Oyeka and Okeh, 2:2
http://dx.doi.org/10.4172/scientificreports.654

Open Access Scientific Reports

Research Article Open Access

The Logistic Regression and ROC Analysis of Diagnostic Tests Results for Gestational Diabetic Mellitus

Oyeka ICA[1] and Okeh UM[2]*

[1]Department of Applied Statistics, Nnamdi Azikiwe University, Awka, Nigeria
[2]Department of Industrial Mathematics and Applied Statistics, Ebonyi State University, Abakaliki, Nigeria

Abstract

This paper proposes a matrix approach for estimating parameters of logistic regression, with a view of estimating the effects of risk factors of gestational diabetic mellitus (GDM). The proposed method, unlike other methods of estimating parameters of non-linear regression, is simpler, and convergence of parameters is quicker. The odds ratio obtained from the logistic regression were used to interpret the effects of these risk factors of GDM, where obesity and family history as risk factors, were positively associated with GDM on application of the proposed method, with data from five randomly selected hospitals in Ebonyi State, Nigeria. The proposed method was seen to compare favorably with other known methods.

Keywords: GDM; Odds ratio; Logistic regression; Dichotomous; Newton-raphson

Introduction

The constant evolution of medicine over the last two decades has meant that statistics has had to develop methods to solve the new problems that have appeared, and has come to play a central part in methods of diagnosis of diseases [1]. A diagnostic method consists of the application of a test with a group of patients, in order to obtain a provisional diagnosis regarding the presence or the absence of a particular disease [2]. In this work, logistic regression has been proposed for the purpose of estimating the effects of various predictors on some binary outcome of interest. Here, logistic regression regresses a dichotomous dependent variable on a set of independent variables, as a way of knowing the effects of these independent variables [3,4].

We, therefore here, propose to develop a matrix approach for solving a system of nonlinear equations, with P+1 unknown parameters. These methods will be applied in estimating the effects of risk factors on the occurrence of gestational diabetic mellitus (GDM) [5-7]. The proposed method will be illustrated using data on gestational diabetic mellitus (GDM), and have been shown to compare favorably with other existing methods in terms of efficiency.

The Proposed Method

The fundamental model for any multiple regression analysis assumes that the outcome variable is a linear combination of a set of predictors, and this is represented as:

$$\left.\begin{array}{l} Y=\beta_0+\beta_1 x_1+\ldots+\beta_k x_{ik}+\varepsilon=\sum_{k=0}^{P}\beta_k x_{ik}+\varepsilon, i=1,2,\ldots,N \\ or\ Y=X\beta \quad by\ matrix\ notation \end{array}\right\} \quad (1)$$

Where β_0 is the expected value of Y, when the x's are set to 0, β_k is the regression coefficient for each corresponding predictor variable, x_{ik}, ε is the error of the prediction. The binary logistic model is based on a linear relationship between the natural logarithm (ln) of the odds of an event, and a numerical independent variable. The form of this relationship is as follows:

$$L=In(o)=In\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{k=0}^{P}\beta_k x_{ik}+\varepsilon \quad (2)$$

The logistic regression model indirectly models the response variable based on probabilities associated with the values of Y. Let π_i be the probability that Y=1 and π_i -1 be the probability that Y=0. These probabilities are represented as:

$$\left.\begin{array}{l} \pi_i=P\left(Y=1\middle|X_1,X_2,\ldots,X_{ik}\right) \\ 1-\pi_i=1-P\left(Y=1\middle|X_1,X_2,\ldots,X_{ik}\right) \\ or \quad =P\left(Y=0\middle|X_1,X_2,\ldots,X_{ik}\right) \end{array}\right\} \quad (3)$$

But, the general form of logistic model is given by

$$\log it\ \pi_i=In\left[\frac{\pi_i}{1-\pi_i}\right]=\log\left[\frac{\pi_i}{1-\pi_i}\right]=e^{\sum_{k=0}^{P}\beta_k x_{ik}} \quad (4)$$

Where i=1,2,....N

And $\frac{\pi_i}{1-\pi_i}$ are the odds of developing any disease for a subject with risk factor. By logit transformation of the inverse of log odds to favour Y=1, we obtain the linear component as

$$\frac{\pi_i}{1-\pi_i}=e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i=e^{\sum_{k=0}^{P}\beta_k x_{ik}}-\pi_i e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i+\pi_i e^{\sum_{k=0}^{P}\beta_k x_{ik}}=e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i\left(1+e^{\sum_{k=0}^{P}\beta_k x_{ik}}\right)=e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

***Corresponding author:** Okeh UM, Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria, E-mail: uzomaokey@ymail.com

Received February 07, 2013; **Published** March 30, 2013

Citation: Oyeka ICA, Okeh UM (2013) The Logistic Regression and ROC Analysis of Diagnostic Tests Results for Gestational Diabetic Mellitus. 2: 654. doi:10.4172/scientificreports.654

$$\pi_i = \frac{e^{\sum_{k=0}^{P}\beta_k x_{ik}}}{1+e^{\sum_{k=0}^{P}\beta_k x_{ik}}} = \frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}$$

Similarly,

$$1-\pi_i = 1-\frac{1}{1+e^{-(\sum_{k=0}^{P}\beta_k x_{ik})}} = \frac{1}{1+e^{(\sum_{k=0}^{n}\beta_k x_{ik})}} = \frac{e^{-(\sum_{k=0}^{P}\beta_k x_{ik})}}{1+e^{-(\sum_{k=0}^{n}\beta_k x_{ik})}}$$

Using the inverse of logit transformation of the natural logarithm of the odds (log odds) to favor Y=1, we equates to the linear component to have:

$$\log it\ \pi_i = In\left[\frac{\pi_i}{1-\pi_i}\right] = \log\left[\frac{\pi_i}{1-\pi_i}\right] = (\sum_{k=0}^{P}\beta_k x_{ik}), i = 1,2,...,N \quad (5)$$

Therefore,

$$\frac{\pi_i}{1-\pi_i} = \frac{\frac{1}{1+e^{-(\sum_{k=0}^{P}\beta_k x_{ik})}}}{\frac{e^{-(\sum_{k=0}^{P}\beta_k x_{ik})}}{1+e^{-(\sum_{k=0}^{P}\beta_k x_{ik})}}} = e^{\sum_{k=0}^{P}\beta_k x_{ik}} \quad (6)$$

Maximum Likelihood Estimation (Mle) for Logistic Regression

We here estimate the P+1 unknown parameters β in Equation 5, with MLE, by finding the set of parameters for which the probability of the observed data is greatest. Since each y_i represents a binomial count in the i^{th} population, the joint probability density function of Y is:

$$f(Y|\beta) = \prod_{i=1}^{N}\frac{n_i!}{y_i!(n_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \quad (7)$$

Where β is from π_i in Equation 3. For each population, there are $\binom{n_i}{y_i}$ different ways to arrange y_i successes from among n_i trials. Since the probability of a success for any one of the n_i trials is π_i, the probability of y_i successes is $\pi_i^{y_i}$. Likewise, the probability of n_i-y_i failures is $(1-\pi_i)^{n_i-y_i}$. The joint probability density function in Equation 7 expresses the values of Y as a function of known, fixed values for β. The likelihood function has the same form as the probability density function, except that the parameters of the function are reversed: the likelihood function expresses the values of β, in terms of known, fixed values for Y. Thus,

$$L(\beta|Y) = \prod_{i=1}^{N}\frac{n_i!}{y_i!(n_i-y_i)!}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \quad (8)$$

The maximum likelihood estimates are the values for β that maximize the likelihood function in Equation 8. Thus, finding the maximum likelihood estimates requires computing the first and second derivatives of the likelihood function. Since the factorial terms do not contain any of the π_i, they are essentially constants that can be ignored. Therefore, maximizing the equation without the factorial terms will come to the same result, as if they were included. By rearranging the terms, the equation to be maximized which is the conditional likelihood can be written as:

$$L(\beta|Y) = \prod_{i=1}^{N}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$$

or

$$\left.\begin{aligned} L(\beta|Y) &= \prod_{i=1}^{N}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)^{n_i} \\ L(\beta|Y) &= \prod_{i=1}^{N}\pi_i^{y_i}\times(1-\pi_i)^{n_i}\times(1-\pi_i)^{-y_i} \\ L(\beta|Y) &= \prod_{i=1}^{N}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)^{n_i} \end{aligned}\right\} \quad (9)$$

Recall from Equation 6 that

$$\frac{\pi_i}{1-\pi_i} = \frac{\frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}}{1-\frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}} = \frac{\frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}}{\frac{e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}} = \frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}} \div \frac{e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}$$

$$= \frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}} \times \frac{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}{e^{-\sum_{k=0}^{P}\beta_k x_{ik}}} = e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\sum_{k=0}^{P}x_{ik}\beta_k} \quad (10)$$

$$\frac{\pi_i}{1-\pi_i} = e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i = e^{\sum_{k=0}^{P}\beta_k x_{ik}} - \pi_i e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i + \pi_i e^{\sum_{k=0}^{P}\beta_k x_{ik}} = e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i\left(1+e^{\sum_{k=0}^{P}\beta_k x_{ik}}\right) = e^{\sum_{k=0}^{P}\beta_k x_{ik}}$$

$$\pi_i = \frac{e^{\sum_{k=0}^{P}\beta_k x_{ik}}}{1+e^{\sum_{k=0}^{P}\beta_k x_{ik}}} = \frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}$$

Which, after solving for π_i (the same thing as the result of Equation 3) becomes,

$$\pi_i = \left(\frac{e^{\sum_{k=0}^{P}x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{P}x_{ik}\beta_k}}\right) \quad (11)$$

Substituting Equation 10 for the first term and Equation 11 for the second term, Equation 9 becomes:

$$L(\beta/Y) = \prod_{i=1}^{N}\left(e^{\sum_{k=0}^{P}x_{ik}\beta_k}\right)^{y_i}\left(1-\frac{e^{\sum_{k=0}^{P}x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{P}x_{ik}\beta_k}}\right)^{n_i} \quad (12)$$

Use $(a^x)^y = a^{xy}$ to simplify the first product and replace 1 with $\frac{1+e^{\sum\infty\beta}}{1+e^{\sum\infty\beta}}$ to simplify the second product. We have:

$$L(\beta/Y) = \prod_{i=1}^{N}\left(e^{\beta_0 y_i + y_i\sum_{k=1}^{P}x_k\beta_k}\right)\left(1+e^{\beta_0+\sum_{k=1}^{P}x_k\beta_k}\right)^{-n_i} \quad (13)$$

This is the kernel of the likelihood function to maximize. We here simplify further by taking its log. Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a

maximum of the log likelihood function, and vice versa. Thus, taking the natural log of Equation 13 yields the log likelihood function:

$$l(\beta)=\sum_{i=1}^{N} y_i\left(\sum_{k=0}^{P} x_{ik}\beta_k\right)-n_i.\log\left(1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}\right) \quad (14)$$

To find the critical points of the log likelihood function, set the first derivative with respect to each β equal to zero. In differentiating Equation 14, note that

$$\frac{\partial}{\partial\beta_k}\sum_{k=0}^{P} x_{ik}\beta_k = x_{ik} \quad (15)$$

Since the other terms in the summation do not depend on β_k, and can thus be treated as constants. In differentiating the second half of Equation 15, take note of the general rule that $\frac{\partial}{\partial x}\log y=\frac{1}{y}\frac{\partial y}{\partial x}$, as well as the fact that $\frac{\partial e^{y}}{\partial x}=\frac{e^{y}\partial y}{\partial x}$. But, $\log(1+e^{\sum_{k=0}^{P} x_{ik}\beta_k})=\left(\frac{\partial\sum_{k=0}^{P} x_{ik}\beta_k}{\partial\beta_k}\right)\left(\frac{1}{1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}}.\frac{\partial e^{\sum_{k=0}^{P} x_{ik}\beta_k}}{\partial\beta_k}\right)$

So that $\log(1+e^{\sum_{k=0}^{P} x_{ik}\beta_k})=x_{ik}\pi_i$. Thus, differentiating Equation 14 with respect to each β_k,

$$\begin{aligned}\frac{\partial l(\beta)}{\partial\beta_k}&=\sum_{i=1}^{N} y_i x_{ik}-n_i.\frac{1}{1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}}.\frac{\partial}{\partial\beta_k}\left(1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}\right)\\&=\sum_{i=1}^{N} y_i x_{ik}-n_i.\frac{1}{1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}}e^{\sum_{k=0}^{P} x_{ik}\beta_k}.\frac{\partial}{\partial\beta_k}\sum_{k=0}^{P} x_{ik}\beta_k\\&=\sum_{i=1}^{N} y_i x_{ik}-n_i.\frac{1}{1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}}e^{\sum_{k=0}^{P} x_{ik}\beta_k.x_{ik}}\end{aligned}$$

Since $\pi_i=\frac{e^{\sum_{k=0}^{P} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}}$

$$\frac{\partial l(\beta)}{\partial\beta_k}=l'(\beta)=\sum_{i=1}^{N} y_i x_{ik}-n_i.\pi_i.x_{ik} \quad (16)$$

Therefore,

$$l'(\beta)=\sum_{i=1}^{N} x_{ik}(y_i-\mu_i)$$

Where

So that the gradient of the log likelihood in matrix form is given as:

$$l'(\beta)=X^{T}(y-\mu) \quad (17)$$

Which is a column vector of length P+1, whose elements are $\frac{\partial l(\beta)}{\partial\beta_k}$.

Let μ be a column vector of length N, with elements $\mu_i=n_i\pi_i$. The maximum likelihood estimates for β can be found by setting each of the P+1 equations in Equation 16 equal to zero, and solving for each $\beta_{k'}$. Each such solution, if any exists, specifies a critical point (either a maximum or a minimum). The critical point will be a maximum if the matrix of second partial derivatives (Hessian matrix) is negative definite; that is, if every element on the diagonal of the matrix is less than zero [8]. It is formed by differentiating each of the P+1 equations in Equation 16, a second time with respect to each element of β, denoted by $\beta_{k'}$. The general form of the matrix of second partial derivatives is

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_{k'}}&=\frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N} y_i x_{ik}-n_i x_{ik}\pi_i\\&=\frac{\partial}{\partial\beta_{k'}}\sum_{i=1}^{N}-n_i x_{ik}\pi_i\\&=-\sum_{i=1}^{N} n_i x_{ik}\frac{\partial}{\partial\beta_{k'}}\left(\frac{e^{\sum_{k=0}^{P} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{P} x_{ik}\beta_k}}\right)\end{aligned} \quad (18)$$

$$\pi_i=\frac{e^{\sum_{k=0}^{P}\beta_k x_{ik}}}{1+e^{\sum_{k=0}^{P}\beta_k x_{ik}}}=\frac{1}{1+e^{-\sum_{k=0}^{P}\beta_k x_{ik}}}$$

The Hessian in a matrix form is given as

$$l''(\beta)=-X^{T}WX \quad (19)$$

Where W is a square matrix of order N, with elements $n_i\pi_i(1\text{-}\pi_i)$ on the diagonal, and zeros everywhere else. To solve Equation 18, we will make use of two general rules for differentiation. First, a rule for differentiating exponential functions:

$$\frac{d}{dx}e^{u(x)}=e^{u(x)}.\frac{d}{dx}u(x) \quad (20)$$

In our case, let $u(x)=\sum_{k=0}^{P} x_{ik}\beta_k$. Second, the quotient rule for differentiating the quotient of two functions:

$$\left(\frac{f}{g}\right)'(a)=\frac{g(a).f'(a)-f(a).g'(a)}{[g(a)]^2} \quad (21)$$

Applying these two rules together allows us to solve Equation 18.

$$\begin{aligned}\frac{d}{dx}\frac{e^{u(x)}}{1+e^{u(x)}}&=\frac{\left(1+e^{u(x)}\right).e^{u(x)}\frac{d}{dx}u(x)-e^{u(x)}.e^{u(x)}\frac{d}{dx}u(x)}{\left(1+e^{u(x)}\right)^2}\\&=\frac{e^{u(x)}\frac{d}{dx}u(x)+[e^{u(x)}]^2\frac{d}{dx}u(x)-[e^{u(x)}]^2\frac{d}{dx}u(x)}{\left(1+e^{u(x)}\right)^2}\\\text{or}\quad&=\frac{e^{u(x)}\frac{d}{dx}u(x)+\left[1+e^{u(x)}-e^{u(x)}\right]}{\left(1+e^{u(x)}\right)^2}\\&=\frac{e^{u(x)}\frac{d}{dx}u(x)}{\left(1+e^{u(x)}\right)^2}=\frac{e^{u(x)}}{\left(1+e^{u(x)}\right)^2}\frac{d}{dx}u(x)\\&=\frac{e^{u(x)}}{1+e^{u(x)}}.\frac{1}{1+e^{u(x)}}.\frac{d}{dx}u(x).\end{aligned} \quad (22)$$

Since

$\frac{du(x)}{dx}=\frac{d}{dx}\sum_{k=0}^{P} x_{ik}\beta_k=x_{ik'}$ while, π_i *and* $1-\pi_i$ clearly defined. Thus, Equation 18 can now be written as:

$$-\sum_{i=1}^{N} n_i x_{ik}\pi_i(1-\pi_i)x_{ik'} \quad (23)$$

Newton-Raphson Iteration Procedure

In finding the roots of Equation 16 using Newton-Raphson method, we generalize the method to a system of P+1 equations. This is done by expressing each step of the Newton-Raphson (NR) algorithm, through letting β^{old} or $\beta^{(0)}$ represent the vector of initial approximations for each β_k, so that the result of this algorithm in matrix notation gives:

$$\beta^{new}=\beta^{old}+\left[-l''(\beta^{old})\right]^{-1}.l'(\beta^{old}) \quad (24)$$

Substituting the values of $l'(\beta)$ *and* $l''(\beta)$ above simplifies the equation to a matrix form, given as

$$\begin{aligned}\beta^{new}&=\beta^{old}+\left(X^{T}WX\right)^{-1}X^{T}(Y-\mu)\\&=\left(X^{T}WX\right)^{-1}X^{T}W\left(X\beta^{old}+W^{-1}(Y-\mu)\right)\end{aligned}$$

$$= \left(X^T WX\right)^{-1} X^T WZ \text{ where } Z = X\beta^{old} + W^{-1}(Y-\mu) \quad (25)$$

Where $Z = X\beta^{old} + W^{-1}(Y-\mu)$ is a vector and W is the diagonal weight vector, with entries $\pi_i(1\text{-}\pi_i)$.

The last equation is called the weighted least square regression, which finds the best least-squares solution to the equation. The equation is called recursive weighted least squares, because at each step, the weight vector W keeps changing (since the β's are changing). Now, Equation 25 can be written:

$$\beta^{(1)} = \beta^{(0)} + \left[X^T WX\right]^{-1} .X^T (Y-\mu) \quad (26)$$

Continue applying Equation 26 until there is essentially no change between the elements of β from one iteration to the next. At that point, the maximum likelihood estimates are said to have converged, and Equation 19 will hold the variance-covariance matrix of the estimates. Because the estimation algorithm for the parameter of the logistic regression model is iterative, parameter estimates based on small samples way fail to converge, or converge to local rather than global, stationary points. This informed the application of large sample in this study. This iterative procedure is handled by SAS software in this work.

Illustrative Example

In estimating the effects of risk factors on GDM, 1000 subjects (pregnant women at risk for GDM) were sampled from the five randomly selected hospitals from January 2010 to December 2011 in Ebonyi State through a retrospective study, out of which 490 (49%) were those less than 28 weeks of their gestational age, and 510 (51%) were those at least 28 weeks of their gestational age. In the total sampled subjects, 530 (53%) were gestational diabetic and 470 (47%) were non-gestational diabetic. Since GDM is a dichotomous variable, it is coded as 0 or 1, and the independent factors considered in this work are Age, Category of pregnant women, Obesity, Income group, Life-style and exercise, F.H of diabetes, Hypertension, and Diet habit are also categorical and coded between 0 and 3. These are presented in table 1.

Results of Analysis

The results are shown in the following tables: Tables 2 and 3

The table 3 shows that three risk factors: Obesity, F.H and Exercise, were significant because for all the above variables p-value was less than 0.05. Since the hospitals where these data were collected are mainly located in the urban areas, it means that by the results

No	Variables	Code number	Coding	Frequency
1	Age	0 if age <30, and 1 for at least 30	0 1	247 753
2	Category of pregnant women	0 if <28 wks of gestational age, and 1 if at least 28 wks	0 1	490 510
3	Obesity	0=non-obsessed and 1 for obsessed	0 1	415 585
4	Income	1=High, 2=Middle, 3=Low	1 2 3	140 390 470
5	Family history	0=Absent,1=Present	0 1	551 449
6	Exercise	0=Sedentary, 1=Light, 2=Moderate	0 1 2	391 472 137
7	Hypertension	0=Non-hypertension,1=Hypertension	0 1	636 364
8	Diet Habit (DH)	0=if absent,1=if present	0 1	652 348

Table 1: Code sheet of concerned independent variables.

Variable	χ^2	Df	P-value	Result	Phi or Creamer's V value
Age	1.350	1	0.245	N.S	-0.037/0.037
Categories of women	0.451	1	0.502	N.S	0.021
Obesity	74.34	1	0.000	S	0.273
Systolic hypertension	1.166	2	0.558	N.S	0.034
Family history	58.357	1	0.000	S	0.242

Table 2: Chi-square analysis of covariates showing significance, after comparison with p and phi-value for the sample.

Variable	$\hat{\beta}$	SE($\hat{\beta}$)	Wald	Df	P-value	Odds ratio	LCL	UCL
Obesity	1.104	0.142	60.597	1	0.000	3.017	2.285	3.984
FH	0.912	0.139	43.170	1	0.000	2.489	1.896	3.267
Constant	-0.709	0.147	23.145	1	0.000	0.492		

Table 3: Results of fitting the Multiple Logistic Regression Model, including O.R and 95% C.I, by using stepwise logistic procedure for the sample.

obtained, it implies that lifestyle of urban area, taking high calories food, less physical activity, invention of remote control equipments and less exercise are the causes of incidence of obesity in the sample data analysized. Moreover, genetical and environmental behaviors are also the reasons of obesity. The reference group for obesity was taken as non-obese persons. The O.R for obesity was 3.017, which shows that an obese person has 3.017 times more chance of getting a significant GDM, as compared to non-obese person keeping all other factors constant. As the O.R for obesity was greater than 1 and the 95% confidence interval for obesity did not include 1, therefore, obesity has a positive association with GDM, and was statistically significant. The reference group for F.H was taken as absent of F.H persons. The O.R for F.H was 2.489, which means that a pregnant woman in Ebonyi State with positive F.H has 2.489 times more chance of getting a significant GDM, as compared to a pregnant woman in which F.H of GDM was absent. Therefore, F.H was significantly different from reference group, and was positively associated with GDM. The reference group for, exercise was sedentary life style. The O.R for exercise was 0.519, which is less than 1 because by general rule, if O.R is less than 1 and chi-square is significant, then there is a protection of exposure against outcome; also 95% confidence interval for exercise did not include 1, therefore, O.R for exercise was significantly different from reference group, and shows that the person who take light exercise have 0.481 probability of protection against GDM. In the light of the above analysis for the 1000 sampled pregnant women, since it turns out that 3 risk factors, obesity, F.H and exercise were significant, that means empirical findings confirm concept and theory of risk factors. So clinicians and public health personal should take appropriate measures to control these risk factors, and prevention programs should be started against GDM. In the remaining 5 risk factors; age, category of women, income, hypertension and D.H, empirical findings do not confirm the concept and theories of risk factors. The theme of every study started with past literature and studies done by experts. According to the literature, these five variables were also the risk factors of diabetes in different regions of the world.

Multivariate Version with Interaction Terms

All the interactions terms were calculated separately and tested for significance at 5% level of significance (Table 4).

In the sample analysis, the main effect factors: category of women, age, obesity and F.H were significant risk factors. Besides the independent factors age was interacted with gender (P=0.005), exercise (P=0.000), and D.H (P=0.016) showed significant effect. Similarly,

Variable and Interactions	β	P-value	OR
category of women	-1.009	0.001	0.365
Age	-1.252	0.025	0.286
Obesity	1.884	0.000	6.582
F.H	0.986	0.000	2.679
Age*Gender	0.926	0.005	2.525
Age*Exercise	-0.770	0.000	0.463
Age*D.H	0.850	0.016	2.339
Obesity*INCM	-0.524	0.008	0.592
Obesity*D.H	0.799	0.012	2.223
D.H*INCM	0.634	0.003	1.886

Table 4: Results of significant main effects and interaction terms of sample.

the factor obesity was interacted with INCM (P=0.008), and D.H (P=0.01) was the significant factor, while the factor "D.H" (P=0.01) was interacted with INCM, and had significant effect. The odd ratio for category of women 0.365 and odd ratio for age 0.286 indicated that those women less than 28 weeks of their gestational age and number of pregnant women less than 30 years of age were protected against this disease. Obese (O.R=6.582, P=0.000) and F.H of GDM (O.R=2.679, P=0.000) indicated that obese pregnant women have 6.582 times of chances of disease, as compared to non-obese pregnant women, while the pregnant women having GDM in their family have 2.679 times of developing disease, as compared to that pregnant women in which F.H of GDM was absent. Exercise was insignificant factor, but when it was interacted with age, it become significant (P=0.000). The interaction of age with category of women (P=0.005) and D.H, (P=0.016), separately were the significant factors. Obesity was also significant when it was interacted with INCM (P=0.008), and with D.H (P=0.012), since obesity has "O.R"=6.582, (P=0.000) in the main effects, but when it was interacted with D.H, the "O.R" decreases to 2.223, (P=0.01); that means by using balanced or proper diet, obesity can be reduced. Some of these interaction terms were very important, while the others were not statistically significant, or explaining no biological relationship for interpretation. For example: in the main effect model, age and category of women showed insignificant effect, but their interaction showed significant effect with odd ratio greater than 1. Similarly, INCM and obesity when interact with each other gave misleading interpretation with O.R=0.592.

Logit Model for Overall Sample with and without Interaction Terms

The model with out interaction terms:

$$\hat{g}(x) = -0.709 + 1.104 * Obes + 0.912 * F.H - 0.656 * Exer$$

The model with interaction terms for the sample is given below

$$\hat{g}(x) = 0.140 - 1.009 \times category\ of\ women - 1.252 \times Age + 1.884 \times Obesity + 0.986 \times F.H + 0.926(Age \times Category\ of\ women) - 0.770(Age \times Exercise) + 0.850 = (Age \times D.H) - 0.524(Obesity \times INCM) + 0.799(Obesity \times D.H) + 0.634(Obesity \times INCM)$$

Summary of Conclusions

We here summarize and conclude as follows:

1. In this hospital base study, ratio of GDM pregnant women is greater than the ratio of non-GDM pregnant women, and the pregnant women from 28 weeks of their gestational age are more liable to diabetes than those less than 28 weeks of their gestational age. The pregnant women entering the hospitals for GDM screening, greater than thirty years of age are three folds than the pregnant women of less than thirty years, concluded that GDM is more common in people above thirty years, and prevalence rate of GDM clearly increased with advancing age. Similarly, obese pregnant women are 1.4 folds than the non-obese pregnant women, and pregnant women with family history of GDM are approximately equal to with out having F.H of GDM in this sample. It is also concluded from the epidemiological study that educated pregnant women have awareness of GDM, and are more careful than the uneducated pregnant women.
2. In the sample analysis, the risk factors: obesity, F.H, were positively associated with GDM, and factor exercise was protection against this disease.

Exercise is protection against this disease, that means pregnant women who take exercise and led a simple life-style are at lesser risk of GDM and other diseases, as compared to those pregnant women who led sedentary lifestyle.

Recommendations

We here recommend on the ROC analysis that a threshold of 177 mg/dl becomes the cutoff value of 50 grams GCT, for screening of GDM in each trimester in GDM risk women, and it is suitable for low BMI or non-obese pregnancy. I also recommend that semi-parametric GLMM method, be used in evaluating the impact of covariates in diagnostic testing programmes, since by comparison it is far better than other methods, in terms producing smooth ROC curves, and compares favorably with other methods. In the second aspect of the analysis, I recommend that since emphasis is on prevalence of GDM, pregnant women with more than thirty years of age, greater number of pregnant women from 28 weeks of gestational age than those less than 28 weeks of gestational age, obesity, F.H and educational level suggests that GDM is not associated to only single risk factor, but it may be associated by more than one risk factor. It is clear from the findings of the study that in overall sample analysis. Obesity and F.H of diabetes are associated risk factors; so a GDM patient or non-GDM pregnant woman must be aware about the consequences of a regular high or low blood sugar level, and amount of cholesterol in blood, so precautionary measures must be taken to control the sugar level. Physical activity is inversely related with BMI, so it is recommended to urgently adopt measures to increase physical activity in these populations. Only a small numbers of pregnant women are aware of the increased genetic susceptibility of their first or second-degree relatives to develop GDM, suggested for weight reduction and regular physical exercise. As a rapidly expanding society problem, GDM requires collective efforts, which must include giving attention to prevention. Consistent with epidemiological concepts, prevention of GDM should be focused by reducing the threat of incidence of the disease, with the help of good nutritional status, physical fitness and regular check up for the individuals of the society; secondary early detection of the disease is necessary. Clinicians should advise the pregnant women, especially to more than 30 years of age, having F.H of GDM for monitoring adequate blood glucose level, or at least urine test for diagnosing GDM. Advice for measuring blood pressure is also very necessary. The doctors or clinicians should arrange staged management programmes. These programmes would be very beneficial and economical for the society. If the probability for getting GDM is high after clinical prediction model, then clinicians should advise the patients for controlling obesity and blood pressure, motivate for exercise, and to use balanced diet. They should arrange seminars at district level. Greater knowledge of risk factors about GDM may help to plan prevention programmes for GDM in future. Government of

Ebonyi State and Health Ministries, with the collaboration of WHO, should arrange the maximum number of seminars and conferences on diabetes. To educate and aware the people against GDM, media should play its significant role. Non-Government Organizations (N.G.O's) can also play their role with the help of well- trained health care team, educating both patients and general public with the consequences and complications of this chronic disease. In rural areas, special arrangements should be made for educating the people about balance diet and about this disease. Further studies are needed to specify the change associated with psychosocial problems in Ebonyi State, and to study the genetic components of individually as well as collectively effect of those risk factors, which are associated to GDM.

References

1. Agresti A (2007) An Introduction to categorical data analysis. (2nd Edn), Wiley, New York, USA.
2. Alonzo TA, Pepe MS (2002) Distribution-free ROC analysis using binary regression techniques. Biostatistics 3: 421-432.
3. Hosmer DW, Lemeshow S (2000) Applied logistic regression. (2nd Edn), Wiley-Interscience Publication, New York, USA.
4. Pepe M (2004) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York, USA.
5. Chou P, Liaq MJ, Tsai ST (1994) Risk factors of diabetes. Diabetes Res Clin Pract 26: 229-235.
6. Jafar TH, Chaturvedi N, Pappas G (2006) Prevalence of overweight and obesity and their association with hypertension and DM in an Indo-Asian population. CMAG 175: 1071-1077.
7. Hagura R, Matsuda A, Kuzuya T, Yoshinaga H, Kosaka K (1994) Family history of diabetic patients in Japan. Diabetes Res Clin Pract S69-S73.
8. Fox J (2005) Maximum-likelihood estimation of the logistic regression model. UCLA/CCPR Notes.