International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 565-571 www.ijaem.net ISSN: 2395-5252

Multivariate Logistic Regression Modeling for Assessing the Risk Factors of Diabetes in Gusau Zamfara State, Nigeria

Gerald Ike Onwuka [1], Wasiu Babayemi [2], and Mubarak Hassan Augie[3]

Faculty of Physical Science, Department of Mathematics, Statistics Unit.
Kebbi State University of Science and Technology, Aliero. Kebbi state, Nigeria.

Submitted: 02-01-2022 Revised: 09-01-2022 Accepted: 12-01-2022

ABSTRACT: This research is based on the development of a multivariate logistic regression model to assess the effect of various risk factors such as age, cholesterol level, smoking, alcohol consumption, gender, marital status, occupation status, hypertension, and diabetes family history on the prevalence of diabetes in Gusau local government of Zamfara State. The calculated odds ratio of hypertension 2.608 (260.8%), alcohol consumption 2.405 (240.5%), smoking 2.261 (226.1%), high cholesterol level 2.240 (224.0%), family history of diabetes 2.153 (215.3%), and age 1.019 (101.9%) showed that people with these underlying conditions have a higher risk of developing diabetes than those who do not. A multiple logistic regression model was built with all of the predictor variables, and the coefficients were tested for significance. On the one hand, the Wald test demonstrates that age, cholesterol level, smoking, alcohol intake, hypertension, and a family history of diabetes are significant predictors. Gender, occupation, and marital status, on the other hand, are not statistically significant predictors.Model selection criterion such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hosmer-Lemeshow goodness of fit test were all employed and hence the model with only significant parameters outperformed the model with all parameters. As a result, the proposed logistic regression model can be an effective statistical tool for identifying the relationship between the most significant risk factors to diabetes and thus promptly alerting to take necessary actions to lower the chance of developing the disease.

KEYWORDS: Logistic regression, Model, Risk Factor, Diabetes, Goodness-of-Fit

I. INTRODUCTION

The International Diabetes Federation (IDF, 2019), defined Diabetes mellitus, or diabetes, as a dangerous, long-term or chronic disorder in which a person's blood glucose levels are elevated due to their body's inability to generate any or enough insulin, or to efficiently use the insulin that it does produce. Insulin is a hormone that is generated in the pancreas. The pancreas produces insulin, which is a necessary hormone that allow blood glucose to reach the body's cells, where it is turned into energy. Insulin is required for protein and fat metabolism. High blood glucose levels (hyperglycaemia), which is the clinical sign of diabetes, are caused by a shortage of insulin or the inability of cells to respond to it.

Long-term insulin deficiency can harm many of the body's organs, resulting in severe and life-threatening health issues including cardiovascular disease, nerve damage, kidney damage (nephropathy), and eye illness (vision loss, and even blindness). These catastrophic problems, on the other hand, can be delayed or avoided entirely if diabetes is well managed (WHO, 2019).

According to the World Health Organization, approximately 45,060 fatalities in Nigeria were due to diabetes in 2016, with 18,150 deaths occurring in those aged 30 to 69 and 26,910 deaths occurring in people aged 70 and more. Many Nigerians have increasing insulin resistance and a higher prevalence of impaired glucose, indicating a critical need for solutions to assist combat the rising trend of diabetes (WHO, 2016).

Several risk factors for diabetes have been identified by experts all around the world. These risk factors belong to one of four categories of non-communicable disease risk factors (NCDs). Many of these risk factors are related with one's lifestyle and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 565-571 www.ijaem.net ISSN: 2395-5252

diet (Danquah et al., 2012).This way, modifiable risk factors like High cholesterol, High blood pressure, harmful use of alcohol, and Smoking factors can be altered in the prevention of the disease. Non-modifiable risk factors like Increased age, Sex/Gender, family history and Socioeconomic status may serve to alert an individual of his or her baseline risk of contracting the disease. An individual's chance of acquiring diabetes is determined by the total presence and mix of modifiable and non-modifiable risk factors (Gudjinu & Sarfo, 2017).

Given this context, quantifying the prevalence of diabetes and how people will be affected by the risk factors currently and in the future is very important to allow rational planning and implication of adequate actions in timely controlling the disease and its associated risk factors.

The prime objective of this study is to test the association between diabetes and risk factors such as age, gender, marital status, hypertension, smoking, family history, alcohol intake, and occupation status. Secondly, to develop a multiple logistic regression model on the prevalence of diabetes based on the risk factors such as age, gender, smoking, occupation status, alcohol intake, cholesterol level, hypertension, and diabetes family history.

II. LITERATURE REVIEW

Pirzado et al., (2021), evaluate the effectiveness of a logistic regression model for analyzing risk factors linked with diabetes mellitus.Results obtained from univariate and multivariate model that the risk of Diabetic Mellitus among the 45+ age group is 2.90 times higher as compared to those respondents in the ≤20 age group. This risk increased to 3.32 after controlling for income, physical activity, maternal status, and body mass index. Overweight and obese respondents whose body mass index is greater than or equal to 45 kg are at higher risk of developing diabetic mellitus especially female with an odds ratio of 4.57.

Safieddine et al. (2020), examines the socioeconomic inequalities of diabetes in employed individuals, non-working spouses and pensioners.Multivariate logistic regression analysis was applied to examine socioeconomic inequalities in diabetes in the three population subgroupsin Lower Saxony, Germany.Results showed that diabetes prevalence was four times higher in male non-working spouses (24.2%) and 2.6 times higher in female nonworking spouses (12.7%) compared to employed men (6.4%) and women (4.7%) respectively, while it accounted for 40% of men and 36% of women in pensioners.

Rastogi and Singh (2019) Developed a multivariate logistic regression model to assess the effect of various risk factors like age, body mass index, meal-regularity and fast-food consumption on the prevalence of diabetes spatially in urban and rural areas of India.The test-of-association showed that age, body mass index and fast-food were significantly associated with diabetes in rural areas ($p<0.05$). While, age and body mass index were found to be associated with diabetes in urban area. The Wald test and Odds ratio (OR>1) showed that age and body mass index were significant predictors of diabetes.

Niyikora et al. (2015) studied the risk factors of developing diabetes using logistic regression modelling. It was found that age, alcohol consumption, cholesterol level, occupation status and hypertension were associated with the outcome of having diabetes while gender of a person, having a family history of diabetes and smoking were not significantly associated with diabetes outcome.

Stella (2012), uses multiple logistic regression modelling to find out if factors like age, gender, occupation status, marital status, smoking, alcohol consumption, hypertensive, contribute to the clinical diagnosis of diabetes. The result obtained showed that alcohol, hypertension, cholesterol level, and occupation were statistically significant at 5% to the outcome of diabetes while marital status, smoking, family history and gender were not statistically significant.

III. METHODOLOGY

The study data came from secondary source which was taken from the department of Endocrinology at the Federal Medical Centre Gusau. Data from 2017 to 2020 was taken as a sample. The following information was taken directly from the folders of individuals who had been diagnosed with cardiovascular disease. Age, Cholesterol level, Smoking behaviour, Alcohol consumption, Gender, Marital status, Occupational status, Hypertension status, Family history of diabetes and Diabetes status were all documented.

Logistic Regression Model

The logistic regression equation is similar to the regression equation in many ways. When just one predictor variable X_1 is present, the logistic regression equation from which the probability of Y is predicted is as follows:

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

In which $\pi(x)$ is the probability of Y occurring, e is the base of natural logarithms, and the other coefficients form a linear combination much the same as in simple regression. This equation can be extended to include many predictors, just like linear regression. When more than one predictor is present, the equation becomes:

$$\pi(x) = P(Y=1/X_1, X_2, ..., X_K) = \frac{\ell^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}}{1+\ell^{\beta_0+\sum_{j=1}^{k}\beta_j X_j+\varepsilon}} \qquad (2)$$

The above equation (1) is the same as the equation (2) when just one predictor is present, except that the linear combination has been expanded to include any number of predictors.

Assessment of Fitted Logistic Regression Model

Following the estimation of the coefficients, several assessment parameters or tests must be carried out in order to examine the appropriateness, applicability, and adequacy of the generated logistic regression model. The statistical test of each predictor variable, goodness-of-fit statistics, and model discrimination are examples of these evaluation parameters.

The Wald Statistic

The Wald Statistic analogous to the t-test in linear regression is used to assess the significant of individual logistic regression coefficients. It is the ratio of the square of regression coefficient to the square of the standard error of the coefficient.

$$W_j = \frac{\widehat{\beta}_j^{\,2}}{SE(\widehat{\beta}_j)^2} \qquad (3)$$

Where $\widehat{\beta}_j$ represents the estimated coefficient of β and $SE(\widehat{\beta}_j)$ is its standard error. Under the null hypothesis $H_0: \beta_j = 0$. The quantity follows a chi-square distribution with one degree of freedom.

Goodness of Fit the Hosmer-Lemeshow Test

Hosmer and Lemeshow (2000) developed a goodness-of-fit test for logistic regression models with binary responses. They proposed grouping based on the value of the estimated probabilities. This test is obtained by calculating the Pearson chi-square statistic from the 2×g table of observed and expected frequencies, where g is the number of groups. The statistic is written as

$$H = \sum_{g=1}^{10} \frac{\left(O_g - E_g\right)^2}{E_g} \qquad (5)$$

Where, O_g and E_g denote the observed events and expected events for the gth risk deciles groups, respectively. This test follows χ^2 distribution with 8 (number of groups-2) degree of freedom.

The Odd Ratio

The odds ratio is a measure of association for 2×2 contingency table (Agresti, 2007). In 2×2 tables the probability of "success" is π_1 in row 1 and π_2 in row 2. Within row 1, the odds of success are defined to be:

$$odds_1 = \frac{p_1}{1-p_1} \quad \text{and} \quad odds_2 = \frac{p_2}{1-p_2} \qquad (6)$$

Agresti (2002) define the odds ratio in two groups of subjects as "the ratio of odds". Thus;

$$OR = \frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \qquad (7)$$

Hence, for logistic regression with a dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and the regression coefficient is $OR = \ell^{\beta i}$

- A 95% CI for log Odds ratio is given by:

log Odds ratio $= \ln(OR) \pm 1.96 \times \{SE\ln(OR)\}$ (8)

Where, $\ln(OR)$ is the sample log odds ratio and $SE\ln(OR)$ is the standard error of the log odds ratio.

- A 95% Confidence interval for Odds ratio is given by:

Odds ratio $= \ell^{\ln(OR) \pm 1.96 \times [SE\ln(OR)]}$ (9)

Alkaike Information Criterion (AIC)

Suppose that we have a statistical set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function).The AIC value of the model is the following.

$$AIC = 2k - 2\ln(Ł) \qquad (10)$$

Where, k = number of estimated parameters in the model.

Ł = Maximized likelihood function for the estimated model

Bayesian Information Criterion (BIC)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 565-571 www.ijaem.net ISSN: 2395-5252

In statistics, the Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models, models with lower BIC are generally preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC). The BIC was developed by Gideon E. Schwarz and published in a 1978 where he gave a Bayesian argument for adopting it. he BIC is formally defined as

$$BIC = k\ln(n) - 2\ln(\hat{L}) \qquad (11)$$

Where

$\hat{L}$ = the maximized value of the likelihood function of the model

n = the number of observations, or equivalently, the sample size

k = the number of parameters estimated by the model.

IV. RESULTS AND DISCUSSION

Descriptive Analysis

The descriptive statistics of study consists of 101 (51.5%) people who were employed and 95 (48.5%) people who were unemployed. 94 (48.0%) people were hypertensive and 102 (52.0%) were not hypertensive. 76 (38.8%) people consumed alcohol and 120 (61.2%) do not consume alcohol. 82 (41.8%) people were smokers and 114 (58.2%) were not. 112 (57.1%) people had a family history of diabetes while 84 (42.9%) people do not have family history of diabetes. There were 117 (59.7%) males and 79 (40.3%) females. 97 (49.5%) people were with high cholesterol level and 99 (50.5%) are with low cholesterol level. 84 (42.9%) people were diabetics while 112 (57.1%) were not. 100 (51.0 %) people are single while 96 (49.0%) are married.

Multiple Logistic Regression Results

Table 1: Model 1 - Analysis of the Maximum Likelihood Estimate of model with all Predictors

Parameter	Estimate	Standard Error	Wald Test	Sig.
Intercept	-2.748	0.707	15.105	0.000
Age	0.017	0.009	3.622	0.041
Alcohol	0.871	0.345	6.375	0.012
Hypertension	1.063	0.348	9.399	0.002
Cholesterol	0.871	0.346	6.337	0.012
Smoking	0.784	0.338	5.369	0.021
Occupation	-0.417	0.342	1.482	0.223
Gender	-0.133	0.347	0.148	0.701
Family History	0.753	0.348	4.685	0.030
Marital status	-0.486	0.348	1.947	0.163

Hence, Considering the Wald and Significance column in the table, it reveals that at 95% confidence interval the predictors i.e. age, alcohol, hypertension, cholesterol level, smoking, and family history of diabetes has significant effect on the occurrence of diabetes ($p < 0.05$). However, no significant difference ($p > 0.05$) was observed with occupation, gender and marital status.The Fitted Multiple Logistic Regression Model with all predictorsis given by:

$$\pi = \frac{\exp\left(\begin{array}{l} -2.748 + 0.017\times age + 0.871\times alcohol + 1.063\times \\ hypertension + 0.871\times cholesterol + 0.784\times smoking \\ -0.417\times occupation - 0.133\times gender + 0.753\times \\ familyhist - 0.486\times maritalstat + \varepsilon \end{array}\right)}{1+\exp\left(\begin{array}{l} -2.748 + 0.017\times age + 0.871\times alcohol + 1.063\times \\ hypertension + 0.871\times cholesterol + 0.784\times smoking \\ -0.417\times occupation - 0.133\times gender + 0.753\times \\ familyhist - 0.486\times maritalstat + \varepsilon \end{array}\right)}$$

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 565-571 www.ijaem.net ISSN: 2395-5252

2: Model 2 - Analysis of the Maximum Likelihood Estimate of only significant Predictors reduced

Parameter	Estimate	Standard Error	Wald Test	Sig.
Intercept	-3.309	0.615	28.899	0.000
Age	0.019	0.009	4.520	0.034
Alcohol	0.878	0.341	6.610	0.010
Hypertension	0.959	0.335	8.179	0.004
Cholesterol	0.806	0.334	5.827	0.016
Smoking	0.816	0.335	5.926	0.015
Family History	0.767	0.342	5.013	0.025

Table 2 shows the maximum likelihood estimates output of the model with only significant predictors. The Fitted Multiple Logistic Regression Model will now become

$$\pi = \frac{\left(\exp\begin{pmatrix}-3.309+0.019\times age+0.878\times alcohol+\\ 0.959\times hypertension+0.806\times cholesterol+\\ 0.816\times smoking+0.767\times familyhist+\varepsilon\end{pmatrix}\right)}{1+\left(\exp\begin{pmatrix}-3.309+0.019\times age+0.878\times alcohol+\\ 0.959\times hypertension+0.806\times cholesterol\\ +0.816\times smoking+0.767\times familyhist+\varepsilon\end{pmatrix}\right)}$$

Table 3: Odds Ratios and Confidence Intervals for the coefficient estimates of the Significant Predictors

Parameter	$Exp(\beta)$	95% confidence Interval for $Exp(\beta)$	
		Upper Bound	Lower Bound
Age	1.019	1.001	1.036
Alcohol	2.405	1.232	4.696
Hypertension	2.608	1.352	5.031
Cholesterol	2.240	1.164	4.312
Smoking	2.261	1.172	4.361
Family Hist.	2.153	1.100	4.211

The results in table 3 above indicate that: patients with hypertension are more susceptible to develop diabetes; An person with lower age is less susceptible to develop diabetes; consuming alcohol increases the susceptibility; a person with hypertension is at high risk of developing diabetes than those with no hypertension; smokers are more liable of developing diabetes than non-smokers; persons with high cholesterol level are more susceptible to develop diabetes than those with low cholesterol level and patients with family history of diabetes are more likely to develop diabetes than those with no family history of diabetes.

Table 4: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Model	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)
Model 1 – Model with All Predictors	233.376	266.157

Model 2 – Model with Significant Predictors Only	227.861	250.808

Table 4 shows the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) of the two models. In selecting the best model for diabetes disease, the calculated AIC and BIC of each model is to be considered. The smaller these values, the better the model fits the data. Hence, the second model which is the model with only significant predictors will be used for prediction.

Table 5 Goodness of Fit Hosmer Lemeshow goodness of fit result.

Step	Chi-square	df	Sig.
Model 2	7.707	8	0.463

The results from the Hosmer-Lemeshow test shows that significance level was greater than 0.05 and hence, we fail to reject the null hypothesis that there is no difference between observed and model predicted values, thus the observed proportion of events were found to be similar to the predicted probabilities.

V. CONCLUSION

In this study, some risk factors that influence the occurrence of diabetes were studied using logistic regression analysis. The risk factors (independent or predictor variables) were age, alcohol, hypertension, cholesterol level, smoking, occupation, gender, family history of diabetes and marital status. The prevalence of diabetes (dichotomous variable) was considered as the dependent variable. The multivariate logistic regression was fitted after fulfilling its various assumptions. The model was then used to estimate the probability of occurrence of diabetes given the risk factors, respectively.

The test of association of diabetes with all the predictor variables was done and it was found that age, alcohol, hypertension, high cholesterol, smoking and family history of diabetes were associated with the outcome of and thus were considered as statistically significant. The significance testing for the logistic coefficients using Wald test showed that the factors (age, alcohol, hypertension, high cholesterol, smoking and family history of diabetes) were significant predictors of diabetes. Hence, the Odds ratio (OR) were also found to be greater than 2 for these predictors.

In conclusion, the study demonstrated that logistic regression can be a prevailing statistical technique for exercise when the dependent variable is binary or dichotomous. The usefulness of the developed logistic model was supported by Wald test and 95% CI, Odds Ratio, Goodness-of-fit (Hosmer-Lemeshow test), Thus, mathematically driven information on these factors is very crucial in order to facilitate change or to make necessary measures to control such factors that are making this problem so extreme.

VI. RECOMMENDATION

Based on the findings from this study, the following recommendations are made to reduce the burden of most disabling disease like diabetes in among individuals.

- The government at all level should give more priority attention to seeking ways to bolster the level of awareness of individuals to understanding different risk factors associated with diabetes disease also to establish a working group to oversee this area and encourage appropriate adjunct studies.
- Access to and utilization of health services in our communities must be expanded, particularly for diabetes and other cardiovascular illnesses. This will allow people to come to hospital sooner, rather than waiting for the disease to reach its peak before going to hospital. We also advocate for increased public awareness and regular screening programs to aid in the early detection of diseases such as diabetes in our communities.
- Secondary data from the Federal Medical Centre in Gusau, Zamfara State, Nigeria, was used in this investigation. As a result, we advise researchers interested in continuing with this topic to acquire primary data directly from patients. Secondary data is unreliable due of its unreliability.
- Finally, more follow-up studies should be conducted to evaluate the benefits of various treatment methods on the control of diabetes risk factors such as high blood pressure and high cholesterol levels in diabetic patients in order to avoid further complications. Especially

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 565-571 www.ijaem.net ISSN: 2395-5252

the focus should be on assessing the effect of the interventions based on healthy lifestyle such as increased physical activity, smoking and alcoholic cessation, healthy dietary pattern.

Areas for Further Research

The following are the areas for further study

i. further risk factors should be identified and studied.
ii. lognormal regression analysis may be applied to discrete factors to compare results obtained with the multivariate approach.

REFERENCES

[1] Agresti, A. (2002). Categorical data analysis. John Wiley & Sons, New York, 2nd edition

[2] Afifi, A., Clark, V. A., & May, S. (2004). Computer- Aided Multivariate Analysis. Fourth Edition, Champman and Hall/CRC.

[3] Pirzado, A., Pirzado, M. S., Khanzada, A. H., Jatoi, I. K., Ahmed, I., & Arain, M. S. (2021). A Logistic Regression Analysis of the Diabetes Meletus DiseaseRisk: A Case Study of District NausheroFeroze, Sindh, Pakistan. Journal of Critical Reviews, 8(2), 440-453.

[4] Stella, A. (2012). Multiple Logistic Regression Analysis to Determine Risk Factors for The Clinical Diagnosis of Diabetes Case Study: Komfo Anokye Teaching Hospital (2008-2009). Kwame Nkrumah University of Science and Technology, Kumasi.

[5] Paulweber, B., Valensi, P., Lindstrom, J., Lalic, N., Greaves, C., & McKee, M. (2010). European evidence-based guidline for prevension of type 2 diabetes. Horm Metab Res. 2010;42(S01): S3 36.

[6] Safieddine, B., Sperlich, S., Beller, J., Lange, K., Epping, J., Tetzlaff, J., Tetzlaff, F., &Geyer, S. (2020). Socioeconomic inequalities in type 2 diabetes in employed individuals, nonworking spouses and pensioners. SSM - Population Health, 11(2020),1-9.https://doi.org/10.1016/j.ssmph.2020.100596

[7] Hosmer, D.W., & Lemeshow, S. (2000). Applied Logistic Regression. Second edition. John Wiley & Sons Inc: New York.

[8] Gudjinu, H., & Sarfo, B. (2017). Risk factors for type 2 diabetes mellitus among out-patients in Ho, the Volta regional capital of Ghana: a case–control study. BMC Res Notes. 2017(10), 324. Http://doi.org/10.1186/s13104-017-2648-z .

[9] International Diabetes Federation (IDF), Diabetes Atlas, 9th Edition. International Diabetes Federation2019. www.idf.org/e-library/epidemiology research/diabetes-atlas/159-idf-diabetes-atlas ninth-edition-2019.html. Accessed 15th August 2021.

[10] Danquah, I., Bedu-Addo, G., Terpe, K., Micah, F., Amoako Y., &Awuku, Y. (2012). Diabetes Mellitus Type 2 in urban Ghana: Characteristics and associated factors. BMC Public Health, 2012;12(1), 210

[11] Rastogi, P., & Singh, B. K. (2019). A Multivariate Binary Logistic Regression Modeling for Assessing Various Risk Factors that affect Diabetes. International Journal of Scientific & Technology Research, 8(8), 589-599.

[12] Niyikora, S., Joseph, M. K., & Marcel. N. (2015). Multiple logistic regression modeling on risk factors of diabetes.Case study of Gitwe Hospital (2011-2013). Jomo Kenyatta University of Agriculture and Technology. International Journal of Mathematics and Physical Science Research, 3(1): 48-61.

[13] World Health Organization. Global report on diabetes 2019. https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf. Accessed 15th August 2021.

[14] Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association. Vol. 72, pp. 851-853.