ECRONICON OPEN ACCESS

EC DIABETES AND METABOLIC RESEARCH

Research Article

Comparison of Logit and Probit Models in the Analysis of Severity of Diabetes

Chrysogonus C Nwaigwe*, Alloy C Onyeka and Sabinus N Nwanneako

Department of Statistics, School of Physical Sciences, Federal University of Technology, Owerri, Imo State, Nigeria

***Corresponding Author**: Chrysogonus C Nwaigwe, Department of Statistics, School of Physical Sciences, Federal University of Technology, Owerri, Imo State, Nigeria.

Received: February 21, 2020; **Published:** May 08, 2020

Abstract

Recently, diabetes is one of the common and killer diseases in the world. The severity of the disease may depend on the treatment as well as some other factors. This severity may be classified into different categories according to the levels of its complications in the patients. This study is an attempt to investigate the significant factors in severity of diabetes using the logit and probit models and compare results from the two models. The study also seeks to identify the effects of sample size on the estimates of the parameters of the models. The results obtained show that the probit model performed a little better with age as the only significant factor identified in the two models among the factors investigated. The results obtained also show that the efficiency of the coefficients of the logit and probit models increases as the sample size increases.

Keywords: *Severity; Diabetes; Logit Model; Probity Model; Sample Size*

Introduction

Diabetes mellitus is one of the most common diseases in the world. It has been observed that it affects a significant percent of the world population. The major sign of diabetes is hyper-glycemia, which means an increased blood sugar concentration. The symptoms include, increased thirst and frequent urination, insatiable appetite, weight loss, fatigue, visual disturbances and slowly healing wounds among others [1]. In 2014, diabetes affects 10.3% of the Nigerian population, and it leads to costly adverse healthcare outcomes [2]. Current practices, relying primarily on the presence of several factors, are not effective in capturing the risk of poor prognosis. Hence, little evidence exists so far to help prioritize care for these patients.

Hypoglycemia can be defined as a case of severe hypoglycemia where the patient seeks medical attention or is admitted to hospital. In the case of moderate hypoglycemia, patients experience symptoms of hypoglycemia and require assistances from relatives or friends. Mild hypoglycemia may be determined from blood glucose measurements (< 2.22 mmol/l; 40 mg/dl in any case, and 2.22 - 2.78 or 50 mg/dl in case of symptoms). The severity of diabetes in patients could be put in categorical form which makes the data to require a statistical analysis method that allows for categorical data. A categorical variable is defined as one that can assume only a limited number of discrete values [3]. The measurement scale for such a variable is unrestricted. It can be "Nominal", which means that the observed levels are not ordered. It can be "Ordinal", which means the observed levels are ordered in some way; or it can be "Interval", which means that the observed levels are ordered and numeric, and that any interval of one unit of the scale of measurement represents the same amount, regardless of its location on the scale [2]. Categorical data arise from observations on multiple subjects having one or more categorical variables observed for each subject. Ordered categorical data arise often in biomedical research. Sometimes categories are the result of grouping a data set from a continuous variable, such as age, or they arise if the measurement is inherently imprecise, so that an interesting series can only be observed on the ordinal scale. However, ordinal data often result from subjective assessments under ordered categories, example, excellent, good, fair and poor.

A categorical distribution is a discrete probability distribution that describes the possible results of a random event that can take on one of K possible outcomes, with the probability of each outcome separately specified. The categorical distribution is the generalization of the Bernoulli distribution for a categorical random variable, i.e. for a discrete variable with more than two possible outcomes, such as the roll of a die, [4].

Previous studies in this area were mainly experimental and descriptive. In a case where the dependent or the outcome variable of a given event is in categories rather than continuous, it is inappropriate to use linear regression because, the response value is not measured on a ratio scale nor the error term normally distributed. The linear regression model can generate predicted values, real numbers ranging from negative to positive infinity, whereas categorical variables can only take a limited number of discrete values with a specific range [5]. The logit and probit models are well-known methods for handling categorical data. Probit regression, also called a probit model, is used to model dichotomous or binary outcome variables. In the probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors [6].

The Proportional Odds Model (POM) is sometimes referred to as the logit model, as described by [7] and it is the most popular model for logit regression. The hallmark of the POM is that the odds ratio for a predictor can be interpreted as a summary of the odds ratio obtained from separate binary logistic regression using all possible cut off points of the ordinal outcome [8]. The POM has also been referred to as the constrained cumulative logit model or cumulative logit model [6]. The model is based on cumulative distribution function [9]. It is used to model categorical response data when the response categories have a natural ordering.

In [10] the effect of type 2 diabetes mellitus (DM) on the presentation and treatment response of pulmonary tuberculosis (TB) was studied. It was observed that DM seems to have a negative effect on the outcome of TB treatment. In [11], a comprehensive study to identify factors that account for heterogeneity of severity in patients treated for type 2 diabetes was undertaken. Clinical factors identified included baseline glycated hemoglobin A1c or fasting plasma glucose (FPG) levels, insulin response or sensitivity, C-peptide, body composition, adipose tissue proteins, lipid profile, plasma albumin levels and duration of disease or insulin treatment. Other factors identified included age, sex, race and socioeconomic status. Data on severity of diabetes are categorical. Often, when data are collected on severity of diabetes, the response appears in multinomial form (in more than two categories). This seems to be the first-time severity of diabetes is being modeled using probity and logit models. Furthermore, previous research works have limited the comparison of the logit and probit models to binary response data, compared the logit model with another model or ignored comparison of models in their approaches. More so, the difference between the results obtained from the models may depend on the nature of both the dependent variable and independent variables. This research seeks to compare the fit of the logit and probit models in modeling the multinomial severity of diabetes as obtained from diabetic patients. It also seeks to identify the significant factors in the severity of diabetes and examine the effects of sample sizes in the fit of the models. The factors whose effects are to be investigated are age, weight, blood pressure and genotype.

Methodology

In this section, the models adopted for the analysis of data are discussed.

The logit model

Let Y be a response variable with an ordinal scale,

$\underline{X}$, a vector of independent or explanatory variables.

$\propto_j$, the jth intercept of the generalized linear model

$\pi_i(\underline{X})$, the probability of the ith subject due to the effect of $\underline{X}$ or the link function of the generalized linear model.

Let J be the number of categories of the dependent or response variable.

Then,

$$P[Y \leq j|\underline{X}\,] = \pi_1(x) + \pi_2(x) + \ \ldots + \pi_j(x), j = 1, 2, \ldots \ldots, J - 1 \ (2.1)$$

$$\text{Logit } P[Y \leq j|\underline{X}] = In\, \frac{P[Y \leq j|\underline{X}]}{1 - P[Y \leq j|\underline{X}]} \quad (2.2)$$

$$\text{Logit } P[Y \leq j|\underline{X}] = In \left\{\frac{\pi_1(x) + \pi_2(\underline{x}) + \ldots\ldots\ldots + \pi_j(\underline{x})}{\pi_{j+1}(x) + \pi_{j+2}(x) + \ \ldots\ldots\ldots + \pi_J(x)}\right\} \quad (2.3)$$

For purpose of parsimony, we use,

$$\text{Logit } P\left[Y \leq j|\underline{X}\right] = \propto_j + \ \underline{\beta}'\underline{X}, j = 1, 2, \ldots \ldots, J - 1 \quad (2.4)$$

Estimation of the parameters of logit model

The parameters of the logit model are estimated by the methods of maximum likelihood. The values of the parameters that maximize the log of the likelihood function of the logit distribution are the maximum likelihood estimates of the parameters.

Cumulative probit model

Suppose we assumed that is a latent variable (unobserved) that underlie Y, for fixed value of explanatory variables $\underline{X}$, see [12], let the link function be given by,

$$\eta(\underline{X}) = \ \underline{\beta}'\underline{X}.$$

Let the cumulative density function of Y* be

$$\text{P}(Y^* \leq y^*|\underline{X}) \ = \ G(y^* - \eta) = G\left(y^* - \underline{\beta}'\underline{X}\right) (2.5)$$

Under this latent variable structure

$$P\left[Y \leq j|\underline{X}\right] = P\left[Y^* \leq \ \propto_i \ (\underline{X})\right] \ = \ G\left(\propto_i - \ \underline{\beta}'\underline{X}\right) (2.6)$$

Therefore, the link function required to obtain a linear predictor is G-1 of the cumulative density function of Y given as,

$$G^{-1}P\left[Y \leq j|\underline{X}\right] = \propto_i - \ \underline{\beta}'\underline{X} \qquad (2.7)$$

Where G is the cumulative density function of the standard logistic distribution with

$$G(\varepsilon) = \frac{\varepsilon^{\varepsilon}}{1+e^{e}} \qquad (2.8)$$

Then G^{-1} is the logit link function.

When the G in (2.6), (2.7) and (2.8) is a cumulative density function of the standard normal distribution, the model in (2.1) becomes probit ordered model given by [12] and [13] as:

Probit $P[Y \leq j|\underline{X}] = \propto_j + \underline{\beta}'\underline{X}, j = 1, 2 \dots\dots\dots.., J-1$

where,

α_j, β', $\underline{X}$ and Y, have the same meanings as given in Equation (2.1).

Estimation of the parameters of the probit model

Using maximum likelihood techniques, we can compute estimates of the coefficients () and their corresponding standard errors that are asymptotically efficient. The values of the parameters that maximize the log of the maximum likelihood are the estimates of the parameters. However, these estimates cannot be interpreted in the same manner that normal regression coefficients are interpreted.

Test for multicollinearity

Multicollinearity is a common problem when estimating linear or generalized linear models, including logit and probit regression models. It occurs when there are high correlations among predictor variables, leading to unreliable and unstable estimates of regression coefficients [14]. Multicollinearity problem makes a significant variable insignificant by increasing its standard error.

The most widely-used diagnostic for multicollinearity is the variance inflation factor (VIF) [15].

It may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the R_i^2 from that regression. The VIF factor for {\displaystyle {\hat {\beta }}_{i}}the estimated coefficient $\hat{\beta}_i$ may be obtained from:

$$VIF_i = \frac{1}{1-R_i^2} \quad (2.9)$$

Where R_i^2 is the coefficient of determination for the regression with X_i on the left hand side, and all other predictor variables on the right hand side.

If VIF value exceeds 4.0, or by tolerance less than 0.2 then there is a problem with multicollinearity [16].

Variance Inflation Factor between 0 and 5 suggests that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

The need to reduce multicollinearity depends on its severity and the primary goal on the model.

Test for goodness of fit

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question.

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

Suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model. Let $\hat{\mathrm{L}}$ be the maximum value of the likelihood function for the model. Then the AIC value of the model, according to [17] is the following,

$$\mathrm{AI}C = 2K - 2In(\hat{\mathrm{L}}) \quad (2.10)$$

When the sample size is small, there is a substantial probability that AIC will select models that have too many parameters, i.e. that AIC will overfit [18]. To address such potential overfitting, AICc was developed: AICc is AIC with a correction for small sample sizes.

The formula for AICc depends upon the statistical model. Assuming that the model is univariate, linear in its parameters, and has normally-distributed residuals (conditional upon regressors), then the formula for AICc is as follows [19]:

$$AIC_c = AIC + \frac{2k^2 + 2k}{n - k - 1} \quad (2.11)$$

where n denotes the sample size and k denotes the number of parameters. Thus, AICc is essentially AIC with an extra penalty term for the number of parameters. Note that as n → ∞, the extra penalty term converges to 0, and thus AICc converges to AIC [16].

If the assumption that the model is univariate and linear with normal residuals does not hold, then the formula for AICc will generally be different from the formula above. For some models, the formula can be difficult to determine. For every model that has AICc available, though, the formula for AICc is given by AIC plus terms that include both k and k^2. In comparison, the formula for AIC includes k but not k^2. In other words, AIC is a first-order estimate (of the information loss), whereas AICc is a second-order estimate [20].

The Likelihood-Ratio test (LR test) is a statistical test used for comparing the goodness of fit of two statistical models, a null model (representing the null hypothesis) against an alternative model (representing an alternative hypothesis). The test is based on the ratio of the likelihoods of the two models, denoted by ; the ratio expresses how many times more likely the data are under one model than the other. This likelihood ratio statistic is computed as:

LR= $-2\log\lambda$ (2.12),

which asymptotically has a χ^2 distribution [21].

The likelihood-ratio test provides the decision rule as follows:

If λ >c, do not Reject H_0

If λ <c, reject H_0

Reject with probability q if λ =c

The values c and q are usually chosen to obtain a specified significance level , via the relation.

$$q.P(\lambda = c \mid H_0) + P(\lambda < c \mid H_0) = \alpha \,. \quad (2.13)$$

The likelihood-ratio test rejects the null hypothesis if the value of this statistic is too small. How small, depends on the significance level of the test [22].

The Pseudo R^2 measures are logical analogs to ordinary least squares R^2, the measures which lie between 0 and 1. McFadden's R^2 is perhaps the most popular Pseudo R^2 in measurement of goodness of fit. The formula of the Pseudo R^2 is given by:

$$R^2 = \frac{G_m}{Dev_0} \qquad (2.14)$$

$$= 1 - \frac{LL_m}{LL_0} \qquad (2.15)$$

$$= 1 - \frac{Dev_m}{Dev_0} \qquad (2.16)$$

where G_m is the Likelihood Ratio, L_0 is the Log Likelihood intercept, L_m is the Log Likelihood Model, Dev_m is the Deviance of the Model and Dev_0 is the deviance of the Intercept [23].

Results and Discussion

In this section, the logit and probit models were applied to data on the severity of diabetic cases as obtained from the medical ward at Federal Medical Centre (FMC) Owerri, Imo State, Nigeria. The variables included severity to treatment, classified into three categories (severe = 1, moderate = 2 and slightly/mild = 3), age in years, weight measured in kilogram (kg), blood pressure measured in millimetres of mercury (mmHg) and genotype. The severity of diabetic cases was used as the dependent/response variable while the other variables were used as the independent variables. The results from the analyses of the data were discussed. The data for analysis are given in table 1.

Age	Weight	Blood pressure	Genotype	Severity
54	81	140	2	2
60	62	146	1	3
45	74	143	3	3
51	95	139	1	2
67	73	144	1	1
53	78	128	3	2
66	90	147	2	1
58	89	146	3	2
70	71	122	3	1
52	79	148	1	2
60	98	140	1	3
58	80	139	1	1
67	81	120	2	2
53	67	145	3	1
64	93	120	3	2
59	57	143	1	2
61	94	140	2	1

42	66	150	1	3
53	95	140	2	3
47	69	120	2	2
44	77	146	3	3
50	100	152	3	1
69	98	150	2	2
48	88	130	1	3
65	84	144	1	2
66	90	152	1	1
68	98	130	2	3
59	74	150	3	1
55	76	144	3	2
60	70	160	3	2

***Table 1:** Data on Severity of cases from Diabetic patients.*
Severity (1- severe, 2 - moderate, 3- mild/slight), Genotype (1- AA, 2- AS, 3- SS).

Logit and probit regression models of severity of diabetes on age, weight, blood pressure and genotype

The results of applications of ordered logit and probit regression models are as shown in table 2. The results in column 2 are for logit model while those in column 3 are for probit model. The results show that only age of the patients was a significant factor in determining the severity of cases of patients with diabetes for both models at 0.05 level of significance.

Variable	Treatment Severity as Dependent Variable	
	Logit Model	Probit Model
Intercept 1	-11.9288* (0.0701)	-7.6201** (0.0501)
Intercept 2	-9.6326 (0.1340)	-6.2393 (0.1014)
Age	-0.1286** (0.0140)	-0.0775** (0.0104)
Weight	0.0198 (0.5446)	0.0112 (0.5421)
Blood Pressure	-0.0296 (0.3873)	-0.0200 (0.3311)
Genotype	-0.4612 (0.2763)	-0.3098 (0.2210)
LR $\left(\chi^{2}\right)$	7.8647 (0.0967)	8.1625 (0.0858)
Log Likelihood	-28.3487	-28.1998
Pseudo R-squared	0.1218	0.1264
AIC	2.2900	2.2800
JB	2.2600	1.5367
Number of Observations	30	30

***Table 2:** Logit and probit regression models of severity cases of diabetic patients.*
*( ) p-value, *: Significant at α = 0.10, **: Significant at α = 0.05.*

Citation: Chrysogonus C Nwaigwe., *et al.* "Comparison of Logit and Probit Models in the Analysis of Severity of Diabetes". *EC Diabetes and Metabolic Research* 4.6 (2020): 01-11.

The values in parentheses in table 2 are the p-values. The intercept of the model is also known as limit point, cuts or threshold. The number of intercepts equals the number of categories of the response variables minus 1. Specifically, since our response variable is categorized into 3 (Mild/Slightly, Moderate and Severe), the number of intercepts will be 2 as evident in table 2.

These intercepts are the estimated ordered logits for the adjacent levels of the response variable, severe verses Mild and Moderate, as well as Severe and Moderate versus Mild. Here the reference category is Moderate.

In the above results, the intercept 1, the estimated log odds for severe treatment response versus moderate and mild treatment response when the predictors (independent variables) are evaluated at zero is -11.9288 (p = 0.0701) and -7.6201 (p = 0.0501) for logit and probit models respectively. Based on these p-values, we infer that the intercepts for the respective models are not significantly different from zero at 0.05 level of significance.

The Intercept 2, the estimated log odds for severe and moderate versus mild, when the independent variables are evaluated at zero are -9.6326 (p = 0.1340) and -6.2393 (p = 0.1014) for logit and probit models respectively. From the p-values we infer that the intercepts for the respective models are not significantly different from zero at 0.05 level of significance.

The estimates of the coefficients for the age of patients in the logit and probit models are $\hat{\beta}_1 = -0.1286(p = 0.0140)\ and - 0.0775(p = 0.0104)$ respectively. The estimate ($\hat{\beta}_1$) is significant (p< 0.05). The estimate, $\hat{\beta}_1 = -0.1286$ for the logit model implies that a one - unit change in the predictor variable, age, is associated with -0.1286 change in the logit of the severity cases of diabetes. In other words, a one - unit increase in age, would lead to 0.1286 decrease in the logit of the severity of cases of diabetes.

Similarly, the estimate $\hat{\beta}_1 = -0.0775$ for the probit model implies that a one - unit change in the predictor variable age, is associated with -0.0775 change in the probability of the severity of cases of diabetes. This implies that a unit increase in age would lead to 0.0775 decrease in the probability of the severity of cases of diabetes.

In order to compare the performances of both models, the Log Likelihood value and Akaike Information Criterion (AIC) for both models were obtained. For logit model, the Log Likelihood value is 28.3487 with AIC = 2.29 while the Log Likelihood value is -28.1998 with the AIC = 2.28 for the probit model. Although there is an indication that the logit model and the probit model almost performed equally for the given data set, but the probit model performed better.

The residual of both models were diagnosed using Jarque Bera. The results obtained were JB = 2.2600 and JB = 1.5367 for logit and probit models respectively, which show that the residual is a white noise and normally distributed with constant variance. Therefore, the fitted models can be represented explicitly thus for logit and probit models respectively as:

$$Logit[\hat{P}(Y \leq j]] = \alpha_j - 0.1286X_1 + 0.0198X_2 - 0.0296X_3 - 0.4612X_4$$

$$probit[\hat{P}(Y \leq j]] = \alpha_j - 0.0775X_1 + 0.0112X_2 - 0.0200X_3 - 0.3098X_4$$

Where,

j=1, 2 (i.e. 1 = intercept for severe category, 2 = intercept for moderate category and mild/slightly is the base or reference category)

X_1 = age, X_2 = weight, X_3 = blood pressure and X_4 = genotype.

From the table 3, the Variance Inflation Factor (VIF) values lie within the cut off limit (0.2 ≤ VIF ≤ 4.0), this shows that there is no multicollinearity among the independent variables.

Predictors	Logit	Probit
Age	1.1417	1.1375
Weight	1.2326	0.3505
Blood Pressure	1.2062	1.1316
Genotype	1.1008	1.1190

Table 3: *Results of VIF test for multicollinearity.*

Following the above results, the insignificant factors were removed from the models. The analyses were repeated with the only significant factor, age. The results are shown in table 3.

From the results in table 4, the estimate of the logit model coefficient is $\hat{\beta}_2 = -0.1119$ for the age. In other words, a one - unit increase in age, would lead to 0.1119 decrease in the logit of the severity cases of diabetes. Similarly, the estimate, $\hat{\beta}_2 = -0.0650,$ for the probit model implies that a unit increase in age would lead to 0.0650 decrease in the probability of the severity of cases of diabetes.

Variable	Treatment Response as Dependent Variable	
	Logit Model	Probit Model
Intercept 1	-7.4202** (0.0113)	-4.3339** (0.0099)
Intercept 2	-5.2608* (0.0554)	-3.0493* (0.0596)
Age	-0.1119** (0.0223)	-0.0650** (0.0216)
LR $\left(\chi^2\right)$	5.6817 (0.0171)	5.5262 (0.0187)
Log Likelihood	-29.4402	-29.5179
Pseudo R-squared	0.0880	0.0856
AIC	2.1627	2.1619
JB	1.7458	0.8633
Number of Observations	30	30

Table 4: *Logit and probit regression analyses of severity of cases of diabetes on age.*
*( ) p-value, *: Significant at α=0.10, **: Significant at α=0.05.*

In order to compare the performance of both models, the Log Likelihood value and Akaike Information Criterion (AIC) for both models were obtained. For the logit model, the Log Likelihood value is -29.44 with AIC = 2.16 while the Log Likelihood value = -29.52 with the AIC = 2.16 for the probit model. This therefore suggests that both the logit model and the probit model almost performed equally for the given data set.

To ascertain the effect of sample size (number of observations) on the model performance, the number of observations was reduced arbitrarily, and the analyses were repeated. The results of the analyses for different number of observations are as shown in table 5.

Variable	Response to Treatment as Dependent Variable			
	n = 25		n = 20	
	Logistic	Probit	Logistic	Probit
Intercept 1	-11.5983 (0.0904)	-7.2918* (0.0691)	-14.8271 (0.0756)	-9.2062 (0.0583)
Intercept 2	-9.1534 (0.1691)	-5.8271 (0.1366)	-12.0648 (0.1339)	-7.5730 (0.1086)
Age	-0.1339** (0.0283)	-0.0797** (0.0208)	-0.1631** (0.0244)	-0.1003 (0.0146)
Weight	0.0480 (0.2064)	0.0273 (0.2005)	0.0609 (0.1650)	0.0352 (0.1716)
Blood Pressure	-0.0384 (0.2865)	-0.0249 (0.2412)	-0.0504 (0.2708)	-0.0312 (0.2535)
Genotype	-0.6523 (0.1754)	-0.4004 (0.1567)	-1.1399* (0.0712)	-0.6760* (0.0576)
LR$\left(\chi^2\right)$	7.8434 (0.0975)	8.0901 (0.0883)	9.0560 (0.0597)	9.2910 (0.0543)
Log Likelihood	-26.7090	-22.6639	-16.8139	-16.6964
Pseudo R-squared	0.1468	0.1514	0.2122	0.2177
AIC	2.3030	2.2931	2.3814	2.3696
JB	1.7713	1.2808	1.6092	1.5131
Number of Observations	25	25	20	20

Table 5: *Logit and probit regression models of cases of diabetes at different number of observations.*
*( ) p-value, *: Significant at α = 0.10, **: Significant at α = 0.05*

From the results in table 5 the coefficients of the models were found to be consistent in sign even as the number of observations reduces. The age coefficients still maintained their negative signs. However, there was gradual reduction in magnitude of the coefficients as the number of observations was reduced, though this did not significantly affect the coefficient as the z-value and p-value had no significant reduction as the number of observations was reduced. It was observed that as the number of observations reduces, the coefficients become less significant, that is to say that the sample size has effect on the efficiency of the estimates. On the measure of the goodness of fit of the model, the Pseudo R^2 was not relatively stable even with the change in the number of observations. However, the Akaike-Information-Criterion (AIC) was found to be increasing as the number of observations reduces, indicating that the models performed better with large sample size. Therefore, the estimates maybe said to be asymptotically efficient. The asymptotic robustness is also evident with the value of the Likelihood-Ratio Chi-Square [LR(χ^2)] which was found to increase with reduction in number of observations.

Conclusion

In this study, the applications of the logit and probit models in the analysis of severity of diabetes were compared. Thirty diabetes patients' records were used. Independent variables such as age, weight, blood pressure and genotype were considered as factors influencing the severity of diabetes. The logit and the probit models almost equally fitted the data having age only as a significant determinant factor but the probit model performed a little better. The AIC for the logit model was 2.29 and the AIC for the probit model was 2.28, showing that the probit model fairly performed better than the logit model. The results obtained show that the risk of diabetes is higher in older people than in younger people. Results obtained also show that the efficiency of the coefficients of the logit and probit models increases as the sample size increases.

Bibliography

1. Nwaigwe CC. "Statistical investigation of the significance of gender in determining diabetes type". *FUTO Journal Series* 3.1 (2017).

2. Adeloye D., *et al*. "An estimate of the prevalence of hypertension in Nigeria: a systematic review and meta-analysis". *Journal of Hypertension* 33.2 (2015): 230-242.

3. Yates Frank. "The Practice of Statistics". Wiley Online Library (2003).

4. Murphy KP., *et al*. "Cerebral Palsy, Neurogenic Bladder and Outcomes of Lifetime Care". *Developmental Medicine and Child Neurology* 54.10 (2012): 945-950.

5. Agresti. "An Introduction to Categorical Data Analysis". 2nd edition. John Wiley & Sons, Inc (2007).

6. Hosmer DW and Stanley L. "Applied Logistic Regression". Wiley InterScience (Online Service) (2000).

7. McCullagh P and Nelder JA. "Generalized Linear Models, 2nd edition". Chapman and Hall/CRC Monographs on Statistics and Applied Probability (1989).

8. Scott SC., *et al*. "Statistical assessment of ordinal outcomes in comparative studies". *Journal of Clinical Epidemiology* 50.1 (1997): 45-55.

9. Bender R and Benner A. "Calculating Ordinal Regression Models in SAS and S-Plus". *Biometrica* 42.6 (2000).

10. Bachti A., *et al*. "The effect of type 2 diabetes mellitus on the presentation and treatment response of pulmonary tuberculosis". *Clinical Infectious Diseases* 45.4 (2007): 428-435.

11. Alatorre Cantrell RA., *et al*. "A review of treatment response in type 2 diabetes: Assessing the role of patient heterogeneity". *Diabetes, Obesity and Metabolism* 12.10 (2010): 845-857.

12. Agresti. "Analysis of Ordinal Categorical Data". 2nd edition. John Wiley & Sons Inc (2010).

13. McCullagh P and Cox DR. "Invariants and Likelihood Ratio Statistics". *Annals of Statistics* 14.4 (1986): 1419-1430.

14. Allison SD. "A trait-based Approach for Modelling Microbial Littre Decomposition". *Ecology Letters* 15.9 (2012): 1058-1070.

15. James G., *et al*. "An Introduction to Statistical Learning (8th edition)". Springer Science+Business Media New York (2017).

16. Hair JF., *et al*. "Multivariate Data Analysis". 7th Edition, Pearson, New York (2010).

17. Akaike H. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* 19.6 (1974): 716-723.

18. Mcquarrie ADR and Tsai C. "Regression and Time Series Model Selection". World Scientific Publishing Company, Singapore (1998).

19. Burnham KP and Anderson DR. "Understanding AIC and BIC in Model Selection". *Sociological Methods and Research* 33.2 (2004): 93.

20. Burnham KP and Anderson DR. "Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach". Springer Link (2002).

21. White H. "Asymptotic Theory for Econometricians". 1st edition. Elsevier (1984).

22. Neyman EL and Pearson L. "On the Problem of the most efficient Tests of Statistical Hypothesis". Breakthroughs in Statistics (1933): 73-108.

23. McCulloch DK., *et al*. "Hypoglycemia (low blood sugar) in diabetes mellitus (Beyond the Basics)". Wolters (2018).

Volume 4 Issue 6 June2020