Major Risk Factors of Low Birth Weight Babies

Taehoon Ha

Abstract Low birth weight can be a major factor related to the health and survival of infants and finding risk factors of low birth weight can be helpful in preventing low birth weight or providing preventative treatments to subpopulations in great risk. This project aims to determine the major risk factors of low birth weight based on a sample size of 189 with numbers of potential risk factors and the identifier of low birth weight using generalized linear models with model selection based on deviance and p-values. Results show that low birth weight is associated with the mother's premature labor history, mother's race, mother's smoking status, mother's weight of last menstrual period, and mother's history of hypertension. Therefore, we should be more aware of mothers who are black and had a low weight before pregnancy and who smokes and have history of premature labor and hypertension.

1. Introduction

Birth weight is a significant indicator used to determine the future health of an infant. Babies with less than 2500 grams weight when born are considered to be low birth weight babies. According to preliminary studies, low birth weight of a baby is greatly associated with its health not only during their infancy but in overall life, including early death or adult diseases such as diabetes, cardiovascular, and renal disease. The maternal health is crucial since the weight of a baby at birth is dependent on the mother's health status. Therefore, this study aims to identify the main risk factors of low birth weight among various factors of the maternal health.

The dataset includes the low birth weight indicator (low), age of the mother (age), mother's weight at the last menstrual period (lwt), mother's race (race), mother's smoking status (smk), mother's history of premature labor (ptl), mother's history of hypertension (ht), mother's presence of uterine irritability (ui), and number of physician visits during the first trimester(ftv). The project aims to develop a model that could help identify the major risk factors that are highly associated with low birth weight.

2. Methods

Our study was based on logistic regression models with on the outcome *low* to identify the main risk factors among eight potential variables. Several steps were performed in order to find the model that best fits our data, in other words, to find the combination of factors or their interactions that best explains our outcome, low birth weight.

Descriptive analysis was performed to take a glance of the dataset. In this step, we generate the summary values of each variable by each group: low birth weight and non-low birth weight. For continuous variables, *age* and *lwt*, we calculate the mean and standard deviance and for each variabl. For categorical variables, *race, smk, ptl, ht, ui, and ftv,* we generate the counts and percentage among each outcome group. We also generate simple statistical test results to see the independence of each variable and outcome using t-test for continuous variables and chi-square or fisher's exact test (when one of the expected cell counts is less than 5). Normality check on continuous variables with histogram will also be performed and transformed if non-normal since normality can affect the quality of our final model. Based on the result of the descriptive analysis, we also explore combining different levels of ordinal categorical variables to see if there is a reduced p-value in the independency test results. Modified variables are used in our next steps.

The second step is the univariate analysis. We fit univariate logistic models for each variable on the outcome *low*. This is used as a pre-filtering process to find which variables we should include in our multivariate analysis. We first filter out the variables that have p-values greater than the significance level which we define as 0.1 at this step since it infers low association between the variable and outcome. Here, we obtain the raw odds ratio of each variable.

In the next step, we perform the multivariate analysis. In this project, we used the forward selection process to select the variables that we want to include in our final model. First, we choose one variable, excluding those filtered out from the step before, with the lowest deviance in their univariate model. Then, we add one other variable from our filtered list and choose the model that has the model with the lowest deviance. Lowest deviance indicates that the model has a good fit as the saturated model (full model) has the deviance of 0, as they consider each observation as a covariate, hence has the best explanation of the data. However, the deviance gets smaller as the number of covariates increases. Hence, we set a threshold of the significance level of 0.1 as well and decide not to add the variable with a p-value greater than 0.1 in the multivariate model even if they have reduced deviance.

From these steps, we obtain the model with main effects with the adjusted odds ratios of each variable. Next, we explore if there are any interactive effects between the selected variables. The interaction model is selected as our final model if the deviance gets lower and the p-value of the interaction term is lower than the significance level.

After the final model is selected, we test the performance of our model using the Deviance goodness-of-fit test. In addition, we fit the model again with the data excluding the influential points to see whether there is any change of the fit and odds ratios. The influential points are determined using the standardized deviance residuals, where influential points lie outside of the interval [-2, 2].

3. Results

3.1 Descriptive Analysis

The data has in total 189 subjects including 130 subjects with non-low birth weight babies and 59 subjects with low birth weight babies. The results can be seen in Table 1.

From Table 1, we notice that *age, race,* and *ftv* does not have significant difference between the two groups ($p>0.05$). In the univariate analysis, this factor will be taken into account when filtering the variables to be included in the model.

Table 1 : Descriptive Analysis Results

Variable	Non-Low birth weight (N=130)	Low Birth Weight (N=59)	Test statistic, p-value
Age	23.66 (5.58)	22.31 (4.51)	t = 1.77, p = 0.0783 *
Weight at Last Menstrual Period (Lwt)	133.3 (31.72)	122.1(26.56)	t = 2.52, p = 0.0131 *
Race			$\chi^2 = 5.0048$, p = 0.0819 **
White	73 (56.15%)	23 (38.98%)	
Black	15 (11.54%)	11 (18.64%)	
Others	42 (32.31%)	25 (42.37%)	
Smoking Status (Smk)			$\chi^2 = 4.9237$, p = 0.0265 **
No	86 (66.15%)	29 (49.15%)	
Yes	44 (33.85%)	30 (50.85%)	
Premature Labor History (Ptl)			p = 0.0003 ***
none	118 (90.77%)	41 (69.49%)	
1	8 (6.15%)	16 (27.12%)	
2	3 (2.31%)	2 (3.39%)	
3	1 (0.77%)	0 (0%)	
Hypertension History (Ht)			$\chi^2 = 4.3880$, p = 0.0362 **
No	125 (96.15%)	52 (88.14%)	
Yes	5 (3.85%)	7 (13.46%)	
Uterine Irritability History (UI)			$\chi^2 = 5.4008$, p = 0.0201 **
No	116 (89.23%)	45 (76.27%)	
Yes	14 (10.77%)	14 (23.73%)	
Number of Physician visits during first trimester (Ftv)			p = 0.2866 ***
None	64 (49.23%)	36 (61.02%)	
1	36 (27.69%)	11 (18.64%)	
2	23 (17.69%)	7 (11.86%)	
3	3 (2.31%)	4 (6.78%)	
4	3 (2.31%)	1 (1.69%)	
6	1 (0.77%)	0 (0%)	

Mean (standard deviation) were reported for age and weight at last menstrual period
Count (Column proportion) were reported for all other variables
* Satterthwaite t-test results ** Chi-square test results *** Fisher's exact test results

For variables *ptl* and *ftv,* the counts of the cells are sparse with low counts in upper levels of the ordinal variables. Hence, we try to see if combining the levels and making the variables binary, i.e. premature labor history yes or no, and physician visits during first trimester yes or no. However, regrouping did not change the results of the independence test, as the p-value were still over 0.05.

For continuous variables, *age* and *lwt*, histograms were plotted to see the distribution. In figure 1, we can see that the two variables are right-skewed and non-normal. For future use, we performed log-transformation to see whether it improves the skewedness. Although the transformation reduced the skewedness, it still does not seem to be normal and since logistic regression does not assume normality of the data, we decide to keep the original form of these variables to minimize data manipulation.

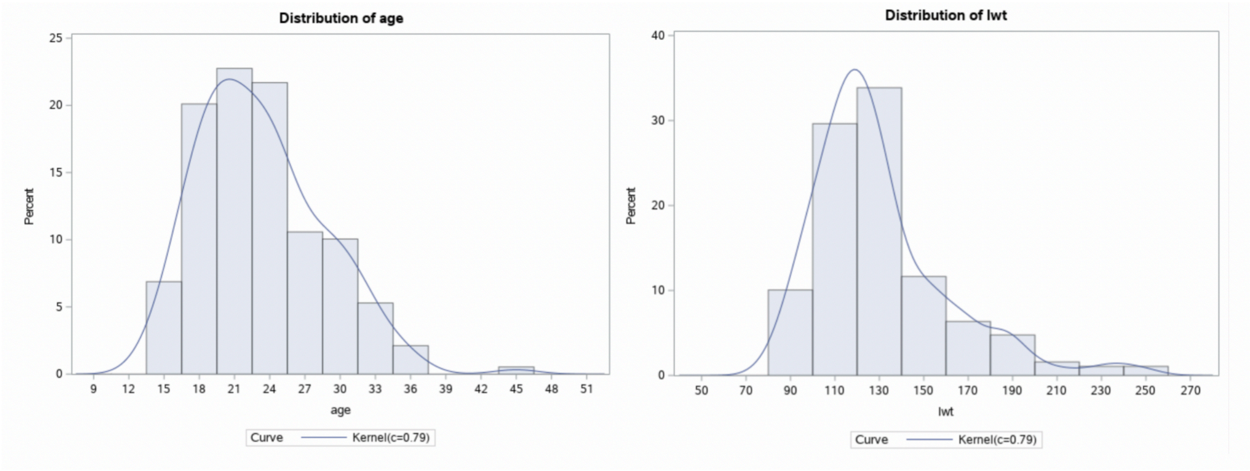

Figure 1 Histogram of *age* and *lwt*

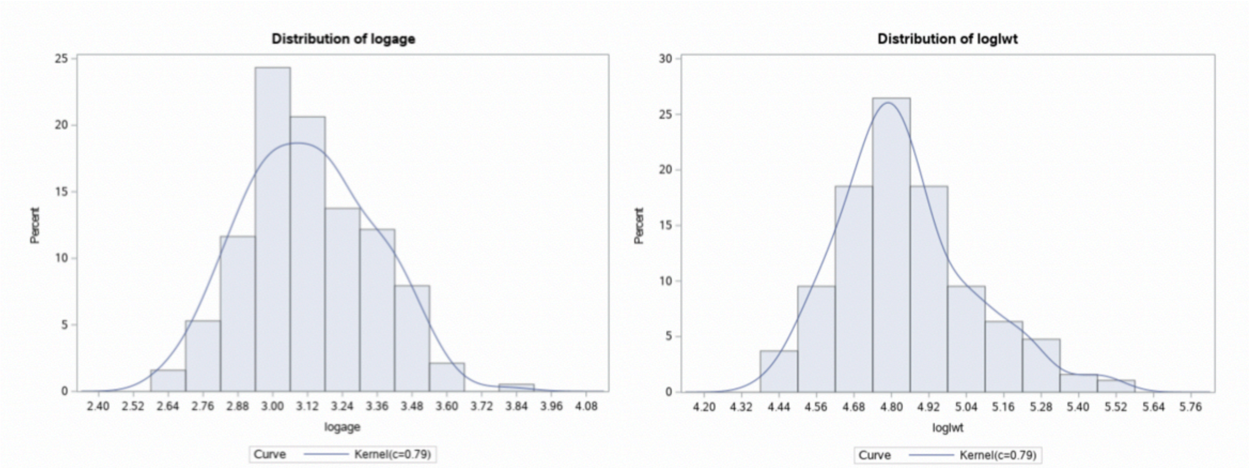

Figure 2 Histogram of *log(age)* and *log(lwt)*

3.2 Univariate Analysis

The univariate logistic models are fit for each variable. It can be seen that the *age* and *ftv* have insignificant association with the outcome and was thus decided to exclude these two variables for our next step in multivariate analysis. Hence, the final list of variables that will be included in the multivariate analysis is *lwt, race, smk, ptl, ht,* and *ui*.

Table 2 Univariate Analyses with Wald Thtest

Variable	Effect	Raw Odds Ratio	95% Confidence Interval	p-value
Age	-0.051	0.950	(0.893, 1.011)	0.1047
Lwt	-0.014	0.986	(0.974, 0.998)	0.0227
Race				
Black vs White	0.845	2.327	(0.938, 5.772)	0.0683
Others vs White	0.636	1.889	(0.955, 3.735)	0.0675
Smk				
Yes vs No	0.704	2.022	(1.081, 3.783)	0.0278
Ptl				
>=1 vs none	1.463	4.317	(1.916, 9.726)	0.0004
Ht				
Yes vs No	1.214	3.365	(1.021, 11.088)	0.0461
UI				
Yes vs No	0.947	2.578	(1.139, 5.834)	0.0231
Ftv				
>=1 vs none	-0.479	0.620	(0.331, 1.159)	0.1339

- p-values are from Wald test results
- Confidence intervals are the intervals of oddsratios

3.3 Multivariate Analysis

According to the results from the univariate analysis, the univariate model with an intercept and *ptl* also has the lowest deviance ($\chi^2 \sim 221.8978$), hence it would be selected as our first step of our multivariate analysis. Based on the model with only *ptl*, we find out that by adding *race* into the model the deviance goes down to $\chi^2 \sim 217.0226$ and the p-value of *race* in the model is smaller than the significance level 0.1 (p=0.0911). From the model with *ptl* and *race, smk* is added with the lowest deviance of $\chi^2 \sim 210.9023$ and a significant p-values(p=0.0149). From the model with three covariates, we add *lwt* and *ht* in each turn with deviance lowered each turn, $\chi^2 \sim 207.0394$ and $\chi^2 \sim 200.4823$ responsively, with p-values lower than the significance level (p=0.0624, p=0.0126). However, although adding the last remaining variable, *ui*, did improve the deviance down to 197.8516, the p-value was insignificant as it was over 0.1 (p=0.1018).

Hence, our main effect model includes *ptl, race, smk, lwt,* and *ht*. Then, we explored the deviance and p-values of the interaction models by adding all possible interaction terms, in total 10, to see whether adding interactions can improve our model. Results show that adding *race**smk* interaction into the model brings down the deviance to 198.9011 but the p-value is noticeably insignificant (p=0.2157). Therefore, we decide not to include any interaction terms in our model.

Our final model is therefore:

$$logit(\Pr(Low = 1)) = 0.0946 + 1.2314ptl + 1.2637race2 + 0.8642race3 + 0.8761smk - 0.0167lwt + 1.7674ht$$

The adjusted odds ratios and the Wald test results of each covariates can be found in Table 3. It can be seen that mothers with hypertension history have 5.856 odds of having low birthweight babies compared to those who never had hypertension. Also, Black and other race mothers have greater odds of having newborns with low weight compared to White mothers. Mothers with smoking habits and history of premature labor also have greater odds than mothers who do not. It can also be seen that as the mother's weight at the last menstrual period is higher, the chances of having a low birthweight baby gets lower.

Table 3. Final Model Results

Variable	Effect	Adjusted Odds Ratio	95% Confidence Interval	p-value
Lwt	-0.017	0.983	(0.970, 0.997)	0.0161
Race				
Black vs White	1.264	3.539	(1.254, 9.986)	0.0170
Others vs White	0.864	2.373	(1.011, 5.567)	0.0470
Smk				
Yes vs No	0.876	2.402	(1.095, 5.267)	0.0288
Ptl				
>1 vs none	1.231	3.426	(1.429, 8.216)	0.0058
Ht				
Yes vs No	1.767	5.856	(1.461, 23.474)	0.0126

- p-values are from Wald test results
- Confidence intervals are the intervals of odds ratios

3.4 Model Checking

After fitting the final model, we perform a goodness-of-fit test using Deviance to see whether our final model fits our data well. The results of Deviance goodness-of-fit test shows deviance of 200.4823 with 181 d.f., and the p-value of 0.1654. This shows that our model is a good fit for our data. We then see if there is any change after removing the influential points in our dataset.

Using the standardized deviance residual, we found that 3 subjects have residual greater than 2: subject 127, 155, and 183. In figure 3, we can see that the standard residuals go over 2 in those three points. After removing these three subjects, we fitted the model again resulting in:

$$
\begin{aligned}
logit(\Pr(Low = 1)) \\
&= 0.3680 + 1.3303ptl + 1.4476race2 \; + \; 0.9753race3 + 0.8971smk \\
&- 0.0203lwt + 1.9630ht
\end{aligned}
$$

The goodness-of-fit improves as we remove the influential points, with deviance of 190.3398 and the p-value of 0.2669.

4. Discussion

Based on our analyses, we have determined that the major risk factors among mother's health status are the history of premature labor, race, smoking status, weight at last menstrual period, and history of hypertension. Based on this findings, pregnant women who are black and low-weighted, with premature labor, hypertension history and smoking habits should be more cautiously observed for preventative care in giving birth to low birth weight babies. Women with interest of conception should be recommended to quit smoking and gain more weight before pregnancy to reduce the chance of low birth weight of their babies.

However, there are some limitations to our study. First, the distribution of race in our data seems to be imbalanced. The proportion of race in both outcome groups are White 50.8%, Black 13.8%, and Others 35.4%. This can be a cause of a misleading result in our result as we have more samples in one group and might have caused bias. Hence, including race as a major risk factor might not be adequate and there should be further studies with bigger sample size with better balance among race groups.

Secondly, our study did not compare the deviance, or goodness-of-fit, between models but only used the numeric deviance as given. This might have resulted in having more variables than are needed since we would like to make our model as simple as possible because we do not want to describe the data but to "summarize" the data. Further analysis can be done using ANOVA test to compare the nested models to see whether the decrease in deviance is significant enough to choose models with more terms.

In addition, there are many unconsidered potential risk factors of low birthweight babies, e.g. mental illness, sleep deprivations, vitamin consumptions, and etc. Further studies should be done with a more high-dimensional dataset.

5. Reference

Kildea, S. V., Gao, Y., Rolfe, M., Boyle, J., Tracy, S., & Barclay, L. M. (2017). Risk factors for preterm, low birthweight and small for gestational age births among Aboriginal women from remote communities in Northern Australia. *Women and Birth,30*(5), 398-405. doi:10.1016/j.wombi.2017.03.003

Tubay, A. T., Mansalis, K. A., Simpson, M. J., Armitage, N. H., Briscoe, G., & Potts, V. (2018). The Effects of Group Prenatal Care on Infant Birthweight and Maternal Well-Being: A Randomized Controlled Trial. Military Medicine. doi:10.1093/milmed/usy361

Zhao, L., McCauley, K., & Sheeran, L. (2017). The interaction of pregnancy, substance use and mental illness on birthing outcomes in australia. *Midwifery, 54*, 81-88. doi:10.1016/j.midw.2017.08.007

Appendix

Table 1 Univariate Analysis Deviance

Model	Deviance
Low ~ intercept	234.6720
Low ~ Lwt	228.6907
Low ~ Race	229.6616
Low ~ Smk	229.8046
Low ~ Ptl	**221.8978**
Low ~ Ht	230.6499
Low ~ UI	229.5959

Table 2 Multivariate Analysis Deviance & Wald test p-values

Model	Deviance	p-value
Low ~ Ptl	221.8978	
Low ~ Ptl + Lwt	217.4968	0.0484
Low ~ Ptl + Race	**217.0226**	**0.0911**
Low ~ Ptl + Smk	219.3289	0.1077
Low ~ Ptl + Ht	217.6619	0.0399
Low ~ Ptl + UI	219.1204	0.0921
Low ~ Ptl + Race + Lwt	212.1795	0.0379
Low ~ Ptl + Race + Smk	**210.9023**	**0.0149**
Low ~ Ptl + Race + Ht	213.2263	0.0519
Low ~ Ptl + Race + UI	214.2788	0.0945
Low ~ Ptl + Race + Smk + Lwt	**207.0394**	**0.0624**
Low ~ Ptl + Race + Smk + Ht	207.1646	0.0545
Low ~ Ptl + Race + Smk + UI	208.3978	0.1103
Low ~ Ptl + Race + Smk + Lwt + Ht	**200.4823**	**0.0126**
Low ~ Ptl + Race + Smk + Lwt + UI	205.1621	0.1674
Low ~ Ptl + Race + Smk + Lwt + Ht + UI	197.8516	0.1018

Table 3 Interaction term variable selection

Model	Deviance	p-value
Low ~ Ptl + Race + Smk + Lwt + Ht	**200.4823**	
Low ~ Ptl + Race + Smk + Lwt + Ht + Ptl*Race	200.2966	0.6659
Low ~ Ptl + Race + Smk + Lwt + Ht + Ptl*Smk	200.1090	0.5426
Low ~ Ptl + Race + Smk + Lwt + Ht + Ptl*Lwt	200.4145	0.7957
Low ~ Ptl + Race + Smk + Lwt + Ht + Ptl*Ht	199.7534	0.9896
Low ~ Ptl + Race + Smk + Lwt + Ht + Race*Smk	198.9011	0.2157
Low ~ Ptl + Race + Smk + Lwt + Ht + Race*Lwt	199.9699	0.4782
Low ~ Ptl + Race + Smk + Lwt + Ht + Race*Ht	200.4655	0.8974
Low ~ Ptl + Race + Smk + Lwt + Ht + Smk*Lwt	199.5918	0.3533
Low ~ Ptl + Race + Smk + Lwt + Ht + Smk*Ht	199.9840	0.4838
Low ~ Ptl + Race + Smk + Lwt + Ht + Lwt*Ht	200.4710	0.9161

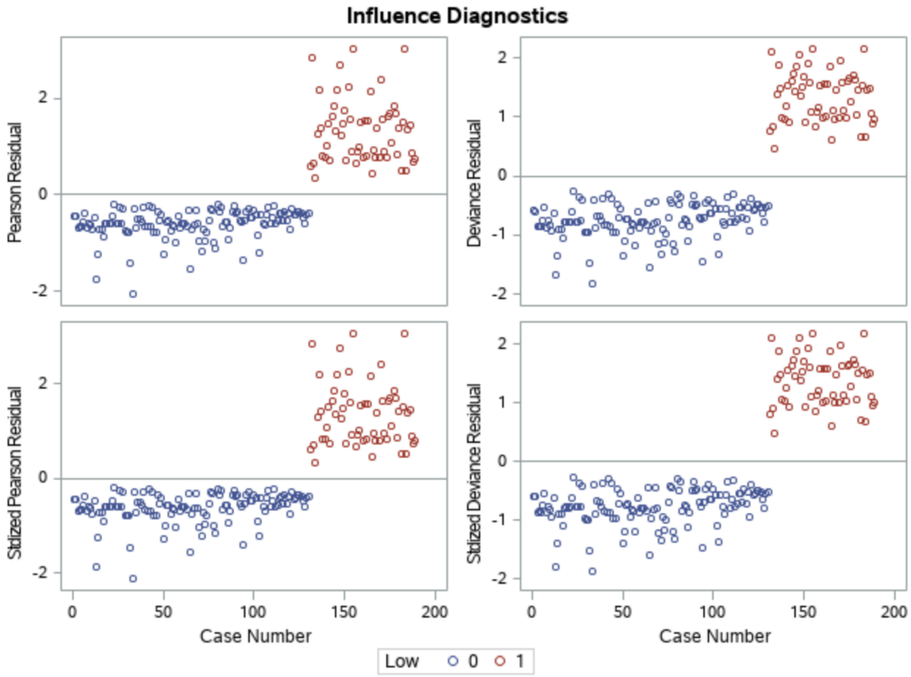

Figure 3 Diagnosis Plot for Influential Points