A Study of Risk Factors Associated With Diadetes Using A Multiple Logistic Regression Modelling

Anthony K. Odior #1,Felix Elugwu *2

[1]Department of Statistics, Delta-State Polytechnic, Otefe- Oghara, Delta State Nigeria.

[2]Department of Computer Science, Delta-State Polytechnic, Otefe- Oghara, Delta State Nigeria.

Abstract

This paper examined empirically the risk factors associated with a common human disorder termed diabetes. The study considers the age of patients (AOP), gender of patients (GOP), occupational status of patients (OSOP) as possible risk factors that occasioned the health challenge of diabetes. The study utilized secondary data captured through the record unit of Central Hospital, Sapele Delta -State. The fitted multiple logistic regression model using AOP, GOP, and OSOP as independent variables revealed that AOP and OSOP were statistically significant in model as influencing risk factors and are associated with a higher probability of causing the human disorder diabetes while GOP was statistically not significant in the model as a contributory risk factor under investigation. Also the estimated logistic regression parameters: $\hat{\beta}_0$=-47.549, $\hat{\beta}_1$=1.142, $\hat{\beta}_2$=0.143 and $\hat{\beta}_3$=-1.208 respectively indicates that independent variables with higher positive value is associated with a higher probability of the risk factor in causing the disorder diabetes. Odd ratio analysis (ODA) revealed that patients of older age are highly susceptible to diabetic occurrence when compared with other risk factors.

Keywords*: Multiple logistic regressions, Diabetes, Risk factors, Independent variables, Dichotomous response variable, odd ratio.*

I. INTRODUCTON

One emerging health challenging confronting the human race is diabetes. Technically refer to as chronic disease that occur either when the pancreas does not produce enough insulin or when the body mechanism cannot effectively use insulin it produces. Today’s emerging diabetes hotpots include countries in the middle East, Western pacific, sub-saharan Africa and south-East Asia where economic development has transformed lifestyle. Statistics released by the [7] suggest that the number of people living with diabetes has almost quadrupled since 1980 to 225 million adults with most of them living in developing countries. In Nigeria, diabetes is growing at the rate of 4.3% with individuals of age 18 years and above mostly affected. The world prevalence rate of diabetes in 2010 among adults aged 20-79 years was estimated to 6.4% affecting 285 million of adults [1].

The prevalence rate of diabetes as a global epidemic has become a major source of concern to health workers and the population at risk. It is characterized by high blood glucose level resulting from defect in insulin production, insulin action or both. Along with the increase of diabetes both individual and societal expectation concerning the management of diabetes have also increased with many reports from the center for disease control (CDC), United States department of health and human services (USDHHS), and the national institute of health (NIH) urging patients to take charge of diabetes and thereby conquer this human monster called diabetes. The management and control of diabetes is intended to improve the quality life that will be almost impossible if the associated risk factors are not properly checked.

Many diseases such as cardiovascular disease, retinopathy, nephropathy, peripheral vascular and peripheral neuropathy are co-morbidities of diabetes. In addition to this serious complication, diabetes often cause life threatening event such as diabetes ketoacidosis and hyperosmolar (nouketotic) coma resulting from biochemical imbalances [5]. Infections such as influenza or pneumonia are also of serious concern for patient with diabetes asthey are more likely to die as a result of infection than individual who do not have diabetes. Diabetes has been associated with increase complication, such as heart disease, stroke, kidney disease, blindness, neural damage. Diabetes is now truly pandemic and its effects are particularly severe in low and middle income countries.

The burden of illness caused by diabetes and the reduction in life expectancy in sub-saharan Africa will continue hinder the region’s economic growth and development.

According to [3] population based study of cardiovascular diseases (CVD) risk factors trends among subject with and without diabetes shows differing trend in favor of those without diabetes Stella (2008) in his previous studies have found appropriate lifestyle intervention and or drug treatments as effective method of preventing both diabetes and its complications.

II. MODEL SPECIFICATION AND DEVELOPMENT

This paper addresses the possible risk factors associated with a common human disorder called diabetes using the multiple logistic regression analysis (MLRA). The statistical tool provides a flexible general-purpose modeling strategy with straightforward interpretation. Its application in research permeates social and biomedical sciences involving the need to predict the probability that an event will occur or not. It is a technique for modeling dichotomous (two category) Y variables but with polychotomous explanatory variables. Therefore, MLRA is usually employed in research studies to model the association between a binary response and a set of explanatory variables. MLRA is intended to obtain the best fitting equation and most parsimonious model to describe the relationship between a categorical out-come variable and the one or more categorical or continuous predictors (independent) variables. MLRA is instead a designation of one of two possible out-comes: alive or dead, success or failure, yes or no, defect or defect free by [6]. Over the years, logistic regression modelling has become applicable in many fields the standard method of analysis in this situation. Generally, in logistics regression, statistical theory as well as in practice the relationship between E(Y) and X_1, X_2…$X_p$can be better specified by a non-linear equation.

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}}$$

…………………….……………………………..(1)

Given that the two values of the dependent variable $\pi(x)$ are coded as 0 or 1, the $\pi(x)$ in equation (1) provides the probability that $Y = 1$ given a particular set of value for the independent variables $X_1, X_2, ...X_p$

The transformation of $\pi(x)$ is pivotal in order to express $\pi(x)$ as linear functions of the regression parameters. The logit transformation can be define as:

$$g(x) = \frac{\pi(x)}{1 + \pi(x)} \text{.......................................(2)}$$

$$g(x) = \ln \left[\frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}}}{1 + \left(\dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}} \right)} \right] \text{......................(3)}$$

$$g(x) = ln$$

$$\left[\frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}} \times \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 .}}{1} \right]$$

$$\pi(x_i)^{y_{ii}} (1 - \pi(x_i))^{1 - y_i} \text{ ....(5)}$$

$$g(x) = ln(e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p})$$

$$g(x) = \log_e e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... = \beta_p x_p}$$

$$g(x) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \text{...........................(4)}$$

The transformed g(x) has many of the desirable properties of a linear regression model. The logit g(x) is linear in its parameters may be continuous and may range from - ∞ to $+\infty$ depending on the range of x. The regression parameters are estimated using the maximum likelihood approach. This can be achieved through the construction of likelihood function. Suppose, Y is coded as 0 or 1, $\pi(x)$ provides the set of parameters. Given that $P(y_i = 0/x)$, thus it follows that the quantity 1-$\pi(x)$ gives the conditional probability that Y is equal to zero given x, $P(y_i = 0/x)$ where $y_i = 1$, the contribution to the likelihood function is $\pi(x)$ where $y_i = 0$, the contribution to the likelihood function is 1-$\pi(x_i)$

The pair of x_i can be expressed as:

In general, odds ratio enables researchers to compare the odds for two different events, since the observations are assumed to be dependent, the likelihood function is obtained as the product of the terms in equation (5).

The likelihood expression of equation (5) yields.

$$\beta = \overset{n}{\pi}\, \pi(X_i)^y \left[(X_i)^{1-y} \right] \text{..........................(6)}$$

$$L(\beta) = \ln(\beta) = \sum_{i=1}^{n} \left[y_i \ln(\pi(X_i)) \right]_i = 1 + (1 - y_i) \ln(1 - \pi(X_i))$$

To obtain the value of β that maximizes L(β), we differentiate L(β) with respect to the parameters partially and set the resulting expression equal to zero. Odd ratios (OR) are often used comparatively to describe the strength of an effect. OR is the ratio of the odds at two different values of X. They provide another excellent way to interpret logic coefficients. The odds in favor of an event occurring is viewed as the probability that the event will occur divided by the probability the event will not occur. In logic regression, the odds in favor of Y=1 can be specified as:

$$odds = \frac{p(y=1/X_{1,}X_{2,}...,X_p)}{p(y=0/X_1,X_2,...,X_p)} = \frac{p(y=1/X_1,X_2,...,X_p}{1-p(y=1/X_1,X_2,...,X_p)_1}$$

Thus OR is estimated by $OR = e^{\hat{\beta}}$. OR measures the impact in the odds of a one- unit increase in only one of the independent variables. In general, odd ratio enables researchers to compare the odds for two different events.

III. DATA ANALYSIS

in order to understand the influence of the selected risk factor on diabetes, a multiple logistic regression model was formulated. The patient is diabetic (PID) as response variable is expressed as a function of select risk factors. The selected risk factors under consideration are: age of patients (AOP), gender of patient (GOP), occupational status of patient (OSOP).

d.

Fig 1 below show the data captured for each of the risk factor under study.

	Response	*Number of patients*
Patient is diabetic	*No*	*66*
	Yes	*109*
AOP	*20-39 years*	*44*
	40-59 years	*55*
	60+	*76*
GOP	*Male*	*105*
	Female	*70*
OSOP	*Employed*	*93*
	Unemployed	*82*

Source: Central Hospital Sapele.

In order to ensure that the parameter estimates, standard error value and the confidence interval provide adequate information about the data and the subsequent statistical inference are reliable, the data was subject to multicollinearly test. One useful way of detecting multicollinearty is to determine the variance inflation factor (VIF).

Fig II shows the result of multi-collinearity using SPSS 15.0

a.

		Collinearity Statistics
	Tolerance	*Variance Inflation factor(VIF)*
AOP	*----------------*	
GOP	*0.971*	*1.030*
OSOP	*0.951*	*1.051*

b.

		Collinearity Statistics
	Tolerance	*VIF*
GOP	*------------*	*--------------------*
AOP	*0.947*	*1.056*
	0.948	*1.055*

c.

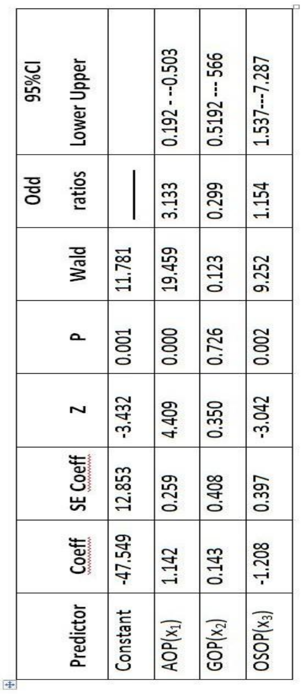
		Collinearity Statistics
	Tolerance	*VIF*
OSOP		*------------------*
AOP	*0.879*	*1.134*
rGOP	*0.898*	*1.114*

The value of VIF that is less than 10 indicates inconsequential collinearity between the independent variable in the model, [4]

Therefore, Multiple logistic regression model is specified as:

$$\pi \;= (\frac{(e^{\beta_0 + \beta_1 AOP + \beta_2 GOP + \beta_3 OSOP})}{(1+e^{\beta_0 + \beta_1 AOP + \beta_2 GOP + \beta_3 OSOP})})$$

Fig III: Partial Logistic Regression Output



Predictor	Coeff	SE Coeff	Z	P	Wald	Odd ratios	95%CI Lower Upper
Constant	-47.549	12.853	-3.432	0.001	11.781	——	
AOP(x_1)	1.142	0.259	4.409	0.000	19.459	3.133	0.192 - --0.503
GOP(x_2)	0.143	0.408	0.350	0.726	0.123	0.299	0.5192 --- 566
OSOP(x_3)	-1.208	0.397	-3.042	0.002	9.252	1.154	1.537---7.287

Therefore, the fitted logistic regression equation can be expressed as follows:

$$\pi = \left(\frac{e^{-47.549 + 1.142x_1 + 0.143x2_2 - 1.208x_3}}{1+e^{-47.549 + 1.142x_1 + 0.143x2_2 - 1.208x_3}}\right)$$

Discussion Results

The estimated logistic regression parameters of β_1 = 1.142, β_2 = 0.143 and β_3 = -0.208 indicate an association of patients with level of probability of having diabetics. In assessing the statistical significance of the independent variables, AOP(x_1) having a Z value of 4.409 and corresponding P value of 0.000. Thus, at the 0.05 level of significance, there is convincing sample evidence to reject $H_o : \beta_1 = 0$. In similar fashion, we accept $H_{o:}\beta_2 = 0$ and reject $H_o : \beta_3 = 0$. Hence, at 0.05 level of significance, AOP and OSOP are the statistically significant while GOP is not statistically significant in this study.

Also, the odds ratio for AOP is 3.133 suggest that AOP has a significant and a positive impact (influence) in the probability of diabetics occurring. The confident interval shows that AOP has a significant effect on the estimated odds ratio since it does contain a value of 1. In the same vain, OSOP has positive impact on the probability of having diabetics since the odds ratios value is greater than 1. However, the odds ratio for GOP is less than 1, an indication that the GOP has no significant impact (influence) on the probability of the patient having diabetics.

IV. CONCLUSION

This study examined common human disorder diabetes (CHDD)and selected risk factor(RF) such as AOP, GOP and OSOP using data generated from a government hospital. The findings revealed that AOP and OSPOP are high risk factors that has the probability of causing diabetes while the gender was not seen as risk factors in the study.

REFERENCES

[1] Alan, A. (2007). An introduction to categorical analysis. Second edition. Department of Statistics, University of Florida.

[2] M. & Kaushik R. (2010). Diabetes; the hidden pandemic and its impact on Sub- Saharan Africa, Prepared for the Diabetes Leadership Forum, Africa, Johannesburg, 30 September and 1 October, 2010.

[3] Bewick, V., L., &Ball, J. (2005). Statistics review 14; Logistic regression. Critical Care (London, England), 9(1) 112-118. http://dx.doi.org/10.1186/cc3045.

[4] Hamiton, (2006).Regression with Graphics: A second course in Applied Statistics. Brooks/Cole Publishing company pacific grove, California.International Diabetes Federation (IDF), Diabetes Atlas, sixt edition 2013.

[5] [5]Scot, A C (2010). Maximum like estimation of logistic regression models: Theory and Shaw, J.E.,& Sicree, R.A &Zimmet, P.Z Global estimates of the prevalence diabetes for 2010 and 2030. Diabetes Res Clin Pract 2010.

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

[7] World Health Organization (WHO, 2014). Fact Sheet No. 312: What is Diabetes? Available at:Http://www.who.iny/mediacentre/factsheets /fs312/en/Accessed on: September 2, 2014.