2021

PH1700 Final Project Report

STATSTICAL ANALYSIS TO IDENTIFY RISK FACTORS FOR INFANT LOW BRITHWEIGHT

XU ZUO

Final Project Report

Xu Zuo, M.S.

Introduction

Birthweight is an important metric for evaluations of infant health (MacDorman & Mathews, 2009). Low birthweight infants are more likely to have multiple diseases and less chance of survival (Bakketeig et al., 2006). Therefore, the identification of potential risk factors that leads to low birthweight of infants is essential for the improvement of infant health.

Massive efforts have been made to investigate the causes of low birthweight. Reichman and Teitler (2006) found that there is a significant association between maternal age and low birthweight of infants born in the US urban areas. Levy and fellow researchers (2005) concluded that maternal anemia has a negative effect on infant birthweight. Delpisheh, Brabin, and Brabin (2006) found that mothers with a history of smoking are more likely to have adverse effects on birth outcomes including low birthweight.

Diverse statistical methods have been used in studies that focus on the analysis of low birthweight infants. Reichman and Teitler (2006) used a multiple logistic regression model to determine the relationship between maternal age and low birthweight. Levy and fellow researchers (2005) used multivariable analysis to determine the association between maternal anemia and low birthweight. Then they used logistic regression to control other confounders such as race and maternal age to quantitatively describe the association.

The objective of this research project is to determine potential risk factors for low birthweight from a series of variables (maternal age, maternal weight, Race, whether mother has smoked during pregnancy, history of hypertension, history of uterine irritability.).

Methods

The dataset used in the study included maternal demographics and medical conditions, as well as the birthweight of infants born at the Baystate Medical Center in Springfield, MA. There are 1,000 observations in total in the dataset. STATA was used for the statistical analysis presented in the study.

The first step of this analysis is to generate descriptive statistics for all independent variables. For continuous variables including Mother's Age and Mother's Weight, I calculated the means and standard deviations for the high birthweight and low birthweight groups. I then evaluate the normality using QQ plots, histograms, and Shapiro-Wilk tests. The hypothesis of the two-sample t test is: H_0: $\mu_{high} = \mu_{low}$, H_1: $\mu_{high} \neq \mu_{low}$. Mother's Weight is also a continuous variable that conforms to the normal distribution. Therefore, I used the same type of hypothesis test to assess the association. Race, Smoking status, History of Preterm Labor, History of Hypertension, History of Uterine Irritability are all categorical variables. I calculated the percentages and counts for each group. The hypothesis of the significant test is: H_0: $p_{high} = p_{low}$, H_1: $p_{high} \neq p_{low}$. I generated the 2x3 contingency table for Race and 2x2 contingency tables for other categorical variables. All expected values of contingency tables are over 5. Therefore, we use chi-square tests to test the hypothesis.

In order to fit the data into an initial multivariate model, I first checked the distributions of dependent variables versus each independent variable to identify skews. Residual analysis for the multivariate regression model was performed after I fitted the initial model. The linearity was determined with QQ plots of raw residuals and standardized residuals. Potential outliers were identified through residual versus fitted plots and residual versus predictor plots. Based on the results of residual analysis, I then finalized the regression model.

Results

Table 1 shows a summary of descriptive statistics for all independent variables. It is important to note that the percentages for all categorical variables are relative percentages. From the descriptive statistics, we can discover that the means of Mother's age and weight are higher for the normal birthweight infants. The normal birthweight infants take up the majority for all races in this study. More mothers of normal birthweight infants have a history of smoking, Preterm Labor, and Hypertension. More mothers of low birthweight infants have a history of uterine irritability. In this project, we assume a significant level of 0.05. The results of hypothesis tests imply that there are significant differences in terms of Mother's Age, Mother's Weight, Race, whether mother has smoked during pregnancy, History of Hypertension, History of Uterine Irritability. For the variable History of Preterm Labor, we cannot reject the null hypothesis. Therefore, we can only conclude that there is no association between the History of Preterm Labor and low versus normal birthweight infants.

Table 1. Descriptive Statistics

Variable	**High Birthweight**	**Low Birthweight**	**P-value**
Mother's Age (in years, M (SD))	22.77 (5.88)	21.86 (4.95)	0.0221
Mother's Weight (in lbs., M (SD))	153.21 (32.80)	144.31 (28.01)	0.0001
Race			0.001
Caucasian (% (n))	76% (405)	24% (126)	
African American (% (n))	66% (82)	34% (43)	
Other (% (n))	66% (226)	34% (118)	
Smoked During Pregnancy (% (n))	62% (242)	28% (149)	0.000
History of Preterm Labor (% (n))	65% (95)	35% (52)	0.053
History of Hypertension (% (n))	59% (34)	41% (24)	0.028
History of Uterine Irritability (%(n))	49% (67)	51% (70)	0.000

Given the STATA output presented in Table 2, we can write the initial multivariate model as follows:

$$bwt = 2668.05 - 1.29 * Mother'sAge + 4.05 * Mother's\ Weight - 237.17$$
$$* Race\ (race = 2) - 304.77 * History\ of\ Smoke\ (smoke = 1) - 79.82$$
$$* History\ of\ Preterm\ Labor\ (ptl = 1) - 351.04$$
$$* History\ of\ Hypertension\ (ht = 1) - 431.30$$
$$* History\ of\ Uterine\ Irritability\ (ui = 1)$$

The t statistic of the regression model is 21.97 and the p-value is 0.0000. Since the p-value is less than the significant level of 0.05, we can reject the null hypothesis and conclude that the regression model is significant.

Table 2. STATA output of the initial multivariate regression model.

```
. regress bwt age lwt i.race i.smoke i.ptl i.ht i.ui

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(8, 991)       =     21.97
       Model |  88605555.1         8  11075694.4   Prob > F        =    0.0000
    Residual |   499563103       991  504100.003   R-squared       =    0.1506
-------------+----------------------------------   Adj R-squared   =    0.1438
       Total |   588168658       999  588757.415   Root MSE        =       710

------------------------------------------------------------------------------
         bwt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -1.285827   4.150464    -0.31   0.757    -9.430534    6.858881
         lwt |   4.051734   .7550875     5.37   0.000      2.56998    5.533488
             |
        race |
          2  |  -237.1685   71.57181    -3.31   0.001    -377.6183   -96.71883
          3  |  -235.8366   50.82055    -4.64   0.000    -335.5648   -136.1083
             |
     1.smoke |  -304.7655   47.29432    -6.44   0.000     -397.574    -211.957
       1.ptl |  -79.81771   64.24414    -1.24   0.214    -205.8879    46.25247
        1.ht |  -351.0429   96.82135    -3.63   0.000    -541.0413   -161.0444
        1.ui |  -431.3039   66.01784    -6.53   0.000    -560.8548   -301.7531
       _cons |   2668.053   141.1664    18.90   0.000     2391.033    2945.072
------------------------------------------------------------------------------
```

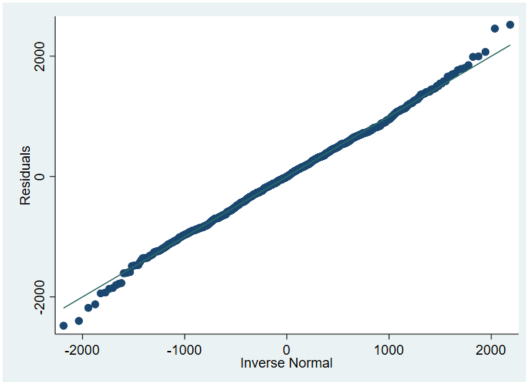

Figure 1. QQ plot of raw residuals.

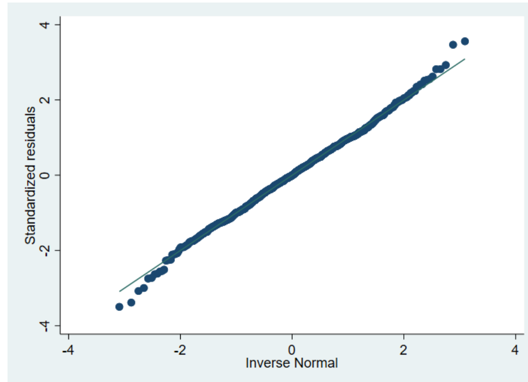

Figure 2. QQ plot of standardized residuals.

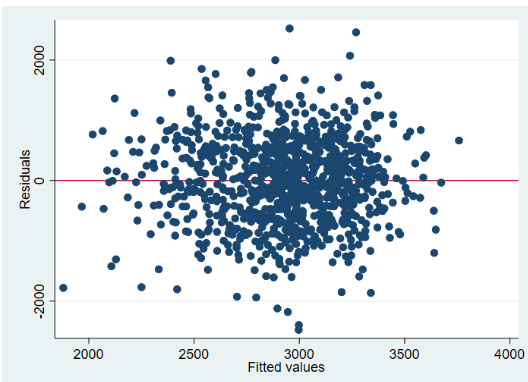

Figure 3. Residual versus fitted plot.

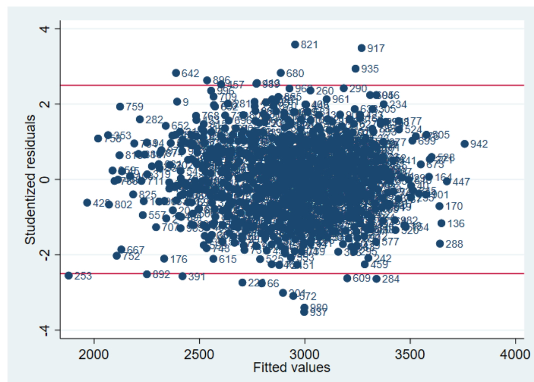

Figure 4. Studentized residual versus fitted plot.

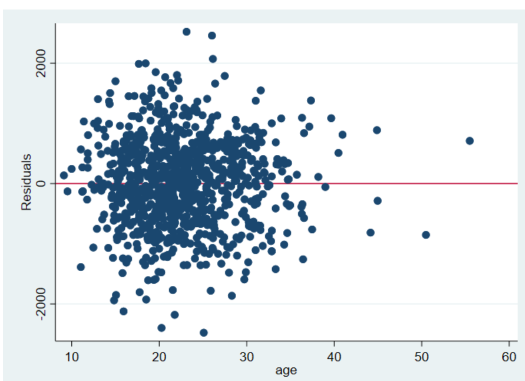

Figure 5. Residual versus Mother's Age

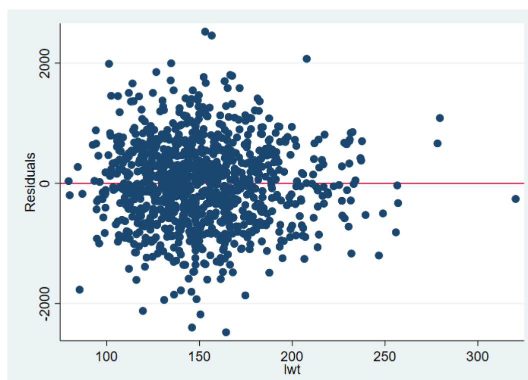

Figure 6. Residual versus Mother's Weight.

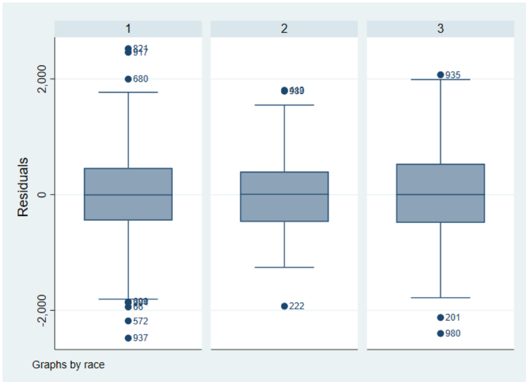

Figure 7. Residual versus Race.

Figure 8. Residual versus History of Smoke.

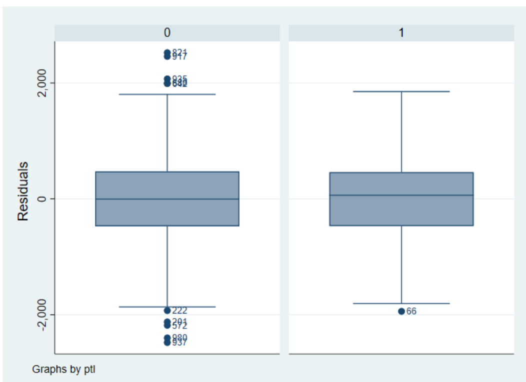

Figure 9. Residual versus History of Preterm Labor.

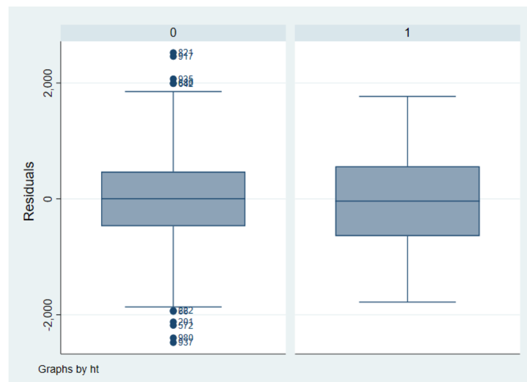

Figure 10. Residual versus History of Hypertension.

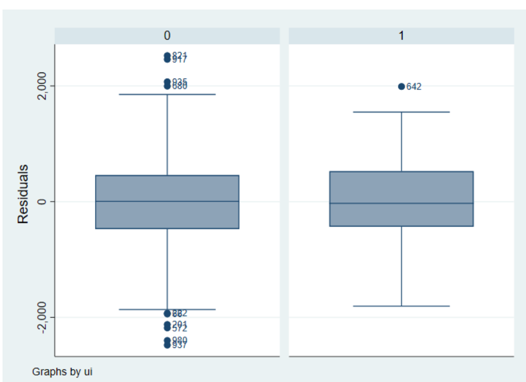

Figure 11. Residual versus History of Uterine Irritability.

The plots generated in the residual analysis are shown in Figure 1 to Figure 11. From both raw residual and standardized residual, we can see little deviation. Given a significant level of 0.05, we can define the outliners as the observations whose absolute value exceeds 2.5 on the studentized residual plot. From the studentized residual plot and residual versus predictor plots, we can discover that there are less than ten outliers. The IDs of those outliners are displayed in plots. Given a population of 1,000, dropping those points from the dataset will not change the regression model significantly.

From t statistic, p-value, and the residual analysis we can conclude that the initial multivariate model describes the relationship between infant birthweight and independent variables. Therefore, it is reasonable to use the initial multivariate model as the final model. The R-squared value is 0.1506, which indicates that 15.06% of the variability of the dependent variable can be explained by the multivariate model.

We can also analyze the coefficient, t statistic, and p-value of each independent variable individually. The t statistic for maternal age is -0.31 and the p-value equals 0.757, therefore maternal age is not a significant risk factor of low birthweight. The t statistic for maternal weight is 5.37 and the p-value equals 0.000. This implies that a unit increase in maternal weight will increase 4.05 grams in infant birthweight. The t statistic for Race (when the race is 2) is -3.31 and the p-value equals to 0.001. This indicates that American Africans are expected to have infants whose birthweight is 237.17 grams less than Caucasians. Mothers who smoked during pregnancy will lead to a reduction of 304.77 grams in infant birthweight. The history of Preterm Labor is not a significant risk factor for low birthweight. The history of hypertension of mothers will lead to a reduction of 351.04 grams in infant birthweight. The history of hypertension of mothers will lead to a reduction of 431.3 grams in infant birthweight.

Discussion

In this study, I investigated potential risk factors that lead to low birthweight of infants using the dataset from the Baystate Medical Center. I first summarized the dataset with descriptive statistics and used different hypothesis tests to determine the association between independent variables and high versus low birthweight groups. Then I fitted the data with a multivariate model. The results show that there is a linear relationship between and all the dependent variables. In addition, we found that maternal age and History of Preterm Labor were not significant risk factors alone. The residual analysis showed that there was no violation of independence and we can use it for further statistical analysis.

References

Bakketeig, L. S., Jacobsen, G., Skjaerven, R., Carneiro, I. G., & Knudsen, L. B. (2006). Low birthweight and mortality: the tendency to repeat low birthweight and its association with early neonatal and infant morbidity and mortality. Paediatric and perinatal epidemiology, 20(6), 507–511. https://doi.org/10.1111/j.1365-3016.2006.00755.x

Delpisheh, A., Brabin, L., & Brabin, B. J. (2006). Pregnancy, smoking and birth outcomes. Women's health (London, England), 2(3), 389–403. https://doi.org/10.2217/17455057.2.3.389

Levy, A., Fraser, D., Katz, M., Mazor, M., & Sheiner, E. (2005). Maternal anemia during pregnancy is an independent risk factor for low birthweight and preterm delivery. European journal of obstetrics, gynecology, and reproductive biology, 122(2), 182–186. https://doi.org/10.1016/j.ejogrb.2005.02.015

MacDorman, M. F., & Mathews, T. J. (2009). The challenge of infant mortality: have we reached a plateau?. Public health reports (Washington, D.C. : 1974), 124(5), 670–681. https://doi.org/10.1177/003335490912400509

Reichman, N. E., & Teitler, J. O. (2006). Paternal age as a risk factor for low birthweight.

American journal of public health, 96(5), 862–866.

https://doi.org/10.2105/AJPH.2005.066324