

Estimating the fisher's scoring matrix formula from logistic model

Okeh UM¹, Oyeka I. C. A.²

¹Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria

²Department of Applied Statistics, Nnamdi Azikiwe University, Awka Nigeria

Email address:

uzomaokey@gmail.com(Okeh UM)

To cite this article:

Okeh UM, Oyeka I.C.A.. Estimating the Fisher's Scoring Matrix Formula from Logistic Model. *American Journal of Theoretical and Applied Statistics*. Vol. 2, No. 6, 2013, pp. 221-227. doi: 10.11648/j.ajtas.20130206.19

Abstract: This paper proposes a matrix approach to estimating parameters of logistic regression with a view to estimating the effects of risk factors of gestational diabetic mellitus (GDM). The proposed method of maximum likelihood estimation (MLE) 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 on GDM where obesity and F.H as risk factors, were positively associated with GDM. The proposed method was seen to compare favorably with other known methods.

Keywords: GDM, Logistic Regression, Dichotomous, Fisher Scoring, Newton-Raphson, Risk factors

1. 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. 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. When the result of the diagnostic test is dichotomous, the accuracy of the test is measured in terms of its sensitivity and specificity. 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 (Hosmer and Lemeshow, 2000).

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). The developed 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.

We here intend to apply a Fisher Scoring method of

maximum likelihood estimation (MLE) of parameters by adopting the matrix approach to solving systems of non-linear equations with the major purpose of estimating the effects of risk factors on the outcome of an event (Knight, 2000; Schworer and Hovey, 2004). These methods are illustrated using data from pregnant women at risk for gestational diabetic mellitus (GDM) [a type of diabetes mellitus defined as glucose intolerance of varying degree of severity with onset or first recognition during pregnancy (Metzger and Coustan, 1998; Setji et al, 2005)] and are shown to compare favorably with other existing methods.

2. The Proposed Method

Given that the general form of logistic model is

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

where $i = 1, 2, \dots, N$

By logit transformation, we have from Equation 1 that

$$\frac{\pi_i}{1-\pi_i} = e^{\sum_{k=0}^P \beta_k x_{ik}} \quad (2)$$

We evaluate Equation 2 to obtain π_i and $\frac{1}{1 - \pi_i}$ as

$$\begin{aligned} \pi_i &= e^{\sum_{k=0}^P \beta_k x_{ik}} - \pi_i e^{\sum_{k=0}^P \beta_k x_{ik}} \\ \pi_i + \pi_i e^{\sum_{k=0}^P \beta_k x_{ik}} &= e^{\sum_{k=0}^P \beta_k x_{ik}} \\ \pi_i \left(1 + e^{\sum_{k=0}^P \beta_k x_{ik}} \right) &= e^{\sum_{k=0}^P \beta_k x_{ik}} \\ \pi_i &= \frac{e^{\sum_{k=0}^P \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^P \beta_k x_{ik}}} \\ \pi_i &= \frac{1}{1 + e^{-\sum_{k=0}^P \beta_k x_{ik}}} \end{aligned} \quad (3)$$

Similarly,

$$1 - \pi_i = 1 - \frac{1}{1 + e^{-\sum_{k=0}^P \beta_k x_{ik}}} = \frac{1}{1 + e^{\sum_{k=0}^P \beta_k x_{ik}}} = \frac{e^{-\sum_{k=0}^P \beta_k x_{ik}}}{1 + e^{-\sum_{k=0}^P \beta_k x_{ik}}} \quad (4)$$

3. Obtaining Parameters of Non-Linear Equations from Binomial Distribution Using Maximum Likelihood Estimation (Mle)

We here estimate the $P+1$ unknown parameters β using MLE according to Czepiel (<http://czep.net/contact.html>) as:

$$f(Y|\beta) = \prod_{i=1}^N \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \quad (5)$$

Where β is incorporated as a regression coefficient in the general form for binary logistic model. The joint PDF in Equation 5 expresses the values of Y as a function of known, fixed values for β . Meanwhile, the likelihood function has the same form as the PDF, 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,

$$L(\beta|Y) = \prod_{i=1}^N \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \quad (6)$$

We here find the maximum likelihood estimates which are the values for β required for computing the first and second derivatives of the likelihood function. We simplify the likelihood equation instead of the difficult task of differentiating with respect to β . Since the factorial terms

do not contain any of the π_i , there are therefore constants and can be ignored. Therefore, the equation to be maximized can be written as:

$$\left. \begin{aligned} L(\beta|Y) &= \prod_{i=1}^N \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \\ \text{or} \\ L(\beta|Y) &= \prod_{i=1}^N \pi_i^{y_i} \times (1 - \pi_i)^{n_i} \times (1 - \pi_i)^{-y_i} \\ \text{or} \\ L(\beta|Y) &= \prod_{i=1}^N \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{n_i} \end{aligned} \right\} \quad (7)$$

Substituting Equation 3 for the first term and Equation 4 for the second term, to obtain:

$$L(\beta|Y) = \prod_{i=1}^N \left(e^{\sum_{k=0}^P x_{ik} \beta_k} \right)^{y_i} \left(1 + e^{\sum_{k=0}^P x_{ik} \beta_k} \right)^{-n_i} \quad (8)$$

But

$$\begin{aligned} \frac{\pi_i}{1 - \pi_i} &= e^{\sum_{k=0}^P x_{ik} \beta_k} \text{ and } 1 - \pi_i = \frac{1}{1 + e^{\sum_{k=0}^P x_{ik} \beta_k}} \\ L(\beta|Y) &= \prod_{i=1}^N \left(e^{y_i \sum_{k=0}^P x_{ik} \beta_k} \right) \left(1 + e^{\sum_{k=0}^P x_{ik} \beta_k} \right)^{-n_i} \end{aligned} \quad (9)$$

Equation 9 is the actual likelihood function to maximize. We here simplify Equation 9 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 9 yields the log likelihood function:

$$l(\beta) = \sum_{i=1}^N y_i \left(\sum_{k=0}^P x_{ik} \beta_k \right) - n_i \cdot \log \left(1 + e^{\sum_{k=0}^P x_{ik} \beta_k} \right) \quad (10)$$

In differentiating Equation 10, note that

$$\frac{\partial}{\partial \beta_k} \sum_{k=0}^P x_{ik} \beta_k = x_{ik} \quad (11)$$

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta_k} &= \sum_{i=1}^N y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^P x_{ik} \beta_k}} \cdot \frac{\partial}{\partial \beta_k} \left(1 + e^{\sum_{k=0}^P x_{ik} \beta_k} \right) \\ &= \sum_{i=1}^N y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^P x_{ik} \beta_k}} e^{\sum_{k=0}^P x_{ik} \beta_k} \cdot \frac{\partial}{\partial \beta_k} \sum_{k=0}^P x_{ik} \beta_k \\ &= \sum_{i=1}^N y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^P x_{ik} \beta_k}} e^{\sum_{k=0}^P x_{ik} \beta_k} x_{ik} \end{aligned} \quad (12)$$

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

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

Therefore

The maximum likelihood estimates for β can be found by setting each of the $P + 1$ equations in Equation 12 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 (Glub and Van,1996). The Hessian matrix also forms the variance-covariance matrix of the parameter estimates. It results from taking the second derivative of Equation 12. The general form of the matrix of second partial derivatives (Hessian matrix) is

$$\begin{aligned} \frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} &= \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^N y_i x_{ik} - n_i x_{ik} \pi_i \\ &= \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^N -n_i x_{ik} \pi_i \\ &= - \sum_{i=1}^N n_i x_{ik} \frac{\partial}{\partial \beta_{k'}} \left(\frac{e^{\sum_{k=0}^P x_{ik} \beta_k}}{1 + e^{\sum_{k=0}^P x_{ik} \beta_k}} \right) \\ \text{where } \pi_i &= \frac{e^{\sum_{k=0}^P \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^P \beta_k x_{ik}}} = \frac{1}{1 + e^{-\sum_{k=0}^P \beta_k x_{ik}}} \end{aligned} \quad (13)$$

To solve Equation 13 we will make use for exponential functions and the rule for quotient of two functions so as to obtain

$$\begin{aligned} \frac{d}{dx} \frac{e^{u(x)}}{1 + e^{u(x)}} &= \frac{(1 + e^{u(x)}) \cdot \frac{d}{dx} u(x) - e^{u(x)} \cdot \frac{d}{dx} u(x)}{(1 + e^{u(x)})^2} \\ &= \frac{e^{u(x)} \frac{d}{dx} u(x) + [e^{u(x)}]^2 \frac{d}{dx} u(x) - [e^{u(x)}]^2 \frac{d}{dx} u(x)}{(1 + e^{u(x)})^2} \\ \text{or } &= \frac{e^{u(x)} \frac{d}{dx} u(x) + [1 + e^{u(x)} - e^{u(x)}]}{(1 + e^{u(x)})^2} \\ &= \frac{e^{u(x)} \frac{d}{dx} u(x)}{(1 + e^{u(x)})^2} = \frac{e^{u(x)}}{(1 + e^{u(x)})^2} \frac{d}{dx} u(x) \\ &= \frac{e^{u(x)}}{1 + e^{u(x)}} \cdot \frac{1}{1 + e^{u(x)}} \cdot \frac{d}{dx} u(x). \end{aligned} \quad (14)$$

Since

$$\frac{du(x)}{dx} = \frac{d}{dx} \sum_{k=0}^P x_{ik} \beta_k = x_{ik},$$

while π_i and $1 - \pi_i$ are clearly defined. Thus, Equation 13 can now be written as:

$$l''(\beta) = - \sum_{i=1}^N n_i x_{ik} \pi_i (1 - \pi_i) x_{ik}. \quad (15)$$

3. Newton-Raphson Iteration Procedure

In finding the roots of Equation 12 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:

$$\beta^{new} = \beta^{old} + [-l''(\beta^{old})]^{-1} \cdot l'(\beta^{old}) \quad (16)$$

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

$$\begin{aligned} \beta^{new} &= \beta^{old} + (X^T W X)^{-1} X^T (Y - \mu) \\ \beta^{New} &= (X^T W X)^{-1} X^T W (X \beta^{old} + W^{-1} (Y - \mu)) \\ \beta^{New} &= (X^T W X)^{-1} X^T W Z \text{ where } Z = X \beta^{old} + W^{-1} (Y - \mu) \end{aligned} \quad (17)$$

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

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, Equations 16 and 17 can be written:

$$\beta^{(1)} = \beta^{(0)} + [X^T W X]^{-1} \cdot X^T (Y - \mu) \quad (18)$$

Continue applying Equation 18 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 17 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.

4. Newton-Raphson Optimization/ Algorithm

Having obtained the two derivatives of the log likelihood function, we now establish the general form for Newton-Raphson approach which is an iterative algorithm method for obtaining the roots of non-linear equations. It converges to a unique maximum of the likelihood function (See

Czepiel (<http://czep.net/contact.html>; Lambers, 2009). The algorithm involved are as follows:

1. Start with an initial guess of the solution (parameters) given as $\beta^{(0)}$
2. Approximate the function to be maximized in the neighborhood of the initial (current) guess by a second degree polynomial. Eg. Let $\beta^{(1)} = \beta^{(0)} - \frac{f(\beta^{(0)})}{f'(\beta^{(0)})}$; $n = 0, 1, \dots$ where given $f(\beta)$, its derivative is $f'(\beta)$.

3. Find the maximum of the polynomial to get better guesses for the parameters

4. Using new estimates go back to step 2 and repeat until convergence

We therefore generalize Newton's method to a system of $P+1$ equations so as to obtain the roots. This is written according to Lauritzen (2009) by letting $\beta^{(0)}$ represent the vector of initial approximations for each β_k , as: $\beta^{(1)} = \beta^{(0)} + [-l''(\beta^{(0)})]^{-1} \cdot l'(\beta^{(0)})$ (19)

Where $l'(\beta^{(0)})$ and $l''(\beta^{(0)})$ represents the first and second derivatives of the log likelihood function while $\beta^{(0)}$ represents the initial guess parameter while $\beta^{(1)}$ is the new estimate

5. Matrix Form for Newton-Raphson Algorithm

We first transform the results of $l'(\beta)$ and $l''(\beta)$ to a matrix form

Let μ be a column vector of length N with elements $n_i \pi_i$. Therefore, Equation 19 becomes

$$l'(\beta) = \sum_{i=1}^N x_{ik} (y_i - \mu_i) = X'(y - \mu) \quad (20)$$

where $\mu_i = n_i \pi_i$ and $X'(y - \mu)$ is the gradient of the log likelihood in matrix form.

Also, let W be a square diagonal matrix $(P+1 \times P+1)$ of order N , with elements $n_i \pi_i (1 - \pi_i)$ on the diagonal and zeros elsewhere. Again, transforming the second derivative to matrix form gives

$$l''(\beta) = -X'WX \quad (21)$$

Where $W = n_i \pi_i (1 - \pi_i)$

Therefore, according to Lauritzen (2009), the Newton-Raphson method in matrix form is written as:

$$\beta^{(1)} = \beta^{(0)} + (X'WX)^{-1} \cdot X'(y - \mu) \quad (22)$$

Simplifying

$$\begin{aligned} \beta^{(1)} &= \beta^{(0)} + (X'WX)^{-1} X'(y - \mu) \\ &= (X'WX)^{-1} \{X'WX \beta^{(0)} + W^{-1}(y - \mu)\} \\ \beta^{(1)} &= (X'WX)^{-1} X'WZ \end{aligned} \quad (23)$$

Where $X' = W^{-1}$, $Z = X \beta^{(0)} + W^{-1}(y - \mu)$

and Z is an $N - \dim$ vector with i th component

Continue applying the general form for Newton-Raphson method 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. Because the estimation algorithm for the parameter of the logistic regression model is iterative, parameter estimates based on small samples may fail to converge. This informed the application of large sample in this study. Another problem with the Newton-Raphson method is its lack of stability. When the initial value β^0 is far from β , it might wildly oscillate and not converge at all (Lauritzen, 2009; Lambers, 2009).

6. Fisher Scoring

This is a similar method to Newton-Raphson but modified to overcome the convergence problem of the Newton-Raphson method (Farbod et al, 2010). It is given by

$$\begin{aligned} B^{(1)} &= B^{(0)} + \left\{ E \left(-\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} \right) \right\}^{-1} \frac{\partial l(\beta)}{\partial \beta_k} \\ \text{or} \\ B^{(1)} &= B^{(0)} - [E \{ l'' B^{(0)} \}]^{-1} l(B^{(0)}) \end{aligned} \quad (24)$$

Whereby given the initial estimate $\beta^{(0)}$, both derivatives are evaluated at $\beta^{(0)}$ and the algorithm updates them to $B^{(1)}$, so that the expectation is evaluated as if $\beta^{(0)}$ were the true parameter values. Secondly, $\beta^{(0)}$ is then replaced by $\beta^{(1)}$ and the updating is repeated until convergence. In subsequent steps where there exists $\beta^{(2)}, \beta^{(3)}, \dots, \beta^{(k)}$, the Hessian matrices of $\beta^{(2)}, \beta^{(3)}, \dots, \beta^{(k)}$ are replaced by their expectations. This is actually what distinguishes it from the Newton-Raphson method. In Fisher Scoring method, the use of $E[H(\beta)]$ rather than $H(\beta)$ has two advantages over the Newton-Raphson method. First because $E \left[\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} \right] = -E \left[\left(\frac{\partial l(\beta)}{\partial \beta_k} \right) \left(\frac{\partial l(\beta)}{\partial \beta_{k'}} \right) \right]$ we need only calculate first, rather than second, derivatives of log likelihood function. A second advantage is that $E[H(\beta)]$ is guaranteed to be positive-definite, thus eliminating the possible non-convergence problems of the Newton-Raphson method. The result obtained in Equation 24 is actually used to obtain the roots of the non-linear equation from where the odds ratio used to interpret the effects of risk factors of GDM were estimated.

7. Illustrative Example

Since GDM is a dichotomous dependent variable such as having present and absent, it is coded as 0 or 1 and its

independent variables considered are Age, Category/level of pregnant women, Systolic Hypertension, Obesity/BMI, Family history of diabetes. There are also categorical and are coded between 0 and 1.

Table 1. Distribution of patients by GDM and risk factors including; Age, Categories of women, Obesity, Systolic Hypertension, F.H, for Overall Sample

Variable	Coding	GDM		Total
		0	1	
Age	0	268	48	316
	1	699	99	798
Categories of women	0	458	36	494
	1	509	111	620
Obesity	0	407	53	460
	1	560	94	654
Systolic Hypertension	0	424	80	504
	1	543	67	610
FH	0	561	69	630
	1	439	45	484

Table 2. Chi-square analysis of covariates showing significance after comparison with p and phi/creamer-value for Overall Sample

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

Variable	$\hat{\beta}$	SE($\hat{\beta}$)	Wald	Df	P-value	Odds ratio	Confidence interval	
							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		

Two risk factors; Obesity and F.H were significant because for all the above variables p-value was less than 0.05. The reference group for obesity was taken as non-obese women. The O.R for obesity was 3.017, which shows that an obese pregnant woman has 3.017 times more chance of getting a significant GDM as compared to non-obese pregnant woman 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 from pregnant women with absent of F.H. The O.R for F.H was 2.489, which means that a pregnant women in Ebonyi State with positive F.H has 2.489 times more chance of getting a

8. Results of Analysis on the Estimation of Effect of Risk Factors for GDM

Contingency tables were constructed and appropriate tests were applied including chi-square, as well as the multiple logistic regression model. The O.R was calculated to describe the association of risk factors with response variable GDM. Women at <24weeks of gestation and those at greater than or equal to 24 weeks of gestation were entered separately. In overall sample analysis, obesity and family history of diabetes were positively associated with GDM. The risk factor, FH was positively associated with GDM and significant.

9. Univariate Analysis

In this section the association of every risk factor was tested with GDM. For this purpose χ^2 test was applied. It was found that the risk factors obesity and F.H were significantly associated with GDM. Obesity has the largest phi/creamer's V value (0.273). So it was highly associated with GDM among all the significant risk factors.

significant GDM as compared to a pregnant women in which F.H of diabetes was absent. Therefore F.H was significantly different from reference group and was positively associated with GDM.

In the light of the above analysis for overall sample since it turns out that two risk factors obesity, and F.H 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 three risk factors: age, category of women and systolic Hypertension, empirical findings do not confirm the concept and theories of risk factors. The theme of every study is always started with

past literature and studies done by experts. According to the literature, these three variables were also the risk factors of GDM in different parts of the world. The factor, category of women was in fact significant in the overall sample analysis but it was excluded from the study because it was insignificant when its association was checked with GDM by using 2×2 chi-square table. Moreover because it is also not a significant factor due to the reason that in Ebonyi State, there is no discrimination in the category of pregnant women affected by this disease and ratio of those at 2nd and 3rd trimesters (≥ 24 weeks of gestational age) is little higher than 1st trimester (< 24 weeks of gestational age) as in this study. In the factor age, only 24.7% pregnant women were less than 30 years and 75.3% were at least 30 years. Perhaps the reason for this disease is that it usually develops more often from age at least 30 years. Similarly in the factor such as history of GDM, although the percentage of the factors was very high in the descriptive analysis but these factors were not significant in the analytic study. The aim of analytic study was to draw out the main important significant factors and predict for the future by using backward elimination method. This means that it discarded the less important factors and only included the much important risk factors in the model. The reason for insignificant of these factors was that the risk factors of any disease cannot absolutely be demonstrated to be correct in testing the hypothesis and there could be other factors that may interact with the risk factors under study, which will result in rejecting hypothesis. These factors may be related to genetic, immune response and environment. That was the reason of different insignificant factors in overall sample.

In the overall 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 category of women ($P = 0.005$), Similarly the factor, obesity was interacted with systolic Hypertension ($P = 0.008$). The odd ratio for category of women is 0.365 and odd ratio for age is 0.286 which indicates that women at < 24 weeks of gestational age and number of women less than 30 years of age do not have the likelihood of having the disease. Obese ($O.R = 6.582$, $P = 0.000$) and F.H of diabetes ($O.R = 2.679$, $P = 0.000$) indicated that obese persons have 6.582 times of chances of having GDM as compared to non-obese women while the pregnant women having diabetes in their family have 2.679 times of developing disease as compared to that pregnant women in which F.H of diabetes was absent. The interaction of age with category of women ($P = 0.005$) were the significant factors. Obesity was also significant when it was interacted with history of GDM ($P = 0.008$), since obesity has " $O.R$ " = 6.582, ($P = 0.000$) in the main effects. 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 systolic Hypertension and obesity when interact with each other gave misleading interpretation with $O.R = 0.592$.

10. Summary and Conclusions

As in this study; emphasis is on prevalence of GDM, pregnant women from thirty years of gestational age, greater number of women at least 24 weeks of gestational age than those from 24 weeks of gestational age, obesity and F.H suggests that GDM is not associated to only single risk factor but it may be associated by more than one risk factor. As a rapidly expanding society problem, GDM requires collective efforts, which must include giving attention to prevention. Consistent with epidemiological concepts, prevention of diabetes 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 pregnant women or patients especially at least 30 years of age, having F.H of diabetes for monitoring adequate blood glucose level or at least urine test for diagnosing diabetes. 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 pregnant women for controlling obesity and blood pressure, motivate for exercise and to use balanced diet. They should arrange seminars at district level. Also arrangements should be made on union council level for screening diabetes or diabetes care teams should go door to door for this purpose. Greater knowledge of risk factors about diabetes may help to plan prevention programmes for diabetes in future. Government of Ebonyi State and Health Ministries with the collaboration of W.H.O. should arrange the maximum number of seminars and conferences on diabetes. To educate and aware the people against diabetes, 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. Second Edition, Wiley, Inc., New York.

- [2] Alonzo, T.A. and Pepe, M.S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics* 3, 421-432.
- [3] Berger H, Crane J, Farine D, Armson A, De La RS, Keenan-Lindsay L, et al. (2002) Screening for gestational diabetes mellitus. *J Obstet Gynaecol Can*; 24: 894-912.
- [4] Cai, T. and C. S. Moskowitz. (2004). "Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test," *Biostatistics*, vol. 5, no. 4, pp. 573-586.
- [5] Chou, P. Liaq, M.J. Tsai, S.T. (1994) Risk factors of diabetes. *Diabetes Res clin Pract*; 26(3): 229-35
- [6] Cox, D.R. (1970). *Analysis of binary data* London: Chapman and Hill
- [7] Farbod, D., Ebrahimpour, M., Ghayournoradi, Z., (2010), Maximum Likelihood Estimation for Distribution Generated by Cauchy Stable Law, *International Journal of Mathematics and Computation*, Vol. 7, No. J10, pp. 23-28.
- [8] Fox John(2005).Maximum-likelihood Estimation of the Logistic Regression Model.UCLA/CCPR.Notes.
- [9] Friedman S, Khoury-Collado F, Dalloul M, Sherer DM, Abulafia O. Glucose challenge test threshold values in screening for gestational diabetes among black women. *Am J Obstet Gynecol* 2006; 194: e46-8.
- [10] Hagura, R. Matsuda, A. Kuzuya, T. Yoshinaga, H. Kosaka, K. (1994) F.H of diabetic patients in Japan; *Diabetes Res Clin Pract Suppl*: S69-73
- [11] Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression*, Second Edition, Wiley, Inc., New York.
- [12] Jafar, T.H Chaturvedi, N. Pappas, G. (2006). Prevalence of overweight and obesity and their association with hypertension and DM in an Indo-Asian population *CMAJ* ;175(9):1071-7
- [13] Khurram, M. Paracha, S.J. Khar, H.T. Hasan, Z. (2006) Obesity related complications in 100 obese subjects and their age matched controls. *J Pak Med Assoc*, 56(2): 50-3
- [14] Lloyd, C.J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *JASA*, 93, No 444, pp. 1356-1364.
- [15] Lambers, J., (2009), www.math.usm.edu/lambers/mat460/fall09/lecture9.pdf
- [16] Mengesha, B. and Qadeer, A. (1997) Heritability of DM. *JPMA*; 74(1): 37-40
- [17] Morris, R.D. Rimm, D.L. Hartz, A.J. Kalkhoff, R.K. Rimm, A.A (1989) Obesity and F.H of NIDDM, A cross-sectional study. *Amj Epidemiol*; 130(1): 112-21
- [18] Miyakoshi K, Tanaka M, Ueno K, Uehara K, Ishimoto H, Yoshimura Y. Cutoff value of 1 h, 50 g glucose challenge test for screening of gestational diabetes mellitus in a Japanese population. *Diabetes Res Clin Pract* 2003; 60: 63-7.
- [19] Pepe, M. S.(2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York, NY, USA.
- [20] Schmidt, M. I. Duncan, L.H. Canani, C. Karohi, L. Chambless (1992) Association of waist -hip ratio with DM, *Diabetes Care* 15: 912-914. 112
- [21] Shafi, S. Rao, M.H. Bukhsh, I. Soomro, M. (2004) The effect of life style and socio-economic factors in the development of obesity in young adults. *Pakistan J. Med. Res.* 43(2):
- [22] Swets, J.A., Dawes, R.M., Monahan, J., 2000. Better decisions through science. *Scientific American* 283, 82-87.
- [23] Tanir HM, Sener T, Gurer H, Kaya M.(2005). A ten-year gestational diabetes mellitus cohort at a university clinic of the mid-Anatolian region of Turkey. *Clin Exp Obstet Gynecol*; 32: 241-4.
- [24] Vitoratos N, Salamalekis E, Bettas P, Kalabokis D, Chrisikopoulos A.(1997).Which is the threshold glycoase value for further investigation in pregnancy? *Clin Exp Obstet Gynecol*; 24: 171-3.
- [25] Zou, K. H. and Hall, W. J. (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* 27:5, 621-631.