International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 3, Issue 1, pp: (48-61), Month: April 2015 - September 2015, Available at: www.researchpublish.com

Multiple Logistic Regressions Modeling on Risk Factors of Diabetes Case Study of GITWE Hospital (2011-2013)

[1]Niyikora Sylvere, [2]Dr. Joseph K. Mung'atu, [3]Dr. Marcel Ndengo

[1]Students at Jomo Kenyata University of Agriculture and Technology (JKUAT/KIGALI CAMPUS),
Master of Science in applied statistics
[2]Lecturer at Jomo Kenyata University of Agriculture and Technology/Kenya
[3]Lecturer at Jomo Kenyata University of Agriculture and Technology/ Kigali

Abstract: **The number of people with diabetes is increasing all over the world. A misconception that diabetes is a disease for urban areas while rural areas are also concerned; this is the motivation of the study. In this paper, a multiple logistic model is used to fit the risk factors of diabetes. A three year period (2011 to 2013) data from Gitwe Hospital are used. The test of independence between the dependent variable (diabetes) and the independent variables is performed. It is found that older age, alcohol consumption, cholesterol level, occupation status and hypertension were associated with the outcome of having diabetes. The predictors like gender; smoking, family history of diabetes had negligible association with having diabetes.**

A multiple logistic regression model containing all the predictor variables is fitted and a test of significance on coefficients is performed. The Wald test reveals that on one hand, the significant predictors are: older age, Occupation status, Alcohol consumption, Cholesterol level and Hypertension. On the other hand, the predictors which are not statistically significant are: Gender, smoking and family history of diabetes.

From the odds ratio results, older age persons, patients who consume alcohol, patients with high cholesterol level and hypertensive persons are highly susceptible for diabetes occurrence.

Finally, a multiple logistic regression with only significant parameters was fitted. Based on their respective Receiver Operator Characteristic (ROC) curve and their overall explanatory strength the conclusion is that the reduced model fits better the data than the model with all predictor variables.

Keywords: **Generalized model, Logistic regression, Diabetes, risk factors.**

1. INTRODUCTION

1.1. Problem statement:

The number of people with diabetes is increasing due to population growth, aging, urbanization, and increasing prevalence of obesity and physical inactivity. Quantifying the prevalence of diabetes and the number of people affected by diabetes, now and in the future, is important to allow rational planning and allocation of resources. (*Sarah et Al., 2004*).

According to *Shaw et al. (2010)*, the world prevalence of diabetes in 2010 among adults aged 20-79 years was estimated to 6.4%, affecting 285 millions of adults. Between 2010 and 2030, there is an expected 70% increase in number of adults with diabetes in developing countries and a 20% increase in developed countries.

Each year more than 231,000 people in the United states and more than 3,96 million people worldwide die from diabetes and its complications (*IDF, 2009*) and this number is expected to increase by more than 50 percent over next decade .

Estimated global healthcare expenditures to treat and prevent diabetes and its complications was at least 376 billion US Dollar (USD) in 2010. By 2030, this number is projected to exceed some 490 billion USD.

Environmental and lifestyle factors are the main causes of the dramatic increase in type 2 diabetes prevalence. Genetic factors probably identify those most vulnerable to these changes. *(IDF, 2010)*.

Diabetes is now truly a pandemic, and its effects are particularly severe in low and middle income countries.

The following table shows the situation of diabetes over the world in 2013 and the projected percentage of increase in people with diabetes in 2035:

Table 1.1: Number of people with diabetes by IDF regions, 2013 and projection in 2035

Region	Number of people with diabetes Year 2013(in millions)	Predicted percentage of increase in 2035
NORTH AMERICA AND CARIBEAN	37	37.3 %
SOUTH AND CENTRAL AMERICA	24	59.8 %
EUROPE	56	22.4%
MIDLE EAST AND NORTH AFRICA	35	96.2%
SUB-SAHARAN AFRICA	20	109.1%
SOUTH EAST ASIA	72	70.6%
WESTERN PACIFIC	138	46%

Source: IDF Atlas 2013, page 11 to page 12

Here are some basic facts about diabetes worldwide, according to *IDF (2013) :*

1. Each year the number of people with diabetes increases by 7 millions in the world.
2. By 2035, about 592 million people will have diabetes , a number which was 382million in 2013
3. More than 79000 children developed type 1 diabetes
4. During 2013, diabetes killed about 5.1 million adults worldwide.
5. Diabetes leads to complications and severe disabilities, including kidney disease, blindness, heart attack, stroke and neural damage leading to amputation and the need for chronic care.
6. The trend in 2013 revealed that there are three new cases every 10 seconds.
7. More than 80% of spending on medical care for diabetes is in the world's richest countries, even though 80% of the people with diabetes live in low and middle income countries, where 76% of the burden lies.
8. The burden of illness caused by diabetes and the reduction in life expectancy in sub-Saharan Africa will hinder the region's economic growth.
9. Diabetes caused at least 548 billion USD in health expenditure in 2013 (it means 11% of the total health spending on adults and this amount is predicted to be 627 USD in 2035.
10. More than 21 million live births were affected by diabetes during pregnancy in 2013

Concerning Sub-Saharan Africa, *Mbanya (2009)* says: "Soon, four out of every five people with diabetes will live in developing countries. And the men and women most affected are of working age – the breadwinners of their families." Diabetes was once considered as a rare disease in sub-Saharan Africa. But in that part of the word, in 2010; 12.1 millions adults were estimated to have diabetes and by 2030, it is estimated that 23.9 million adults in sub-Saharan Africa will have diabetes.

Data of 2010 on the condition of people with diabetes in sub-Saharan Africa and the complications of diabetes that they suffer is very scarce. According to *Ayesha Motala et al.(2010*), it was estimated that at least:

1. 4.51 million people had eye complications.
2. 2.23 million people needed dialysis because of kidney damage.
3. 907,500 people had cardiovascular disease.

4. 423,500 people were blind because of diabetes.
5. 399,300 people had cerebrovascular disease.
6. 169,400 people had lost a foot because of amputation.

Concerning Rwanda, the number of deaths due to diabetes in 2013 was estimated to be 5464.

In that same year, the prevalence in adults (20-79 years) was 4.38% and the total number of people living with diabetes was estimated to be 234000. (*IDF, 2013*)

1.2. Research objectives and hypothesis:

This study has the following objectives:

1. To test for the association between the risk factors (older age, gender, smoking, occupation status, alcohol consumption, Cholesterol level, hypertension, and family history of diabetes) and diabetes.

2. To fit a multiple logistic model on the incidence of diabetes given the risk factors (older age, gender, smoking, occupation status, alcohol consumption, Cholesterol level, hypertension, and family history of diabetes)

The following hypotheses were formulated in order to achieve the above objectives:

$\boldsymbol{H_0}$**:** There is no association between having diabetes and risk factors like age, gender, smoking, occupation status, alcohol consumption, Cholesterol level, hypertension, and family history of diabetes.

To test those hypotheses, the chi-square test of independence is used.

And

$\boldsymbol{H_0}$**:** $\beta_i = 0$ (it means the coefficient β_i in the fitted multiple logistic regression is not statistically significant)

To test those hypotheses, the Wald test is used.

2. METHODOLOGY

2.1. GENERALISED LINEAR MODELS (GLMs) AND LOGISTIC REGRESION:

The logistic regression model is an example of a broad class of models known as Generalized Linear Models (GLMs). For example, GLMs also include linear regression, ANOVA, Poisson regression, etc.

There are three components to a Generalized Linear Model:

-Random Component: The *random component* of a Generalized Linear Model identifies the response variable Y and selects a probability distribution for it. Denote the observations on Y by $(Y_1, Y_2, \ldots, Y_n)$. Standard GLMs treat $Y_1, Y_2, \ldots, Y_n$ as independent.

-Systematic Component: The *systematic component* of a GLM specifies the explanatory variables. These enter linearly as predictors on the right-hand side of the model equation. That is, the systematic component specifies the variables that are the $\{x_j\}$ in the expression:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

This linear combination of the explanatory variables is called the *linear predictor*.

-Link Function: Let us denote the expected value of Y, the mean of its probability distribution, by $\mu = E(Y)$.

The third component of a GLM, the *link function*, specifies a function $g(\cdot)$ that relates μ to the linear predictor as:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

The link function $g(\cdot)$ connects the random and systematic components.

2.2. LOGISTIC REGRESSION MODEL:

2.2.1. Introduction:

In general, the logistic regression model is used to model the outcomes of a categorical dependent variable.

Logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.

The logistic regression is the most popular multivariable method used in health science *(Tetrault et al., 2008).*

2.2.2. Binary Logistic Regression with single independent variable:

Many categorical response variables have only two categories. Denote a binary response variable by *Y* and its two possible outcomes by 1 ("success") and 0 ("failure").

The distribution of *Y* is specified by probabilities:

$P(Y = 1) = \pi$ of success and $P(Y = 0) = (1 - \pi)$ of failure. Its mean is $E(Y) = \pi$.

For n independent observations, the number of successes has the binomial distribution specified by the index n and parameterπ. Although Generalized Linear Models can have multiple explanatory variables, let us start by introducing only one independent variable$x$.

The value of π can vary as the value of x changes, and π is will be replaced by $\pi(x)$ to describe that dependence π on x.

Relationships between *π(x)* and *x* are usually nonlinear rather than linear. In the logistic regression model, the random component for the (success, failure) outcomes has a *binomial distribution*. The link function is the logit function $ln[\pi/(1 - \pi)]$ of π, which is defined as the log of odds of success and symbolized by "logit(π)." Logistic regression models are often called *logit models*. Whereas π is restricted to the range [0,1], the logit can be any real number.

The model:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X \qquad (2.1)$$

From equation (2.1), we deduce:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 X}$$

$$\pi(x) = e^{\beta_0+\beta_1 X} - \pi(x)e^{\beta_0+\beta_1 X}$$

$$\pi(x)\left(1 + e^{\beta_0+\beta_1 X}\right) = e^{\beta_0+\beta_1 X}$$

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad (2.2)$$

2.2.3. Interpretation of regression coefficients:

Consider the case in which the dependent variable may take only the values 1 (for success) and 0 (for failure) and a single independent variable x.

In this case, the logistic regression equation is:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \quad \text{as given in equation (2.1)}$$

Now, suppose we consider an impact of a unit increase in x. The logistic regression equation becomes:

$$ln\left(\frac{\pi'(x)}{1-\pi'(x)}\right) = \beta_0 + \beta_1(x + 1)$$

$$ln\left(\frac{\pi'(x)}{1-\pi'(x)}\right) = \beta_0 + \beta_1 x + \beta_1 \qquad (2.3)$$

Subtracting equation (2.1) from (2.3) we get:

$$ln\left(\frac{\pi'(x)}{1-\pi'(x)}\right) - ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x$$

to arrive at:

$$\beta_1 = \ln\left(\frac{\frac{\pi'(x)}{1-\pi'(x)}}{\frac{\pi(x)}{1-\pi(x)}}\right)$$

$$\beta_1 = \ln\left(\frac{Odds'}{Odds}\right) \quad (2.4)$$

That is, β_1 is the log of the ratio of the odds at $x+1$ and x.

Which may be also written as:

$$e^{\beta_1} = \frac{Odds'}{Odds} \quad (2.5)$$

The regression coefficient β_1 is interpreted as the log of the odds ratio comparing the odds after a one unit increase in x to the original odds.

2.2. 4. Multiple Logistic Regression:

2.2.4.1. The model:

Let us consider the general logistic regression model with multiple explanatory variables. Denote the k predictors for a binary response Y by $X_1, X_2, \dots, X_k$.

We use $\pi(x)$ to represent the probability that $Y = 1\ for\ success$, and $1-\pi(x)$ to represent the probability that $Y = 0$.

These probabilities are written in the following form:

$$\pi(x) = P(Y = 1/X_1, X_2, \dots, X_k) \quad (2.6)$$

$$1-\pi(x) = P(Y = 0/X_1, X_2, \dots, X_k) \quad (2.7)$$

The model for the log odds is:

$$logit\left(\pi(x)\right) = \ln\frac{P(Y=1/X_1,X_2,\dots,X_k)}{P(Y=0/X_1,X_2,\dots,X_k)} = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_k + \varepsilon$$

$$\therefore\ ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_k + \varepsilon$$

$$\therefore\ ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{j=1}^{k}\beta_jX_j + \varepsilon \quad (2.8)$$

which yields to:

$$\pi(x) = P(Y = 1/X_1, X_2, \dots, X_k) = \frac{e^{\beta_0 + \sum_{j=1}^{k}\beta_jX_{j+\varepsilon}}}{1+e^{\beta_0 + \sum_{j=1}^{k}\beta_jX_{j+\varepsilon}}} \quad (2.9)$$

The parameter β_j refers to the effect of X_j on the log odds that $Y = 1$, controlling the other predictor variables. For example, $exp(\beta_j)$ is the multiplicative effect on the odds of a one-unit increase in X_j, at fixed levels of the other predictor variables.

Thus we have constructed a logistic regression model that bounds the conditional mean between 0 and 1.

2.2.4.2. The Parameters estimation:

The goal of logistic regression is to estimate the $K+1$ unknown parameters $\beta = (\beta_0, \beta_1, \dots \beta_k)$ in equation 2.9. This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. The maximum likelihood equation is derived from the binomial distribution of the dependent variable.

For a set of observations in the data $(x_i\ ;\ y_i)$, the contribution to the likelihood function is $\pi(x_i)$, where $y_i = 1$, and $1-\pi(x_i)$, where $y_i = 0$. The following equation results for the contribution (call it $\varphi(xi)$) to the likelihood function for the observation $(x_i\ ;\ y_i)$:

$$\varphi(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (2.10)$$

The equation 2.10 accounts for only one set of observations. The observations are assumed to be independent of each other so we can multiply their likelihood contributions to obtain the complete likelihood function. The result is given in equation (2.11):

$$l\,(\beta) = \prod_{i=1}^{k}\varphi(x_i) = \prod_{i=1}^{k}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

$$\therefore \quad l(\beta) = \pi(x_i)^{\sum_{i=1}^{k} y_i}[1-\pi(x_i)]^{k-\sum_{i=1}^{k} y_i}$$

$$\therefore \quad l(\beta) = \pi(x_i)^{\sum_{i=1}^{k} y_i}[1-\pi(x_i)]^{k}[1-\pi(x_i)]^{-\sum_{i=1}^{k} y_i}$$

$$\therefore \quad l(\beta) = \left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]^{\sum_{i=1}^{k} y_i}[1-\pi(x_i)]^{k} \qquad (2.11)$$

Note that the equation (2.8) and (2.9) give respectively:

$$\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j} \quad \text{and} \quad \pi(x_i) = \frac{e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}{1+e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}$$

This leads us to write equation (2.11) as:

$$l(\beta) = \left(e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}\right)^{\sum_{i=1}^{k} y_i}\left(1 - \frac{e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}{1+e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}\right)^{k}$$

$$\therefore l(\beta) = \left(e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k}\beta_j X_j}\right)\left(\frac{1+e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j} - e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}{1+e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}\right)^{k}$$

$$\therefore \quad l(\beta) = \left(e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k}\beta_j X_j}\right)\left(\frac{1}{1+e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}}\right)^{k}$$

$$\therefore \quad l(\beta) = \left(e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k}\beta_j X_j}\right)\left(1 + e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}\right)^{-k} \qquad (2.12)$$

In the equation (2.12), β is the collection of parameters $\beta_0, \beta_1, \ldots \beta_k$, and $l(\beta)$ is the likelihood function of β. The Maximum likelihood estimates (MLE's) $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ can be obtained by calculating the β which maximizes $l(\beta)$. However, to simplify the mathematics, let us take the logarithm of equation (2.12). As shown in equation (2.13), $L(\beta)$ denotes the log likelihood expression.

$$L(\beta) = \ln(l(\beta)) = ln\left[\left(e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k}\beta_j X_j}\right)\left(1 + e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}\right)^{-k}\right]$$

$$\therefore L(\beta) = \ln(l(\beta)) = \left(\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k}\beta_j X_j\right) - k\, ln\left(1 + e^{\beta_0 + \sum_{j=1}^{k}\beta_j X_j}\right) \qquad (2.13)$$

The critical points of a function (maxima and minima) occur when the first derivative equals 0. If the second derivative evaluated at that point is less than zero, then the critical point is a maximum. Thus, finding the maximum likelihood estimates requires computing the first derivative of the log likelihood function $L(\beta)$.

Thus, differentiating equation (2.13) with respect to β_0, we get:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k} y_i - k\frac{e^{\beta_0 + \sum_{k=0}^{K}\beta_j x_j}}{1+e^{\beta_0 + \sum_{k=0}^{K}\beta_j x_j}}$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k} y_i - k\pi(x_i)$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k}[y_i - \pi(x_i)] \qquad (2.14)$$

Also, differentiating equation (2.13) with respect to β_j, we get:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} y_i \sum_{j=1}^{k} x_j - k\sum_{j=1}^{k} x_j \frac{e^{\beta_0 + \sum_{k=0}^{K}\beta_j x_j}}{1+e^{\beta_0 + \sum_{k=0}^{K}\beta_j x_j}}$$

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} y_i \sum_{j=1}^{k} x_j - k\sum_{j=1}^{k} x_j \pi(x_i)$$

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} x_{ik}[y_i - \pi(x_i)] \qquad (2.15)$$

The maximum likelihood estimates $\hat{\beta}_0$ *and* $\hat{\beta}_j$ for β_0 *and* β_j can be found by setting each of the equations respectively (2.14) and (2.15) equal to zero and solving for each β_j .

It means, solving

$$\sum_{i=1}^{k}[y_i - \pi(x_i)] = 0 \qquad (2.16)$$

and

$$\sum_{i=1}^{k} x_{ik}\,[y_i - \pi(x_i)] = 0 \qquad (2.17)$$

The solving of these likelihood equations requires special statistical software packages.

2.3. Sample size and sampling procedure:

Gitwe Hospital and the three years 2011, 2012 and 2013 were purposively selected according to the objectives of the study.

The target population of the study includes in total 311 patients from Gitwe Hospital (2011-2013) dispatched in the following six different sectors as:

Table 2.1: Number of patients by sector

Sector	Ruhango	Kabagali	Mukingo	Kinihira	Bweramana	Busoro	**TOTAL**
Number of patients	18	46	43	35	166	3	**311**

Source: Researcher, March 2015.

The sample size for patients is determined using the *Yamane* (1967) formula which is:

$$n = \frac{N}{1+N(e)^2} \qquad (2.18)$$

Where N is the population size and e is the precision level.

Concerning our case study; the total number of patients' folders (N) is 311.Then, by the equation (2.18), the sample size is given as $n = \frac{311}{1+311*(0.05)^2} = 174.9 \approx 175$

Table: 2.2: Calculation of sample size by Sector

PERCENTAGE FOR EACH SECTOR	SAMPLE SIZE BY SECTOR
RUHANGO: $\frac{18*100}{311} = 6\ \%$	$n_{Ruhango} = \frac{175*6}{100} = 10.5 \approx 11$
KABAGALI: $\frac{46*100}{311} = 15\%$	$n_{Kabagali} = \frac{175*15}{100} = 26.25 \approx 26$
MUKINGO: $\frac{43*100}{311} = 14\ \%$	$n_{Mukingo} = \frac{175*14}{100} = 24.5 \approx 25$
KINIHIRA: $\frac{35*100}{311} = 11\ \%$	$n_{Kinihira} = \frac{175*11}{100} = 19.25 \approx 19$
BWERAMANA: $\frac{166*100}{311} = 53\ \%$	$n_{Bweramana} = \frac{175*53}{100} = 92.75 \approx 93$
BUSORO: $\frac{3*100}{311} = 0.9\ \%$	$n_{Busoro} = \frac{175*0.9}{100} = 1.5 \approx 1$
TOTAL SAMPLE SIZE	**n=175**

Source: Researcher, March 2015

However, systematic random sampling has been used to select the patients' folders to be included in the sample size of each Sector.

3. RESULTS PRESENTATIONS

3.1. Chi-square test of association between the dependent and independent variables:

Table: 3. 1:Chi-square test results

Factor	p-value	Conclusion
Older age	.000	There is statistical evidence between age and the outcome of diabetes.
Gender	.899	No statistical evidence between gender and the outcome of diabetes.
Occupation status	.001	The outcome of diabetes is statistically associated with the occupation status.
Smoking	.679	The outcome of diabetes is not statistically associated with the smoking.
Alcohol consumption	.027	The outcome of diabetes is statistically associated with the alcohol consumption.
Cholesterol level	.001	There is statistical evidence of the association between the outcome of diabetes and the cholesterol level.
Hypertension	.000	The outcome of diabetes is statistically associated with the hypertension.
Family history of diabetes	.289	The outcome of diabetes is not statistically associated with the family history of diabetes.

Source: Researcher, March 2015

The table 3.1 reveals that the risk factors with statistically significance association to the outcome of diabetes are older age, Occupation status, alcohol consumption, cholesterol level and Hypertension.

Other factors like gender, smoking and family history of diabetes have no statistical significance to the outcome of the disease.

3.2. Multiple logistic regression model fitting:

3.2.1. The fitted model with all the predictor covariates:

Table 3. 2: The Estimated coefficients, their standard error, and Wald test for the full model

Parameter	B	Std. Error	Wald	Sig.
Intercept	-47.549	13.853	11.781	.001
Age of patient	1.142	.259	19.459	.000
Gender of patient	.143	.408	.123	.726
Occupation Status	-1.208	.397	9.252	.002
Smoking	-.640	.678	0.891	.345
Alcohol consumption	.818	.387	4.466	.035
Cholesterol level	.991	.394	6.324	.012
Hypertension	1.028	.391	6.914	.009
Family history of diabetes	.482	.458	1.106	.293

Source: Researcher, March 2015

The table 3.2 displays parameter estimates in the B column, the standard error and the Wald test.

Thus, using the estimates of the parameters in table 3.2, we get the following model:

$$\pi = \frac{\left(\exp(-47.549+1.142*age+0.143*gen-\ 1.208*Occ.Status-0.640*smok\ +0.818*Alcoh.+\ 0\ .991*Chol\ +\ 1.028*Hyper\ +0.482*Famhist+\varepsilon\ )\right)}{1+\left(\exp(-47.549+1.142*age+0.143*gen-\ 1.208*Occ.Status-0.640*smok\ +0.818*Alcoh+\ 0\ .991*Chol\ +\ 1.028*Hyper\ +0.482*Famhist\ +\varepsilon)\right)} \quad (3.1)$$

3.2.2. Testing for the significance of the individual parameters in the model:

To test the hypothesis:

H_0 : $\beta_j = 0\ \ (for\ the\ individual\ parameter\ \beta_j)$

Versus

$H_1 : \beta_j \neq 0\ \ (for\ the\ individual\ parameter\ \beta_j)$

Consider Wald and Sig. column of the table 3.2. The information given by the table reveals that the significant predictors are: Older age (p-value = 0.000 < 0.05), Occupation status (p-value = 0.002 < 0.05), Alcohol consumption (p-value=0.035 < 0.05), Cholesterol level (p-value=0.012 < 0.05) and Hypertension (p-value=0.009 < 0.05).

On the other hand, the predictors which are not statistically significant are:

Gender (p-value = 0.726 > 0.05), smoking (p-value = 0.345 > 0.05) and family history of diabetes (p-value = 0.293 > 0.05).

3.2.3. Signs of coefficients analysis:

The sign of the coefficients of the estimated logistic function in Table 3.2 above gives an explanation of the explanatory variables used, as given in Table 3.3.

Table: 3. 3: The sign analysis

Covariate	Codes	Sign	Explanation
Older age	1 old 0 young	Positive	Older age increases the probability of having diabetes.
Gender	1 Male 0 Female	Positive	Male increases the probability of having diabetes.
Occupation status	1 Employed 0 Unemployed	Negative	To be employed decreases the probability of having diabetes.
Smoking	1 No 0 Yes	Negative	Not Smoking decreases the probability of having diabetes.
Alcohol consumption	1 Yes 0 No	Positive	Consumption of alcohol increases the probability of having diabetes.
Cholesterol level	1 High 0 Law	Positive	High cholesterol level increases the probability of having diabetes.
Hypertension	1 Yes 0 No	Positive	Being hypertensive increases the probability of having diabetes.
Have a Family History of diabetes	1 Yes 0 No	Positive	Having a Family History of diabetes increases the probability of getting the disease.

Source: Researcher, March 2015

3.2.4. The odds ratio results:

The Exp(B) column contains the exponential of parameter estimates. These values represent odds ratios for the corresponding predictor variables. In the table 3.4 bellow, the 95% Wald confidence limit shows the confidence interval (CI) for the odds ratio.

Table 3.4: Odds Ratios and 95% Confidence Intervals for Covariates

Variable	Exp(B)	95% Confidence Interval for Exp(B)	
		Lower Bound	Upper Bound
Age of patient	3.133	.192	.530
Gender	1.154	.519	2.566
Occupation status	0.299	1.537	7.287
Smoking	0.527	.140	1.991
Alcohol consumption	2.266	1.061	4.840
Cholesterol level	2.694	1.244	5.832
Hypertension	2.796	1.299	6.017
Have a Family History of diabetes	1.619	.660	3.976

Source: Researcher, March 2015

From Table 3.4 , it is evident that patients of older age, patients who consume alcohol, persons with high cholesterol level and hypertensive persons are highly susceptible for diabetes occurrence.

3.2.5. The full model assessment:

Table: 3.5: Likelihood ratio test

	Model Fitting Criteria	Likelihood Ratio Tests		
MODEL	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept only	204.670			
Final	144.209	60.461	8	.000

Source: Researcher, March 2015

The table 3.5 displays the Likelihood Ratio test.

The -2 log likelihood for the constant only model obtain by fitting the constant only model is 204.670; and the -2 log likelihood for the overall model was 144.209.

Thus the value of the likelihood ratio test is;

$$G = 204.670 - 144.209 = 60.461$$

The null hypothesis is:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_8 = 0.$$

$$H_1 : \exists\, \beta_j \neq 0, j = 1, 2, \ldots, 8$$

The results show that at least one of the predictors ' regressions coefficient is not equal to zero because of the small p-values =0.000 which is less than 0.05. This would lead us to reject H_0 in favor of H_1 and we conclude that at least one and perhaps all beta's coefficient are different from zero.

Table 3.6: Classification table for the model with all predictor variables.

	Predicted		
Observed	No	Yes	**Percent correct**
No	43	23	**65.2%**
Yes	16	93	**85.3%**
Overall percentage	33.7%	66.3%	**77.7%**

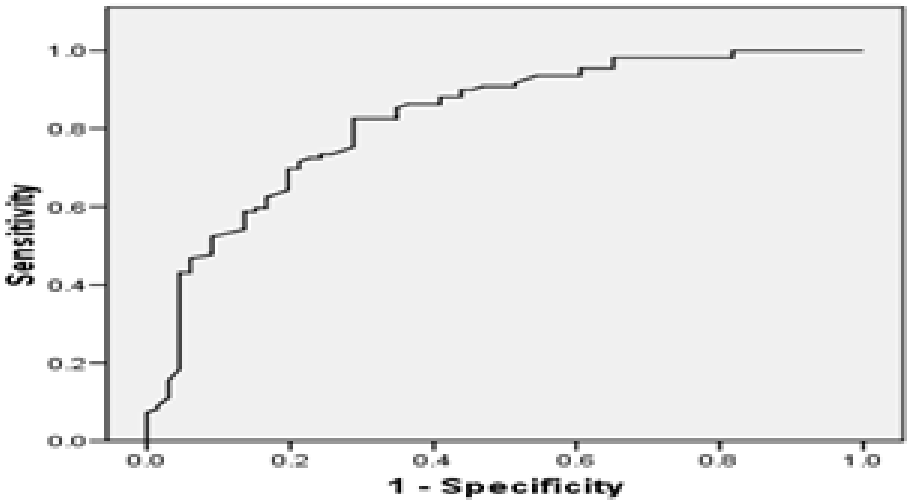
Source: Researcher, March 2015

From table 3.6, we conclude that:

65.2% of all patients who do not have diabetes are correctly classified and 34.8% are incorrectly classified.

85.3% from all patients who have diabetes are correctly classified and 14.7% are incorrectly classified.

The overall correct percentage was 77.7% which reflects the model's overall explanatory strength.



Source: Researcher, March 2015

Figure 3. 1: Receiver Operating Characteristic (ROC) curve for the full model

By use of the ROC curve in *figure 3.1* for the classification accuracy, it is found that the area under the ROC curve, which ranges from 0 to 1 provides the measure of the model's ability to discriminate between those subject who experience the response of interest versus those who do not. The area under the ROC curve for the full model is 0.825 which may be considered as reasonable discrimination.

Table 3.7: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.037	8	.533

Source: Researcher, March 2015

By Hosmer and Lemeshow test, the table 3.7 gives the output from SPSS 15.0 .

Our *Hosmer-Lemeshow statistic* has a significance of 0.533 which means that it is not statistically significant and we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level.

3.2.6 .The model with significant parameters only:

The following step is the fitting of model with statistically significant parameters only (age, occupation status, alcohol consumption, cholesterol level and hypertension).

Results are summarized in Table 3.8.

Table 3. 8: Summarized results for the reduced model

Parameter	B	Std. Error	Wald	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
						Lower Bound	Upper Bound
Intercept	-44.679	9.170	23.742	.000			
Age of patient	1.089	.249	19.217	.000	2.971	.207	.548
Occupation Status	-1.215	.390	9.685	.002	0.297	1.568	7.241
Alcohol consumption	.825	.381	4.682	.030	2.282	1.081	4.818
Cholesterol level	.962	.386	6.223	.013	2.616	1.229	5.570
Hypertension	1.043	.386	7.318	.007	2.838	1.333	6.041

Source: Researcher, March 2015

From Table 3.8, the reduced model is written as follows:

$$\pi = \frac{\exp(-44.679+1.089*age-1.215*Occ.Stat + 0.825*Alcoh + 0.962*chol + 1.043*Hyper)}{1+\exp(-44.679+1.089*age-1.215*Occc.Stat + 0.825*Alcoh + 0.962*chol.+ 1.043*Hyper)} \quad \textbf{(3.2)}$$

The results above indicates that: patients with older age are more susceptible to develop diabetes; An employed person is less susceptible to develop diabetes; consuming alcohol increases the susceptibility; persons with high cholesterol level are more susceptible than those with low cholesterol level and hypertensive patients are more likely to develop diabetes than those who are not hypertensive.

The exponent (Exp (B)) in Table 3.8 is the odds ratio. Thus, for example:

-The odds for patients who consume alcohol to those patients who do not take it to develop diabetes is 2.282.

-The odds for patients with high cholesterol level to patients with low cholesterol level to develop the illness is 2.616.

-The odds for hypertensive person to that one who is not hypertensive to develop diabetes is 2.838.

Table 3. 9: Classification table for the reduced model

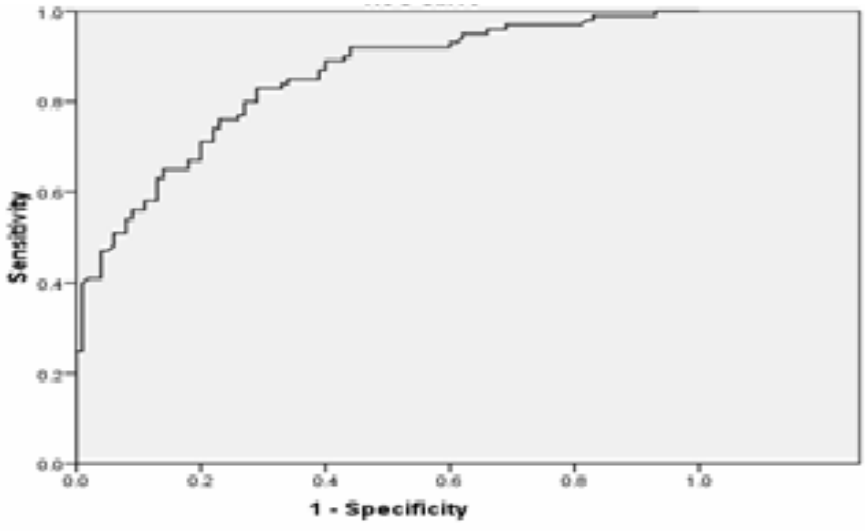
		Predicted		
		The Patient is diabetic		Percentage Correct
Observed		No	Yes	No
The Patient is diabetic	No	41	15	62.1%
	Yes	11	98	89.9%
Overall Percentage		29.7%	64.6%	79.4%

Source: Researcher, March 2015

Table 3.9 gives the classification table. The information from the same table is that observations are classified as follows:

- 62.1% of all patients who do not have diabetes are correctly classified, and 37.9% are incorrectly classified.
- 89.9% from all patients who have diabetes are correctly classified, 10.1% are incorrectly classified.
- The overall correct percentage was 79.4%, which reflects the model's overall explanatory strength.

Plotting sensitivity versus (1–specificity) over all possible cut-points is shown in the Figure 3.2 below .The area under the ROC curve for the full model is 0.843 this is considered reasonable discrimination.



Source: Researcher, March 2015

Figure: 3.2: Receiver Operating Characteristic (ROC) curve for the reduced model

Comparing the two models (model with all predictor covariates and the reduced model), area under the ROC curve has become a particularly important measure for evaluating models' performance because it is the average sensitivity over all possible specificities. The larger the area, the better the model performs. (Bradley,1997).

We conclude that the reduced model (which has the area under the ROC curve of 0.843 and its overall explanatory strength is 79.4%) fits better the data than the model with all predictor variables (which has the area under the ROC curve of 0.825 and its overall explanatory strength is 77.7%).

4. CONCLUSION AND RECOMMENDATIONS

In this study, risk factors of developing diabetes using logistic regression model were studied. The binary logistic regression model is used to estimate the probability of having diabetes. Firstly, the chi-square test of association between diabetes and all the predictor variables showed that older age, occupational status, alcohol consumption, cholesterol level and hypertension are statistically significant.

Secondly, significance testing for the logistic coefficients using Wald test show that factors like older age, occupational status, alcohol consumption, cholesterol level and hypertension are significant as predictor variables of diabetes. The model fitted showed that getting diabetes does not depend significantly on the gender of a person, having a family history of diabetes and smoking. Instead, there is an increased risk of getting the diabetes as a person gets older. To assess the fitness of the model the maximum likelihood test and Hosmer and Lemeshow test are used.

Based on the findings from this study, the following recommendations are formulated in order to give our contribution in fighting the most disabling disease like diabetes in people:

-The people of rural areas would be aware of the diabetes and know that it is no longer a disease for rich persons or for elders but it has been common for all social classes and of all ages.

- Continue the good habit of doing physical activities that many researchers have shown that the risk of diabetes and associated insulin resistance can be reduced significantly by trying to lose weight, especially for those who are severely obese (BMI > 35 kg/m^2).

- The nutrition should also play a vital positive role on health. By eating sufficient fruits and vegetables, one gets access to several health benefits, due to an assumed complex interaction of containing biological active compounds.

- To go to healthcare centers regular tests for the occurrence of diabetes in the body.

- To cease bad habits like smoking and alcohol consumption.

- More follow up studies should be done to assess the benefits of different treatment modalities on control of cardiovascular risk factors such as blood pressure and lipids in diabetes patients to prevent serious complications in Rwanda. Especially, assessing the effect of the interventions based on healthy lifestyle such as increased physical activity, smoking cessation, weight loss and a healthy dietary pattern, and the rural area should be focused on.

REFERENCES

[1] Afifi, A., Clark, V.A., & May, S. (2004). Computer-aided multivariate analysis. Forth edition. Champman and Hall/CRC.

[2] Alan, A. (2007). An introduction to categirical analysis.Second edition. Gainesville, Florida: Departement of Statistics, University of Florida.

[3] Ayesha, M., & Kaushik, R. (2010). Diabetes, the hidden pandemic and its impact on sub-saharan Africa. Johannesburg: Diabetes leadership forum.

[4] Belsley, D.A, Kuh, E., & Welsch, R.E. (1980). Regression diagnostics: identifying influential data and sources of collinearity. New York: wiley.

[5] Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression.Critical care. London,England.: http//dx.doi.org/10.1186/cc3045.

[6] Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition.

[7] Chao-Ying, J. (2002). Logistic regression analysis and reporting.Departement of canseling and educational psychology. Iindiana University-Bloominton.

[8] Elisa, T.L., & John, W.W. (2003). Statistical method for survival data analysis. Third edition. Wiley interscience.

[9] Federation, I. D. (2009). Diabetes Atlas. Brussels.

[10] Ferri, C., Flach, P., & Hernandez-Orallo, J. (2002). Learning decision trees using the area under the ROC curve. Morgan Kaufmann.

[11] Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1995). Multivariate data analysis, 3rd edition. New York: Macmillan.

[12] Hosmer, D.w., & Lemeshow, S. (2000). Applied logistic regression. New York: John Wiley& sons Inc.

[13] IDF. (2009). Diabetes Atlas. 4th edition. Brussels.

[14] IDF. (2013). Diabetes Atlas. sixth edition. Brussels.

[15] James, E. (2013). Theory of statistics. Virginia: Fairfax county.

[16] Jennings, D. (1986). Judging inference adequacy in logistic regression. Journal of an the American Statistical Association. 81,471-476.

[17] John, A. (1995). Mathematical statistics and data analysis.second edition. International Thomson publishing.

[18] Katz, M. (1999). Multivariable analysis: Apractical guide for clinicians. Cambridge University press.

[19] Lemeshow, S., & Hosmer, D.W. (1983). Estimation of odds ratios with categirical scaled covariates in multiple logistic regression analysis. American Journal of Epidemiology, 119,147-151.

[20] Mbanya, J. (2009). Making a difference to global diabetes. Diabetes voice, 54.

[21] Mccullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. . Journal of the American Statistical Association,81,104-107.

[22] Morris, J.A., & Gardner, M.J. (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. British medical Journal(clinical reasearch Ed.),296(6632),1313-1316.

[23] Neter, J., Wasserman, w., & Kutner, M.H. (1989). Applied linear regression models. Homewood,IL:Irwin.

[24] Ramachandran, K.M., & Chris, P.T. (2009). Mathematical statistics with applications. Elsevier Academic Press.

[25] Ronald, C. (1990). Log-linear models and logistic regression. Second edition. springer.

[26] Sahoo, P. (2013). Probability and mathematical statistics. Departement of statistics,. University of Lousville, KY40292 USA.

[27] Sarah, W. (2004). Global prevalence of Diabetes: estimates for the year 2000 and the projections for 2030.

[28] Scott, A. (2010). Maximum likelihood estimation of logistic regression models: Theory and implementation. http://czep.net/contact.html.Accessed on Sept 21, 2014.

[29] Shaw,J.E., Sicree, R.A., & Zimmet, P.Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Clin Pract 2010.

[30] Stella, A. (2012). Multiple regression analysis to determine risk factors for the clinical diagnosis of diabetes. Case study: Komfo Anokye teaching hospital (2008-2009).

[31] Tetrault, J.M., Sauler, M., Wells, C.K., & Concato,J. (2008). reporting of multivariable methods in the medical literature. Journal of investigate medicine,56.

[32] WHO. (2011). Global Atlas on cardiovascular disease prevention and control. Mendis,S., Puska, P., Norrving, B. editors.(in collaboration with the World Heart Hedereration and World Stroke Organisation). Geneva.

[33] Yamane, T. (1967). Statistics: An introductory Analysis, 2nd Ed. New York: Harper and Row.