Journal of Advances and Scholarly Researches in Allied Education
Vol. XV, Issue No. 1, April-2018, ISSN 2230-7540

A Statistical Study on Risk Factors for Low Birth Weight Using Logistic Regression & Roc Curve

Sharmy Ann James*

Department of Statistics, Madras Christian College, Chennai, India

Abstract – The Focus of the study is to identify mother's behavioral characteristics that has significant effect on child's birth weight. Low birth weight is the leading cause of infant and child mortality and contributes to several poor health outcomes. Proper knowledge of risk factors of low birth weight is important for identifying those mothers at risk and thereby for planning and taking appropriate actions. The study proposes to evaluate maternal periodontal parameters to predict preterm delivery and low birth weight delivery to correctly classify between low birth weight cases and non-cases. The discriminatory performance of binary logistic regression model is measured using two approaches. The first approach is the use of fitted binary logistic regression model to correctly predict the subjects that are cases and non-cases. The alternative approach is based on receiver operating characteristic (ROC) curve for the fitted binary logistic regression model and then determining the area under the curve (AUC) as a measure of discriminatory performance. The data is abstracted from Baystate Medical Centre Springfield, Massachusetts. This data set contains information on 189 births to women seen in the obstetrics clinic and 59 of these births were low birth weight. The goal of the study is to determine whether these variables were risk factors in the clinic population being studied by Baystate Medical Centre. The logistic regression model is built in R-language.

The present analysis identifies that mother's smoking status (p=0.024), presence of hypertension (p=0.032), presence of Uterine cancer (p=0.097) and history of premature labor (p=0.004) are statistically significant on her neonate birth weight. However, to reduce the infant mortality due to low birth weight this study suggests that a mother should be non-smoker, free of hypertension, free of Uterine cancer and without any premature labor.

----------X----------

INTRODUCTION

The objective of the research is to study about the behavioural characteristics of mother that significantly effects birth weight of the infant. The logistic regression model and receiver operating characteristic curve (ROC) are used for better classification and prediction. The neonatal low birth weight data for this study is collected at Baystate Medical Centre, Springfield. Logistic regression model was used to identify the significant factors which contribute for low birth weight and classify them based on probability concept. ROC curve evaluates the performance of the classification model.

Birth weight is a potent predictor and indicator of infant growth and existence. One of the commonest causes of neonatal mortality in the world is prematurity and low birth weight. Low Birth Weight defined by WHO as a birth weight less than 2500 g, since below this value birth-weight-specific infant mortality begins to rise rapidly. Epidemiological research often seeks to identify a causal relationship between the risk factors and the disease. Mechanisms of mother lifestyle characteristics on her neonate weight are intricately complicated. The primary reason of low birth weight is premature birth (birth before 37 weeks gestation). The present study analyses the relationship of neonate birth weight (response) to the mother's lifestyle explanatory variables. This analysis considered the following factors such as mother weight at last menstrual period, her race, age, smoking status during pregnancy, number of Physician visit during the first trimester, history of premature labor, history of hypertension and presence of Uterine cancer.

METHODOLOGY

Data collection:

Secondary data is used to study the behavioural characteristics of mother. Data was collected as part of a large study at Baystate Medical Centre in Springfield, Massachusetts during 1986. Data was collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. This is a complete data set from Applied Logistic regression written by Hosmer and Lemeshow. The goal of the current study was to

determine whether these variables were risk factors in the clinic population being studied by Baystate Medical Centre.

Source: Hosmer and Lemeshow (2000) Applied Logistic Regression; Second Edition. These data are copyrighted by John Wiley & sons Inc. and must be acknowledged and used accordingly. Data were collected at Baystate Medical Centre, Springfield, Massachuselts, during 1986.

Study variables:

The outcome variable y was defined as a binary response variable conforming to the risk of an infant born with low birth weight (LBW). That is,

$$Y = \begin{cases} 1, Low\ birth\ weight\ infant \\ 0, otherwise \end{cases}$$

There are two sets of independent variables, qualitative and quantitative. Six independent variables (coded low birth weight, mother race, her smoking status during pregnancy, history of premature labour, history of hypertension, presence of Uterine cancer) are qualitative, two are continuous (mother age and her weight at the last menstrual period) and one is discrete (number of physician visits during the first trimester) variables. The present study has neglected the coded low birth weight as an independent variable, as the original neonate birth weight is treated as the response variable.

Models and techniques:

Logistic Regression model is used to identify the significant factors which contributes for LBW by assessing the odds ratios (OR) and their 95% confidence interval (CI). The goal of model is to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$\text{Odds} = \frac{p}{1-p} = \frac{\text{probability of presence of characteristic}}{\text{Probability of absence of characteristic}}$$

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

The Receiver Operating Characteristics (ROC) plot is a popular measure for evaluating classifier performance. The ROC plot is a model-wide evaluation measure that is based on two basic evaluation measures (1-specificity) and sensitivity. Specificity is a performance measure of the whole negative part of a dataset, whereas sensitivity is a performance measure of the whole positive part. The ROC plot uses (1-specificity) on the x-axis and sensitivity on the y-axis. False positive rate (FPR) is identical with 1-specificity, and true positive rate (TPR) is identical with sensitivity. A ROC point is a point with a pair of x and y values in the ROC space where x is 1-specificity and y are sensitivity. A ROC curve is created by connecting all ROC points of a classifier in the ROC space. Two adjacent ROC points can be connected by a straight line, and the curve starts at (0.0, 0.0) and ends at (1.0, 1.0).

When we are dealing with logistic regression, there are two classes coded as 1 and 0. Then we can compute probabilities that given some explanatory variables an individual belongs to the class coded as 1. If we choose a probability threshold and classify all individuals with a probability greater than this threshold as class 1 and below as 0, In most of the cases we will make some errors because usually two groups cannot be discriminated perfectly. For this threshold we can now compute our errors and the so-called sensitivity and specificity. If we do this for many thresholds, we can construct a ROC curve by plotting sensitivity against 1-Specificity for many possible thresholds. The area under the curve comes in play if we want to compare different methods that try to discriminate between two classes. After fitting a model to the observed data, one of the next essential steps is to investigate how well the proposed model fits the observed data. One method which is used to determine the suitability of the fitted logistic model is a goodness of fit test statistic. A model is said to be fit poorly if either the model's residual variation is large, systematic, or does not follow the variability postulated by the model. In the case where important predictor variables or interaction terms are omitted from the postulated model, the resulting logistic model fit would be poor due to an incorrect linear component. If the predicted values produced by the logistic model accurately reflect the observed values, then the logistic model may be a good fit for the given data.

RESULTS & DISCUSSION

Table 1: Univariate Analysis

Variables	Low birth weight		Normal weight		P – Value
	n	%	n	%	
Age (Mean ± SD)	22.3±4.51		23.6±5.59		0.105
Race					
White	23	39.0	73	56.2	0.082
Black	11	18.6	15	11.5	
Other	25	42.4	42	32.3	
Smoking					
No	29	49.2	86	66.2	0.026
Yes	30	50.8	44	33.8	
Pt1					
None	41	69.5	118	90.8	<0.001
One and More	18	30.5	12	9.2	
Hypertension					
No	52	88.1	125	96.2	0.052
Yes	7	11.9	5	3.8	
UI					
No	45	76.3	116	89.2	0.020
Yes	14	23.7	14	10.8	
FTV					
None	36	61.0	64	49.2	0.281
One and More	23	38.9	66	50.8	

Table 2: Logistic Regression Multivariate Analysis

Variables	Odds ratio	95.0% CI. for OR		P value
		Lower	Upper	
Age	0.95	0.88	1.02	0.171
Smoking				
Yes	2.45	1.12	5.34	0.024
No	1.00			
Race				
White	0.38	0.16	0.90	0.028
Black	1.01	0.37	2.72	0.983
Others	1.00			
Hypertension				
Yes	3.91	1.12	13.66	0.032
No	1.00			
Presence of Uterine Cancer				
Yes	2.15	0.87	5.33	0.097
No	1.00			
History of premature Labour				
None	1.00			0.004
one or more	3.73	1.52	9.14	

In Table 1: Univariate Analysis

There were 189 births to women seen in the obstetrics clinic. 39% of the baby had low birth weight. Low birth weight babies mothers mean age is 22.3(4.51). Table 1 shows that univariate analysis for socio demographic variables and clinical variables results. Out of them 50.8% were smoker and 49.2% were non-smoker had low birth weight, which is statistically significant (p=0.026). 30.5% of the mother had history of premature Labor one or more and 69.5% of who did not had history of premature Labor, which is statistically significant (<0.001). Which are the variables being significant in the univariate analysis at 0.1 levels were included in the multivariate analysis.

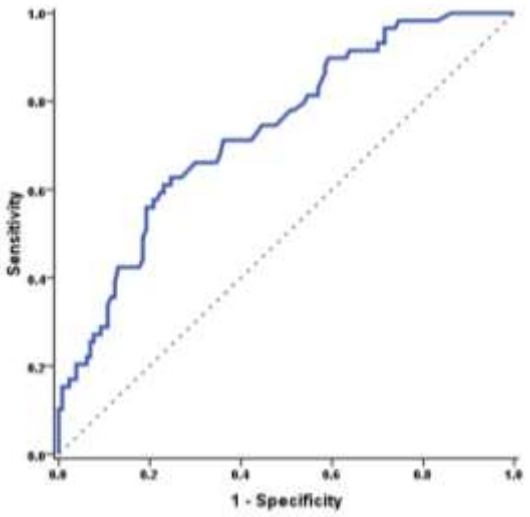
In Table 2: Logistic regression Multivariate Analysis

The variables that were significant at the univariate analyses were considered as potential variables for multivariable logistic regression analyses. In the multivariate analysis, as compared to non-smoker, smoker mother had 2.45(95%CI: 1.12 -5.34) times higher risk for getting low birth after adjusting for other variables (p=0.024). Similarly, mother who had presence of hypertension 3.91(95%CI: 1.12-13.66) times higher risk for getting low birth weight babies as compared to patients with absence of hypertension (p=0.032). Who had Race, white and black had less risk as compared to others after adjusting for other variables, which was statistically significant(p<0.051). mother who had presence of Uterine Cancer 2.15(95%CI: 0.87-5.33) times higher risk for getting low birth weight babies as compared to patients with absence of Uterine Cancer (p=0.097). Similarly, mother who had History of premature Labor 3.73 (95%CI: 1.52-9.14) times higher risk for getting low birth weight babies as compared to patients with absence of history of premature Labour, which is statistically significant (p=0.004).

ROC Curve:

The ROC curves were plotted based on the model Predicted values with low birth weight (Low/Normal). ROC for Model based Sensitivity was (74.5, 95%CI: 61.6 – 85.0, Specificity (52.3, 95%CI: 43.4 – 61.1). AUC was 0.73 (95%CI: 0.66, 0.81)

Figure: ROC Curve



CONCLUSION

Low birth weight is a significant public health concern that is linked to multiple factors. To reduce the infant mortality due to low birth weight this study suggests that a mother should be non-smoker, free of hypertension, free of Uterine cancer and without any premature labor. Therefore, counselling and vital assortment should be carried out for the pregnant women during pregnancy to prevent and reduce preterm deliveries and births of low birth weight infants. The fitted logistic regression model could be implied to categorize the high-risk group of pregnant mothers in the future aspects and some strategies and effective efforts can be emphasized to control the percentage of delivery in low birth weight infants.

LITERATURE CITED

A. M. Ferreira, S. Roy, D. D. Motghare, F. S. Vaz, and M. S. Kulkarni, (2009), "Maternal determinants of low birth weight at a tertiary care," The Journal of Family Welfare, vol. 55, pp. 79–83.

A. Matin, S. K. Azimul, A. K. M. Matiur, S. Shamianaz, J. H. Shabnam, and T. Islam, (2008), "Maternal socioeconomic and nutritional determinants of low birth weight in urban area of Bangladesh," Journal of Dhaka Medical College, vol. 17, no. 2, pp. 83–87.

Hosmer R. Lemeshow J. Applied logistic regression.2nd edition New York: John Wiley & sons Inc; 2000

Jian Bi and Hye-Seong Lee, (2012)," Statistical analysis of receiver operating characteristic curves for the ratings of the A-Not A and the same-different methods", journal of sensory studies, vol.28, no.1.

K. Agarwal, (2011), "Prevalence and determinants of low birth weight among institutional deliveries" Indian journal of low birth weight, vol. 5, no. 2, pp.48-52

Murat KORKMAZ, Selami GUNEY, Sule Yuksel YIGITER, (2012)," The importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields". vol.12, no.2, pp.25-36.

Seong Ho Park, Jin Mo Goo, Chan-Hee Jo, (2004)," Receiver operating characteristic curve: Practical review for radiologists", Korean journal of radiology, vol.5, no.1, pp.11-18.

Corresponding Author

Sharmy Ann James*

Department of Statistics, Madras Christian College, Chennai, India