JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

****

MODELLING THE PROBABILITY OF CREDIT CARD DEFAULT IN BANKS.

A project proposal submitted in the partial requirement for the awarding of a Bachelor of Science degree in Financial Engineering at the Jomo Kenyatta University of Agriculture and Technology

College of Pure and Applied Sciences

School of Mathematical Sciences

Department of Statistics and Actuarial Sciences

DECLARATION

We declare that the work found in this proposal belongs to us and has not been presented for the award of any degree at any university. There is accurate citation of all other research included in this project that has been used as reference.

Signed:............................................. Date:...............................

**Grace Mwonga - .SCM214-9270/2015**

Signed:............................................. Date:...............................

**Eileen Muga - SCM214-8876/2015**

Signed:............................................. Date:...............................

**Thomas Mbashu - SCM222-0223/2016**

Signed:............................................. Date:...............................

**Micheal Maina - SC283-3636/2014**

This proposal report has been submitted for examination with my approval as the University Supervisor.

Signed:............................................. Date:...............................

**Ms. Charity Wamwea**

ACKNOWLEGEMENTS

We would like to start by thanking our supervisor, Madam Charity Wamwea, who has walked with us every step of the way and whose insightful advice has enabled us to make progress at every juncture of this research. We would also like to thank our parents and our siblings whose support and encouragement gave us the morale to keep moving forward. Above all we would like to thank the Good God Almighty, who gave us the strength, wisdom, knowledge and understanding to be able to come this far in our research. All Glory and Honor be unto Him.

TABLE OF CONTENTS

[DECLARATION 2](#_Toc23357800)

[ACKNOWLEGEMENTS 3](#_Toc23357801)

[TABLE OF CONTENTS 4](#_Toc23357802)

[CHAPTER ONE: 6](#_Toc23357803)

[INTRODUCTION 6](#_Toc23357804)

[1.1 Background of the study 6](#_Toc23357805)

[1.2 Statement of the problem 8](#_Toc23357806)

[1.3 Objectives 9](#_Toc23357807)

[1.3.1 Main objective 9](#_Toc23357808)

[1.3.2 Specific objectives 9](#_Toc23357809)

[1.4 Justification of the study 9](#_Toc23357810)

[1.5 Scope 9](#_Toc23357811)

[CHAPTER TWO: 10](#_Toc23357812)

[LITERATURE REVIEW 10](#_Toc23357813)

[2.1 Introduction 11](#_Toc23357814)

[2.2 Theoretical review 11](#_Toc23357815)

[2.2.1 Credit Risk Theory 11](#_Toc23357816)

[2.2.2 Determinants of Credit Default 11](#_Toc23357817)

[2.2.2.1 Gender and Credit Card Default 11](#_Toc23357818)

[2.2.2.2 Age and credit card default 11](#_Toc23357819)

[2.2.2.3 Level of income and credit card default 12](#_Toc23357820)

[2.3 Empirical study 13](#_Toc23357821)

[2.3.1Empirical study on credit card default in developed countries 13](#_Toc23357822)

[2.3.2 Empirical evidence on credit card default in Kenya 14](#_Toc23357823)

[CHAPTER THREE: 14](#_Toc23357824)

[RESEARCH METHODOLOGY 14](#_Toc23357825)

[3.1 Introduction 15](#_Toc23357826)

[3.2 Research design 15](#_Toc23357827)

[3.3 Data collection 15](#_Toc23357828)

[3.4 Descriptive analysis 15](#_Toc23357829)

[3.4.1 Logistic Regression Model 15](#_Toc23357830)

[3.4.1.1 Sigmoid Function 16](#_Toc23357831)

[3.4.1.2 Log odds (Logit Function) 17](#_Toc23357832)

[3.5 The logistic regression equation 19](#_Toc23357833)

[3.5.1 Parameter Estimation 20](#_Toc23357834)

[3.5.1.1 Maximum Likelihood Estimation 20](#_Toc23357835)

[3.6 Assumptions of The Logistic Regression Model. 22](#_Toc23357836)

[3.7 Hypothesis tests 23](#_Toc23357837)

[3.8 Why Logistic Regression? 24](#_Toc23357838)

[3.9 Study Site 24](#_Toc23357839)

[WORK PLAN 24](#_Toc23357840)

[BUDGET 26](#_Toc23357841)

[REFERENCES 27](#_Toc23357842)

CHAPTER ONE:

INTRODUCTION

# 1.1 Background of the study

A credit card is a financial instrument that allows the cardholder to obtain funds at interest from a financial institution at his or her own discretion, up to some limit (Edward and Robert, 1997). A credit card loan or credit card debt refers to the money one borrows when he or she uses his or her credit card. Credit card default is the failure to make credit card debt repayments by the due date for six months in a row.

The use or issuance of credit cards has been growing exponentially in the financial sector. The Federal Reserve Payments Study 2016 estimates that the credit card payment value in the US grew from $2.55 trillion in 2012 to $3.16 in 2015 at a rate of 7.4% per annum. Capgemini 2016 annual report noted that the trends in the US are observed worldwide. Such growth has catapulted credit cards to the leading noncash payment system.

Credit cards being an example of a revolving product has made modern banks take a different approach in credit risk mitigation unlike the one employed in traditional loan facilities. Unlike the traditional loans, with credit card loans the actual borrowing decision is solely at the discretion of the customer after receiving a fixed line of credit. Moreover, credit card loans do not require any collateral hence posing a greater risk to banks.

Since Ausubel (1991), the study of credit card default has gained a lot of traction among researchers. Ausubel’s empirical study found that high profits and interest rates existed in the market despite their being a competitive structure with over 6,000 credit issuers in the US. According to Brito and Hartley (1995), it is rational for credit card holders to hold positive credit card balances despite the high interest rates due to the aspect of high liquidity presented by the credit cards hence saving consumers the opportunity cost of holding cash for payment.

Mester (1994) pointed out a glaring fact that despite the high interest rates charged on credit card loans, information problems on the part of credit card banks would ensure that no irrationality on the part of the consumers usage of credit cards. The phenomenon of high interest rates charged on credit card loans is further expounded by Stavins (1996) and Park (1997) who concluded that the open-ended nature of credit cards and the high credit risk presented by credit card loans forced banks to charge such high interest rates. A strong positive correlation between credit card default and personal bankruptcy filings was a common finding by Ausubel (1997) and Domowitz and Sartain (1999). These findings clearly proved that credit card default posed a threat to the general state of the US economy and led to many researchers taking a keen interest in the credit card default issue.

The current research on credit card default provides vital information about the trends in the credit cards market but lack of detailed data has limited the understanding of consumer behavior and motivation in the use of cards and moreover in the understanding of credit card default, especially in developing countries. As such the study on credit cards default is gaining more momentum in developing countries such as Kenya.

The concept of credit cards is not a new phenomenon in Kenya having first been introduced in the country in 1967 by the Diners Club Africa. The franchise was later on taken over by the Royal Card. Currently all the major banks in Kenya including the Kenya Commercial Bank, the Equity Bank Kenya Ltd, the Barclays Bank of Kenya (which will rebrand to Absa Kenya in 2020), and the Co-operative bank of Kenya. The number of credit cards in Kenya as at May 2016 was 221,050.

Kenya’s banking sector is very competitive and banks in the country are currently trying to have a competitive edge in the credit card service due to the improved technology experienced in the country. The low adoption of credit cards in Kenya is therefore a concern to the banks as they seek to grow their banking operations as they endeavor to also provide a maximum customer satisfaction. However, an increase in the usage of credit cards in the country will increase the number of people who default and therefore banks will suffer more negative consequences. As such banks have to stay adept with the ever-evolving customer attributes in an effort to optimize their credit card service while also seeking a competitive edge in the financial sector.

In this study we will focus on significant customer attributes, namely, age, gender, and the level of income, that influence the ability to promptly pay the credit card loan and show their contributions towards the probability of credit card debt default. We are going to model the probability of credit card default in Kenya using the logistic regression model which is very efficient with categorical variables.

# 1.2 Statement of the problem

Credit card debt is a matter of great concern to the lending banks since an increase in the credit card usage has led to an increase in credit card defaults. Credit card defaults end up becoming bad debts which are then written off by the lending banks hence affecting the banks net profit. However, there has been very scanty research of credit cards default in developing countries due to the slower growth of credit cards in such countries compared to the developed countries.

Credit card debts do not require customers to post collateral hence placing a greater credit risk on the lending bank. In 2017, the International Monetary Fund estimated the average inflation rate in Kenya to amount to 7.99 percent. Such high inflation rates have led to an increase in the cost of living coupled with the never ending cases of employee retrenchment in both public and private services for instance the recent staff layoffs at Telkom Kenya, East African Breweries, and East African Portland Cement. To this end consumer credit is steeply rising hence increasing the default rate since the level of income is either stagnating or taking a downward spiral for the worst. Hence, credit card loans have become more significant due to the need to maintain consumer consumption at the same level.

This has prompted the lending banks to carry out thorough risk analysis, both qualitative and quantitative, on the potential borrower so as to determine their ability to pay back their outstanding credit card loan. This may involve checking the financial history of the potential borrower on the Credit Reference Bureau Kenya, conducting an interview, or even confirming the source of their regular income. As such, the lending banks will be able to effectively model the probability of credit card debt default since default is only observed once the customer is unable to payback their outstanding credit card debt.

Locally, a series of studies concerning credit cards have been conducted including Rotich (2006), Ondieki (2011), and Mucheru (2008) but none of them has modelled the probability of credit card default in banks. This research attempts to fill the gap by proposing a model of the probability of credit card default in banks. Moreover, this study is guided by the research question: What is the probability of credit card default in commercial banks?

We will use the logistic regression model to model the probability of consumer credit card default. This is because we will use two discrete classes namely, default and no default hence the logistic regression model becomes more appropriate than the linear regression model.

# 1.3 Objectives

#### 1.3.1 Main objective

To model the probability of credit card default.

#### 1.3.2 Specific objectives

1. To fit a Logistic Regression Model.
2. To determine the significant variables for predicting credit card default.
3. To estimate the probability of credit card default.
4. To determine the predictive power of the fitted model.

# 1.4 Justification of the study

This study will give the lending banks a clear understanding of the probability of credit card loan default based on different categorical variables such as the level of income hence the lending banks will be able to revert extreme trends of credit card loan default by putting in place appropriate and adequate measures. Moreover, this research is intended to inform the banks’ decision on the reasonable credit card limit that they should allow for each card holder. The model will enable banks to effectively monitor their loan portfolio risk, thus, mitigating credit risk.

The field of academia is always growing and as such this study will be a useful point of reference to researchers interested in credit cards as an area of study.

# 1.5 Scope

This study is carried on the model of the probability of credit card loan default. We have selected three variables which we intend on investigating their statistical significance. The variables are namely: age, gender, and the level of income.

CHAPTER TWO:

LITERATURE REVIEW

# 2.1 Introduction

This chapter summarizes information from previous researched work in the same area of interest. The specific areas covered are theoretical review, empirical review, conceptual framework and lastly the summary of the literature review.

# 2.2 Theoretical review

#### 2.2.1 Credit Risk Theory

Merton (1974) introduced the credit risk theory also called structural theory which suggested the default events derive from a firm’s asset evolution modelled by a diffusion processes with constant parameters.

Basel I accord (1988) which was a set of regulatory requirements on banking institutions in the European Countries. It came up as a way to mitigate risks faced by banks. Then they only considered on credit risk. Banks had to maintain a minimum capital requirement of 8% of risk weighted assets. These regulations were the same throughout the countries.

#### 2.2.2 Determinants of Credit Default

#### 2.2.2.1 Gender and Credit Card Default

According to Abdul-Muhmin and Umar (2007), the tendency to default on credit card loans is significantly higher in males. However, recent studies note that the probability of credit card ownership in Saudi Arabia is higher among females hence the female demographic in Saudi Arabia is at a higher probability of default. An analysis by Pirog and Robert (2007) on credit card misuse scores to determine the effects of demographic variables concluded that the mean credit card scores were essentially the same for men and women. The above previous research suggests that research on gender differences is inconclusive.

#### 2.2.2.2 Age and credit card default

Wickramasinghe and Gurugamage (2009) found that age is one of the significant and socioeconomic characteristic in describing consumer credit card practices. The usage of credit cards based on age has led researchers to suggest both positive and curvilinear relationships. On the curvilinear relationship, evidence suggests that credit cards usage intensity is higher among the middle-aged than the lower- and old-aged demographic (Abudl-Muhmin and Umar, 2007). According to Joireman, Kess, and Sprott (2010), credit cards are very problematic to young adults. According to Mae (2005), the average college student in the US will graduate with more than $2,800 in credit card debt and up to one-fifth carry a credit card of $10,000 or more.

A study conducted by Pirog and Robert (2007) established that the correlation between age and credit card scores to be significant. According to Hamilton and Khan (2001), people age under 35 were significantly more likely to become revolvers and the older one gets, the less likely they are to revolve. They, Hamilton and Khan (2001), conducted their research using Linear Discriminant Analysis and Logistic Regression on a sample of 27,681 bank credit card holders who had held and used their cards in the 14 month sample period to identify the characteristics of active card holders with the great propensity to revolve, that is, pay interest.

#### 2.2.2.3 Level of income and credit card default

The significant shift of consumer debt from installment debt to credit card debt has made consumers’ debt burden to be more sensitive to changes in the level of income. The higher the consumers’ level of income, the greater their likelihood of paying their credit card bills in full. The lower the level of income or a decline in the consumers’ level of income leads to a greater likelihood of late repayments or minimum payments on their credit cards bills hence their debt burden increases leading to a very a high probability of credit card default.

According to White (2007), although credit cards allow consumers to smooth consumption when their incomes fall, the cost of doing so is extremely high and may cause some debtors to enter a state of ongoing financial distress. Unpredictable expenses, such as illness, may lead people in to credit card debt. When people take on debt, credit card issuers react by doing at least one of the following: (1) charge penalties or fees; (2) increase interest rates; and (3) increase credit limits. Such actions increase the likelihood of credit card default. Trying to change consumers’ attitudes toward over-consumption does not necessarily solve the problem of credit card default because in most instances people are simply trying to exist in the prevailing tough economic conditions. According to Scott (2007), over-borrowing among a majority of credit card holders is due to lack of sufficient income.

Stauffer (2003) found that a majority of credit card users have relatively high propensities to consume but limited monetary assets, otherwise, they would not continue to pay high interest rates on unpaid balances and even defaulting completely in the end. Recent studies show that the ownership and use of credit cards by low-income households has increased and credit card holders have become riskier. Credit card companies have taken greater risk to earn abnormal returns and credit card debt is related to bankruptcy filings (Kidane and Mukherji, 2004).

# 2.3 Empirical study

#### 2.3.1Empirical study on credit card default in developed countries

The fact that credit cards do not require consumers to post collateral, unlike traditional banks, places a greater risk on lenders. Stigliz and Weiss (1981) studied the traditional loan market theoretically using the tools of asymmetric information and adverse selection.

The growth of credit cards debt in the US economy hassled researchers to increasingly turn their attention to the various aspects of credit card debt. The empirical study carried out by Ausubel (1991) found that abnormally high profit and sticky interest rates exist in the industry in spite of its competitive structure of over 6,000 credit card issuers. He speculated that search costs and a type of irrational consumer behavior might be involved in such market outcomes. Brito and Hartley (1995) responded to Ausubel’s argument by introducing the aspect of the liquidity service of credit cards which save the consumers the opportunity cost for holding money for payment, thus, arguing that it is rational for consumers to hold positive balances even in the face of high interest rates.

According to Mester (1994), high and sticky interest rates could exist without irrationality on the part of consumers because of information problems for the credit card banks. The situation is explained by Park (1997) who refers to the open-ended nature of credit card loans and the high risk involved with this for banks. Stavins (1996) found that defaulters had higher interest elasticities and this could induce banks to keep their interest rates high.

Mester and Colem (1995) found that the default in credit card debt was due to the fact that card holders with higher balances have higher probabilities of default. Revolving products, such as credit cards, information about borrowers’ repayment ability plays a crucial role in their determining their credit limits. Asymmetric information between borrowers and lenders and lack of collateral to mitigate the informational asymmetric are mainly responsible for credit rationing in some credit markets, hence, banks refuse credit to some borrowers. Stigliz and Weiss (1981) note that credit bureau reports help banks improve the quality of loan supply decision.

#### 2.3.2 Empirical evidence on credit card default in Kenya

According to Kegode (2006), married customers are more credit worthy than single ones, the longer a consumer stays in employment the more credit worthy they were, savings accounts holders were more credit worthy than the current accounts holders, consumers with house telephone defaulted twice as much as those with none, and the highest default rate was among those earning between 50,000/= and 70,000/=.

Currently in Kenya, credit cards are increasingly becoming an essential tool since a credit card offers a cardholder convenience safety and higher purchasing power. According to Mbijiwe (2005), screening out credit risky customers is a crucial step in card application acceptance process.

CHAPTER THREE:

RESEARCH METHODOLOGY

# 3.1 Introduction

The general objective of this study is to model the probability of default of credit card holders. Therefore, in this chapter we shall describe the methods used to achieve the specific objectives of the study which will aid in achieving the general objective wholesomely. The chapter therefore describes the methods used by describing the research design, data collection procedure, pilot test and finally processing the data and analyzing it.

# 3.2 Research design

Research design is defined as a framework of methods and techniques chosen by a researcher to combine various components of research in a reasonably logical manner so that the research problem is efficiently handled.(QuestionPro, 2019). In this case the study took on an analytic research design. According to Kothari, when conducting analytical research, the researcher uses facts and information already available and analyses this to make a critical evaluation of the material (Kothari, 2004). In this case the facts and available information obtained from the Financial Access Survey was the number of credit card defaulters. From this data we chose age, level of income and gender of the client to be the independent variables, while the dependent variable was the probability of default.

# 3.3 Data collection

This study used secondary data from the Financial Access Survey 2016 -International Monetary Fund (IMF). The data entailed general details of credit card defaulters from different banks compiled by the Developers of Financial Access Survey presented in the form of tables and graphs. The data used was from the years 2012 to 2016.

# 3.4 Descriptive analysis

In this study we shall use the logistic regression model to fit the data then analyze using R.

#### 3.4.1 Logistic Regression Model

The Logistic Regression model seeks to estimate that a given event will occur for a randomly selected set of observations against the probability that the event will not occur. This model is best used when the problem being handled is a classification problem. A classification problem is one where the independent variables are continuous in nature while the dependent variable takes on a categorical nature. A categorical variable is also sometimes referred to as a nominal variable and it is one which has two or more categories but has no intrinsic ordering of these categories (UCLA Institute for Digital Research and Education, 2019). These categorical variables take on two values that is success or failure. Take for instance a case of tossing a two faced coin with the outcome being either head or tail, in such a case we can model the outcome say Y as either **0** or **1.** In the case of this study the outcome Y will either be Probability of default or not defaulting and then the probability of this outcome is then predicted on the basis of one or more independent variables which in our case we have chosen as age, level of income and gender. These are then the xi s. Hence the aim when using the logistic regression credit scoring model is to be able to identify the conditional probability of each credit card holder to belong to one class. This implies that a “good” or a “bad” customer would be evaluated based on the values of the predictor variables of each credit card applicant.

The logistic regression has similarities with a linear regression model however, this model uses the **Sigmoid function** to deal with outliers which the linear regression model couldn’t handle if those outliers fall outside of the set threshold point or value. It also uses the **Logit function** to show the nonlinear relationship between the predictor variables and the response variable in the cases where it is not linear. In such cases an “S” curve is obtained and although there are other functions such as probit function, the logit function is preferred because its results are relatively easy to interpret. These two functions are explained further below.

#### 3.4.1.1 Sigmoid Function

In order to explain the logistic regression model, it will be in order to first define the standard logistic function which is a function that models the exponential growth in a population taking into consideration factors like carrying capacity. The function is in its self is a common sigmoid curve and is generally defined as

Equation 1: Sigmoid Function

{\displaystyle f(x)={\frac {L}{1+e^{-k(x-x\_{0})}}}}where

* *e* = the [natural logarithm](https://en.wikipedia.org/wiki/Natural_logarithm) base (also known as [Euler's number](https://en.wikipedia.org/wiki/E_(mathematical_constant))),
* *t*0 = the *x*-value of the sigmoid's midpoint,
* *L* = the curve's maximum value, and
* *k* = the logistic growth rate or steepness of the curve.

( *n(t )* can also be defined as *p(x)*as will be used in the final equation)

The standard logistic function as mentioned above is a standard choice for the sigmoid function which can be defined as

Equation 2: Standard Logistic Function

Taking t to be a linear function in a univariate regression model such that

Equation 3:Linear Regression Equation

Then the logistic equation becomes:

Equation 4:Logistic Equation

#### 3.4.1.2 Log odds (Logit Function)

Before describing a Logit function, it is important to describe what odds and odd ratios are.

**Odds** refer to the ratio of success to ratio of failure (Toward Data Science , 2018) while **Odd ratio** represents the group that has better odds of success and is obtained by calculating the odd ratios of each group. (Toward Data Science , 2018) The logit function is a link function in the logistic regression model. A link function is the mean of the response variable Y that we use as the response instead of Y itself. (The Analysis Factor, 2008) In this case the logit function is the natural log of the odds that Y equals one of the categories. It can be defined as

Equation 5:Logit Function

Recall that the outcomes are modeled as either **0** or **1.** Here **P** is the probability that **Y=1.**

It’s possible to see that the logit function is the natural log of the odds that Y equals one of the categories. In this study K=3 such that x1=age, x2=level of income and x3= gender.

**NOTE:** As mentioned above, we desire to compute the probability that an observation belongs to class 1 where in our case class one is the probability of defaulting then using the Logistic Response Function therefore from

Equation 6:Logistic Response Function

The coefficient beta 0, beta 1 and all other beta Ks are selected to maximize the likelihood of predicting a high probability for observations belonging to class 1 and also for predicting a low probability for observations actually belonging to class 0.

This can simply be defined as logarithmic odds such that

Equation 7

Hence the logit which is the log of odds is simply

Equation 8

# 3.5 The logistic regression equation

From the above we can now understand how the logistic regression comes to formation. In this study we define it in equation (i) as;

**= …………...(i)**

where;

* *pi* = probability of the outcome of the event
* *i* = function of the explanatory variables
* *x1*=age, *x2*=level of income and *x3*= gender
* β0 = intercept
* βj= coefficients associated with the corresponding predictor variables χi , for j =1,2,3 and i=1,2,3

and as defined in the logit function,

* represents the default event Yi.
* ϵi = error term

We then take the antilog of equation (i) and hence obtain an equation that gives the probability of Yi occurring and in order to obtain the critical points of this log likelihood function, we will set the first derivative of each Beta to be 0.

Differentiating equation (i) we obtain

**…....... (ii)**

Hence the probability of an event

**.......... (iii)**

Where **Pi** is the probability of the outcome of the event.

#### 3.5.1 Parameter Estimation

The Maximum Likelihood method of parameter estimation is used to estimate the parameters in the logistic regression model.

#### 3.5.1.1 Maximum Likelihood Estimation

The outcomes Yi are binomially distributed that is Yi~Bin(n,p) with n being the number of successes and p the probability of success. This implies that every Yi represents a binominal count in the ith population and hence the joint probability function also called the likelihood function of Y is;

**………. (iv)**

The Maximum Likelihood Estimators (MLE) are the values of β that maximizes the likelihood function of equation (iv).

Rearranging the equation:

**…........ (v)**

Using the logistic regression model, we are able to equate the logit transformation to the log odds of the probability of success. That is:

**……......... (vi)**

Taking the exponent of both sides of equation (vi)yields:

**…................ (vii)**

Solving for p we obtain

**pi = …............... (viii)**

We then substitute the first term of equation (v) with equation (vii) and the second term we substitute with equation (viii)

**….............(ix)**

Next we simplify the first product equation and then replace equation (i) with in the second product, to obtain:

**…...................(x)**

Equation (x) is the kernel of the likelihood function that is to be maximized.

**NOTE:** The logarithm is a monotonic function, and therefore the maximum of the likelihood function will also be a maximum of the log likelihood function and the contrariwise will be true too.

Now, we can take natural log of equation (ix) in order to simplifying its differentiation and then get hence obtain the log likelihood function as

**……..............(xi)**

In order to be able to find the critical points of the log likelihood function, we set the first derivative with respect to each *β* to be equal to zero and then differentiate equation (xi).

However, take note that:

**…...............(xii)**

And as the other terms in the summation do not depend on βj, then they can be treated as constants.

Differentiating the second half of the equation (xii), we take note that the general rule that

And proceed to differentiate equation (xii) with respect to each βj*.* such that

**=**

**=**

**= -**

# 3.6 Assumptions of The Logistic Regression Model.

The following are the assumptions of the logistic regression model.

##### Assumption of Appropriate Outcome Structure

One of the main assumptions of logistic regression is the assumption that the outcome variable takes an appropriate structure such that binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal

##### Assumption of Observation Independence

The logistic regression model requires that each observation be independent from the other implying that the observations used should not come from repeated measurements or data that has been matched.

##### Assumption of The Absence of Multi-collinearity among the Independent variables

The correlation between the independent variables shouldn’t be too high.

##### Assumption of Linearity of Independent Variables and Log Odds

In as much as the linear regression model does not require for there to be a linear relationship between the response variable and the predictor variables, it does require for the predictor variables to be linearly related with the log odds.

##### Assumption of A Large Sample Size

For the logistic regression model to work accurately it generally requires a large sample size. A conventional way of choosing the sample size is selecting a minimum of 10 cases with the least frequent outcome for each independent variable in the model. For instance, if you have five independent variables then, and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10\*5 /10)

# 3.7 Hypothesis tests

3.7.1 Significance of independent variables

In order to test for the significance of the independent variables, we will use a one sampled t-test.

Our hypothesis is:-

Vs

For

On performing the test, we expect to get a p-value, denoted by Pr {|t|}. For all p-values greater than 0.05 that is **p-value > 0.05** we fail to reject the null hypothesis. For all **p-values** less than or equal to 0.05 that is **p-value ≤ 0.05** we reject the null hypothesis.

# 3.8 Why Logistic Regression?

The main reason we chose the logistic regression model is that the model is designed to use dependent variables that are categorical which our dependent variables are making the logistic regression model the best choice. More so the logistic regression is easy to implement and efficient to train. In the training stage is where we will use 80% of the data collected as the training set and the remaining 20% as the testing set. In the testing stage we will also check for the validity of the individual predictors. To assess the relative importance of individual predictors in the model, we will look at the absolute value of the t-statistic for each model parameter.

Also the logistic regression model is a machine learning model and machine learning models do not require tuning as they have a default set of hyper parameters and this increases the accuracy of the hyperparameters as they aren’t obtained through a trial and error process of changing the hyperparameters ,running the algorithm on the data set again then compare its performance on the validation set in order to determine which set of hyperparameters result in the most accurate model.

# 3.9 Study Site

We intend to examine the logistic regression model by using the data we obtained from the Financial Access Survey 2016 then fit data into the model and the analyze it in R to be able to determine the probability of default of credit card holders.

# WORK PLAN

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ACTIVITY** | **SEPT** | **OCT** | **NOV** | **DEC** | **JAN** | **FEB** | **MARCH** |
| Topic search and selection |  |  |  |  |  |  |  |
| Proposal writing |  |  |  |  |  |  |  |
| Proposal presentation |  |  |  |  |  |  |  |
| Collection of data |  |  |  |  |  |  |  |
| Data analysis |  |  |  |  |  |  |  |
| Formulating conclusion and project compilation |  |  |  |  |  |  |  |
| Project submission and presentation |  |  |  |  |  |  |  |

# BUDGET

|  |  |  |
| --- | --- | --- |
| **NUMBER** | **ITEM** | **TOTAL COST (KSH.)** |
| 1 | Printing and binding | 1050 |
| 2 | Browsing | 3000 |
| 3 | Stationery and miscellaneous expenses. | 450 |
| **GRAND TOTAL** |  | 4500 |

# REFERENCES

Ausubel, Lawrence (1991), “The Failure of Competition in the Credit Card Market,” *American Economic Review*, 81(1), March 1991, 50-81.

Ausubel, Lawrence M. (1997), “Credit Card Default, Credit Card Profits, and Bankruptcy,” *American Bankruptcy Law Journal*, Spring 1997, 71, 249-270.

Brito, Dagobert L., and Peter R. Hartley (1995) “Consumer Rationality and Credit Cards,” *Journal of Political Economy*, 103, 400-33.

Calem, P. and L. Mester (1995) “Consumer Behavior and the Stickiness of Credit Card Interest Rates,” *The American Economic Review*, vol.85 (5, December), p. 1327-1336.

Cohen, M. J. (2005) “Consumer Credit, Household Financial Management, and Sustainable Consumption,” *International Journal of Consumer Studies*, 1470-643.

Rotich, B. (2006) “Analysis of Factors Influencing Credit Card Default in Kenya: A case study of Post Bank,” *School of Business and Economics-Research and Publications.*

Stiglitz, Joseph E., and Andrew Weiss (1981), “Credit Rationing in Markets with Imperfect Information,” *American Economic Review, 71, 393-410.*

Park, Sangkyun (1997), “Effects of Price Competition in the Credit Card Industry,” *Economics Letters, Elsevier, vol.57, pages 79-85, November.*