**Factors of Birth Weight**

Rebekah Morales

INFO 523

University of Arizona

**Introduction**

Are there any factors associated with a baby’s birth weight? Can we control any of these variables to prevent low birth weight? After having two kids myself and having a family history of low birth weight I wanted to examine the factors that have an association with a baby’s weight. On a personal level I am interested in if I can control any of the factors that have an association in the hopes that I can help prevent low birth weight for myself in the future. I would also like to identify if any variables are more significant than others, i.e. cause a larger shift in the data than the other variables.

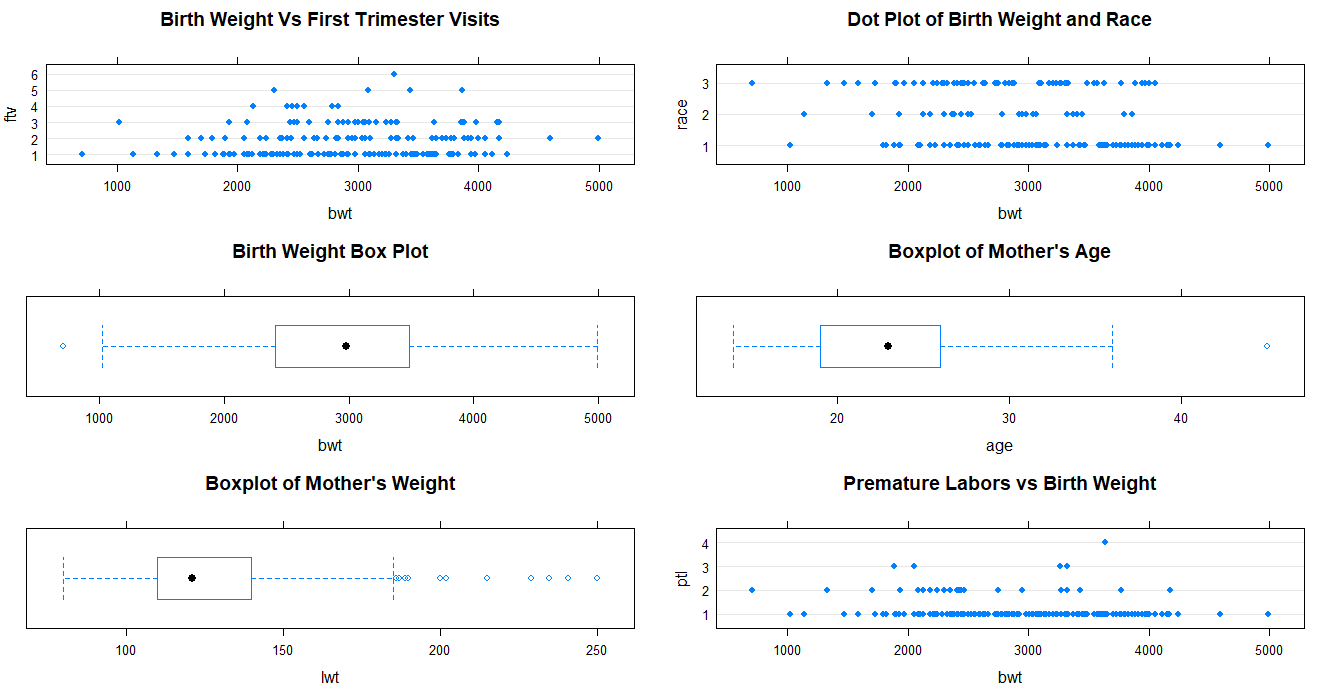
**Data Collection and Exploratory Data Analysis**

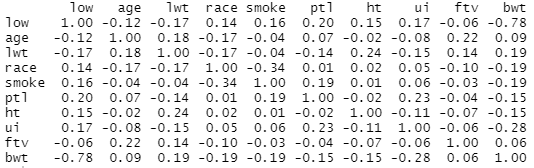
I will be using a dataset from Baystate Medical Center in Springfield Massachusetts during 1986. There are 189 cases and 10 total variables in my data set. My variables include low, age, race, ht, ptl, lwt, smoke, ui, ftv, and bwt. Bwt is the birth weight in grams and is our response variable. Low, is an indicator variable that marks if the baby’s weight is low. All other variables I will be looking at as possible explanatory variables.

|  |  |
| --- | --- |
| **Bwt:**  Birth weight in grams | **Low:**  Indicator of birth weight less than 2.5kg |
| **Age:**  Mother’s age in years | **Race:**  Mother’s race (1=White, 2=Black, 3=Other) |
| **Ht:**  History of hypertension | **Ptl:**  Number of previous premature labor(s) |
| **Lwt:**  Mother’s weight in pound at last menstrual period | **Smoke:**  If the mother smoked during pregnancy |
| **UI:**  Presence of uterine irritability | **Ftv:**  Number of first trimester visits |

Link to Dataset: https://vincentarelbundock.github.io/Rdatasets/datasets.html

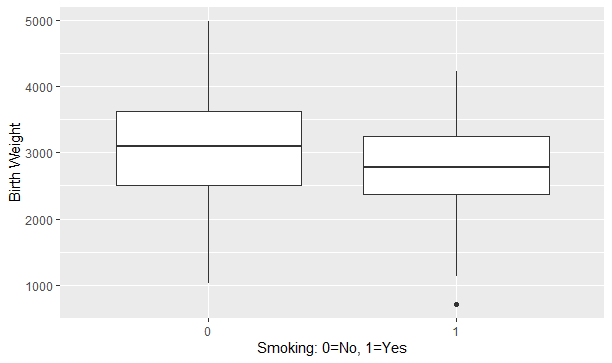
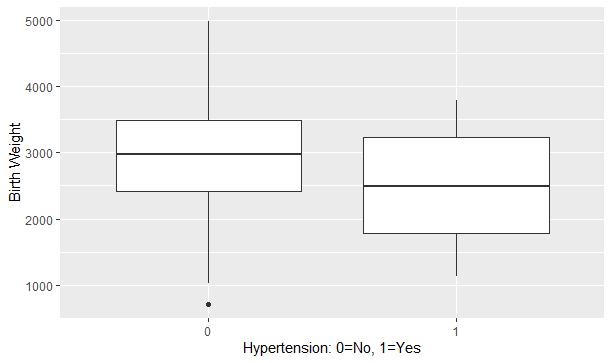
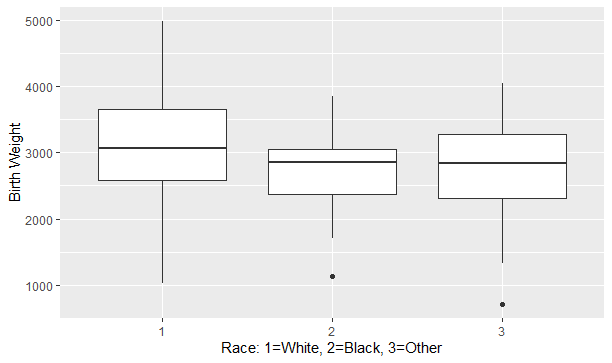
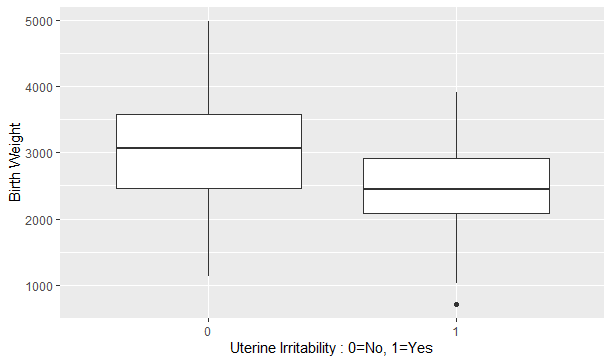
When initially looking at my data I wanted to compare each variable back to the birth weight while excluding low. As low only indicates if the response variable was below 2.5kg I will not be considering it in any of my models. (Please see correlation matrix that supports my removal below)





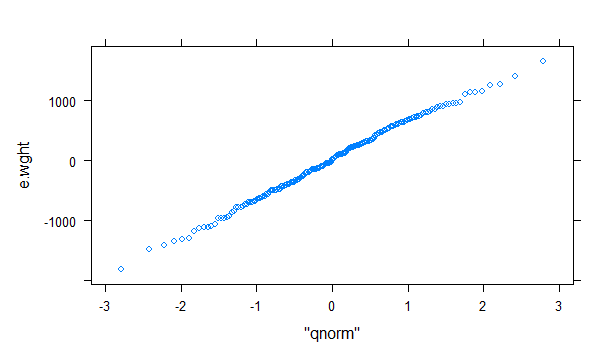
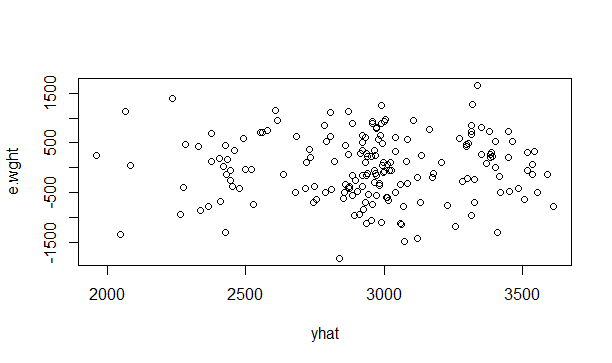
**Correlation Matrix**

When looking at some basic plots to get an idea of what my data holds, I noticed that the levels ftv was spread evenly over the birthweights and may not have an association. There was also a difference in the levels of race and made a note that I may want to look at it closer. I also noticed outliers in birthweight that may be explained by the outliers in age or lwt. When looking at a correlation matrix of my data to check the only two variables outside of a 0.7 threshold are bwt and low. I expected this in the data as low is used as an indicator flag if the bwt falls below 2.5kg. I was still left wondering about my categorical variables (Race, Ht, Ui, Smoke), boxplots for each gave me a better overall sense of the data. Later, if significant, I will look at the confidence intervals for these variables to accurately gather an idea of how these can effect each other with the rest held constant.



**Regression Analysis**

To start reducing my model down to only variables with an association I looked at the full model first. I started by checking my assumptions. There were less data points on the right but overall it met constant variance, there appears to be no pattern, and they seem to add to 0. Normality was reasonably normal and with 189 observations I am not concerned with the 1 or 2 on the top right that seem to shift/deviate away.

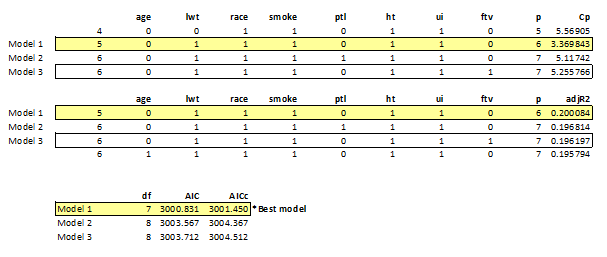


I then also ran the residuals for each individual variable, this way I could see if any might be the cause for the deviation away from the line. I did find that age had an outlier that may account for the deviation.

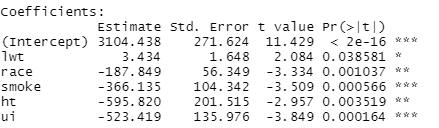
Chart, scatter chart

Description automatically generated

I then looked at the summary of model to see if any of the variables stood out that they could possibly be removed. With variables all over the place I decided to leave all the variables in for now and evaluate using Cp’s and R2adj values to narrow down my variables faster. The table below shows my lowest Cp and the Cp’s closest to p. I highlighted the two best models. I then compared those to the values of my R2adj , and I highlighted the highest value there. These two best models happened to be the same. There were two other models that appeared in both sets of data too. All three models contained lwt, race, smoke, ht, and ui.



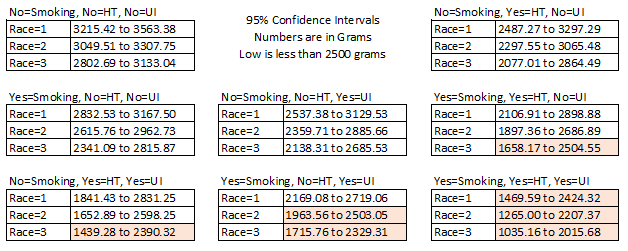
I then reran my assumptions on the data and found that all my assumptions were met as there was constant variance, no pattern, residuals added to 0, and normality was met.



My estimated regression function is now:

Bwt = 3.434(lwt) -187.849(race) -366.135(smoke) -595.82(ht) -523.419(ui)

To look at the effect that each variable has on the baby’s birth weight I looked at the 95% confidence intervals with other variables held constant. As you can see there is quiet a difference between the three races even with no other risk factors. The dramatic change is when all the variables are compounded. No matter what race all have ranges far below the low indicator.



**Results and Conclusion**

Overall, I found that only five of my variables had a significant association. Race, mother’s weight, smoking, hypertension, and uterine irritability were my last remaining variables. Ht, for hypertension had the most significant drop by itself when all other variables remained constant. Ui, for uterine irritability even had a larger drop than smoking when all other variables remained constant. However, you cannot control your race, or developing hypertension or uterine irritability. The only variable that each person can control was smoking. While none of the other variables are within our control there are things each person can do to minimize their risk of having a baby of low birth weight.