STAT2402: Analysis of Observations
Technical Report

Predicting Low Birth Weight Using Logistic Regression and Variable Selection Techniques

Cedrus Dang

October 14, 2024

Executive Summary

This report aims to determine the relationship between low birth weight status (Less than 2.5kg) (LBW) in children and other maternal variables by modelling them. The used dataset contained 187 observations, including the mother's record of age, weight, race, smoking status during pregnancy, previous premature labours, hypertension, uterine irritability, physician visits, and birth weight in grams, alongside the target variable and was collected at Baystate Medical Center, Springfield, Mass during 1986. The final model includes interaction and then is treated with AIC backward selection. This model showed that previous premature labour, maternal weight at last menstrual period, hypertension, race as black and not white, smoking status, and uterine irritability were significant predictors of LBW. The key findings are that except the mother's weight can reduce the odds of LBW in the child, other factors, as mentioned, can increase the odds of LWB from as minimal of 2.4 times if the mother is smoking during pregnant to a maximum of 6.39 times if the mother has hypertension on birth. In the analyses, outliers and potential influential points were investigated, and due to their contributed valuable information, they were still retained in the dataset. Diagnostic checks for the final models reasonably fit the data, though they may not fit with extreme values.

I. Introduction

LBW is an important indicator of a child's vulnerability to the risk of childhood illness and chances of survival. Furthermore, LBW is also a significant public health indicator for many nations and organizations, including UNICEF and WHO. [1] [3] [4] However, LBW remains a global concern. The LBW estimation in 2019 by UNICEF indicated a global prevalence of 14.6% (20.5 million newborns) in 2015, making the World Health Assembly conduct the Reducing Low Birth Weight Program between 2012 and 2025 to reduce 30% of LBW cases [4].

Due to the highly negative impact of LBW, the modelling to analyse the relationship between it and other maternal attributions becomes an attractive objective [4]. Various models were applied and reported in several papers. In a research conducted in India in 2015 [1], the authors used ROC analysis to determine the influence of maternal characteristics on birth weight. Further, they investigated using logistic regression analysis by calculating the odd ratios for LBW. They found out that preterm babies (gestational age < 37 weeks) had an 8.9 times higher risk of being LBW compared to full-term babies, and first-time mothers or those with only one child had a 4.5 times higher chance of being LBW compared to those born to mothers with more children. During their work, they have to face multiple challenges from small variables, multicollinearity and a high number of non-significant predictors.

In another research conducted in Sudan in 2008 [3], Logistic regression models were developed to investigate the probability of LBW and maternal characteristics such as age, weight, height, and mid-arm circumference. In the final model, the authors discovered that maternal height and birth order were significant predictors of LBW. The target variable, LBW, indicates that higher birth order and maternal height will lower the risk of LBW. This study found negative coefficients between two explained variables, birth order and maternal height. Their simplified logistic regression model was:

p(BW ≤ 2500g) = 1/(1 + e^-z) with z = 6.434 - 0.625 * PAR - 0.054 * MH. The authors have to face over-fitting Risk, Sensitivity and Specificity trade-offs and assume independence of observations in logistic mode, which may not always hold.

In other papers, methods like Artificial Neural Networks (ANN) [5] and Quantile Regression [6] also been used to predict birth weight based on maternal factors.

In this article, we estimate the factors influencing low birth weight using a Logistic Regression Model. In the next section, we describe the statistical methodology used, followed by the Results section. This is followed by a discussion of the findings, which are compared with the relevant literature discussed in the Introduction.

II. Methodology

The first step is to explore the data set, separated by the type of the variables, numerical and categorical, to understand the attributes of each variable and handle any special phenomena in the data. Only valid variables will be used to for modelling.

After that, five logistic regression models will be fitted following this order:

1. **Null model.**
2. **Model 1:** Model with all valid variables.
3. **Model 2:** Model 1 reduced via Akaike Information Criterion (AIC) backward selection
4. **Model 3:** Add interactions and reduced by AIC backward selection.
5. **Model 4:** Add interactions and reduced by AIC forward selection.

This setup, from simple to complex models with interaction and applying the AIC selection method to reduce the insignificant variables and interactions. By doing this, it will be ensured that every potential relationship between the variables and interactions is captured.

Model selection was based on the Chi-Square test and AIC for comparison, while Wald's test was used to assess variable significance. The Pearson Chi-Square Goodness-of-Fit Test evaluated the overall fit, and diagnostic plots (e.g., residuals vs. fitted, Cook's distance, and leverage) were used for model diagnostics. The optimal model will then be interpreted to explain the relationship between low birth weight status and other variables in the dataset.

All statistical analysis will be conducted in the R statistical environment, and statistical significance will be taken at $\alpha = 0.05$ (5%). The analyses were conducted using glm() for model fitting, step() for AIC selection, summary() for Wald tests, confint() for confidence intervals, anova(,test = "Chisq") for Chi Squard test in model comparison, residuals() for diagnostic checks, residuals(,type= "pearson") for Pearson's test, and plot() for diagnostic plots.

III. Results

The description and summary of the variables in the data set are given in Table 1.

Variable	Description	Summary
low	Indicator of birth weight less than 2.5 kg	Yes = 58, No = 129
age	Mother's age (Years)	Min: 14, Median: 22, Mean: 23.1, Max: 36, SD: 5.0685
lwt	Mother's weight at last menstrual period	Min: 80, Median: 121, Mean: 129.9, Max: 250, SD: 30.7307

	(pounds)	
race	Mother’s race	White = 95, Black = 26, Other = 66
smoke	Smoking status during pregnancy	Yes = 73, No = 114
ptl	Number of previous premature labours	Min: 0, Median: 0, Mean: 0.1925, SD: 0.4922
ht	History of hypertension on the birth	Yes = 12, No = 175
ui	Presence of uterine irritability	Yes = 27, No = 160
ftv	Number of physician visits during the first trimester	Min: 0, Median: 0, Mean: 0.7968, SD: 1.0633
bwt	Birth weight (Grams)	Min: 1021, Median: 2977, Mean: 2946, Max: 4593, SD: 698.6465

The summary tables show that birth weight (bwt) is highly related to LBW (low) since the status was classified using this variable. Therefore, variable bwt will be excluded from the training data to avoid perfect separation, a phenomenon when a variable perfectly predict the outcome, leading to extreme fitted probability, which will causes the model estimates to be unstable and unreliable.

The data distribution is also highly unbalanced in all categorical and some numerical variables, suggesting low significant level in those variables to the model. Although the unbalance can create limitation in the analysis, it is acceptable as the number of observation is limit and the variables that have high unstable are also uncommon status, such as ht and ui variables.

After fitting all five models, model 3 diagnostic plots indicate that there was perfect separation in the interaction terms, excluding this model from the comparison. With AIC comparison, model 4 has the best AIC value (216.44), followed by model 2 (AIC = 217.15). The Chi-square test and Residual Deviance Test of all four models shows model 4 do fit better compare to all other models. However, this improvement is insignificant in Chi-square test. Pearson’s test shows all three models fit the data reasonably well, with significant level at 5%.

Using information from the comparison, model 4 is the optimal model to answer the target question. After remove insignificance variables from the model using Wald’s test, this is the final equation of model 4 (round up to 6 decimals):

$$\log\left(\frac{\mathrm{P(low=1)}}{1-\mathrm{P(low=1)}}\right) = -0.13444 + 0.93032 * \mathrm{ptl} - 0.01541 * \mathrm{lwt} + 1.85538 * \mathrm{ht1} + 1.22197 * \mathrm{race2} + 0.87373 * \mathrm{smoke1} + 1.11795 * \mathrm{ui1}$$

IV. Discussion

Based on the final model equation, the significant variables are ptl, lwt, ht1, race=2, smoke=1, ui =1. Their effects on the status of the child with low birth weight are as follows:

- Number of previous premature labors (ptl): Each additional time of ptl increases the likelihood of having a low birth weight child by 2.54 times

- Mother’s weight at last menstrual period in pounds(lwt) : An increase in the mother's weight decreases the likelihood of low birth weight by 1.53% for every increase of 1 pound.

- History of Hypertension on the birth (ht1): Mothers with a history of hypertension are more likely to have low birth weight children by a factor of 6.39 times and is the strongest predictors of low birth weight in this model.

- Race is black not white (race2): Black mothers are 3.39 times more likely to have low birth weight children compared to White mothers.

- Smoking During Pregnancy (smoke1): Mothers who smoked during pregnancy have 2.40 times higher odds of having a low birth weight child compared to non-smokers.

- Uterine Irritability (ui1): The presence of uterine irritability increase in the likelihood of low birth weight, with the odds increasing by 3.06 times.

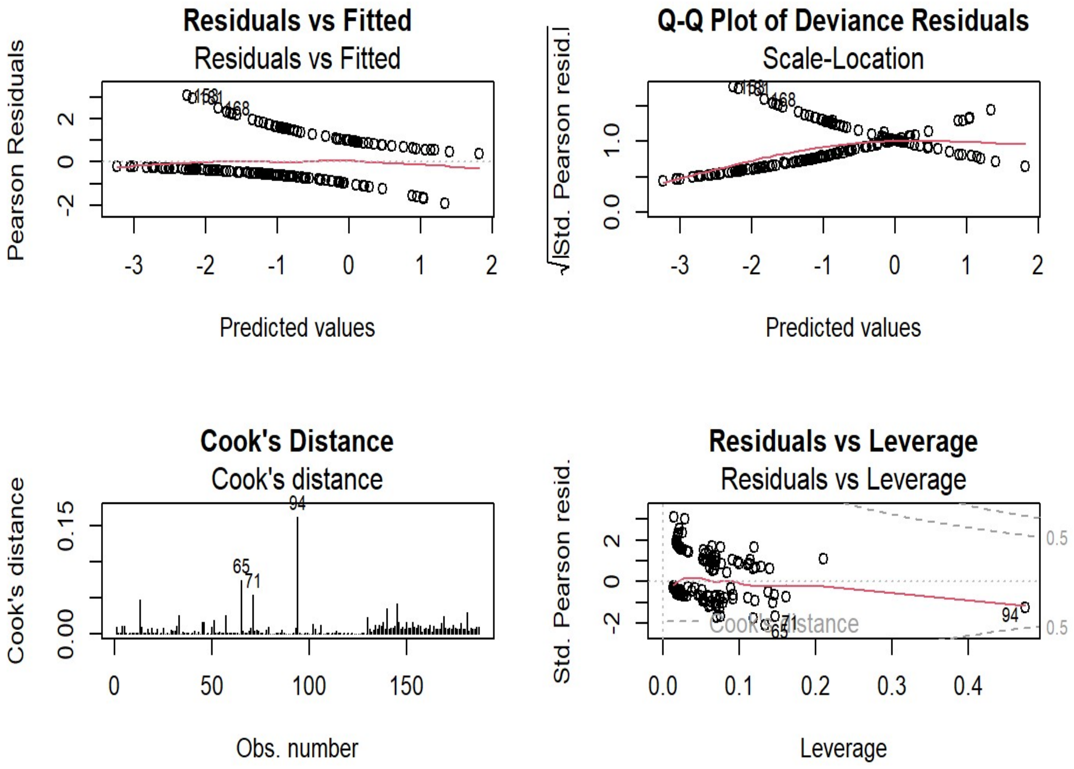

Figure.1: Model 3

Model 3 shows no firm evidence of a systematic pattern in Pearson residuals vs fitter plots. Q-Q Plot of Deviance Residuals shows the model fit well. However, the tails have several outliers, meaning the model might not perfectly capture the extreme values. Three points might be potential, influential outliers; however, after investigating them, there are rare cases and removing them makes some variables lose significance due to the reduction of rare cases. When investigating the outliners, the outlier cases are reasonable and provide valuable information that extends the model’s ability to generalize to a broader range of observations.Therefore,these outliers should be retained.

References

[1] T. Joshi, N. Noor, M. Kural, D. Pandit, and A. Patil, “Study of maternal determinants influencing birth weight of newborn,” *Archives of Medicine and Health Sciences*, vol. 3, no. 2, p. 239, Jan. 2015, doi: 10.4103/2321-4848.171912.

[2] J. F. Monsreal, M. Del Ruby Tun Cobos, J. R. H. Gómez, and L. E. Del Socorro Serralta Peraza, “Risk factors for low birth weight according to the multiple logistic regression model. A retrospective cohort study in José María Morelos municipality, Quintana Roo, Mexico,” *Medwave*, vol. 18, no. 01, p. e7143, Jan. 2018, doi: 10.5867/medwave.2018.01.7143.

[3] E. M. Elshibly and G. Schmalisch, “The effect of maternal anthropometric characteristics and social factors on gestational age and birth weight in Sudanese newborn infants,” *BMC Public Health*, vol. 8, no. 1, Jul. 2008, doi: 10.1186/1471-2458-8-244.

[4] J. Krasevec *et al.*, “Study protocol for UNICEF and WHO estimates of global, regional, and national low birthweight prevalence for 2000 to 2020,” *Gates Open Research*, vol. 6, p. 80, Jul. 2022, doi: 10.12688/gatesopenres.13666.1.

[5] M. Kirişci, “Comparison of artificial neural network and logistic regression model for factors affecting birth weight,” *SN Applied Sciences*, vol. 1, no. 4, Mar. 2019, doi: 10.1007/s42452-019-0391-x.

[6] G. Verropoulou and C. Tsimbos, "Modelling the effects of maternal socio-demographic characteristics on the preterm and term birth weight distributions in Greece using quantile regression," Journal of Biosocial Science, vol. 45, no. 1, pp. 107-119, 2013. doi: 10.1017/s0021932012000430.