Modelling Time Default in University Fee Payment Using Cox Proportional Hazard Model

Winnie Jeptoo Rotich[1], Joel Cheruiyot Chelule[2], Ayubu Anapapa[3], Herbert Imboga[4]

[1, 2, 4]Jomo Kenyatta University of Agriculture and Technology, Department of Statistics and Actuarial Science, Nairobi, Kenya
[3]University of Eldoret, Department of Mathematics and Computer Science, Eldoret, Kenya

Abstract: ***Default is the failure to pay interest or other money that is owed on time. Time default is amount of time taken to clear the given debt past the set time line. Fee payment is guided by the fee payment policy produced by all universities to guide their students on the time they are expected to clear their fee balances. The analysis of the time it takes students to pay their fee balances using Cox Proportional Hazard Model (CPHD) will use the university data collected from the students from the previous semester on how long it took them to pay their fee. The technique is non parametric, independent and robust thus allows a wide range of data. This study is inspired by (Mariusz, 2016).The main aim of this study is to find out the time it takes for students to pay their fee and their reasons using the Cox Proportional hazard Model (CPHD).The general objective of this study is to model Time Default in University Fee Payment using the Cox Proportional Hazard Model. Survival curves are drawn for comparison among different characteristics that affect the fee payment default. Analysis was performed using R and other mathematical modules will be discussed. From the analysed data several students are aware of the Fee Payment Policy and that Age_group,level of Education, Gender, and Employment do not affect fee payment timelines. This report is usefulto the Universities and other institutions of higher learning, to mitigate the problem of fee payment default hence leading to enhanced normalcy in running of the institution.***

Keywords: Cox Proportional Hazard Model, Time Default, Fee Payment Policy

1. Introduction

Universities like any other institution need funds to run its daily activities. This funds are collected from the student's fee and thus payment is essential. Several students take time to pay their fees and others pay theirs on time. When the student fails to pay their fee on time this is known as defaulting. Most researches that have been done show that banks suffer mostly from customers who default in paying their loan thus hindering the interest earned and development of the financial sectors.

During the admission of students in the University, several consideration are taken and their details are also captured. Variables such as age, gender, education level, marital status, occupation are recorded about the student. The student is also given a fee structure that will guide them in paying their fee and on time.

In several institution, default has become a major problem and thus solution need to be determined to ease the daily activities of the institution. In banks, loan default has become alarming as said by (Adedapo, 2007) that credit associations owe it as a duty to secure depositors' funds therefore credit organisations attempt to prevent loan delinquency and default because if the loan is not paid, the lender's capital is lost and the institution will no longer be sustainable. Operation and expansion of business and commercial activities depend a great deal on the availability of loans from commercial banks.

Several factors have contributed to defaulting by individuals in different organisation. These factors as discussed in bank loan default also affect the payment of University school fees. According (Asongo & Idama, 2014) they outlined several reasons that contribute to loan defaulting which include high staff turnover and clients dropouts, non-supervision of some customers on their loan funds utilization, non-reminder of some customers concerning their repayment, multiple borrowings by the customers, lack of penalty to some defaulters, lack of job experience by the staff and failure to comply to loaning policies by the staff. The factors that affect commercial bank loan defaulting include age, education level, marital status and gender as said by (Kiarie & Nzuki, 2013).

2. Review of Socio-Demographic Characteristics

According (Nawai & Mohd, 2012) age, business experience, total sales, income, gender, formal education, and distance from the lender office affect bank loan repayment. This is due to reluctance and lack of concern by the lenders to follow up what is happening hence causing the default rates to increase in commercial banks. They also found out that when total sales and income increase the loan repayment also increases and default is catered for. The study wanted to establish which methods will be used in reminding students that their fees are due for payment in order to reduce default.
When students are admitted in the universities several variables are captured which include age, gender, marital status and education level. These variables resemble those of loan repayments cited by (Kiarie & Nzuki, 2013)where they claimed that factors that affect commercial bank loan defaulting include age, education level, marital status and gender. According to their study age is a significant factor as the younger generation have a higher rate of defaulting compared to the older generation. They also found out that education level did not really influence greatly in defaulting of the bank loan while gender had it being equally likely to default. This study was conducted to establish which age brackets of students are likely to default in payment of fees.

2.1 Review Survival Curves and Models

Survival curve is a graph showing the proportion of a population living after a given age, or at a given time after contracting a serious disease or receiving a radiation dose. A survival curve shows the survival likelihood and they are based on a survival graph as the survivor function S(t) and hazard function h (t). S(t) as the Survivor function expresses the probability that an individual will survive up to and including time t. S (t) = P (T>t) Where, t is a given point in time, T is a random variable denoting the time of an event, P is the probability (Altman, 1991) .Survival function is the probability that the time of an event such as death occurs later than some specified time. The hazard function h (t) expresses the probability that an individual experiences the event instantaneously after a time t given the subject has not yet experienced the event. It can increase, decrease, or remain constant over time. A higher hazard rate is associated with lower survival (Kirkwood & Sterne Jonathan, 2003).

Survival curve is a graph that shows the relationship between the survival and age of a population of a particular species. Survival can be represented as the percentage of individuals of the original population that have survived after a specified time. Survival curve is a statistical representation of the survival knowledge of some patients in the form of a graph showing percentage of surviving versus time (www.cancerguide.org).

Survival curves under Kaplan-Meier estimation can be used to create graphs of the observed while in the log-rank test can be used to relate curves from different groups (Brandon, 2014).

2.2 Review of Time Default

The time it takes for the customers to default is more than three times out of the four times the customer takes the loan (Rajeev & Pankaj, 2015).According to (Sudhamathy, 2016)it is difficult to determine which customer will default and which will not causing a lot of difficulty in many banks and therefore a model should be determined in order to analysis the customers characteristics and get to find out the time it will take for the customers to default. It is therefore difficult to determine which student defaulted thus this possess danger in the existence of the University.

The use of traditional method to determine credit worthiness of a customer will not help the competitive banks know their liquidity which is important for the development of the banks, therefore a comprehensive method should be developed to enable them learn when the bank goes to liquidity thus developing a model to determine time to defaulting is very crucial because when the customers do not pay their dues on time then the profitability of the banks go down thus the model developed will enable them to know the customers that do not pay their dues on time (Irena & Laura, 2008).Therefore developing a system that the University uses is crucial for the development of the institution.

Firms run by partners who have stronger financial support and own working assets exhibit lower hazards of default. This then shows that non-financial information and macroeconomic indicators measured together with financial accounting data will significantly progress the estimating performance of default models (Alnoor, 2009)

(Dermine & Neto, 2005)Stated that the recovery on bad and doubtful bank loans and the distribution of increasing recovery rates and their economic factors using direct costs incurred by that bank on recoveries on bad and doubtful loans have empirical results relating to the timing.

(Mariusz, 2016)Suggested that the conditional expected time to default (CETD) measure has a strong explanation and can be applied in a straightforward way in analysis of loan performance in time. It offers a straight evidence on the timing of a potential loan default under some stress scenarios compared to probability to defaulting. Novel method was applied to compute CETD using Markov probability transition matrices which is a common method in survival analysis literature because it is a time-dependent factor and allows adjustments of time-dependent factors such as age that affect bank loan default. Thus this can be used to analyse time to default by students in paying fee.

2.3 Partial likelihood Method

The partial likelihood method is also referred to as Cox's method. The partial likelihood function is flexible to introduce time-dependent explanatory variables and handle censoring of survival times. Assume independent censoring, conditional on *xi*, *Ti* and *Ci* are independent.
Then the PHM

$$\lambda(t;xi)=\lambda_0(t)e^{\beta_1 x_{i1}} + \cdots + \beta_{p x_{ip}} = \lambda_0(t)e^{\beta xi} \quad (2.4.1)$$

Parameter estimate in the Cox PH model are obtained by maximizing the partial likelihood as opposed to the likelihood. The partial likelihood is given by

$$L(\beta) = \prod \frac{\exp\,(X_i\beta)}{\sum \exp\,(X_j\beta)} \qquad (2.4.2)$$

The partial log-likelihood is given by

$$L(\beta) = log L(\beta) = \sum\{X_i\beta\text{-log}\,[\sum \exp\,(X_j\beta)]\} \qquad (2.4.3)$$

Treating the partial log-likelihood as ordinary log-likelihood, maximum likelihoods of β were derived and used to estimate hazard ratios and the confidence intervals. The hazard ratio (HR) is defined as

$$\text{HR}=\frac{\hat{h}(t,X^*)}{\hat{h}(t,X)} \qquad (2.4.4)$$

Where $X^*=(X^*_1, X^*_2\ldots\ldots X^*_p)$ and $X=(X_1, X_2,\ldots\ldots\ldots Xp)$.Thus showing the hazard for one individual divided by the hazard for a different individual. Usually HR≥ 1 that is

$$\hat{h}(t,X^*) \geq \hat{h}(t,X) \qquad (2.4.5)$$

Such that X* is the group with the larger hazard and X was the group with the smaller hazard. The HR was simplified such that

$$\text{HR}=\frac{\hat{h}(t,X^*)}{\hat{h}(t,X)} = \frac{\hat{h}(t)\exp\,(\sum_{i=1}^{p}\hat{\beta}_i X_i^*)}{\hat{h}(t)\exp\,(\sum_{i=1}^{p}\hat{\beta}_i x_i)} \qquad (2.4.6)$$

$$=\exp\,(\sum_{i=1}^{p}\hat{\beta}_i(X_i^* - X_i)) \qquad (2.4.7)$$

$$=\exp\,(\hat{\beta}_i(X_i^* - X_i)) \qquad (2.4.8)$$

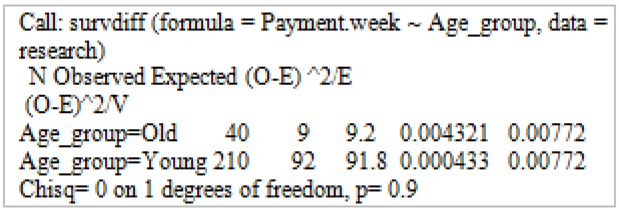
$$=e^{\hat{\beta}_i} \qquad (2.4.9)$$

2.4 Application to Analysis and Time Default

This study analyses Time Default in definite settings that are in real life situations. Cox Proportional Hazard Model is chosen because of its robustness, independence and proportionality. The collected data is then used to illustrate the time it takes for the students to clear their fee balances using the model.

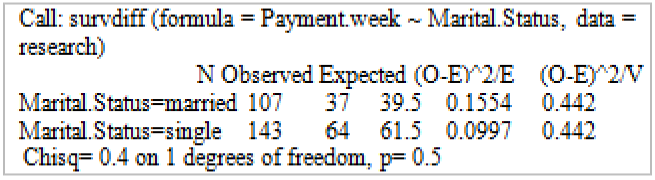
The model is applied to the socio-demographic characteristics collected from the students and the results yielded shows that variables in the model contributed to 86% of the total variability thus forming a good proportion and therefore the model is good for the study. The p value was also greater than 0.05 thus the covariates used were good for the study.

Interpretation of the variables produced the following results.



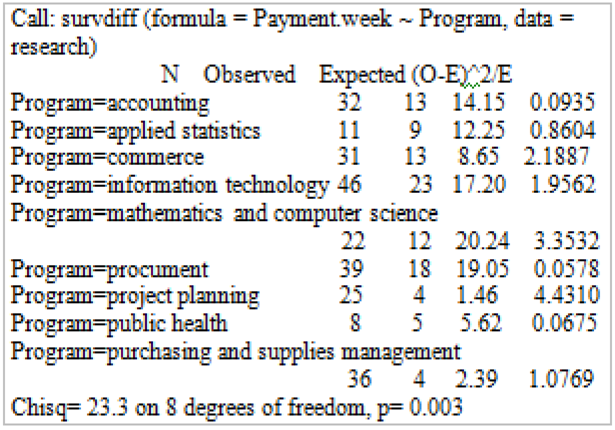
```
Call: survdiff (formula = Payment.week ~ Age_group, data =
research)
 N Observed Expected (O-E) ^2/E
 (O-E)^2/V
Age_group=Old      40      9      9.2   0.004321   0.00772
Age_group=Young 210     92     91.8  0.000433   0.00772
Chisq= 0 on 1 degrees of freedom, p= 0.9
```

The p-value = 0.9 is not significant therefore we will fail reject the null hypothesis - The survival curves are statistically equivalent and conclude that the Age_group is independent of the default time.



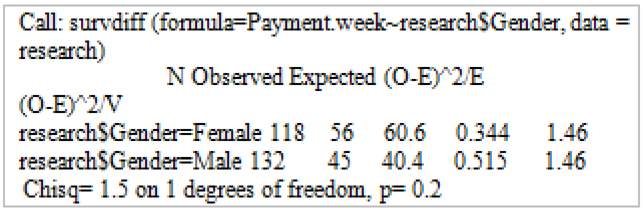
```
Call: survdiff (formula = Payment.week ~ Marital.Status,  data =
research)
                            N Observed Expected (O-E)^2/E   (O-E)^2/V
Marital.Status=married  107     37    39.5   0.1554    0.442
Marital.Status=single    143     64    61.5   0.0997    0.442
 Chisq= 0.4 on 1 degrees of freedom,  p= 0.5
```

The p-value = 0.5is not significant therefore we will fail reject the null hypothesis - The survival curves are statistically equivalent and conclude that the marital status is independent of the default time



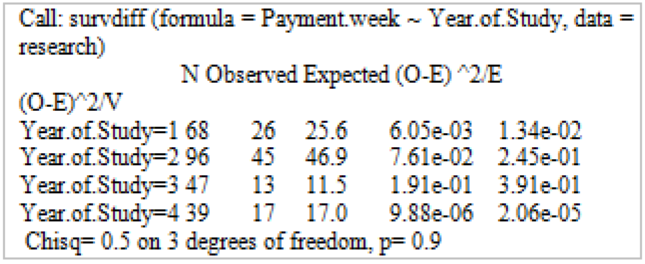
```
Call: survdiff (formula = Payment.week ~ Program, data =
research)
                         N    Observed    Expected (O-E)^2/E
Program=accounting                    32     13   14.15   0.0935
Program=applied statistics             11      9   12.25   0.8604
Program=commerce                      31     13    8.65   2.1887
Program=information technology 46     23   17.20   1.9562
Program=mathematics  and computer science
                                       22     12   20.24   3.3532
Program=procument                     39     18   19.05   0.0578
Program=project planning              25      4    1.46   4.4310
Program=public health                  8      5    5.62   0.0675
Program=purchasing and supplies management
                                       36      4    2.39   1.0769
Chisq= 23.3 on 8 degrees of freedom, p= 0.003
```

The p-value = 0.003 is significant therefore we will reject the null hypothesis - The survival curves are not statistically equivalent andconclude that programs are dependent of the default time.



```
Call: survdiff (formula=Payment.week~research$Gender, data =
research)
                    N Observed Expected (O-E)^2/E
(O-E)^2/V
research$Gender=Female 118    56     60.6    0.344     1.46
research$Gender=Male 132      45     40.4    0.515     1.46
 Chisq= 1.5 on 1 degrees of freedom, p= 0.2
```

The p-value = 0.1 is not significant therefore we will fail reject the null hypothesis - The survival curves are statistically equivalent and conclude that Gender is independent of the default time.



```
Call: survdiff (formula = Payment.week ~ Year.of.Study, data =
research)
                    N Observed Expected (O-E) ^2/E
(O-E)^2/V
Year.of.Study=1 68      26    25.6      6.05e-03   1.34e-02
Year.of.Study=2 96      45    46.9      7.61e-02   2.45e-01
Year.of.Study=3 47      13    11.5      1.91e-01   3.91e-01
Year.of.Study=4 39      17    17.0      9.88e-06   2.06e-05
 Chisq= 0.5 on 3 degrees of freedom, p= 0.9
```

The p-value = 0.9 is not significant therefore we will fail reject the null hypothesis - The survival curves are statistically equivalent and conclude that the Year of Study is independent of the default time.

From the above variables, Age, Marital Status, Gender, Level of Education and Year of Study are independent of time. This means that time default from this variable is less likely as compared to the programs taken that are dependent on time, meaning that several programs lead to default.

3. Conclusions and Recommendations

Discussions, analysis and proposals have been discussed from the results obtained from the model .The cox proportional hazard model results shows that the variables used were effective and that time default is independent of time. Further research may be done using other statistical techniques and the causes of default be analysed.

References

[1] Adedapo, K. (2007). Analysis of Default Risk of Agricultural Loan by some selected Commercial Banks in Osogbo,Osun State,Nigeria. *International Journal of Applied Agriculture and apliculture research*, 24-29.

[2] Mariusz, G. (2016). Measuring expected time to default under stress conditions for corporate loans. *narodowy Bank Polski Education and Publishing Department*, 1.

[3] Kiarie, K., & Nzuki, D. (2013). Influence of Socio-Demographic Determinants on. *trgsfsf*, 29-67.

[4] Asongo, A., & Idama, A. (2014). The causes of Loan Default in Micro-finance Banks:The experience of Standard Micro-Finance Bank,Yola,Adamawa,Nigeria. *international journal of Business and Management*, 74-81.

[5] Nawai, N., & Mohd, M. (2012). Factors affecting loan repayment performance in microfinance programs in Malasyia. *social and behavioural sciences*, 806-811.

[6] Brandon, G. (2014). Survival analysis and regression models. *J Nucl Cardiol*, 686-694.

[7] Irena, M., & Laura, I. (2008). The Evaluation Model of a Commercial Bank Loan Portifolio. *Journal of Business Economics and management*, 269-277.

[8] Sudhamathy, G. (2016). Credit Risk analysis and prediction Modelling of Bank Loans Using R. *DOI*, 5.

[9] Alnoor, B. (2009). *The role of financial,macroeconomic and non-financial information in bank loan default timing prediction.*

[10] Dermine, J., & Neto, C. (2005). *Bank Loan losses Given Default.*

[11] Rajeev, S., & Pankaj, G. (2015). Credit Risk Assessment for Mortgage Lending. *international Journal of Research in Business Management*, 13-18.