Data Analysis Assignment II

Dean Huang

9/6/2020

Lab Report

Question 1:

Fit a regression model for predicting the interval between eruptions from the duration of the previous one, to the data, and interpret your results.

For each additional minute of eruption duration, the subsequent eruption duration will increase by 11 minutes. According to the intercept of the model after centering, the average subsequent eruption duration is 71 minutes when eruption duration is 3.5 minutes. The adjusted R square of 0.73 means 73% of variation in the response variable is explained by the regression fit.

% latex table generated in R 4.0.1 by xtable 1.8-4 package % Tue Sep 15 19:02:04 2020

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.0000	0.6460	109.90	0.0000
durc	10.7410	0.6263	17.15	0.0000

Include the 95% confidence interval for the slope, and explain what the interval reveals about the relationship between duration and waiting time.

We are 95% confident the slope will fall within the range of 69.7 and 72.3. In another words, we are 95% confident for each additional minute of eruption duration, the subsequent eruption duration will increase in the range between 70 and 72 minutes.

% latex table generated in R 4.0.1 by xtable 1.8-4 package % Tue Sep 15 19:02:04 2020

	2.5 %	97.5 %
(Intercept)	69.72	72.28
durc	9.50	11.98

Describe in a few sentences whether or not you think the regression assumptions are plausible based on residual plots (do not include any plots).

1) Linearity: According to the residual plot for duration, the relationship between duration and interval appears to be linear because the plot seems to be random. However, there are not enough data points between the duration of 2.46 and 3.46 to tell if the graph is completely random.
2) Normality: According to qqplot, the model appears to be normal because all points appear to be clustered around the 45 degree line.
3) Equal Variance & Independence: The points in the residual vs fitted plot look "roughly" random and "roughly" equally spread out around zero. Therefore, no violation to the independence and equal variance assumption. However, there are not enough data points for interval between 55 to 70 to tell if the graph is completely random.

Fit another regression model for predicting interval from duration and day. Treat day as a categorical/factor variable. Is there a significant difference in mean intervals for any of the days (compared to the first day)? Interpret the effects of controlling for the days (do so only for the days with significant effects, if any).

There is no significant difference in mean intervals for any of the days because the p-values are all larger than 0.05. Adding the day predictor variable will not help with the prediction of eruption interval.

% latex table generated in R 4.0.1 by xtable 1.8-4 package % Tue Sep 15 19:02:04 2020

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.8770	3.0672	10.72	0.0000
duration	10.8813	0.6622	16.43	0.0000
day2	1.3275	2.7173	0.49	0.6263
day3	0.7825	2.6994	0.29	0.7725
day4	0.1625	2.6461	0.06	0.9511
day5	0.2463	2.6459	0.09	0.9260
day6	1.9918	2.6580	0.75	0.4554
day7	-0.1700	2.7020	-0.06	0.9500
day8	-0.6944	2.6957	-0.26	0.7973

Perform an F-test to compare this model to the previous model excluding day. In context of the question, what can you conclude from the results of the F-test?

The p-value of F-test is p = 0.7837 which is greater than the significance level 0.05. In conclusion, adding day variable as one of our predictor variables wil not help us make more accurate prediction.

% latex table generated in R 4.0.1 by xtable 1.8-4 package % Tue Sep 15 19:02:04 2020

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	105	4689.01				
2	98	4620.16	7	68.85	0.21	0.9828

Using k-fold cross validation (with k=10), compare the average RMSE for this model and the average RMSE for the previous model excluding day. Which model appears to have higher predictive accuracy based on the average RMSE values?

The RMSE for model 1 (with no day variable) and model 2 (with day variable) are 6.51 and 6.49 respectively. Therefore, we can conclude the model that include day as the predictor variable has slightly higher accuracy. However, the difference is negligible. Therefore, we can conclude adding day predictor variable will not help with the prediction of interval.

Question 2:

Summary

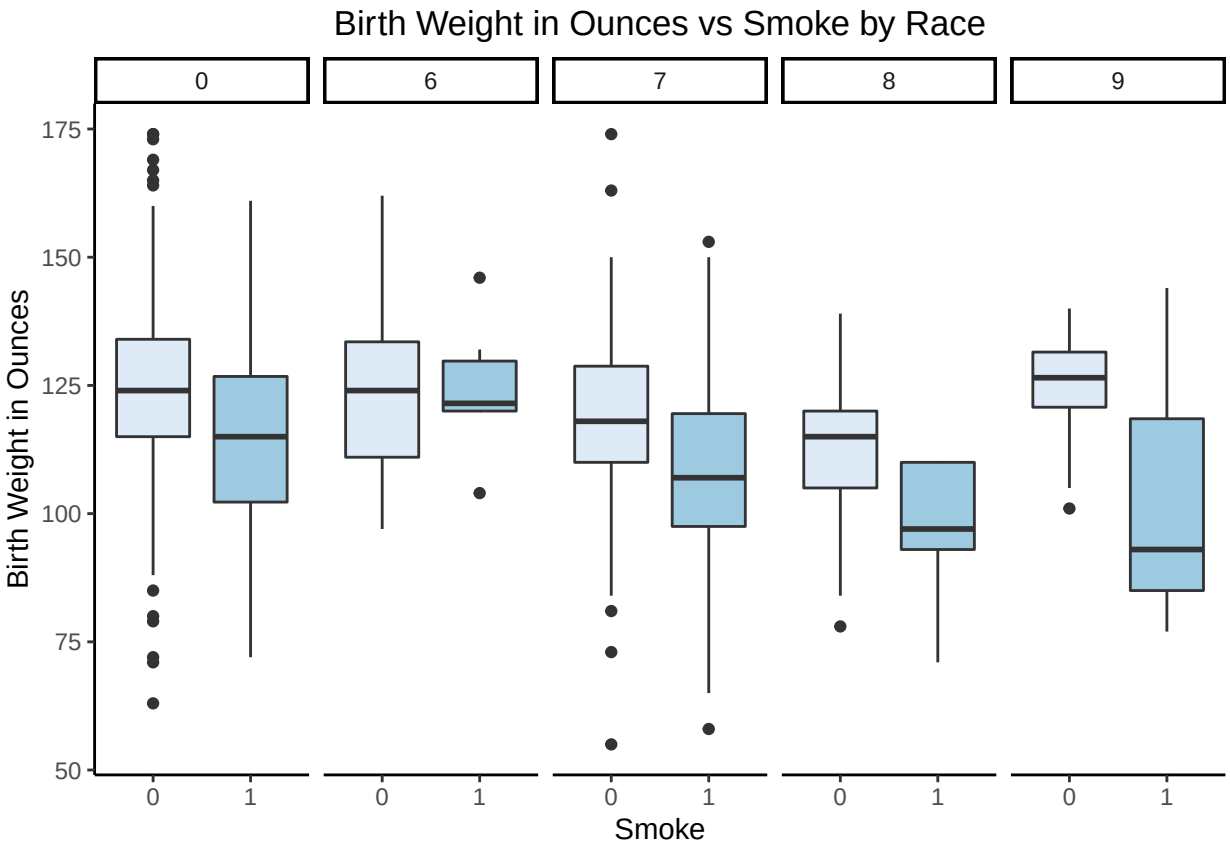
The main question we are trying to address is "Do mothers who smoke tend to give birth to babies with lower birth weights than mothers who do not smoke?". In addition, we are trying to see if the association between smoking and birth weight differs by mothers race. Through careful analysis of our final model, we will provide an estimate range for the difference in birth weights for smokers and non-smokers, and highlight interesting association with birth weight that are worth mentioning. We will begin the study by conducting EDA to check the association of predictor variables and response variable, and highlight the preliminary concerns we have with the response and predictor variables. Next, we will explore potential interactions of the model to see if there is a difference in data trend for different population groups.Last but not least, we will perform data modeling and model assessment to find the appropriate model for prediction and answer our inferential questions. The outcome of the study shows mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke. However, we do not have enough evidence to conclude that the association between smoking and birth weight differs by mother's race.

Introduction

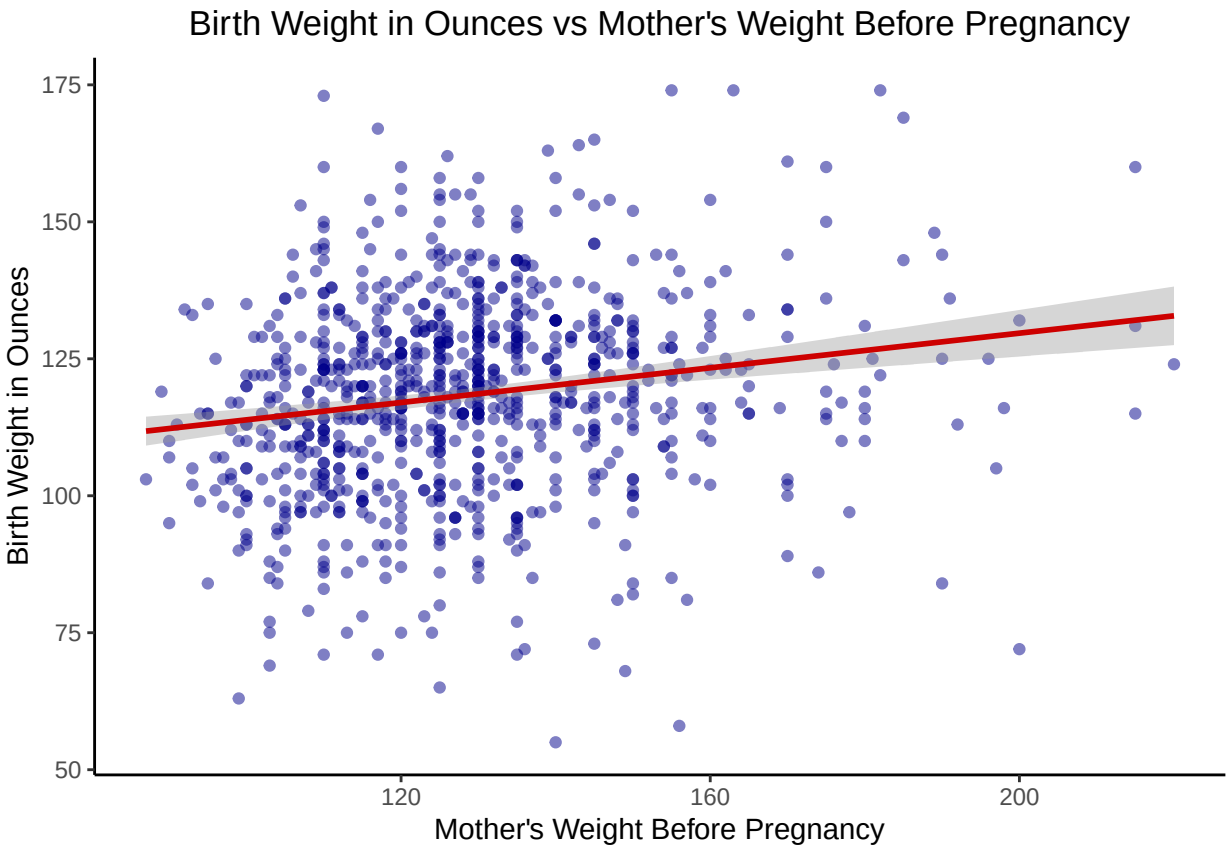
Since we are only interested with the association between the smoking status of mother and birth weight, we will remove the variables that are either irrelevant or insignificant to our question. These variables include id, birth, gestation, drace, ded, dht, dwt, marital, time, number, and premature.The response variable we are interested in looking to is bwt.oz, and the predictor variables we are interested in looking to are parity, mrace, mage, med, mht, mpregwt,income, and smoke. To facilitate the EDA, data modeling, and model assessment processes, we will remove data rows with one or more missing values. We will use boxplots for categorical variable and scatterplots for continuous variables to not only find the association between predictor variable and response variable but also possible interactions. Next, we will employ backward selection and F-test to find our final model. Last but not least, we will check if our final model fulfill the linear assumptions, and do not have multicollinearity and outliers.

Data

According to the histogram of the bwt.oz, the distribution appears to be normal. Hence, transformation of response variable is not needed for now. Since there are no unknown values for parity, we will treat parity as a discrete variable when perform EDA. The scatter plot of parity vs bwt.oz appears to have a weak or close to zero correlation with the fitted line close to horizontal. We might drop parity depending on the results of data modeling and modeling assessment. Since mrace is categorical variable, we will start by factoring the variable then draw a boxplot to observe the association between mrace and bwt.oz. The box plot shows there is a minor difference in distribution of bwt.oz for different race grops. Since mage is a discrete variable, we will draw a scatter plot to observe its association with bwt.oz. The scatter plot of mage vs bwt.oz appears to have a close to zero or weak correlation with the fitted line close to horizontal. Since med is categorical variable, we will start by factoring the variable then draw a boxplot to observe the association between med and bwt.oz. The box plot shows there is a minor difference in distribution of bwt.oz for different mother's education levels. Since mht is a discrete variable, we will draw a scatter plot to observe its association with bwt.oz. The scatter plot of mht vs bwt.oz appears to have a weak positive correlation. Since mpregwt is a discrete variable, we will draw a scatter plot to observe its association with bwt.oz. The scatter plot of mpregwt vs bwt.oz appears to have a weak positive correlation. Since inc is categorical variable, we will start by factoring the variable then draw a boxplot to observe the association between family income and bwt.oz. The box plot shows there is a minor difference in distribution of bwt.oz for different levels of family income. Since smoke is a categorical variable, we will start by factoring the variable then draw a boxplot to observe the association between smoking status of mother and bwt.oz. The box plot shows there is a minor difference in distribution of bwt.oz for difference in smoking status (the distribution for smoker is slightly lower than the distribution for nonsmoker). Next, we will explore the interactions of predictor variables. The first interaction we would like to explore is birth weight in ounces vs mother's pre-pregnancy weight in pounds by mother's height. The reason we picked this interaction to explore is because there is a possibility that bwt.oz vs mregwt has a different distribution for mother with different heights because scientifically a mother's height does affect a mother's pre-pregnany weight. The trend of bwt.oz vs mpregwt appears to be different for height group of 59 inches, and this might be due to the lack of data in this particular height group (interaction might be needed for these two predictor variables). The second interaction we would like to explore is birth weight in ounces vs smoking by mother's race. The main reason we picked this interaction to explore is address the questions on the possibility of bwt.oz vs smoking having a different distribution for mothers in different ethnic groups. The trend of bwt.oz vs smoke appears to be the same for all racial groups. The last interaction we would like to explore is birth weight in ounces vs smoking by family income. The main reason we picked this interaction to explore is to address the questions on the possibility of bwt.oz vs smoking having different distributions for different family income groups. The trend of bwt.oz vs smoke appears to be the same for all levels of family income.




```
## `geom_smooth()` using formula 'y ~ x'
```



Model

For this report, we will be implementing backward selection to find the model that has the lowest BIC. The reason we picked BIC over AIC is because BIC generally places a heavier penalty on modes with more than 8 variables. The single predictor variables we will include for our full models are parity, mrace, mage, med, mht, mpregwt,income, and smoke. Besides the single predictor variables, we will include seven interactions including 1) smoke and parity 2) smoke and mage 3) smoke and race 4) smoke and mht 5) smoke and income 6)mpregwt and mht.Before performing backward selection, we will center the variables mht and mpregwt to make the interpretation of intercept more meaningful. After performing backward selection, the final model ended up having four predictor variables: mrace, mht, mpregwt, and smoke. These four predictor variables match our findings from EDA. However, all interactions are dropped from the model. In order to address the question on the association between smoking and birth weight differs by mother's race, we will have to perform model assessment for the interaction, race and smoke, to see if including the interaction is helpful for the prediction of bwt.oz. In addition, through our observations from EDA, we identify one potential interaction, mpregwt and mht, that appears to have different trends for different groups. Therefore, we will also perform model assessment for this interaction.

We will use F-test to determine if there is a need to add the interaction to our model. According to the summary table of our model including the interaction of mpregwt and mht, the p-value is larger than 0.05. In addition, the result of f test shows including this interaction has a high p-value compare to excluding this interaction. Therefore, we will not include this interaction in our final model. According to the summary table of our model including the interaction of smoke and mrace, the p-value is larger than 0.05. In addition, the result of f test shows including this interaction has a high p-value compare to excluding this interaction.Therefore, we will not include this interaction in our final model. The next step will be to check the multicollinearity of the model. According to our results, the vif value for mht is 1.3, the vif value for mpregwt is 1.27, and the vif value for smoke is 1.01, which are acceptable. For the linearity assumption, the residual plots for mht vs bwt.oz and mpregwt vs bwt.oz appear to be linear because the plot seems

random.The residual fitted plot looks random and "roughly" equally spread out around zero. Therefore, no violation to the independence and equal variance assumption. Most points appear to cluster around the 45 degree line of the qq-plot with some points at both end of tails deviate from the 45 degree line. Overall, the model satisfies the normality assumption. Next, we will check if there are any outliers, leverage points or influential points. There appears to be no influential points and outliers according to the graph of cook's distance. However, there are some leverage points that are not influential.

$$y_i = \beta x_i + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

This is the equation of our final model.y_i is the birth weight in ounces for observation i, and x_i is the vector containing the corresponding values for mother's pre-pregnancy weight in pounds, mother's height in inches, and smoke.

% latex table generated in R 4.0.1 by xtable 1.8-4 package % Tue Sep 15 19:02:05 2020

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	124.6282	0.9408	132.47	0.0000
mhtc	0.9306	0.2611	3.56	0.0004
mpregwtc	0.1179	0.0319	3.70	0.0002
sm1	-9.5640	1.3415	-7.13	0.0000
racef6	0.1915	3.9669	0.05	0.9615
racef7	-8.9203	1.9945	-4.47	0.0000
racef8	-6.3040	3.5440	-1.78	0.0756
racef9	0.7701	4.9211	0.16	0.8757
sm1:racef6	14.5600	7.9768	1.83	0.0683
sm1:racef7	1.6336	2.9235	0.56	0.5765
sm1:racef8	-6.6474	6.6361	-1.00	0.3168
sm1:racef9	-12.3835	10.8830	-1.14	0.2555

The table above shows the summary of model including the interaction of smoke and race.

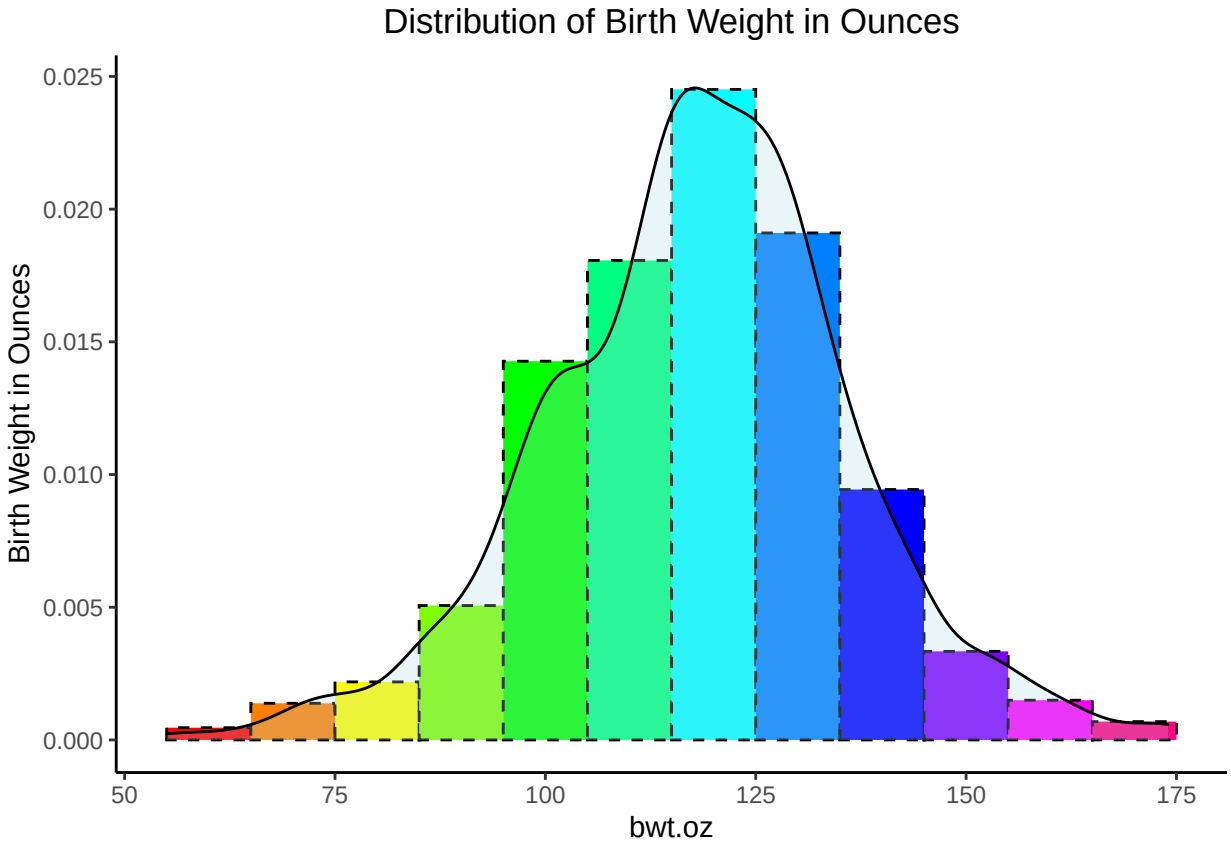
Conclusion

The intercept of our final model shows mother who does not smoke with average height and average pre-pregnancy weight will give birth to a child with 124.5 ounces. As the height of mother increases by one inch, the weight of the child will increase by 0.88 ounces given all other variables are constant. As the pre-pregnancy weight of the mother increases by one pound given all the other variables are constant, the weight of the child will increase by 0.11 lbs given all other variables are constant. Given all the other variables are constant, the birth weight of a child will decreases by 9.06 when the mother smokes compare to mother who does not smoke. With 95% confidence, the range for the difference in birth weights for smokers and non smokers is between -11.35 and -6.77. Therefore, mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke. The adjusted R square of 0.11 means 11% of variation in the response variable is explained by the regression fit. Since the p-value for the F-test for including interaction of mrace and smoke is bigger than 0.05, we do not have enough evidence to conclude that the association between smoking and birth weight differs by mother's race. However, the p-value for mother with the ethnicity of black who smokes is smaller than 0.05. Therefore, it appears to be some association between smoking and birth weight for mothers in the ethnicity group of black. One potential limitation we have for this study is that we do not look at the the impact of smoking on gestation age. Gestation age is important because the higher the gestational age, the higher the birth weight for children. The second limitation is that there are not enough data for some population groups like mothers with an education less than 8th grade to accurately calculate the association of the corresponding predictor variables to bwt.oz.

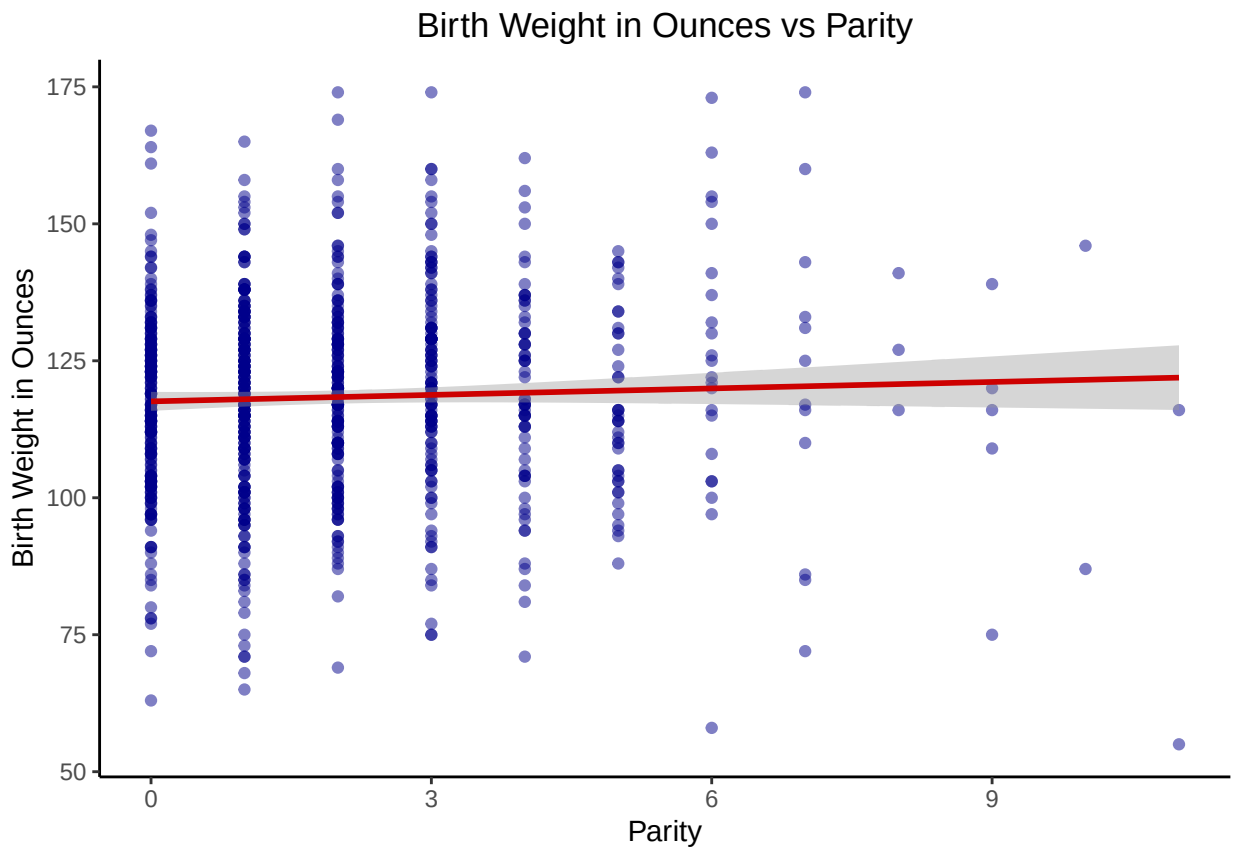
% latex table generated in R 4.0.1 by xtable 1.8-4 package % Tue Sep 15 19:02:05 2020

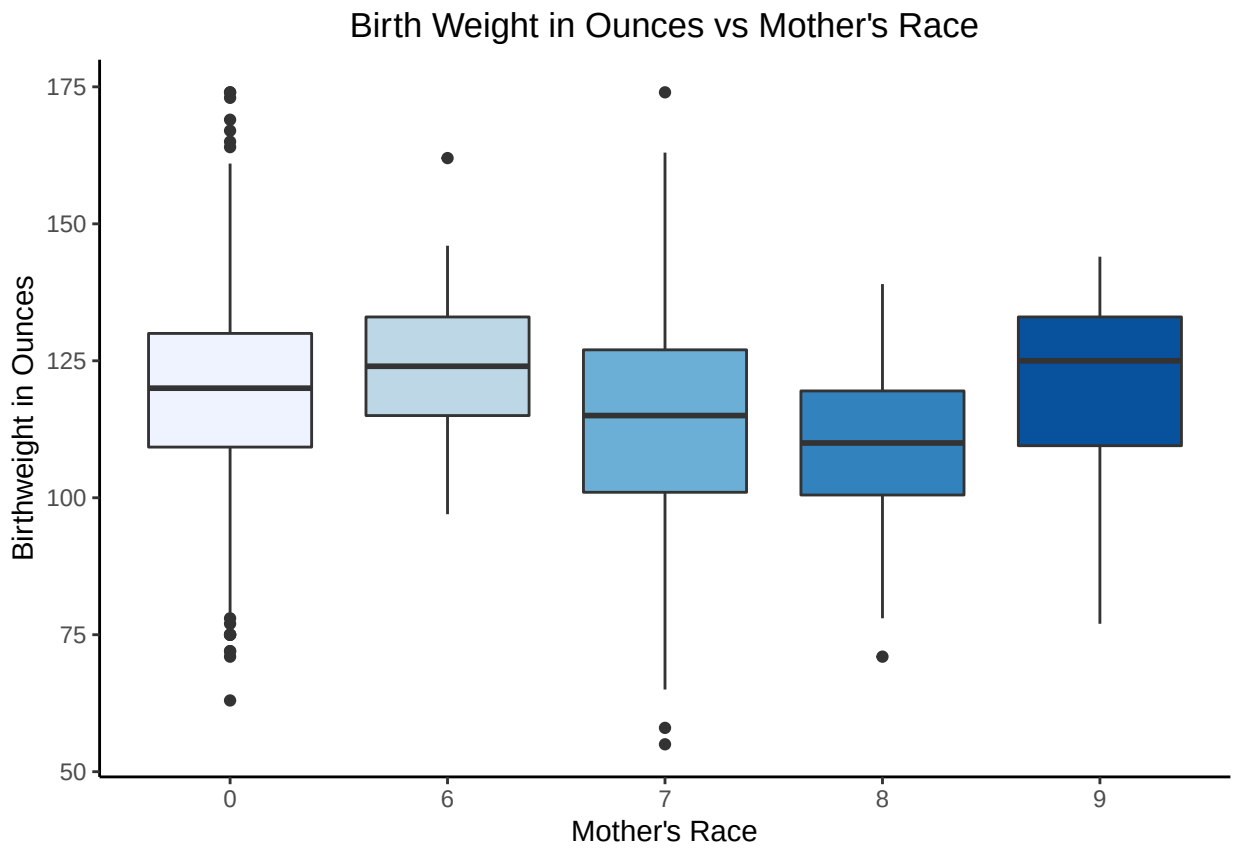
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	124.4977	0.8802	141.45	0.0000
racef6	3.6286	3.4707	1.05	0.2961
racef7	-8.1932	1.4894	-5.50	0.0000
racef8	-8.1468	3.0396	-2.68	0.0075
racef9	-1.6670	4.3927	-0.38	0.7044
mhtc	0.8757	0.2598	3.37	0.0008
mpregwtc	0.1179	0.0319	3.70	0.0002
sm1	-9.2744	1.1539	-8.04	0.0000

Appendix

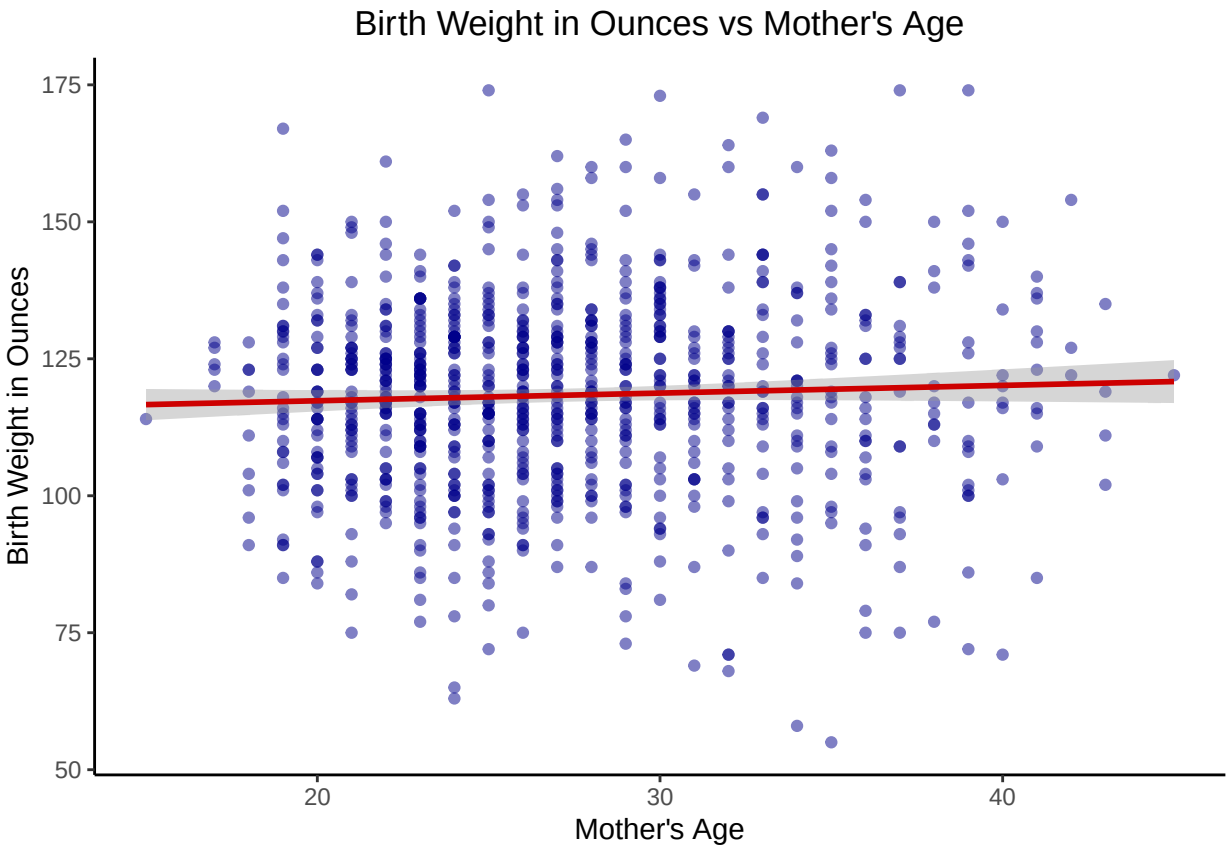


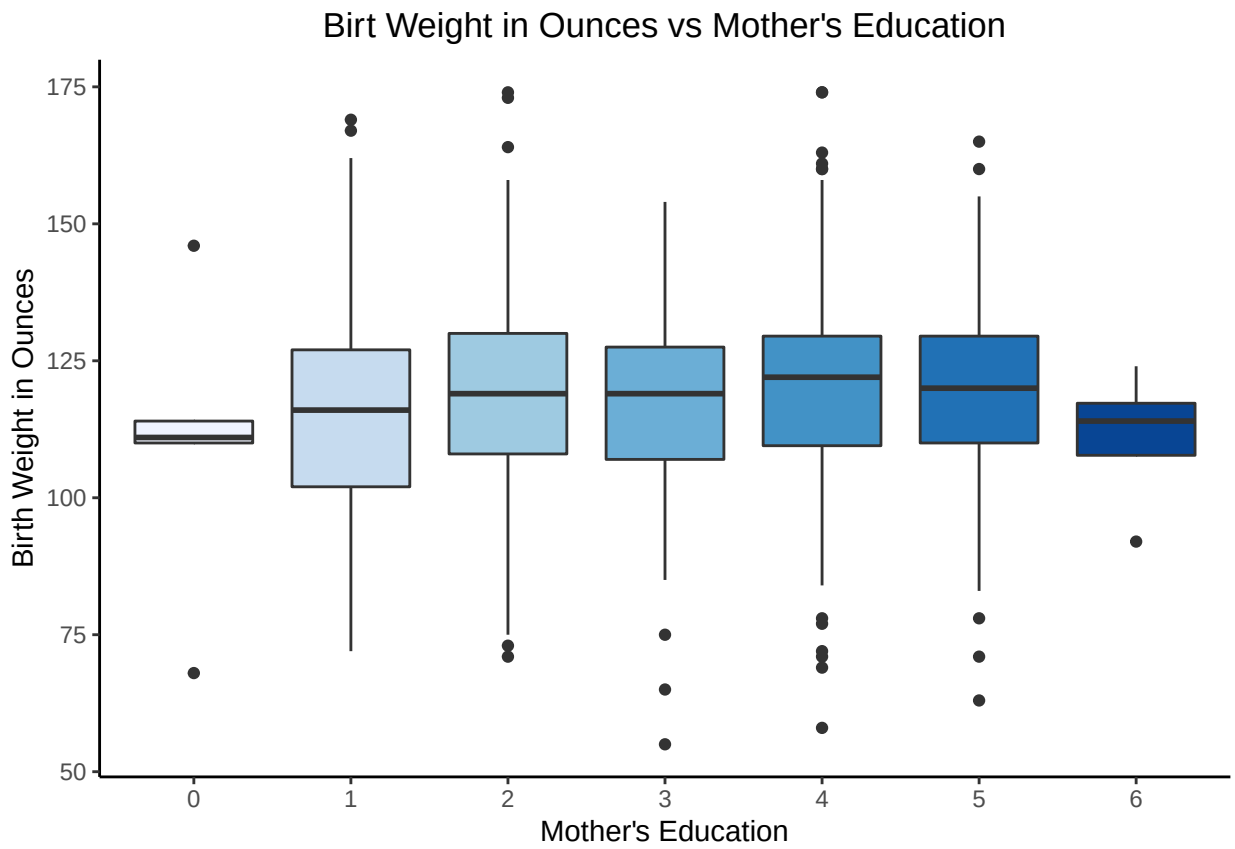

```
## `geom_smooth()` using formula 'y ~ x'
```


Birth Weight in Ounces vs Parity
Birth Weight in Ounces
175
150
125
100
75
50
0
3
6
9
Parity

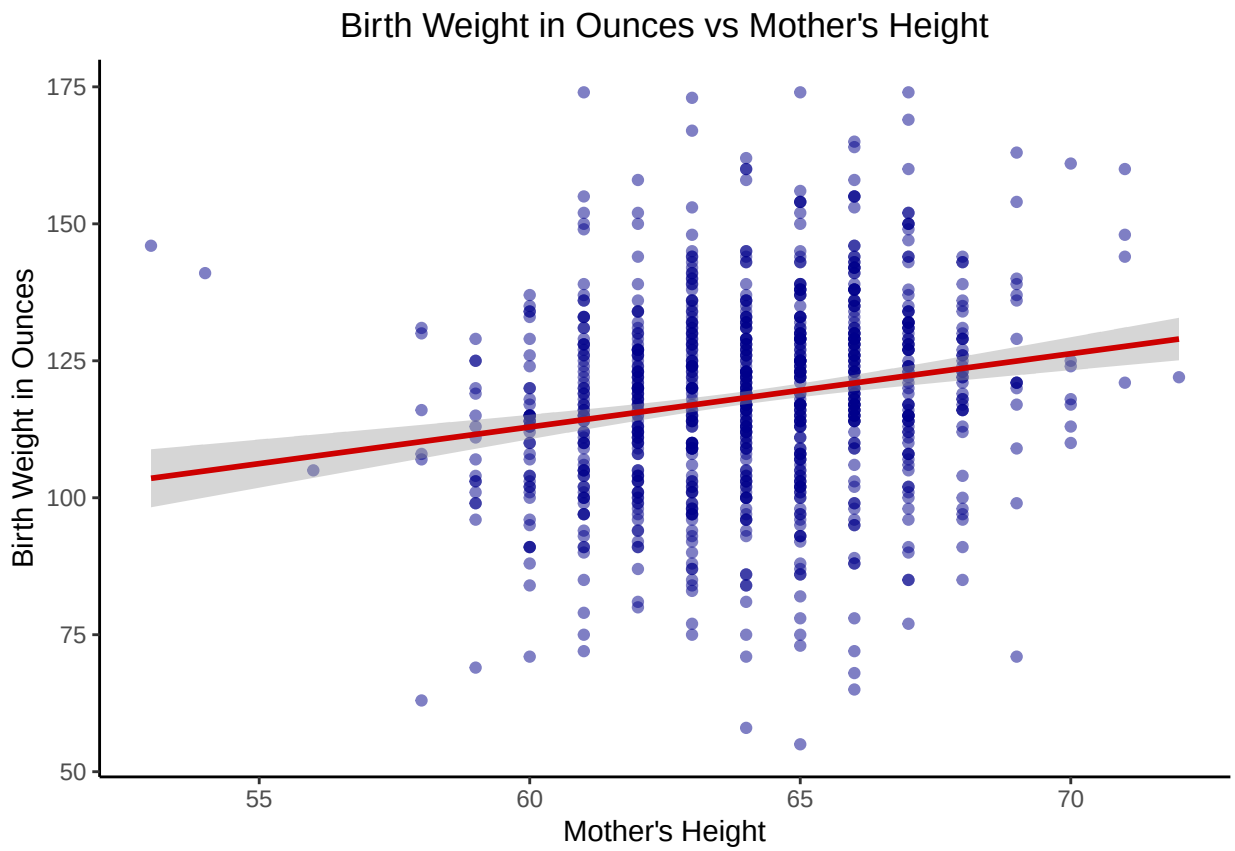



```
## `geom_smooth()` using formula 'y ~ x'
```

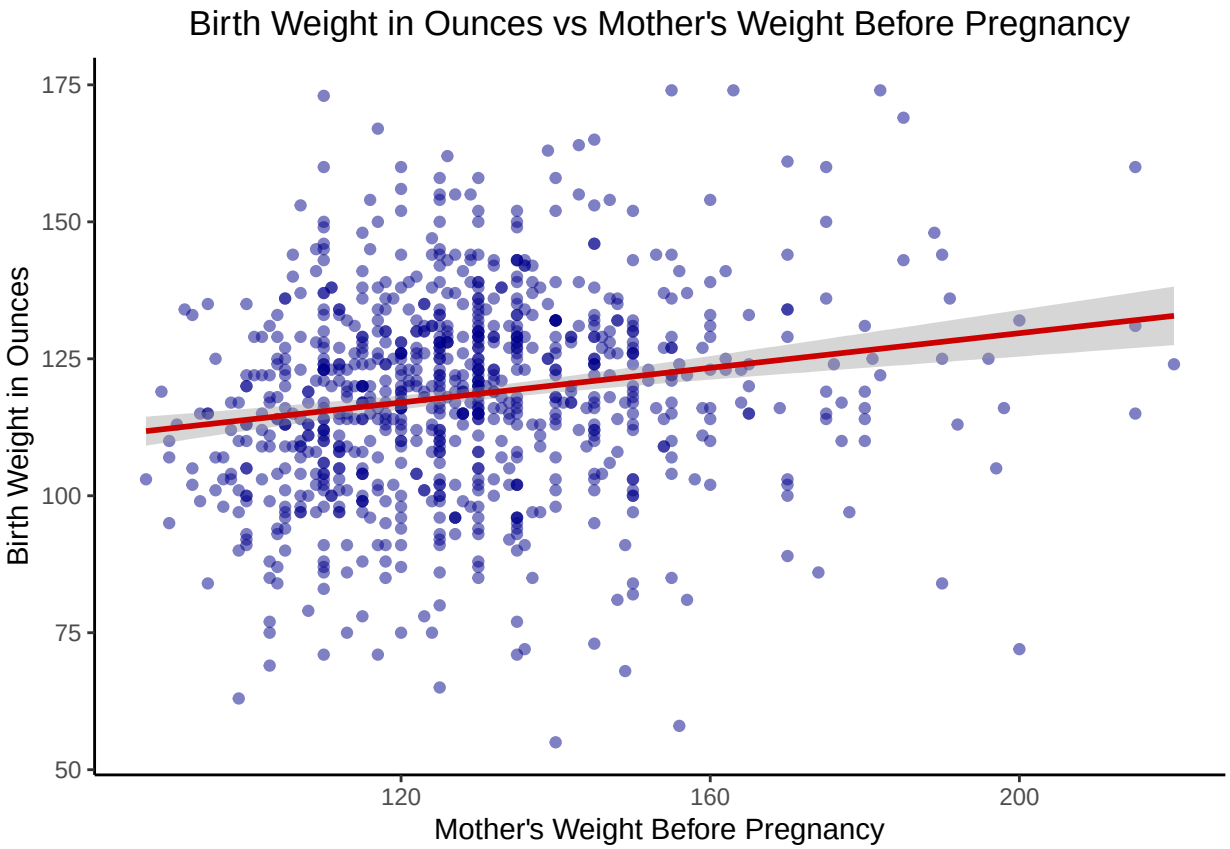

Birth Weight in Ounces vs Mother's Age
Birth Weight in Ounces
175
150
125
100
75
50
20
30
40
Mother's Age



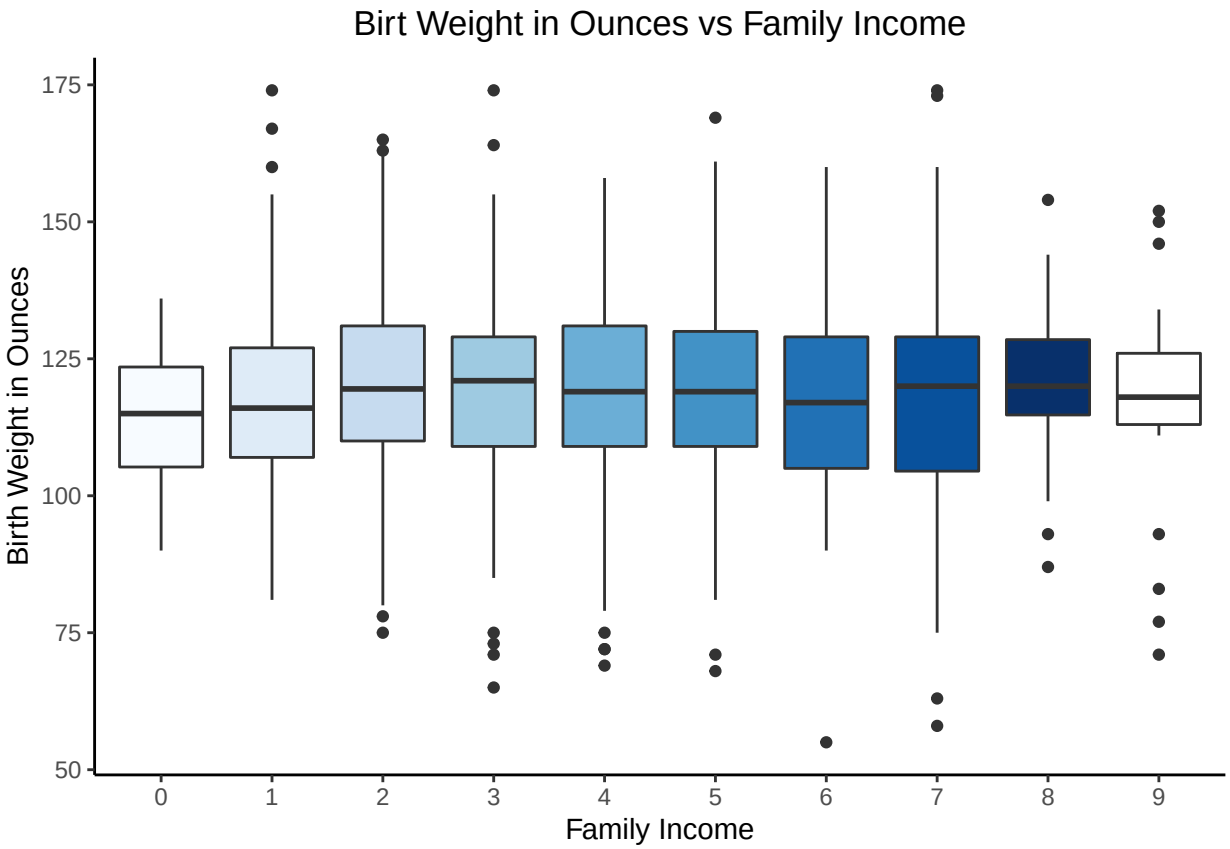

```
## `geom_smooth()` using formula 'y ~ x'
```

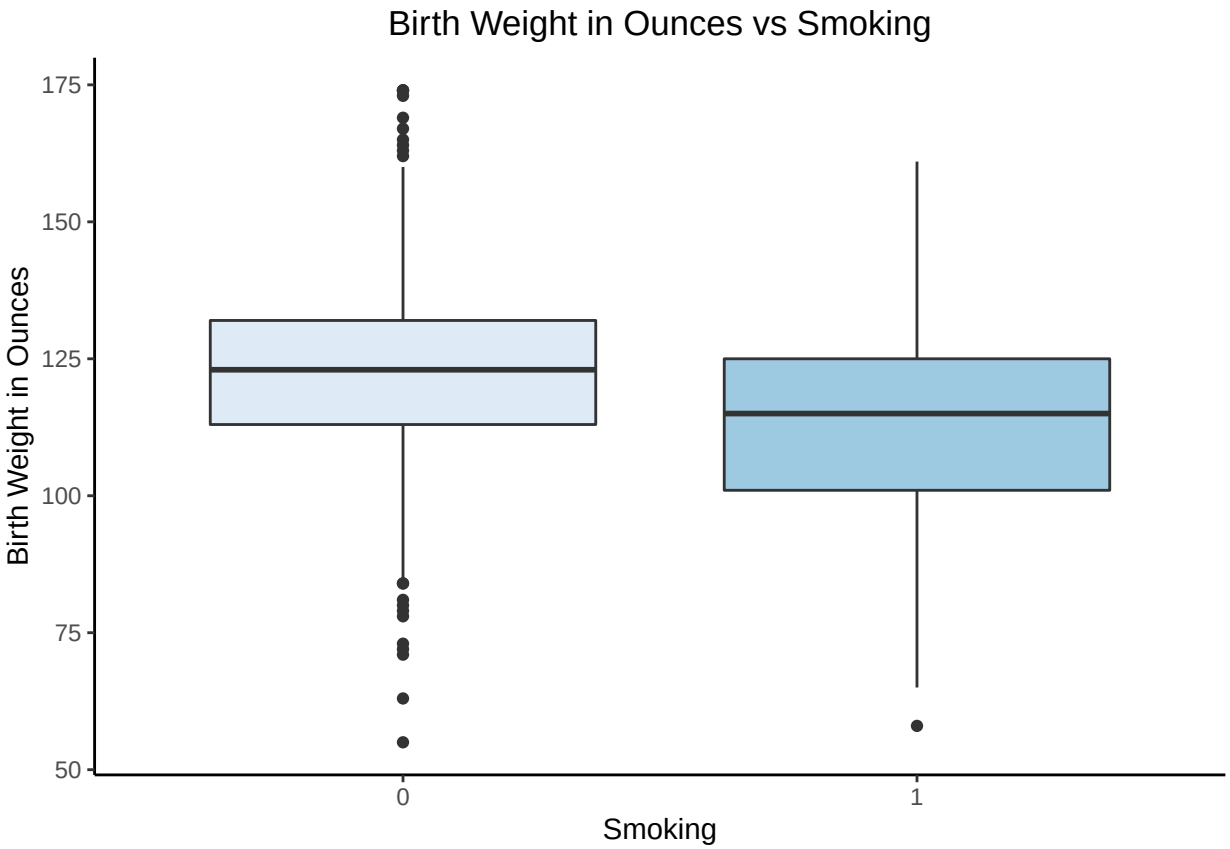


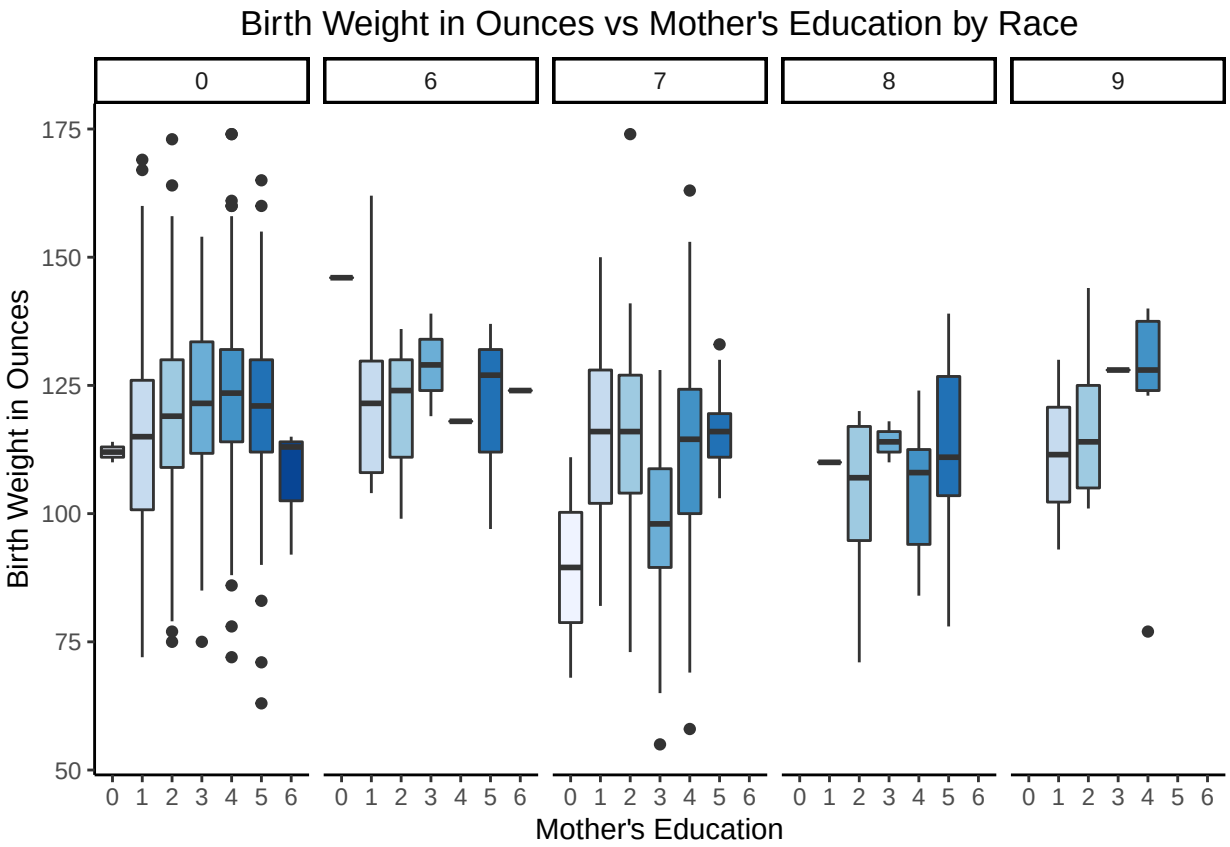

```
## `geom_smooth()` using formula 'y ~ x'
```



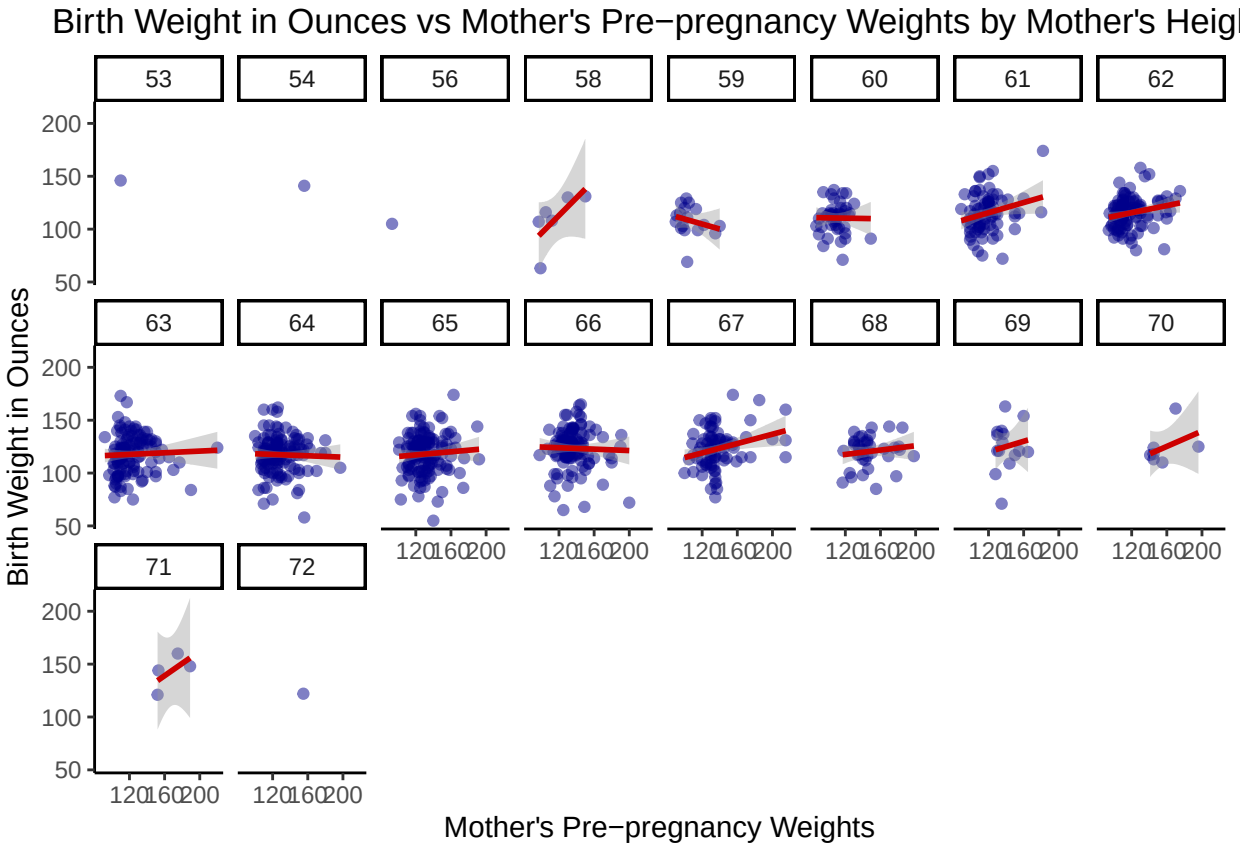

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Blues is 9
## Returning the palette you asked for with that many colors
```

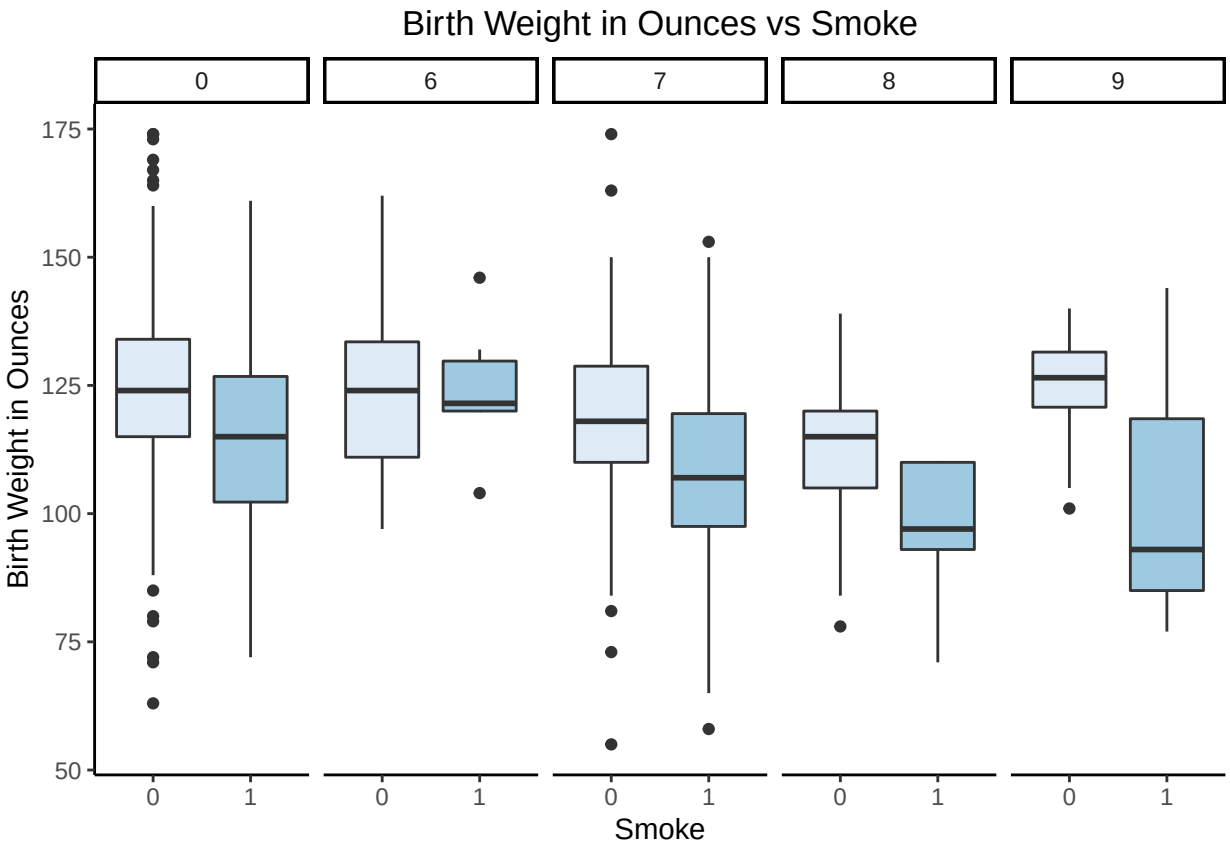

Birt Weight in Ounces vs Family Income
Birth Weight in Ounces
175
150
125
100
75
50
0
1
2
3
4
5
6
7
8
9
Family Income

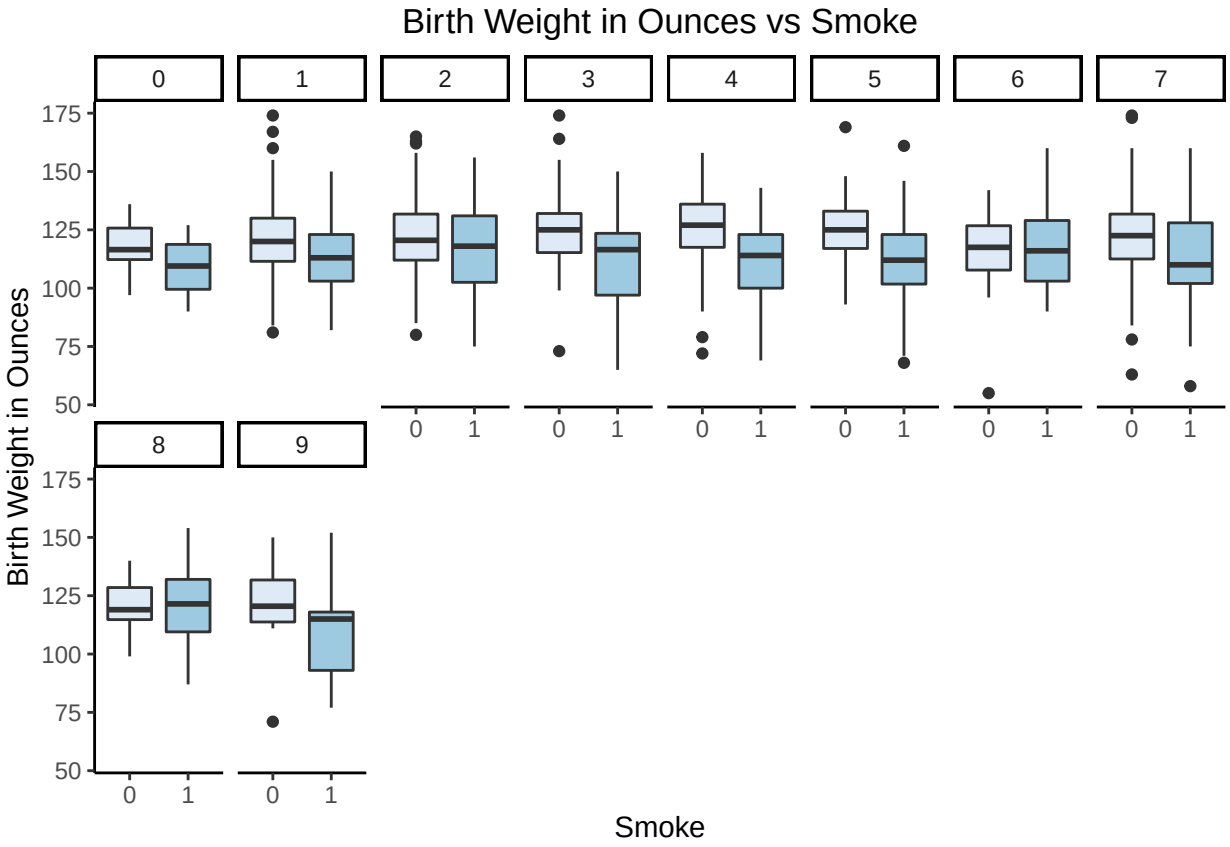

Birth Weight in Ounces vs Smoking
Birth Weight in Ounces
175
150
125
100
75
50
0
1
Smoking

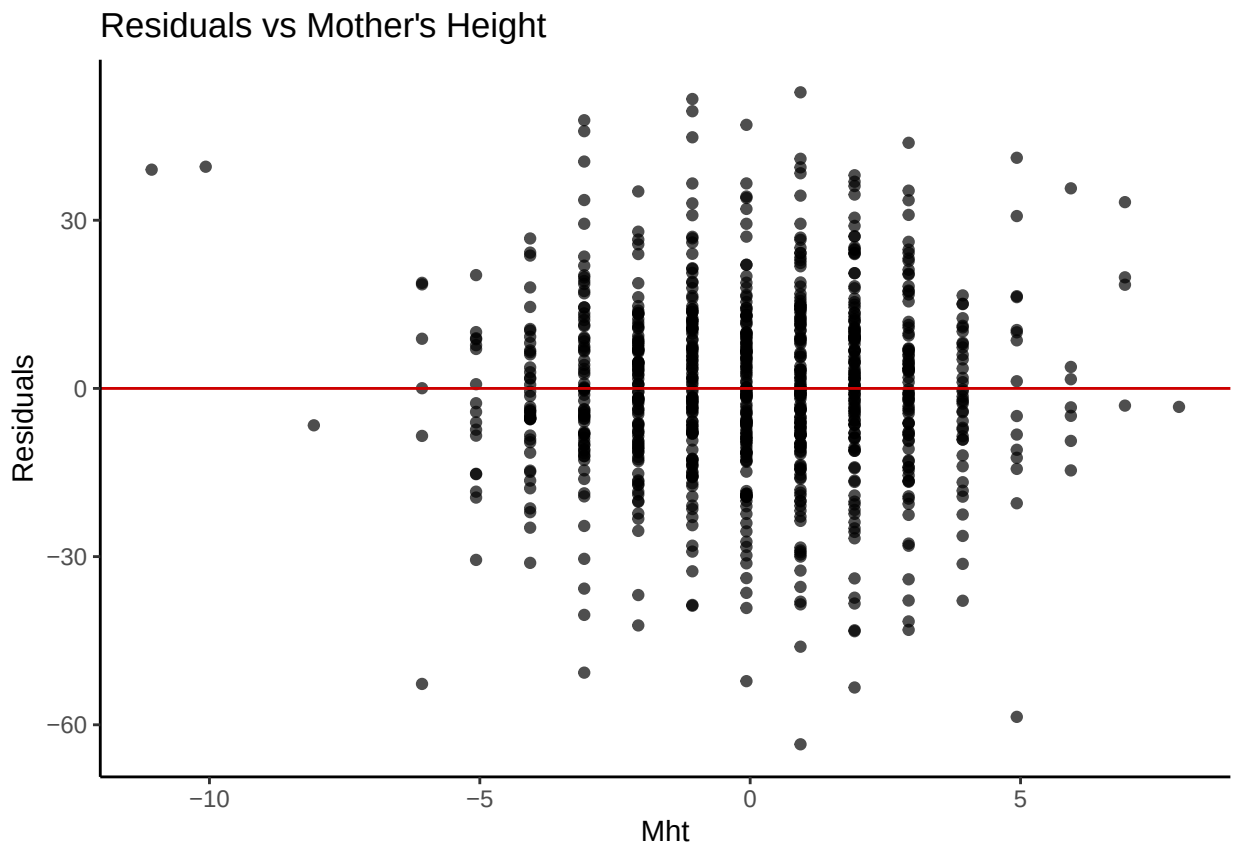


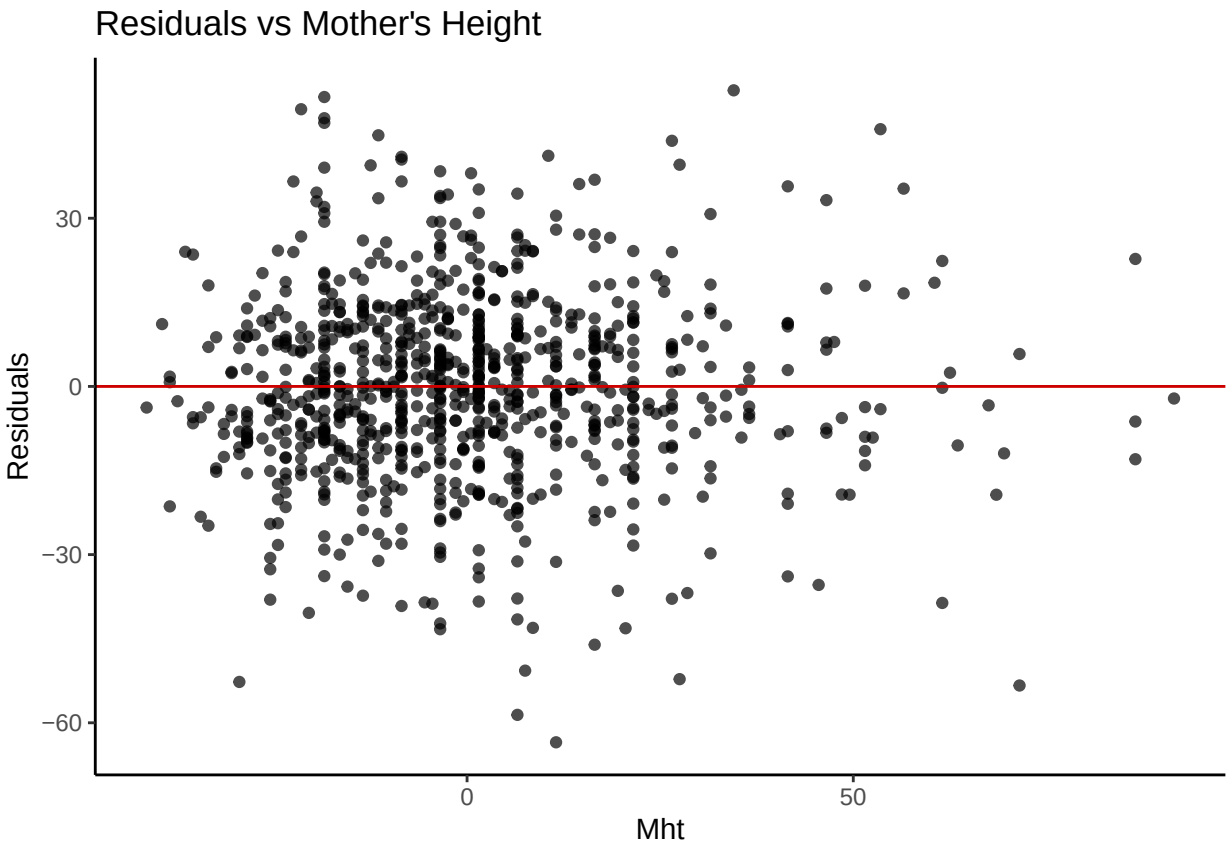

```
## `geom_smooth()` using formula 'y ~ x'
```

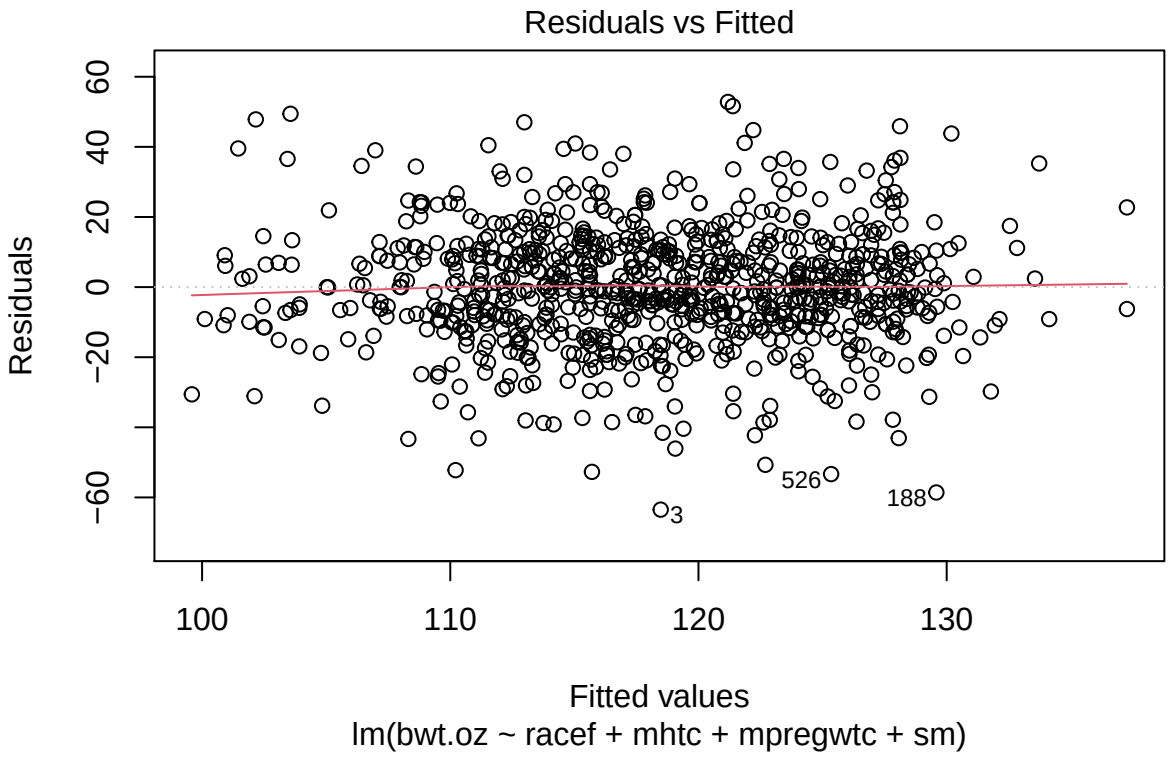

Birth Weight in Ounces vs Mother's Pre−pregnancy Weights by Mother's Heig
53
54
56
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
200
150
100
50
120
160
200
Birth Weight in Ounces
Mother's Pre−pregnancy Weights

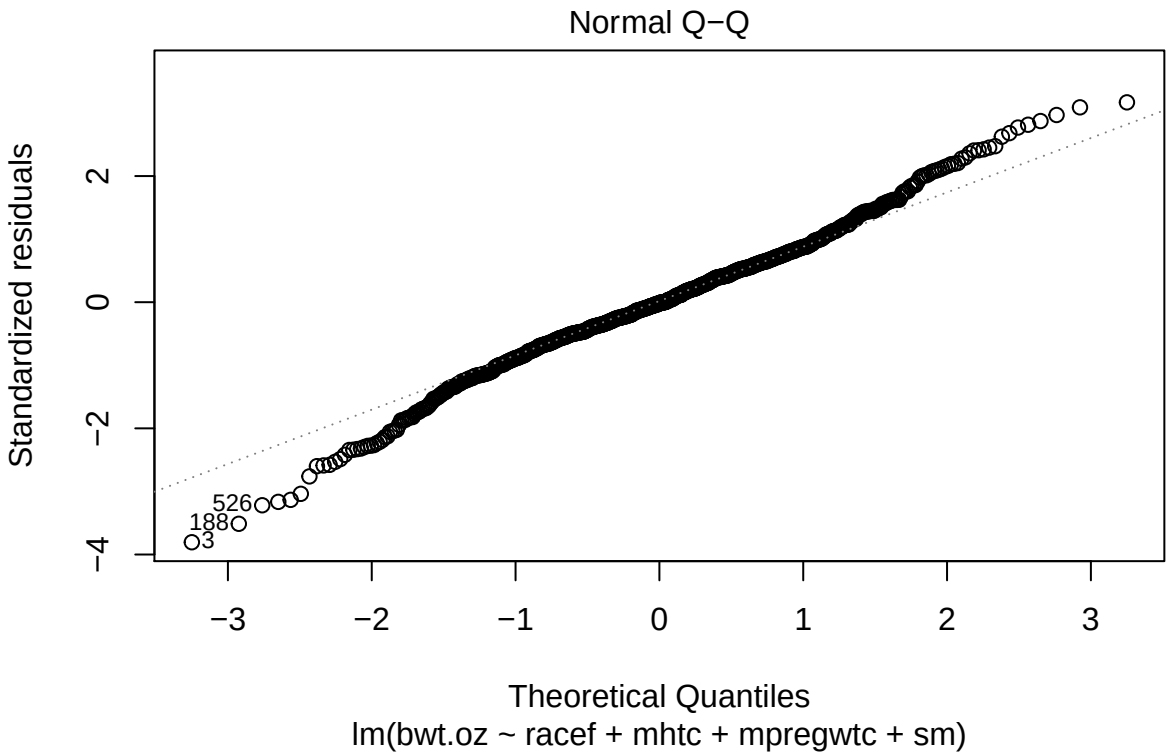

Birth Weight in Ounces vs Smoke
0
6
7
8
9
175
150
125
100
75
50
Birth Weight in Ounces
0
1
Smoke


Birth Weight in Ounces vs Smoke
0
1
2
3
4
5
6
7
8
9
175
150
125
100
75
50
Birth Weight in Ounces
0
1
Smoke

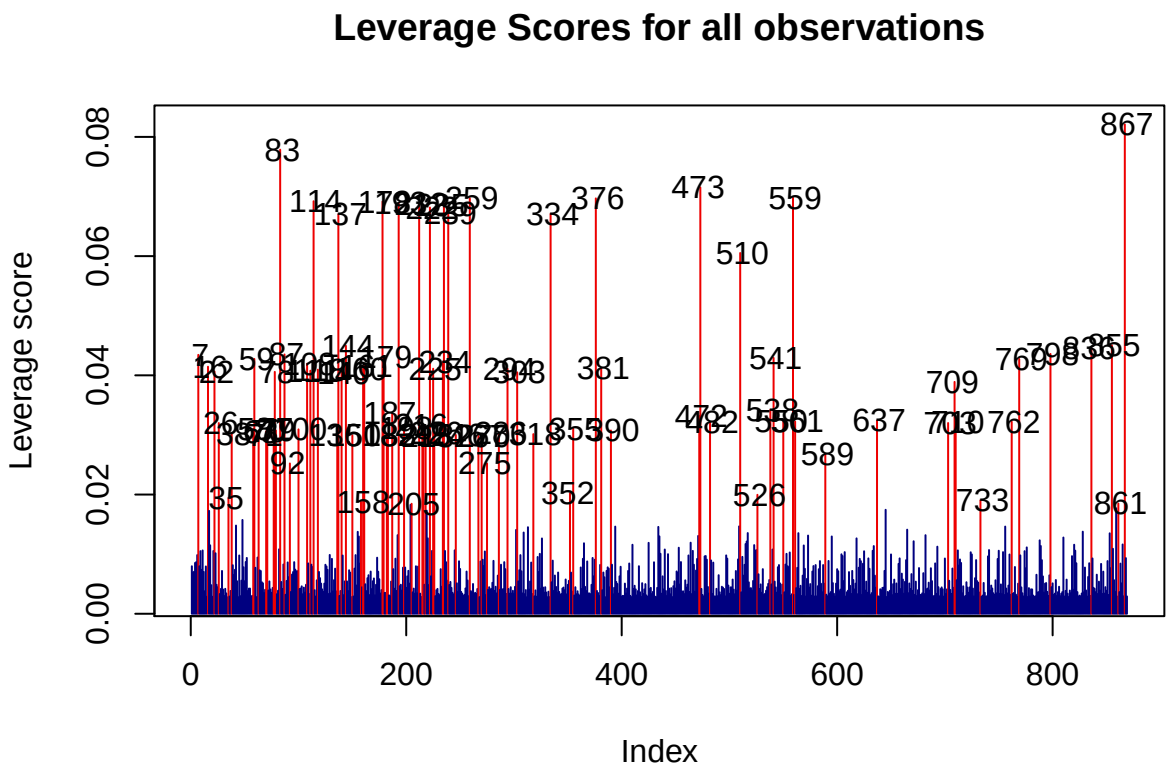

Residuals vs Mother's Height
Residuals
30
0
−30
−60
−10
−5
0
5
Mht

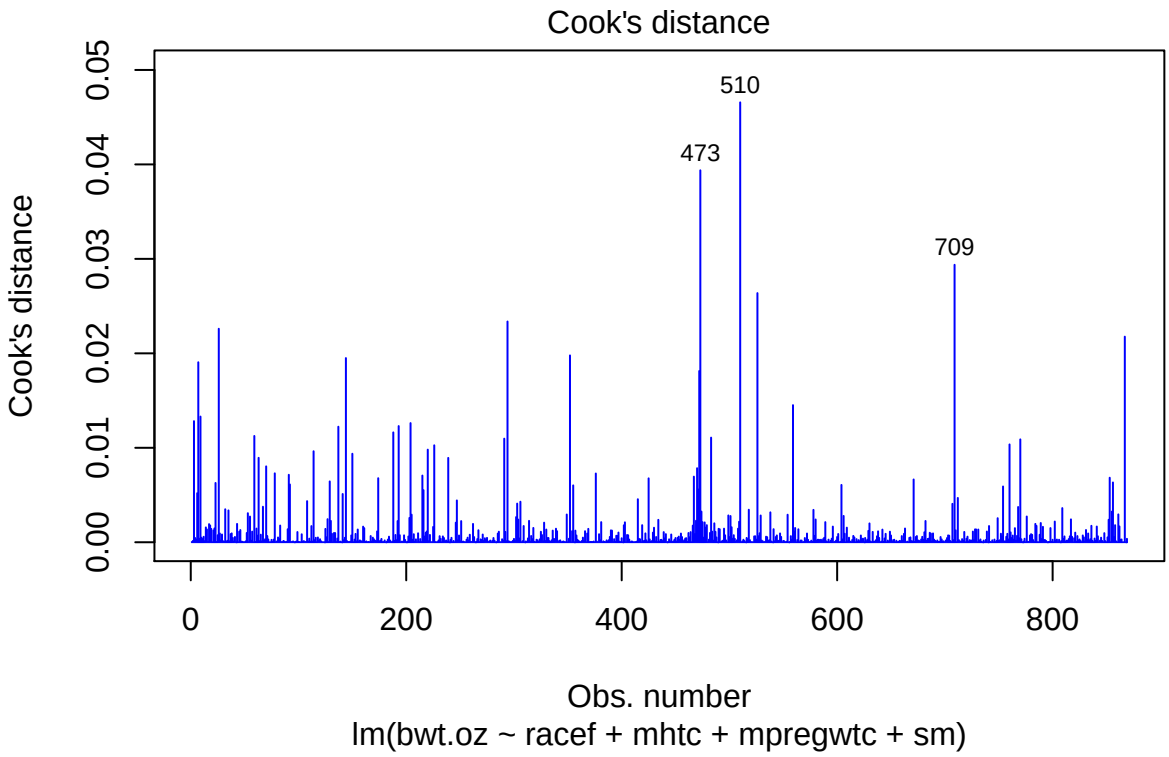

Residuals vs Mother's Height
Residuals
30
0
−30
−60
0
50
Mht

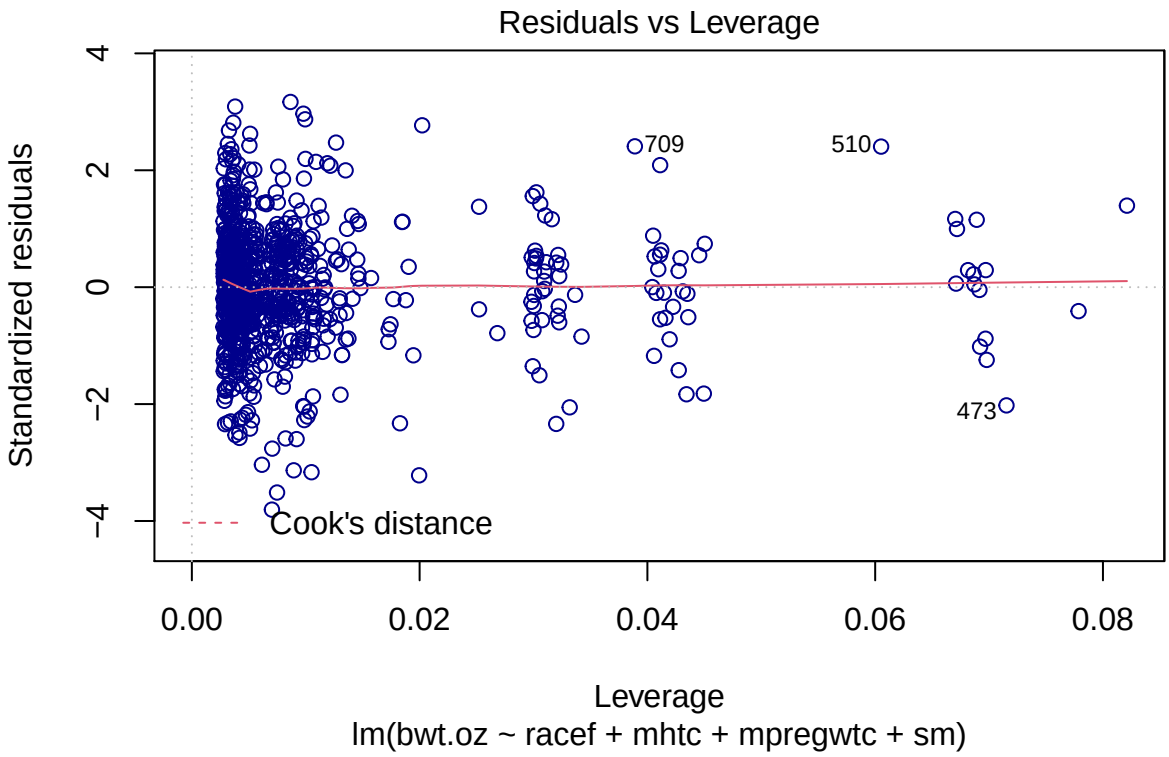

Residuals vs Fitted
Residuals
60
40
20
0
−20
−60
526
188
3
100
110
120
130
Fitted values
lm(bwt.oz ~ racef + mhtc + mpregwtc + sm)




```
## Warning in c(1:n)[lev_scores > (2 * p/n)] + c(rep(2, 4), -2, 2): longer object
## length is not a multiple of shorter object length
```


Leverage Scores for all observations
Leverage score
0.00
0.02
0.04
0.06
0.08
0
200
400
600
800
Index
867
83
473
559
376
359
334
510
541
381
709
762
589
526
733
861
352
275
92
35


Cook's distance
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
510
473
709
0
200
400
600
800
Obs. number
lm(bwt.oz ~ racef + mhtc + mpregwtc + sm)


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
709
510
473
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
lm(bwt.oz ~ racef + mhtc + mpregwtc + sm)