ORDINAL LOGISTIC REGRESSION VERSUS MULTIPLE BINARY LOGISTIC REGRESSION MODEL FOR PREDICTING STUDENT LOAN ALLOCATION

D. K. Muriithi, J. Kihoro and A. Waititu
Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology
E-mail: kamuriithi@yahoo.com

Abstract
This paper examines two different methodologies to a classification problem of higher education loan applicants. The paper looks into the allocations made by the Higher Education Loans Board (HELB) relative to the economic status of the applicant. In this article, we modeled Higher Education Loans Board (HELB) loan application data from three public universities to determine whether the loan was allocated based on the needs of the respective applicants. The data was classified into two natural categories of those not allocated the loan (0) and those allocated the loan (1). This paper classified further to consider the amounts awarded by the HELB. This was possible since we observed that HELB loans were awarded in distinct categories (Kshs 0, Kshs 35,000, Kshs 40,000, Kshs 45,000, Kshs 50,000), Kshs 55,000 Kshs 60,000). In this study, we used ordinal logistic regression and multiple binary logistic regressions in classifying the applicants into the identified categories. The models were generated that included all predictor variables that were useful in predicting the response variable. This study found that HELB allocate a loan amount to Kshs 40,000 but anything behold Kshs 40,000 is based on information provided by an applicant. The study revealed that the loans were not awarded based on the need of respective applicants. This has led to mis-classification when allocating loan. The study found that wealth and amount of fees paid for siblings were other factors that could be considered to identify needy applicants. This results show that an ordinal regression model gives accurate estimates that can enable HELB make a viable awarding decision. It is expected that proper determination of the most accurate model will go a long way in minimizing the number of
mis-classifications when awarding HELB loan. The study raises questions on the criteria used by HELB in loan allocation but further studies may be commissioned to confirm or disapprove our findings.

Key words: regression, logistic, binary, ordinal; higher education, loan

1.0 Introduction

The history of the Higher Education Loans Board dates back to 1952 when the then colonial government awarded loans under the then Higher Education Loans Fund [HELF] to Kenyans pursuing university education in universities outside East Africa notably Britain, the USA, the former USSR, India and South Africa. The Government then introduced the University Students Loans Scheme (USLS), which was managed by the Ministry of Education since there was an increase in the number of students going for university education. Under the scheme, Kenyan students pursuing higher education at Makerere, Nairobi and Dar-es Salaam universities received loans to cover their tuition and personal needs, which they would repay on completion of their education, http://www.helb.co.ke

The USLS lacked the legal basis to recover matured loans from loanees. The Higher Education Loans Board was then established by an Act of Parliament and as one of its roles was to grant loans. The statute known as The Higher Education Loans Board Act, 1995 was legally established as Act number 3 of 1995. It came into existence on the 21st day of July 1995 through Kenya Gazette Supplement (Cap 213A). The Board applies a Means Testing instrument in order to identify deserving students, http://www.helb.co.ke.

Joint Admission Board (JAB) is an organization in Kenya that manages admission of government sponsored students. The students apply for the HELB loan upon receiving admission at the university. Once the student applies for the loan, HELB goes through a process of checking the forms filled and depending on several factors they are able to determine who will be awarded a loan and what amount will be awarded. In the case of the HELB loans we can classify the loans into two natural categories of those not allocated the loan (0) and those allocated the loan (1). We could as well classify further to consider the amounts awarded by the HELB.

2.0 Literature Review

2.1 Ordinal Regression

In recent years, there has been quite some work on ranking relations, which in the literature is often referred to as ordinal regression, McCullagh[8]. Since binary classification is much more studied than ordinal regression, a general framework to systematically reduce the latter to the former can introduce two immediate benefits. Well-tuned binary classification approaches can be readily transformed into good ordinal regression algorithms, which save immense efforts in design and implementation. Next, new generalization bounds for ordinal regression can be easily derived from known bounds for binary classification, which saves tremendous efforts in theoretical analysis, ling and lin[7].

The use of a regression tree learner by mapping the ordinal variables into numeric values was investigated by Kramer[6]. However there might be no principled way of devising an appropriate mapping function. An ordinal regression problem was converted into nested binary classification problems that encode the ordering of the original ranks by Frank[2], then the results of standard binary classifiers can be organized for prediction. A constraint classification approach for ranking problems based on binary classifiers was proposed by Har-Peled[3]. The principle of structural risk minimization, Vapnik[11] was applied by Herbrich[4] to ordinal regression leading to a new distribution-independent learning algorithm based on a loss function between pairs of ranks. Shashua [10] generalized the formulation of support vector machines to ordinal regression and the numerical results they presented showed a significant improvement on the performance.

2.2 Logistic Regression

Outcome prediction studies in recent years have become the main topic in many areas of health care research, for instance, prediction of mortality in head trauma based on initial clinical data. But acceptable models for outcome prediction have been difficult to develop. A predictive model must be simple to calculate, have an apparent structure and should be tested in independent data sets with evidence of generality, Chacha [1] Logistic regression applies maximum likelihood estimation after transforming the dependent variable into a logit variable (the natural log of the odds of the dependent variable occurring or not). Logistic regression estimates the probability of a certain event occurring. The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most appropriate model. A model is created that includes all predictor variables that are useful in predicting the response variable. In stepwise regression, variables can be entered into the model in the order specified by the researcher. Logistic regression can test the fit of the model after each coefficient is added or deleted, Hosmer[5].

In a study, Morrell[9], discussed that discriminant analysis and logistic Regression were both suitable ways to model the outcome of binary dependent variable. Logistic Regression was preferable since it was more capable of handling several dummy variables simultaneously and did not assume normality. A predictive model must be simple to calculate, have an apparent structure and should be tested in independent data sets with evidence of generality Chacha[1]. Logistic regression provides knowledge of the relationships and strengths among variables (e.g. a student who is an orphan and from a poor background is expected to be offered the highest loan compared to one with better status).

3.0 Methodology

3.1 Ordinal Regression

Ordinal means researcher can rank the values but the real distance between categories is unknown. For example, student performance can be bad, good,

better or best in that order. In our case the reason that we use this method of classification is that the amount allocated by HELB is ordinal in nature. The Ordinal Regression procedure (referred to as PLUM in the syntax) allows one to build models, generate predictions, and evaluate the importance of various predictor variables in cases where the dependent (target) variable is ordinal in nature. In this study the amount allocated by HELB is ordinal in nature; therefore the ordinal model is applicable. The model below will be used. The loan allocation categories are Kshs 0, Kshs 35,000, Kshs 40,000, Kshs 45,000, Kshs 50,000, Kshs 55,000 and Kshs 60,000.

$f(\Pr(Y_{ij} \leq y_{ij})) = \emptyset_j - [\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}]$ ..... (1)

where f() is the link function.

y_{ij} is the cumulative probability of the j^{th} category for the i^{th} case.

$\emptyset_j$ is the threshold for the j^{th} category of which we have six categories.

We have 0, 35,000, 40,000, 45,000, 50,000 and 55,000.

$\beta_1 \ldots \beta_p$ are the regression coefficients

p is the number of regression coefficients. .

$x_{i1} \ldots . x_{ip}$values of the predictors for the i^{th} case.

These are;

x_1 - total wealth (computed from property owned by the applicant's parents)

x_2- house wealth (computed from the total materials of the house they live in)

x_3- amount of fees (the cost of education of the siblings of the applicants each year).

This was done for those in secondary school and college only, on the case of x_3.

In ordinal logistic regression, the event of interest is observing a particular allocation or less. We model the following odds:

Pr_1 = prob(allocation of 1)/prob(allocation greater than 1)

Pr_2 = prob(allocation of 1 or 2)/prob(allocation greater than 2)

Pr_3 = prob(allocation of 1,2 or 3)/prob(allocation greater than 3)

Pr_4 = prob(allocation of 1,2,3 or 4)/prob(allocation greater than 4)

Pr_5 = prob(allocation of 1,2,3,4 or 5)/prob(allocation greater than 5)

The last category does not have an odds associated with it, since the probability of scoring up to and including the last allocation is 1. All of the odds are of the form:

Pr_{35} = prob(allocation≤35000) / prob(allocation > 35000) etc.

3.2 Logistic Regression

Modeling the relationship between explanatory and response variables is a fundamental activity encountered in statistics. Logistic regression is used to predict a categorical (usually dichotomous) variable from a set of predictor variables.

Logistic regression has been especially popular with medical research in which the dependent variable is whether or not a patient has a disease.
For a logistic regression, the predicted dependent variable is a function of the probability that a particular subject will be in one of the categories (for example, the probability that Suzie Cue has the disease, given her set of scores on the predictor variables). Multiple linear regression may be used to investigate the relationship between a continuous (interval scale) dependent variable, such as income, blood pressure or examination score. The model can then be used to derive estimates of the odds ratios for each factor. In logistic regression, the dependent variable is a logit, which is the natural log of the odds.

For a single exposure variable E, the model takes the form

$$ln\ \frac{p}{1-p} = a + bx \quad \text{----------(2)}$$

Where *p* denotes the probability of occurrence of the outcome D and *x* is the value of an exposure E. The equation can be inverted to give an expression for the probability of *p* as,

$$P(D) = \frac{1}{1+\exp(-a-bx)} \quad \text{----------(3)}$$

The risk of the outcome given the exposure will thus be obtained by putting *x*=1 in the equation (3), we obtain

$$P(\frac{D}{E}) = \frac{1}{1+exp(-a-b)} \quad \text{----------(4)}$$

while the risk of the outcome given no exposure (*x*=0) we obtain

$$P\left(\frac{\mathrm{D}}{\bar{\mathrm{E}}}\right) = \frac{1}{1+\exp(-a)} \quad \text{----------(5)}$$

The relative risk is the ratio of these two expressions. We will use the odds and odds ratio.
The odds of the outcome given exposure are, from equation (4),

$$\frac{P(\frac{D}{E})}{P(\frac{\bar{D}}{E})} = \frac{P(\frac{D}{E})}{1-P(\frac{D}{E})} = \frac{\frac{1}{1+\exp(-a-b)}}{1-\frac{1}{1+exp(-a-b}} \quad \text{----------(6)}$$

which reduces to *exp*(*a*+*b*). Finally obtain the odds ratio as

$$OR = \frac{exp(a+b)}{exp(a)} = exp(b) \quad \text{----------(7)}$$

This means that the parameter *b* in the model is the natural logarithm of the odd ratio.

3.3 Multiple Binary Logistic Regression Model

If there are p predictor variables x_1, x_2,...,x_p, the general form of multiple logistic regression model is as follows;

$$P(D) = \frac{1}{1+\exp(-a-\sum_{j=1}^{p} b_j x_j)} \quad \text{---------- (8)}$$

Parameters b_1,...,b_p, were estimated using the maximum likelihood method. The parameter should give the significance of each independent variable to the outcome D. The estimated parameter forming the model was used to classify the remaining part of the data into either of the two groups. The outcome of the classification of the model was compared with the already known outcome. Finally the percentage of the correctly classified data was obtained. The percentage performance was used for comparison.

4.0 Survey Data Collection

This method of sampling is applied where the population embraces a number of distinct categories. The frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. This sampling method is used because sampling problems may differ remarkably in different parts of the population. In this article, strata were HELB loan applicants from Jomo Kenyatta University of Agriculture and Technology (JKUAT), Kimathi University College of Technology (KUCT) and Kenyatta University (KU). Each stratum was treated as an independent population and proportional sampling approaches were applied to different strata. The stratum population size was the number of students graduating or admitted per academic year to the university.

The data that was captured by questionnaire included; gender, year of birth, if one has both parents, father's year of birth and his occupation, mother's year of birth and occupation, guardian's year of birth and occupation, how many siblings one has, year of birth of the first and last born, the level of education of the siblings, category of high school attended, the amount of fees paid per year and who exactly did the payment. Also captured: the year one was admitted to the university, the year of study they are in, what amount is spent on fees, food, clothes, rent, medical bills and travel per year, what amount was awarded by HELB, and in which years did he receive the loan, what is HELB in the context of transparency, fairness, courtesy, competence, and if it works as a team. Also it captured: what the parents own in terms of motor vehicles and domestic animals, what land acreage they have, what are the sources of income of the family, what kind of house they are living in and whether rented or owned. For the owned

houses, information was recorded on the roofing materials, wall materials, floor materials and how many rooms there are in total. For the ones in rented accommodation, interest is on how much they pay as rent. Finally, the data was analysed using Statistical Package for Social Scientist (SPSS). This information was used to come up with a conclusion and probably a recommendation.

5.0 Empirical Results

The questionnaire was administered physically. The data has the consideration that all the respondents were awarded or not awarded a loan on application.
Data coded: 0-No 1-Yes

In this analysis the following variables were under consideration:
x_1 -Wealth (sum of property owned by the parents of an applicant). This was computed from lorry/bus, Matatu, motorbike, bicycle, cars, cows, goats, poultry, land and Posh mill. Each was allocated a fixed value and this was used to multiply with the number of units of the property.

x_2-House Worth (computed from the total materials of the house they live in). In the survey questionnaire applicants were to indicate what kind of house they live in, whether in rented or their own. We also inquired on what materials were used and we had roof, floor, walls, and how many rooms. We valued all the materials and computed how many are required for each room. Then the total value of the house was the sum total of the materials by the number of rooms.

x_3-Amount of fees (the cost of education of the siblings of an applicant each year). This was done for those in secondary school and college. We inquired on the number of siblings the respondents had and who were still in school. This was intended to know what other expenses the parents had. A constant value for those in secondary school and college per year was given.

5.1 Ordinal Regression Results

Table 1: Classification results

		Estimate	Std. Error	Wald	Df	Sig.
Threshold	[HELB loan = 0]	-2.353	0.337	48.652	1	0.001
	[HELB loan = 35,000]	-1.722	0.311	30.689	1	0.001
	[HELB loan = 40,000]	-0.558	0.283	3.895	1	0.048
	[HELB loan = 45,000]	1.277	0.315	16.404	1	0.001
	[HELB loan = 50,000]	1.906	0.365	27.303	1	0.001
	[HELB loan = 55,000]	2.482	0.437	32.233	1	0.001
Location	x_1	-0.018	0.008	4.874	1	0.027
	x_2	-0.001	0.001	1.683	1	0.195
	x_3	-0.038	.0.017	.4.881	1	0.027

Link Function: Logit
Pr_{35} = prob(allocation≤35000) / prob(allocation > 35000)
Pr_{40} = prob(allocation≤40000) / prob(allocation > 40000)
Pr_{45} = prob(allocation≤45000) / prob(allocation > 45000)
Pr_{50} = prob(allocation≤50000) / prob(allocation > 50000)
Pr_{55} = prob(allocation≤55000) / prob(allocation > 55000)
From Table 1, we have six thresholds estimates (λ_1, λ_2, λ_3, λ_4 λ_5 *and* λ_6) which are statistically significant and are useful for the model below.

$$f(Pr(Y_{ij} \leq y_{ij})) = \lambda_j - [-0.018x_1 - 0.038x_3] \quad \text{---------- (9)}$$

From Table 1; under the location parameters we observe that all variables have negative coefficients indicating that the presence of these variables increases the likelihood of smaller value of the response. Looking at the analysis, we note that the house worth is not statistically significant thus not included in equation (9) while wealth and amount of fees paid for siblings are statistically significant. This shows that the wealth and amount of fees paid for siblings improve the model positively. Indeed, wealth reduces the chances of getting a loan. From the parameters estimated we develop the following models.

$$f(\lambda_1) = -2.353 - [-0.018x_1 - 0.038x_3] \quad \text{.....(10)}$$

$$f(\lambda_2) = -1.722 - [-0.018x_1 - 0.038x_3] \quad \text{.....(11)}$$

$$f(\lambda_3) = -0.558 - [-0.018x_1 - 0.038x_3] \quad \text{.....(12)}$$

$$f(\lambda_4) = 1.277 - [-0.018x_1 - 0.038x_3] \quad \text{.....(13)}$$

$$f(\lambda_5) = 1.906 - [-0.018x_1 - 0.038x_3] \quad \text{.....(14)}$$

$$f(\lambda_6) = 2.482 - [-0.018x_1 - 0.038x_3] \quad \text{.....(15)}$$

In this case, we reject the first (there is no relationship between wealth and amount awarded to an applicant) and the third (there is no relationship between amount of fees paid for siblings and amount awarded to an applicant) null hypotheses and conclude that there is a relationship between amount allocated and wealth and amount of fees paid for siblings at 5% level of significance. The Wald statistic tests the unique contribution of each predictor in the context of the other predictors.

5.2 Logistic Regression Results

This section presents results of a k-1 fitted binary logistic regression model.

Table 2: First model Results

		B	S.E.	Wald	df	Sig.	Exp(B)
	x_1	0.019	0.10	3.891	1	0.049	1.019
	x_2	0.002	0.001	5.040	1	0.025	1.002
	x_3	0.008	0.024	0.109	1	0.741	1.008
	Constant	-2.416	0.439	30.216	1	0.001	0.089

Table 2 shows the logistic regression coefficient, Wald test, and odds ratio for each of the predictors. Employing a 0.05 criterion of statistical significance, x_1 and x_2 variables had significant effects when no loan is awarded. The exponentiated coefficients in the last column of the output are interpretable as multiplicative effects on loan. Thus, for example, holding all other variables constant, an additional unit of wealth increases the likelihood of being awarded a loan by a factor of 1.019 on average. We observed that the significant values of the first two variables (x_1 and x_2) were less than 0.05 (0.049 < 0.05, 0.025 < 0.05 and 0.0741 > 0.05) meaning they are statistically significant and the other one is not, hence not included in the model below. The Wald statistic tests the unique contribution of each predictor in the context of the other predictors.
The prediction equation in this article is thus:

$$ln\ \frac{p}{1-p} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad \text{---------- (16)}$$

where p is the probability of being allocated the loan. Hence from the Table 2, we develop the following model.

$$ln\ \frac{p}{1-p} = -2.416 + 0.019\,x_1 + 0.002\,x_2 \quad \text{---------- (17)}$$

This result shows that there exist a relationship between the amount allocated, wealth and house worth variables. The significance of the model is less than 0.05 hence the model is statistically significant. Predicted probability is 8.4%.

Table 3: Second model results

		B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
	x_1	-0.003	0.017	0.032	1	0.857	0.997
	x_2	-0.003	0.002	3.759	1	0.053	0.997
	x_3	0.017	0.033	0.260	1	0.610	1.017
	Constant	-1.653	0.486	11.565	1	0.001	0.191

Employing a 0.05 criterion of statistical significance, x_1 and x_3 variables had no significant effects when awarding a loan of Kshs 35,000 while x_2 is marginally significant. The exponentiated coefficient for x_3 indicates that when holding all other variables constant, a loan of 35,000 is 1.017 times more likely to be awarded to a respondent with siblings than been denied a loan. For all the variables considered by the researcher none was significant in this study, hence there is a need to consider more variable. From the Table 3, we develop the following model without x_1 and x_3 which are not statistically significance.

$$ln\ \frac{p}{1-p} = -1.653 - 0.003x_2 \quad \text{----------} \quad (18)$$

This equation was used to predict probability results shown in Table 8.

Table 4: Third model results

	B	S.E.	Wald	df	Sig.	Exp(B)
x_1	-0.001	0.011	0.014	1	0.907	0.999
x_2	0.000	0.001	0.000	1	0.996	1.000
x_3	0.041	0.021	3.701	1	0.054	1.042
Constant	-1.407	0.355	15.74	1	0.001	0.245

Employing a 0.05 criterion of statistical significance, x_1 and x_2 variables have no significant effects when awarding a loan of Kshs 40,000. Cost of education of the siblings (x_3) is marginally significant when allocating the same. The exponentiated coefficient for x_3 indicates that when holding all other variables constant, a loan of 40,000 is 1.042 times more likely to be awarded to respondent with siblings than been denied. In this case, a constant is highly statistically significant, hence it reduces the chances of been awarded a loan of Kshs 40,000. Also we observe that coefficient of the second variable is zero with a probability value of 0.996 thus totally insignificant. The prediction equation is thus,

$$ln\ \frac{p}{1-p} = -1.407 + 0.041x_3 \quad \text{……} \quad (19)$$

The above equation (19) predicts the probability results shown in table 8

Table 5: Fourth model Results

	B	S.E.	Wald	df	Sig.	Exp(B)
x_1	-0.019	0.013	2.038	1	0.153	0.982
x_2	0.000	0.001	0.001	1	0.982	1.000
x_3	-0.010	0.021	0.217	1	0.641	0.990
Constant	-0.452	0.323	1.957	1	0.162	0.636

Employing a 0.05 criterion of statistical significance, all variables have no significant effects when allocating a loan of Kshs 45,000. Besides, the constant has no significant effects when allocating a loan of the same amount. The exponentiated coefficients in the last column of the output are interpretable as multiplicative effects on loan. Thus, for example, holding all other variables constant, an additional unit of wealth reduces the likelihood of been awarded a loan by a factor of 0.982 on average that is 1.8%. Since none of the regression coefficients and constant were significant, we end up without a predication model.

$$ln\ \frac{p}{1-p} = 0 \text{----------} (20)$$

The above equation does not predict any probability results as shown in table 8.

Table 6: Fifth model results

	B	S.E.	Wald	df	Sig.	Exp(B)
x_1	-0.096	0.076	1.592	1	0.207	0.908
x_2	-0.001	0.002	0.154	1	0.695	0.999
x_3	-0.079	0.059	1.815	1	0.178	0.924
Constant	-1.623	0.704	5.307	1	0.021	0.197

Employing a 0.05 criterion of statistical significance, x_1, x_2 and x_3 variables have no significant effects when awarding a loan of Kshs 50,000. Constant is statistical significant when allocating the same. The exponentiated coefficient for x_3 indicates that when holding all other covariates constant, an additional unit increase in x_3 reduces likelihood of been awarded the loan of Kshs 50,000 by a factor of 0.924. With the above results we develop a model with only a constant.

$$ln\ \frac{p}{1-p} = -1.623 \text{.....} (21)$$

Table 7: Sixth model results

	B	S.E.	Wald	df	Sig.	Exp(B)
x_1	0.025	0.014	3.050	1	0.008	1.025
x_2	0.000	0.002	0.034	1	0.854	1.000
x_3	-0.183	0.117	2.471	1	0.016	0.832
Constant	-3.069	0.906	11.474	1	0.001	0.046

Employing a 0.05 criterion of statistical significance, x_2 variable had no significant effects when awarding a loan of Kshs 55,000 while x_1 and x_3 are highly statistically significant to the same effect. The exponentiated coefficient for x_3 indicates that when holding all other covariates constant, an additional unit of x_3 reduces the likelihood of being awarded the loan of Kshs 55,000 by a factor of 0.832. Using results of the Table 7, we construct the following model:

$$ln\ \frac{p}{1-p} = -3.069 + 0.025\,x_1 - 0.183x_3 \quad (22)$$

The above equation (22) predicts probability results shown in table 8.

5.3 Comparison of the Two Models

Table 8: Table of predicted probabilities

Amount Allocated (Kshs)	Ordinal Regression	Binary logistic model
0	0.091	0.084
35,000	0.067	0.160
40,000	0.218	0.203
45,000	0.445	0.500
50,000	0.085	0.015
55,000	0.049	0.038

As shown in Table 8 we observe that the likelihood of being correctly allocated Kshs 45,000 loan depending on the variables used for the binary model is 0. This implies that awarding of the amount is not based on any of the independent variables considered. In general, awarding of a loan of Kshs 45,000 is by chance. Also we observe that probability of being allocated categories on the extreme sides is less than 10% for both models as shown in the table above. Most of the covariates in the binary models have their significance level not less than 0.05 hence the models are not statistically significant. This means that there is need to bring on board more covariates for consideration. Also the ordinal model shows that the likelihood of being allocated a loan of Kshs 45,000 is above 42 %. The Ordinal Regression models have the significance being less than 0.05 hence the model is statistically significant. In general, we observe that some values are much closed to each other. Therefore, mean square error will be computed to determine the best model for recommendation in this study.

6.0 Conclusion and Recommendations

This study was aimed at investigating whether the HELB loan is allocated based on the need of the respective applicants. It was guided by the following objectives; collection of data from more than one university polish issues of both in data collection tool and target population, fit the data to ordinal logistic model following K approach and fit binary logistic model equivalent to ordinal logistic model and compare the result.

The study found that the loan was not awarded based on the need of respective student. This has led to mis-classification when allocating loans. The study revealed that wealth and amount of fees paid for siblings were other factors that could be considered to identify needy student.

The Ordinal Regression model discussed above has the significance being less than 0.05 hence it is statistically significant. We observed that wealth and amount of fees paid for siblings were statistically significant in this model. The above Ordinal Regression model can predict about 70% of the cases under study. The results show fairly good predicted probabilities that can be used for loan allocation.
The binary logistic regression model was applied in each of the categories in this study. From the results in section 5.0.2, we observe that some of the predictors were not statistically significant in many categories and yet believed to have greater impact in terms of loan awarding. The results shows fairly predicted probabilities that can be used for loan allocation but the model is not statistically significant for loan allocation.

Using mathematical tools, we calculated mean square error for both models where the ordinal regression model had a mean square error of 0.006805479 while binary regression model had mean square error of 0.02064224. In this case, the ordinal logistic regression model was the most appropriate model since it had a less mean square error, that is $0.006805479 < 0.02064224$.

In conclusion, we recommend use of ordinal regression model for HELB loan allocation to determine the amount of loan, if any, to be awarded to an applicant. This will minimize the number of mis-classification when awarding HELB loan.

6.1 Applications and Further Research

As earlier discussed we used three predictors for the analysis. We recommend use of more than three factors in order to get precise and more accurate results. Also, we recommend that Multinomial Logistic Regression be applied since it is useful for situations in which you want to be able to classify subjects based on values of a set of predictor variables. This type of regression is similar to logistic regression, but it is more general because the dependent variable is not restricted to two categories.

References

Chacha P. *Logistic regression versus neural networks in classification of binary data*. Master's thesis, Jomo Kenyatta University of Agriculture and Technology, 2007.

Frank M. H. E. *A simple approach to ordinal classification.* European Conference on Machine Learning, 2001. Lecture Notes in Computer Science.

Har-peled S., R. D. Z. D. Constraint classification: a new approach to multi-classification and ranking. *Advances in Neural Information processing systems* (2003).

Herbrich R, G. T. O. K. *Large margin rank boundaries for ordinal regression*. In Advances in Large Margin Classifiers, 2000.

Hosmer L. *Predictive probability model for a Merican civil war fortifications*. Master's thesis, California Institute of Technology., 1989.

Kramer S., G. P. B. D. M. Prediction of ordinal classes using regression. *Fundamental Informaticae*, **47** (2001), pp 1–13.

Ling L. Lin, H. Ordinal regression by extended binary classification. Tech. rep, California Institute of Technology, 2007.

Mccullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society,* **47** (1980)**,** 109 – 142.

Morrell K., L. C. J. A. J. W. A. Mapping the decision to quit: A refinement and test of the unfolding model of voluntary turnover. *Applied Psychology,* **57** (2008), 128 – 150.

Shashua A. Levin, A. Ranking with large margin principle: two approaches. *Advances in Neural Information Processing Systems,* (2003), pp.937 – 944.

Vapnik V. The nature of statistical learning theory. New York: *Springer-Verlag,* (1995).