Investigating the Etiology of Low Infant Birth Weight: An Exploration of Risk Factors

Abstract

This study investigates the association between maternal factors - weight at the last menstrual period, smoking status, race, history of hypertension - and the risk of low infant birth weight, considering interactions with maternal race and smoking status. The analytical framework utilized mosaic displays, log-linear models, logistic regression, and statistical tests. The logistic regression model revealed that an increase in maternal weight reduces the likelihood of low birth weight, while black mothers and smokers faced higher risk, and mothers without hypertension showed significantly lower risk. Mosaic displays and log linear models validated the model's specification, and influence plots ensured that the model is not affected by outlier cases. The study illuminates crucial factors contributing to low birth weight in infants, emphasizing the importance of considering individual factors and their interactions in risk assessments. Potential interventions targeting these risk factors could minimize low birth weight prevalence and improve maternal and child health outcomes.

Future research could include more variables, such as a mother’s socioeconomic status, access to healthcare, dietary habits, marital status, and education level, to bolster the model's accuracy and generalizability. Further exploration of these risk factors and potential determinants of low birth weight is needed for comprehensive understanding and effective interventions.

1. Background and Significance

Low birth weight (LBW) is a major public health concern across the world since it is associated with an increased risk of newborn death and morbidity. Previous research in healthcare institutions around the world has looked into maternal and sociodemographic risk factors for LBW in newborns[1][3] with findings indicating that history of hypertension, timing and number of antenatal care visits, iron and calcium supplementation, maternal education, illnesses during pregnancy, and hypertension are all significant predictors of LBW.

Our study aims to understand why some babies are born with low birth weights, which can impact their health and development. We're focusing on factors such as the mother's weight before her last menstrual cycle[1], whether she smoked during pregnancy, if she has a history of hypertension, and her race[2]. We want to see how all these elements interact to influence the chance of a baby being born with low weight. Our goal is to identify the key risks for low birth weight and provide a foundation for more research and potential solutions in the future.

These study issues are important because they might lead to a better understanding of the variables that contribute to LBW and guide targeted efforts to avoid it. The findings might also have an impact on public health policies and initiatives focused at improving mother and child health outcomes. Overall, this initiative has the potential to give useful insights into the risk factors for LBW and to assist the development of effective prevention methods.

2. Methods

The birthwt dataset, collected in 1986 at Baystate Medical Center in Springfield, Massachusetts, consists of 189 rows and 10 columns. We will focus on predicting low birth weight using the other variables in the dataset. The variables are defined as follows:

- low: an indicator of birth weight less than 2.5 kg.
- age: mother's age in years.
- lwt: mother's weight in pounds at last menstrual period.
- race: mother's race (1 = white, 2 = black, 3 = other).
- smoke: smoking status during pregnancy.
- ptl: number of previous premature labors.
- ht: history of hypertension.
- ui: presence of uterine irritability.
- ftv: number of physician visits during the first trimester.
- bwt: birth weight in grams.

To analyze the data, we will use data manipulation and a combination of exploratory and graphical methods, as well as statistical modeling and diagnostics. Exploratory methods will be used to identify trends and patterns in the data, while graphical methods will be used to visualize the relationships between the variables. We will then fit statistical models to the data to identify significant predictors of low birth weight, and assess the model's performance using various diagnostics. Overall, our aim is to identify the most important risk factors for low birth weight and develop an accurate prediction model that can be used to guide targeted efforts to prevent it.

3. Results

3.1 Data Manipulation

The original ‘birthwt’ dataset consisted only of integer data and it is preprocessed where several variables into categorical factors with levels and labels. The focus is not on ‘bwt’ in this report, hence excluding it from making any changes, analysis or interpretations. The

continuous variables(age, lwt) are converted into ordinal categories where the ‘age’ is split into four categories (<20, 20-30, 30-40, >40) and ‘lwt’ is split into (<120, 120-150, >150). The visualizations are easier to create and understand with ordinal data.

A copy of the original dataset is created,’birthwt_new’, where just the categorical variables (low, race, smoke, ht, ui,ptl, ftv)are identified and converted to factors, while defining their levels and labels. The continuous variables are not changed, and this dataset is used for logistic regression and its visualizations.

3.2 General Observations

3.2.1 Summary of report 2

Our study investigated a variety of maternal and pregnancy-related factors that might contribute to low birth weight in infants. An analysis revealed significant association between maternal smoking and infant birth weight, with a p-value of 0.0395 suggesting a substantial association. Infants born to mothers who smoked during pregnancy had higher odds of being low weight. The contingency coefficient and Cramer's V indicated a weak association between these variables. Conversely, we found no significant association between maternal race and low birth weight, with p-values exceeding the standard 0.05 threshold. Likewise, the mother's weight at the last period and low birth weight also showed no notable relationship, as evidenced by the p-values obtained from both likelihood ratio and Pearson's methods.

Further examination disclosed an interesting relationship between low birth weight and hypertension during pregnancy. Although the chi-squared test didn't show a significant association, the odds ratio suggested that the risk of low birth weight was considerably higher when hypertension was present during pregnancy. Similarly, there appeared to be a significant association between uterine irritability and low birth weight, with the odds of low weight being higher when uterine irritability was present. Both the contingency coefficient and Cramer's V suggested a moderate association between these factors.

An in-depth analysis also shed light on the association between the number of previous premature labors (ptl) and low birth weight. The chi-squared test provided strong evidence against the null hypothesis of independence between these two variables, suggesting that the number of previous premature labors could significantly influence the likelihood of low birth weight in infants. Interestingly, the odds of low birth weight appeared to decrease for mothers with more than one previous premature labor.

On the other hand, we found no significant association between the mother's age and low birth weight, with p-values from both the likelihood ratio and Pearson's methods exceeding the standard threshold of significance. Similarly, no substantial relationship was identified between low birth weight and the number of physician visits during the first trimester. These results underline the complexity of factors influencing infant birth weight and the need for further research in this domain.

3.2.2 Summary of report 3

In our previous report, we have worked on two research questions, each with a different set of predictors. The research questions are as follows -

- **RQ1** : What is the relationship between history of hypertension (ht) and risk of low infant birth weight (low) after controlling for the interaction between uterine irritability(ui) and the number of previous premature labors(ptl)?

- **RQ2** : How do the interactions among maternal race(race), smoking habits(smoke), mother's weight at the last menstrual period(lwt), and the number of previous premature labors(ptl) contribute to the risk of low infant birth weight(low)?

Table 3.1 is created for RQ1 and Table 3.2 for RQ2, which shows the analysis of multiple mosaic displays and log linear models for each research question. For many models, the p-value from the Mosaic plot suggests a relatively good fit. In general, a p-value greater than 0.05 is typically used as the threshold for a good fit. However, it's interesting to note that the loglm models returned NaN for all the combinations. This indicates a problem either with the data or the model specification for these combinations of variables.

S.No	Statistical Notation	P-value from Mosaic	P-value from loglm model
1	[AB][AC][AD]	0.636	NaN
2	[ABC][AD]	0.3287	NaN
3	[AB][ACD]	0.7604	NaN
4	[ABD][AC]	0.9966	NaN
5	[ABCD]	1	1

Table 3.1 *P-values from mosaic, log linear models for RQ1 where A = low. B = ptl, C = ht, D = ui*

S.No	Statistical Notation	P-value from Mosaic	P-value from loglm model
1	[AB][AC][AD][AE]	0.6471	NaN
2	[ABC][AD][AE]	0.6857	NaN
3	[ABD][AC][AE]	0.7693	NaN
4	[ABE][AC][AD]	0.6926	NaN
5	[AB][ACD][AE]	0.8257	NaN
6	[AB][ACE][AD]	0.9760	NaN
7	[AB][AC][ADE]	0.7965	NaN
8	[ABCD][AE]	0.86	NaN
9	[AB][ACDE]	0.9994	NaN
10	[AC][ABDE]	0.7988	NaN

11	[AD][ABCE]	0.9995	NaN
12	[ABCDE]	1	1

Table 3.2 *P-values from mosaic and log linear models for research for RQ2. Here A = low, B = lwt, C = smoke, D = ptl, E = race*

The model that includes all variables interacting with each other has a p-value of 1 in both the Mosaic plot and loglm model indicating a perfect fit to the data, often referred to as a saturated model. The issues in the log-linear model calculations suggest there might be problems with these models for these combinations of variables. Therefore, we chose to reassess our approach and identify a different collection of predictors to estimate the probability of low birthweight. The revised analysis is conducted in the following section.

3.3 Research Question

3.3.1 How do the factors of a mother's weight at last menstrual period(lwt), her smoking status(smoke) during the crucial period of gestation, history of hypertension (ht), and her racial identity(race), together determine the likelihood of low birth weight - a critical indicator of infant health?

In Figure 3.1, we see a Generalized Pairs plot that provides a visual snapshot of our dataset. Bar plots are used in diagonal panels to show how different variables spread out. Scatter plots shows how two variables relate to each other. Box plots illustrate how the variable 'lwt' spreads out across different categories. Mosaic plots depict how one variable behaves given the behavior of another variable.

Looking at Figure 3.1, we notice several trends. More instances of normal birth weight, most mothers in the age group of 20-30, and the majority of mothers weighing between 100 and 150 pounds. More mothers are 'white' or 'other' races than 'black'. Non-smoking mothers are more common, and most have no history of premature labor, hypertension, or uterine irritability. We also see that most mothers have had 1 or 2 visits to the physician before. Fewer outliers are found in 'age', as seen in the box plots in the 2nd column. In contrast, 'lwt' has more outliers as shown in the scatter plot and box plots in the 3rd column. In terms of associations, when low birth weight is 'Yes', there is a strong positive relationship with no previous premature labors (0), and a strong negative association with non-smokers. When 'low' is 'Yes', there seems to be a small positive relationship with 'black' race mothers, smoking mothers, those with a history of hypertension, and those without uterine irritability. The mosaic plots between 'low' and 'ftv' show a small part of the data, as represented by thin, line-like rectangles, indicating a weak association. The plot also shows a strong positive association between 'other' race and non-smokers, and 'white' race and smokers. On the other hand, a strong negative relationship appears between 'other' race and smokers, and a weaker negative association between 'white' race and non-smokers.

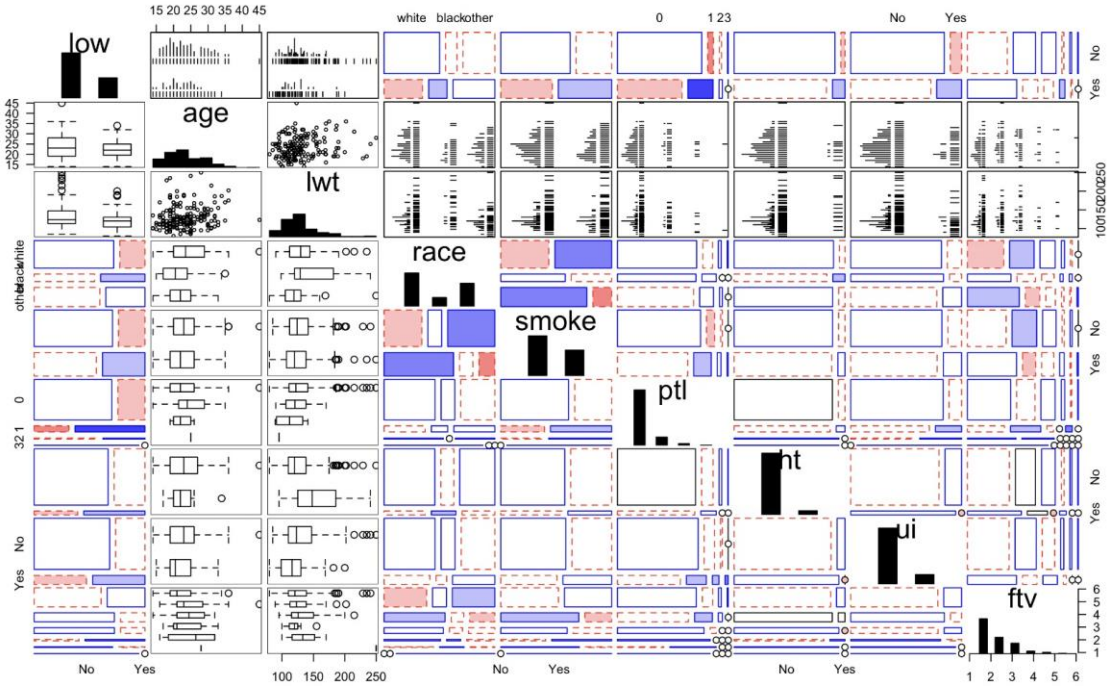

Figure 3.1 *Gpairs plot of all predictor variables*

To answer our research question, the predictors - *lwt, smoke, ht, race* are selected where the response variable is *low*. A double decker plot has been created to visualize the dependence of one categorical variable(low) with other categorical variables.

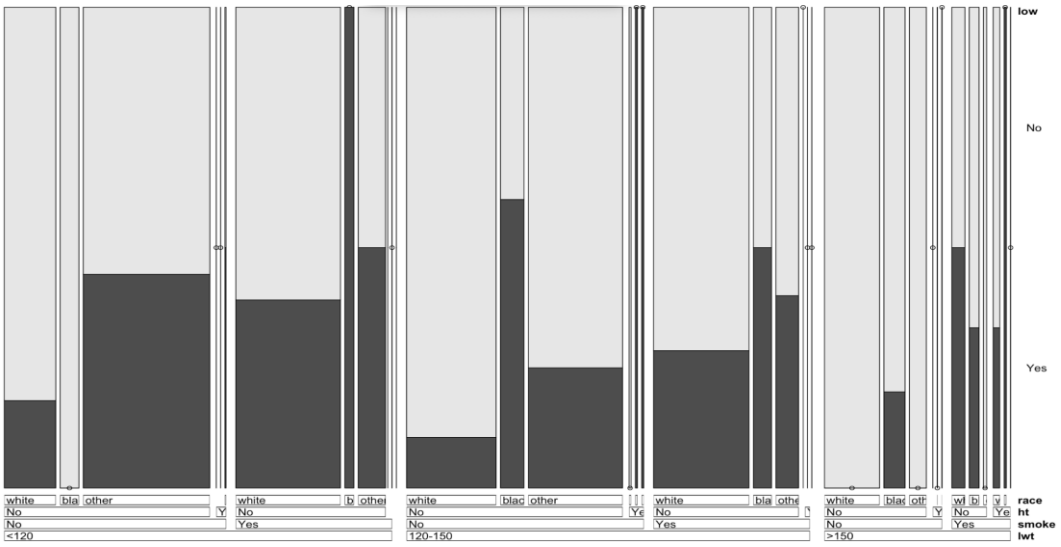

Figure 3.2 *A double decker plot of low(response) and predictors(race, ht, smoke, lwt)*

The double decker plot reveals that for mothers weighing less than 120 pounds, there is a higher number of low birth weight among mothers of races other than white. Interestingly, there are very few instances of low birth weight among mothers who smoke and have hypertension, regardless of their race or weight during the last menstrual cycle. Across all weight categories, black and other race mothers tend to have a higher prevalence of low birth weight, especially when they smoke. Among white mothers, the highest number of low birth weight cases occurs within the weight range of 120-150 pounds. Conversely, for mothers weighing more than 150 pounds during their last menstrual cycle, the occurrence of

low birth weight is lower when they neither smoke nor have a history of hypertension. There is not enough data to interpret about mothers weighing greater than 150 pounds who have a history of hypertension and smoke, and belonging to all races.

After visualizing the categorical variables, we proceeded to create several log linear models and mosaic displays to determine the best fit. In our analysis, we considered one response variable (low) and three predictors (lwt, smoke, race). Table 3.1 below shows the statistical notation of numerous log-linear models, p-values from mosaic displays and log linear models respectively .

Out of the 12 models we could create, ensuring each predictor interacts with the response variable, three models are identified based on p-values as they are greater than 0.05. Model 6,9 and 11 emerged as the most favorable with p-values of 0.0804, 0.1476 and 0.08, respectively. For convenience, a condensed notation (A = low, B = lwt, C = smoke, D = ht, E = race) is used in Table 3.1

S.No	Statistical Notation	P-value from Mosaic	P-value from loglm model
1	[AB][AC][AD][AE]	7.20E-05	7.20E-05
2	[ABC][AD][AE]	1.02E-05	0.0003
3	[ABD][AC][AE]	0.0208	0.0208
4	[ABE][AC][AD]	3.00E-04	0.0003
5	[AB][ACD][AE]	2.80E-05	2.80E-05
6	[AB][ACE][AD]	0.0804	0.0804
7	[AB][AC][ADE]	1.00E-04	0.0001
8	[ABCD][AE]	0.0076	NaN
9	[AB][ACDE]	0.1476	NaN
10	[AC][ABDE]	0.0036	NaN
11	[AD][ABCE]	0.08	NaN
12	[ABCDE]	1	1

Table 3.3 *A table showing p-values of multiple log linear models for the given set of predictors. (A = low, B = lwt, C = smoke, D = ht, E = race)*

From Table 3.3, the mosaic displays for model 6,9 and 11 are shown in Figure 3.3 and Figure 3.4, each with p-value of 0.0842, 0.1476 and 0.0803 respectively. But, in model9 and model11, the p-value for loglinear models is not-a-number,i.e, NaN. It highlights that there might be problems with combinations of variables in their models. Hence, we chose to proceed with model6 which is a good fit. In other words, ***[low lwt][low smoke race][low ht]*** chosen to perform logistic regression.

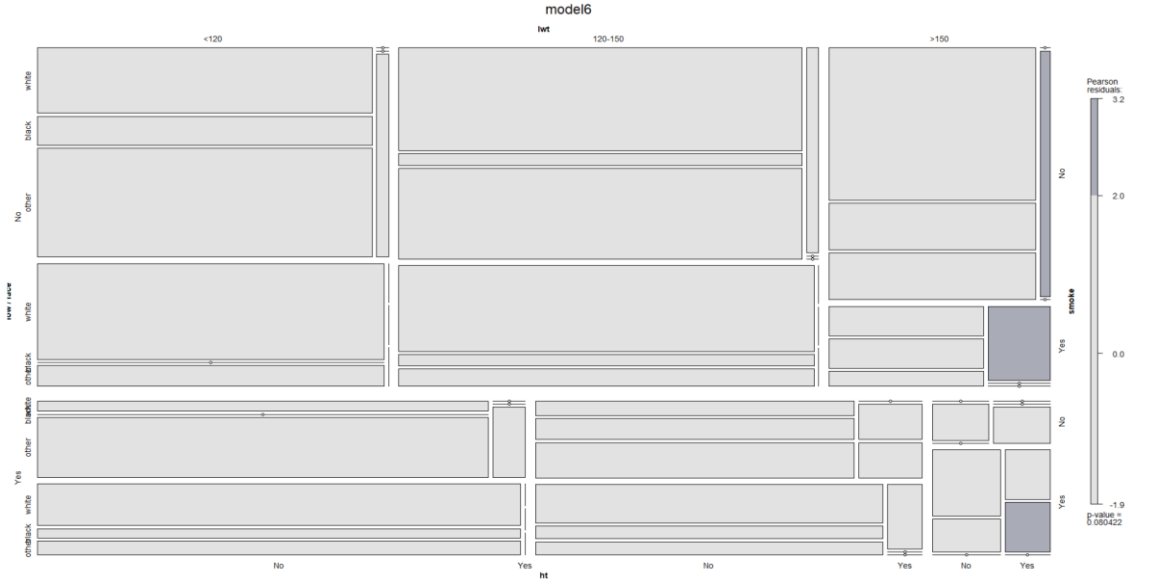

Figure 3.3 *A mosaic display for model6 [AB][ACE][AD], where A = low, B = lwt, C = smoke, D = ht, E = race*

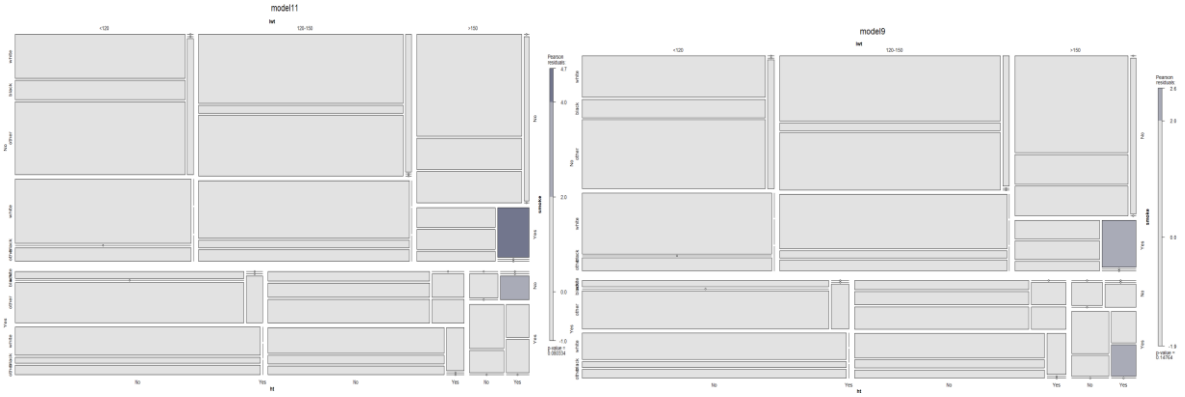

Figure 3.4 *Mosaic displays. Left : model 11 -> [AD][ABCE]; Right : model9 -> [AB][ACDE]. where A = low, B = lwt, C = smoke, D = ht, E = race*

After analyzing the mosaic displays, we are moving on to build logistic regression models. The goal of logistic regression is to examine the relationship between the response variable "low" with a set of predictors(lwt, smoke, race, ht).

Firstly, we're examining our model for both non-linearity and interactions. To test for non-linearity, we've constructed a separate model and compared it with our chosen one using anova. This secondary model includes a non-linear transformation of 'lwt' using a spline function. A p-value of 0.6526 was obtained, which exceeds the typical threshold of 0.05. This implies that the non-linear transformation doesn't significantly enhance the original model. Hence, it appears that 'lwt' and 'low' doesn’t exhibit any non-linear relationship in the given data.

To test for interactions, the main model is compared with a second model using anova, that checks for interaction effects among variables 'lwt', 'smoke', 'race', and 'ht'. However, the output doesn't yield a p-value and the deviance doesn't decrease, but instead increases by a very small amount, suggesting that the interactions among these variables do not significantly improve the model fit. This indicates that the effects of 'smoke', 'race', and 'ht' on 'low' are not dependent on each other or on 'lwt'. Hence, a model that includes main effects (individual effects of each predictor) appears to be a good fit for the data.

So, the finalized logistic model is:

η = logit[P(Y = Yes | X1, X2, X3, X4)] = -0.0563 - 0.0173*lwt + 1.5267*smokeYes + 1.5502*raceblack + 1.4158*raceother + 1.6767*htYes - 0.2948*smokeYes:raceblack - 1.1548*smokeYes:race

where Y = low, X1 =lwt, X2 = smoke, X3 = race, X4 = ht respectively.

Variable	Estimated regression coefficient
Intercept	-0.0563
lwt	-0.0173
smokeYes	1.5267
raceblack	1.5502
raceother	1.4158
htYes	1.6767
smokeYes:raceblack	-0.2948
smokeYes:raceother	-1.1548

Table 3.3 *Output of coeftest(model6), showing the estimated coefficients of model6.*

The interpretation of estimated coefficients as shown in Table 3.3 is as follows:

- Intercept (α) = -0.0563: For nonsmoking non-black mothers with average weight and hypertension, the log-odds of having a low birth weight baby is -0.0563. The odds of having a low birth weight baby are exp(-0.0563) = 0.9454.
- lwt (β1) = -0.0173 : For each unit increase in mother's weight, the log-odds of having a low birth weight baby decreases by -0.0173. This corresponds to the odds of having a low birth weight baby being multiplied by exp(-0.0173) = 0.9829, or roughly a 1.7% decrease, holding other variables constant.

- smokeYes (β2) = 1.5267 : The log-odds of having a low birth weight baby for mothers who smoke is 1.5267 higher than for mothers who don't smoke, holding other variables constant. This corresponds to the odds of having a low birth weight baby being exp(1.5267) = 4.6051 times higher for mothers who smoke compared to those who don't smoke, holding other variables constant.
- raceblack (β3) = 1.5502 : The log-odds of having a low birth weight baby for black mothers is 1.5502 higher than for white mothers, holding other variables constant. This corresponds to the odds of having a low birth weight baby being exp(1.5502) = 4.7115 times higher for black mothers compared to white mothers, holding other variables constant.
- raceother (β4) = 1.4158 : The log-odds of having a low birth weight baby for mothers of other races is 1.4158 higher than for white race mothers, holding other variables constant. This corresponds to the odds of having a low birth weight baby being exp(1.4158) = 4.1206 times higher for mothers of other races compared to white race mothers, holding other variables constant.
- htYes (β5) = 1.6767 : The log-odds of having a low birth weight baby for mothers who have a history of hypertension is 1.6767 higher than for those who don't have hypertension. This corresponds to the odds of having a low birth weight baby being exp(1.6767) = 5.3507 times higher for mothers with hypertension compared to mothers without it.
- smokeYes:raceblack (β6) = -0.2948 : The log-odds of having a low birth weight baby decreases by -0.2948 for smoking black mothers compared to the combined effect of being a black mother and a smoking mother, holding other variables constant. This corresponds to the odds of having a low birth weight baby being exp(-0.2948) = 0.7446 times that for non-smoking black mothers, holding other variables constant.

There are two hypotheses to be tested for this model, one which compares the chosen model with a null model(Type1) and with a saturated model(Type 2) to see which model is a good fit. The results of Type1 and Type2 tests are as follows :

For Type 1 test :

- From the output obtained from the anova() function and as per the Type 1 hypothesis, the inclusion of individual predictors significantly reduces the residual deviance, leading to the rejection of the null hypothesis.
- Specifically, the addition of 'lwt' and 'smoke' decreases the deviance to 228.69 and 224.34, respectively, supported by significant p-values (0.0144 and 0.0370). Similarly, the inclusion of 'race' and 'ht' results in further reduction to 215.01 and 208.25, with significant p-values of 0.0094 and 0.0092, demonstrating the importance of these variables in enhancing model fit. However, the interaction term 'smoke:race' does not significantly improve the model as it only decreases the deviance only by 1.6992

For Type 2 test:

- In this case, the null hypothesis is that the fitted model is correctly specified. The output of LRstats() function shows AIC score of 222.55, BIC score of 248.48, and a Likelihood Ratio Chi-square (LR Chisq) of 206.55 with 181 degrees of freedom. The p-value associated with this test statistic is 0.0935.
- In terms of interpretation, the p-value is greater than the common significance level of 0.05, suggesting that we do not have enough evidence to reject the null hypothesis that the fitted model is correctly specified.

In this next step, we used an all-effects plot to visualize the logistic regression model. Figure 3.5 shows three plots: the lwt effect plot, ht effect plot, and smoke*race effect plot. Here the

y-axis shows the likelihood of the response variable as values in x-axis(or predictor) changes.

In the lwt effect plot, we see that as the mother's weight during the last menstrual cycle increases, the chances of low birth weight decrease. The highest likelihood of low birth weight happens when the mother weighs less than 100 pounds, and it decreases almost to zero once the weight goes beyond 200 pounds.

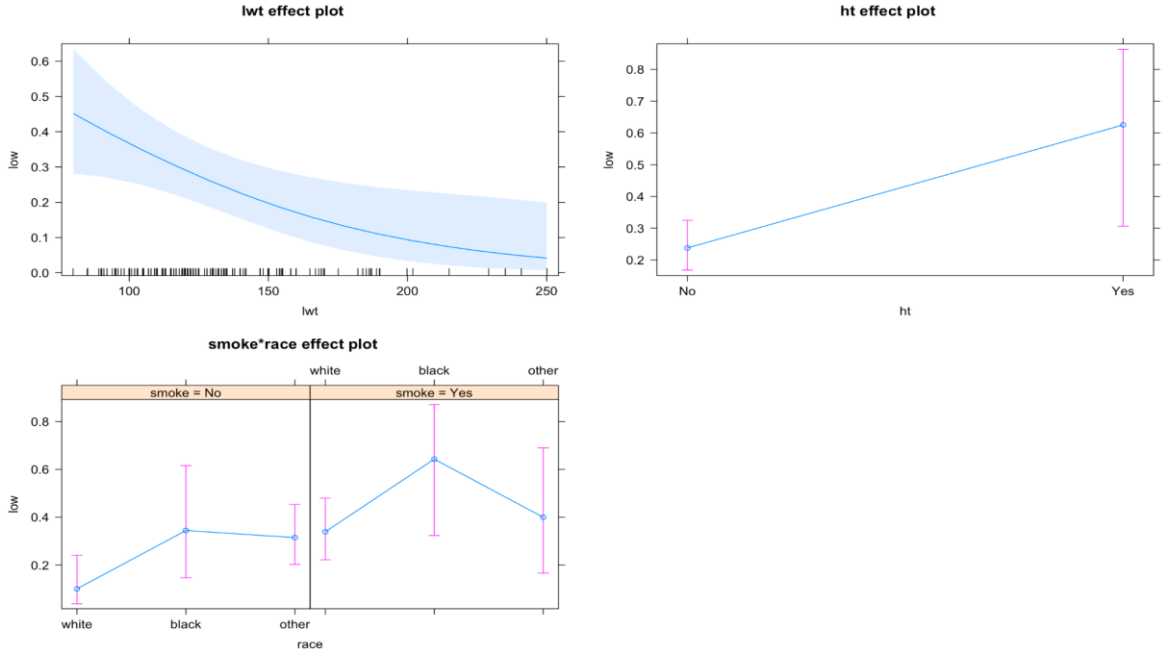

Figure 3.5 *All effects plot for logistic regression, where response variable is low and its predictors are lwt, ht, smoke, race*

Looking at the ht effect plot, the probability of low birth weight is very low (slightly above 0.2) for mothers without a history of hypertension. However, it increases to slightly above 0.6 for mothers with hypertension. The confidence interval is wider for hypertension because there are fewer observations in that category.

In the smoke*race effect plot, we find that the likelihood of low birth weight is higher for black mothers compared to white and other races, regardless of smoking status. Among smoking mothers, the likelihood is around 0.6. For white mothers, the likelihood is relatively low for non-smokers (below 0.2), but slightly higher for smokers (just below 0.4). For other race mothers, the likelihood of low birth weight is below 0.4 for non-smokers and close to 0.4 for smokers.

After interpreting the results of effect plots, we wanted to assess the impact of individual cases that affect the fitted model. Hence, we plotted influence plots and index-influence plots which highlight the outliers. A case is considered influential when it is poorly fit and has unusual predictor values. The test for influential cases is done below.

An influence plot is created as shown in Figure 3.6 which highlights five outliers from the model. The observations from the dataset are - 28,36,98,138 and 202. Now to decide for influential cases, we have to look at cook's distance, studentized residual and hat-value scores for each outlier. An outlier is considered influential if cook's distance is greater than 1, hat-values is greater an 2k/n or 3k/n (where k is number of coefficients, and n are the

number of observations), or when studentized residuals are close to extreme values(here it refers to the points which are close to or greater than 2 or -2).

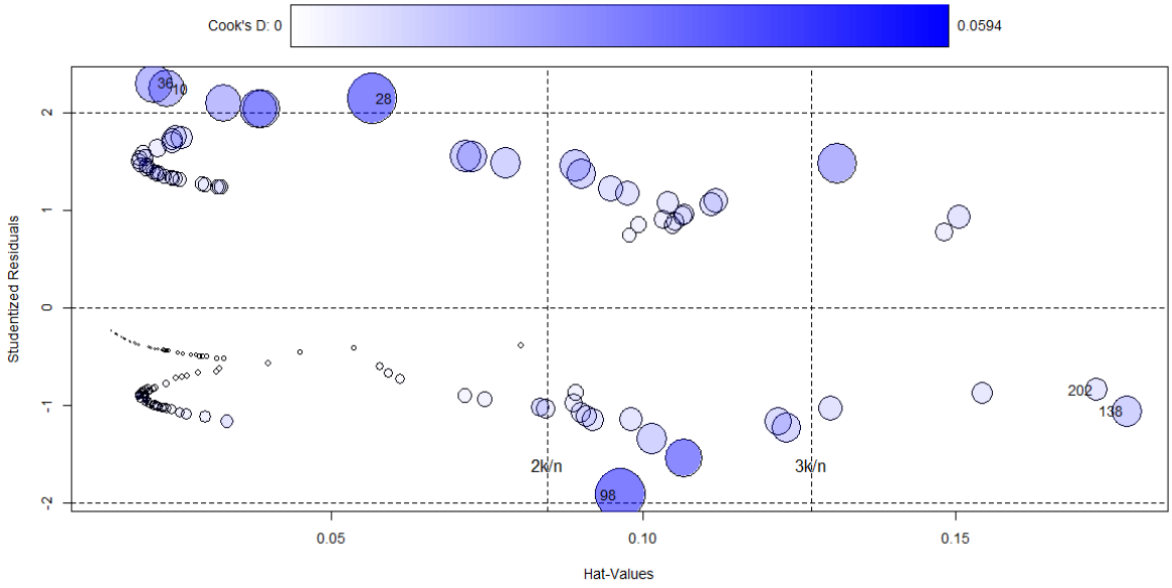

Figure 3.6 *Influence plot for the finalized model*

From the influence plot output in Figure 3.6, there are several cases (98, 10, 28, 36) with Studentized residuals close to or exceeding the typically used threshold of 2, suggesting potential outliers. These do not appear to be unduly influential on the model when considering the other diagnostic measures. The Cook's distances for these observations are all well below the commonly used threshold of 1, with the highest value being 0.0594 for observation 98. This indicates that these cases do not influence the model fit.

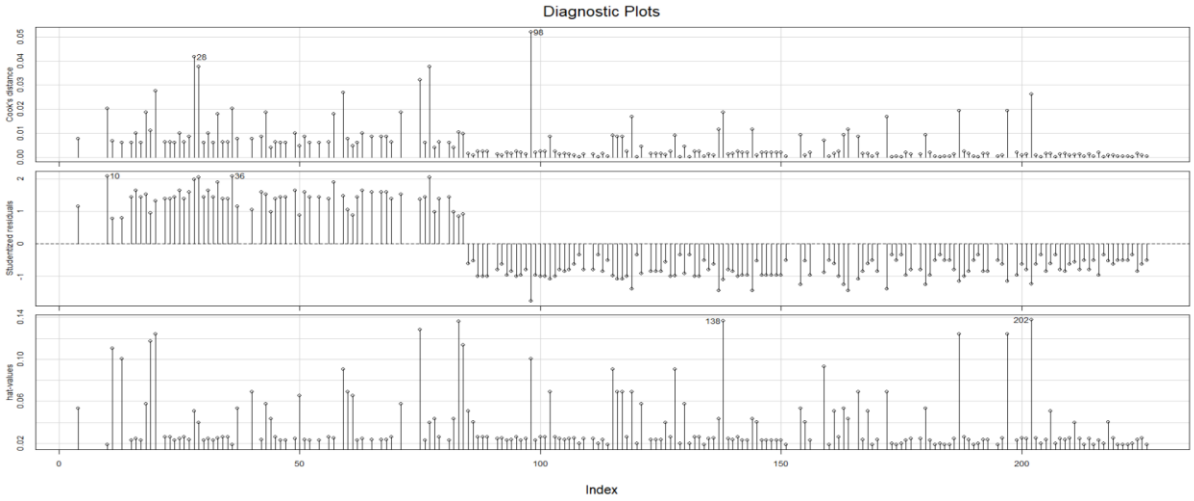

Figure 3.7 *Influence Index plot for the finalized model*

Another diagnostic plot called an influence index plot is created in Figure 3.7 to observe outliers in each metric separately. Based on this plot, observations 138 and 202 stand out as outliers in the hat-values graph. But since their cook's distance is still under 1, they are not influential. We can conclude that there are no influential cases in this model.

4. Conclusion

In this analysis, a logistic regression model was employed to assess the relationship between a response variable "low" and four predictors - lwt, smoke, race and ht. After considering twelve possible log-linear models and their mosaic displays, only the sixth model was chosen for further exploration.

The chosen logistic regression model suggested that the predictors significantly influence the odds of a 'low' outcome. We see that as the mother's weight during the last menstrual cycle increases, the chances of low birth weight decreases. The probability of low birth weight is very low for mothers without a history of hypertension. However, it increases for mothers with a history of hypertension. We also found that the risk of low birth weight varies considerably by race and smoking habits. It was observed that Black mothers have a relatively high risk regardless of smoking status. Among those who smoke, the risk is significantly higher. For white mothers, the risk is generally low, but it increases for those who smoke. For mothers of other races, the likelihood of low birth weight is comparatively lower, but for smokers in this group, the risk is marginally higher.

While testing for non-linearity and interactions, we found that 'lwt' and 'low' does not exhibit a non-linear relationship and the predictors are not dependent on each other or on 'lwt'. The validity of this model was further confirmed through Type 1 and Type 2 hypothesis tests suggesting the model was correctly specified. The model was further visualized using an all-effects plot which proved these findings, revealing clear influence of predictors on the response variable. Lastly, influence plots and index-influence plots were created to identify influential cases that could potentially skew the model's outcomes. Despite a few outliers, no case was considered influential. Hence, we can conclude that this model is doing a good job in finding the likelihood of low birthweight with the given data.

For future work, considering additional variables like socioeconomic status, access to healthcare, dietary habits, education level could enhance the model's accuracy and generalizability.

References

[1] YadavD. K., ChaudharyU., & ShresthaN. (2011). Risk Factors Associated with Low Birth Weight. Journal of Nepal Health Research Council. https://doi.org/10.33314/jnhrc.v0i0.266

[2] Dabbagh, O., Al Talakchi, F., Ajeena, A., Al-Fahdawi, S., & Al-Kubaisy, W. (2015). Birth weight and associated factors in full term newborns: A cross-sectional study in northern Iraq. International Journal of Pediatrics, 2015, 807373. doi: 10.1155/2015/807373

[3] Joshi, H.S.1,; Subba, S.H.2; Dabral, S.B.1; Dwivedi, S.1; Kumar, D.1; Singh, S.1. Risk Factors Associated with Low Birth Weight in Newborns. Indian Journal of Community Medicine 30(4):p 142, Oct–Dec 2005.