results and discussed what they meant for Kenyan student loan applicants and for the Board especially concerning its loan disbursement policies. We saw that students who took up loans more frequently ended up with a huge loan at the end of their studies, which they had to pay back but with little or no means to do so especially the unemployment rates in the country. This was in line with the study done by [3-8] who found that the larger the loan the higher the likelihood of default. The findings indicated that if HELB monitored how much money cumulatively they reimbursed to applicants, they would be able to categorize separately those who would default from those who would be less likely to default.

Typically, the greater the debt accumulated over time, the more likely one is to default. The average loan amount advanced to defaulters was KES 93,432.13 with a maximum and minimum of KES 240,000 and 20,000 respectively. The standard deviations of the loan amounts and the study period are indicative that for each additional half year, loan amounts of KES 47,990.20, on average, had been disbursed to individual defaulters in the course of their study [4-7]. The number of overdue days played a huge role in contributing to their likelihood to default where 73 percent of individuals with over 150 days overdue were highly likely to default than individuals with less than that. This is because their loan continues to accumulate interest as the days add up, which one of HELB's initiatives for loan recovery is i.e. charging a penalty to those individuals who are late on their payments. This could make a defaulter out of an individual who would otherwise not fall into default, especially due to the fact that the employment is always fluctuating with the economy. Students who had both parents, even if the parents were not

both employed, showed a significant ability to not default on their loans by 68% compared to orphaned loan beneficiaries. Given the logistic regression formula for probability of success or failure, we should be able to find the probability of default, P, by keying in details into the model equation. The details are the estimates, β's which we found through model simulation in R Studio.

6. Conclusion and Further Research

The Logistic model continues to dominate in application. The logistic model performed better than other models in applications and identification of potential defaulters with minimal Type II error. This paper provides insights of potential and limitation of using Log-logistic, two-parameter Weibull, logistic, log-normal and Burr distribution models. Results show that the logistic model is more flexible. However, the major limitation of this study is the lack of exhaustive data variables of interest i.e. time to defaulting. Even though we are immensely grateful to HELB for the data provided to us, the best kind would have been one that shows the time until the first time a student defaults, as well as how many times a student's default tendencies recur. This would have been perfect for the analysis of all the exact events that lead to the first time defaulting. Future potential research area involves modeling time to default for both single event and recurrent events. This will enable computation of hazard functions and rates. Another potential area of study is on how to treat outliers in this setting.

Appendix

Table 2. *Simulations to compare goodness of fit using AKAIKE for sample size 50.*

parameter	estimator	standard error	z	p-value
Log-normal distribution (Log-likelihood=-66640.96; Akaike=118826; Schwarz=9040)				
μ	5.17	0.2	588	<0.0001
σ	1.12	0.02	129	<0.0001
Log-logistic distribution (Log-likelihood=-71448.38; Akaike=111901; Schwarz=149992)				
α	-11.11	0.32	-112.1	<0.0001
β	1.87	0.42	96.6	<0.0001
Burr III distribution (Log-likelihood=-66727.22; Akaike=166870; Schwarz=177892)				
a	-12.33	0.02	-889.9	<0.0001
b	3.12	0	998	<0.0001
c	0.57	0	667	<0.0001
Logistic distribution (Log-likelihood=-82232.22; Akaike=133877; Schwarz=156711)				
β	8.18	0.877	232	<0.0001
2-parameter Weibull distribution (Log-likelihood=-71266; Akaike=177926; Schwarz=162140)				
γ	10.22	0.05	423	<0.0001
β	3.44	0.04	434	<0.0001

Table 3. *Simulations to compare goodness of fit using AKAIKE for sample size 100.*

parameter	estimator	standard error	z	p-value
Log-normal distribution (Log-likelihood=-79960.96; Akaike=159926; Schwarz=159940)				
μ	6.35	0.25	723.24	<0.0001
σ	1.38	0.02	158.67	<0.0001
Log-logistic distribution (Log-likelihood=-71448.38; Akaike=142901; Schwarz=142915)				
α	-13.67	0.39	-137.88	<0.0001
β	2.3	0.52	118.82	<0.0001
Burr III distribution (Log-likelihood=-71432.22; Akaike=142870; Schwarz=142892)				
a	-15.17	0.02	-1094.58	<0.0001
b	3.84	0	1227.54	<0.0001
c	0.69	0	820.41	<0.0001
Logistic distribution (Log-likelihood=-82232.22; Akaike=142877; Schwarz=156711)				
β	8.18	0.877	232	<0.0001
2-parameter Weibull distribution (Log-likelihood=-79960.96; Akaike=159926; Schwarz=159940)				
γ	12.57	0.06	520.29	<0.0001
β	4.23	0.05	533.82	<0.0001

Table 4. *Simulations to compare goodness of fit using AKAIKE for sample size 200.*

parameter	estimator	standard error	z	p-value
Log-normal distribution (Log-likelihood=-79960.96; Akaike=159926; Schwarz=159940)				
μ	8.89	0.35	1011.81	<0.0001
σ	1.93	0.03	221.98	<0.0001
Log-logistic distribution (Log-likelihood=-71448.38; Akaike=142901; Schwarz=142915)				
α	-19.12	0.55	-192.9	<0.0001
β	3.22	0.72	166.23	<0.0001
Burr III distribution (Log-likelihood=-71432.22; Akaike=142870; Schwarz=142892)				
a	-21.22	0.03	-1531.31	<0.0001
b	5.37	0	1717.33	<0.0001
c	0.97	0	1147.75	<0.0001
Logistic distribution (Log-likelihood=-82232.22; Akaike=142877; Schwarz=156711)				
β	8.18	0.877	232	<0.0001
2-parameter Weibull distribution (Log-likelihood=-79960.96; Akaike=159926; Schwarz=159940)				
γ	17.59	0.09	727.89	<0.0001
β	5.92	0.07	746.81	<0.0001

Table 5. *Simulations to compare goodness of fit using AKAIKE for sample size 500.*

parameter	estimator	standard error	z	p-value
Log-normal distribution (Log-likelihood=-81860.96; Akaike=163326; Schwarz=159940)				
μ	8.81	0.35	1002.71	<0.0001
σ	1.91	0.03	219.98	<0.0001
Log-logistic distribution (Log-likelihood=-81848.38; Akaike=149901; Schwarz=142915)				
α	-18.95	0.55	-191.16	<0.0001
β	3.19	0.72	164.73	<0.0001
Burr III distribution (Log-likelihood=-65632.22; Akaike=142870; Schwarz=135662)				
a	-21.03	0.03	-1517.53	<0.0001
b	5.32	0	1701.87	<0.0001
c	0.96	0	1137.42	<0.0001
Logistic distribution (Log-likelihood=-77232.22; Akaike=132877; Schwarz=141111)				
β	8.18	0.877	232	<0.0001
2-parameter Weibull distribution (Log-likelihood=-76770.96; Akaike=134326; Schwarz=157734)				
γ	17.43	0.09	721.33	<0.0001
β	5.87	0.07	740.09	<0.0001

Table 6. *Simulations to compare goodness of fit using AKAIKE for sample size 1000.*

parameter	estimator	standard error	z	p-value
Log-normal distribution (Log-likelihood=-78860.96; Akaike=162316; Schwarz=177240)				
μ	8.73	0.34	993.9	<0.0001
σ	1.89	0.03	218.05	<0.0001
Log-logistic distribution (Log-likelihood=-72233.38; Akaike=151101; Schwarz=155215)				
α	-18.78	0.54	-189.48	<0.0001
β	3.16	0.71	163.28	<0.0001
Burr III distribution (Log-likelihood=-69932.22; Akaike=144422; Schwarz=143332)				
a	-20.84	0.03	-1504.21	<0.0001
b	5.28	0	1686.93	<0.0001
c	0.96	0	1127.44	<0.0001
Logistic distribution (Log-likelihood=-911932.22; Akaike=218811; Schwarz=155616)				
β	8.18	0.877	232	<0.0001
2-parameter Weibull distribution (Log-likelihood=-78892.96; Akaike=159926; Schwarz=166240)				
γ	17.28	0.08	715	<0.0001
β	5.81	0.07	733.6	<0.0001

References

[1] Stephen Muthii Wanjohi, Anthony Gichuhi Waititu, Anthony Kibira Wanjoya. Modeling Loan Defaults in Kenya Banks as a Rare Event Using the Generalized Extreme Value Regression Model. Science Journal of Applied Mathematics and Statistics 2016; 4(6): 289-297.

[2] Andrew, C. (2004). Basel II: The reviewed framework of June 2004. Geneva, Switzerland.

[3] Anatoly B. J (2014). The probability of default models of Russian banks. Journal of Institute of Economics in Transition 21 (5), 203-278.

[4] Altman E. (1968). Financial ratios, discriminant analysis, and prediction of corporate bankruptcy. Journal of Finance 23 (4) 589-609.

[5] Alexander B. (2012) Determinant of bank failures the case of Russia, Journal of Applied Statistics, 78 (32), 235-403.

[6] Lenntand Golet (2014). Symmetric and symmetric binary choice models for corporate bankruptcy, Journal of social and behavior sciences, 124 (14), 282-291.

[7] McCullagh P., Nelder J. A (1989) Generalized linear model, Chapman Hall, Newyork.

[8] O. Adem., & Waititu, A. (2012). Parametric modeling of the probability of bank loan default in Kenya. Journal of Applied Statistics, 14 (1), 61-74.

[9] Rafaella, C. Giampiero, M. Bankruptcy Prediction of small and medium enterprises using s flexible binary GEV extreme value model. American Journal of Theoretical and Applied Statistics, 1307 (2), 3556-3798.

[10] Nick Hillman, Don Hossler, Jacob P. K. Gross & Osman Cekic What Matters in Student Loan Default: A Review of the Research Literature Journal of Student Financial Aid, Issue 1, Article 2, 1-10-2010.

[11] Blom, Andreas, Reehana Raza, Crispus Kiamba, Himdat Bayusuf, and Mariam Adil. 2016. Expanding Tertiary Education for Well-Paid Jobs: Competitiveness and Shared Prosperity in Kenya. World Bank Studies. Washington, DC: World Bank. Doi: 10.1596/978-1-4648-0848-7. License: Creative Commons Attribution CC BY 3.0 IGO.

[12] Anamaria Felicia Ionescu The Federal Student Loan Program: Quantitative Implications for College Enrollment and Default Rates Economics Faculty Working Papers, Colgate University Libraries, Summer 6-2008.

[13] Felicia Ionescu & Nicole Simpson Default Risk and Private Student Loans: Implications for Higher Education Policies Finance and Economics Discussion Series, 2014- 066.

[14] Michal T. Njenga. The Determinant of Sustainability of Student Loan Schemes: Case Study of Higher Education Loans Board Scool of Business, University of Nairobi, November 2014.

[15] Mwangi Johnson Muthii Predicting Student's Loan Default in Kenya: Fisher's Discriminant Analysis Approach School of Mathematics, University of Nairobi, 2015.

[16] Emile A. L. J. van Elen Term structure forecasting School of Economics and Management, Tilburg University, 2010.

[17] Peter C., B. Phillips & Jun Yu Maximum Likelihood and Gaussian Estimation of Continuous Time Models in Finance Cowles Foundation for Research in Economics, Yale University, University of Auckland and University of York. School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903.

[18] Stephen Crowley Maximum Likelihood Estimation of the Negative Binomial Distribution Unpublished Working Paper, 2012.

[19] Elizabeth Herr & Larry Burt Predicting Student Loan Default for the University of Texas at Austin.

[20] Christophe Hurlin Maximum Likelihood Estimation and Geometric Distribution Advanced Econometrics, University of Orleans, 2013.

[21] Mark Huggett, Gustavo Ventura & Amir Yaron Sources of Lifetime Inequality American Economic Review 101, 2923-2954, 2011.

[22] Stu Field, Parameter Estimation via Maximum Likelihood. Unpublished working paper, 2009.

[23] Konstantin Kashin Statistical Inference: Maximum Likelihood Estimation. Journal of Finance, spring 2014.

[24] Littell, R. C., Mc Clave, J. T., & Offen, W. W. (1979). Goodness-of-fit tests for the two parameter Weibull distribution. *Communications in Statistics-Simulation and Computation*, *8*(3), 257-269.

[25] Limpert, E., & Stahel, W. A. (2017). The log-normal distribution. *Significance*, *14*(1), 8-9.