Question 2

Mohammad Anas

Summary

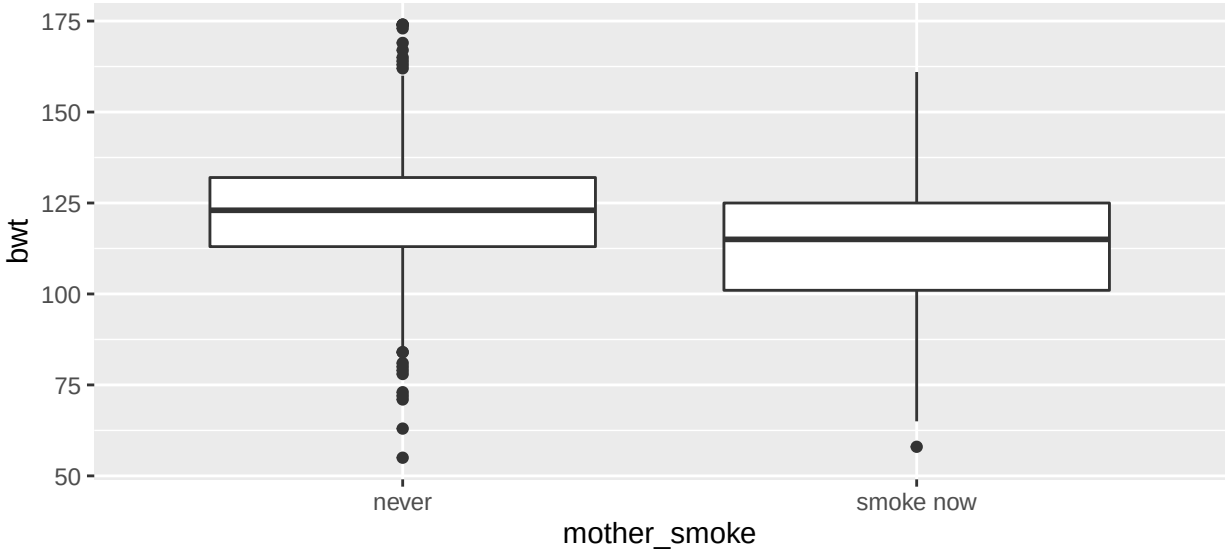
The data set provided to us contains the information about 869 pregnant mothers and their biological and behavioral attributes. The variables of interest in this data are the smoke variable, which indicates whether a mother smoked or not and the bwt.oz variable which tell us the weight of the child in ounces at the time of the birth. We will try to address the following questions throughout this report.

1. Do mothers who smoke tend to give birth to children with low weight as compared to mother who dont smoke? If yes, what is the range in difference of birth weight.
2. Does race(ethnicity) of the mother affect the difference in the birth weight of the child for mother who smoke as compared to mothers who dont smoke.
3. Lastly, we will try to find if there are any other variables that have an interesting association with the birth weight of child.

We conducted a linear regression analysis to answer the above questions and found out that the mothers who smoke do give birth to children with lower birth weight and the difference was found to be statistically significant. However, the race of mothers did not seem to make any changes to our findings for most races. Mother's height and her weight at pregnancy also seemed to have significant affect on the birth weight.

Introduction

During our EDA the difference between birth weight for children whose mothers' smoke as compared to those whose mothers' did not smoke was evident. This can be simply seen through the simple box plot shown below.



To solidify our results we moved on to regression analysis. Through regression we notice that keeping all else constant, a mother who smokes tend give birth to children whose birth weights are on average 16.21 ounces less as compared to children whose mothers don't smoke. According to our 95%, percent confidence interval, we can see that this difference can range from 3.45 ounces to 28.97 ounce (meaning that we can assert with 95% confidence that the true effect of smoking on the birth weight of the new born baby as compared to not smoking lies between this range). We also notice that mother's weight at pregnancy and mother's height also have a statistically significant affect. We note here in our model that all the interaction terms of the smoker and the race variable are statistically insignificant except Mexican. We note an interesting thing here. As compared to Asian- non smoking mothers the Mexican mothers tend to have healthier babies. This positive affect on child's birth weight is further enhanced for the Mexican mothers who smoke. One reason for these counter intuitive results can be the lack of data for Mexican mothers. The below table shows the confidence intervals of standard estimates of all the variables.

Table 1: 95 percent CI of Significant Variables

	2.5 %	97.5 %
(Intercept)	26.02	91.38
mother_smokesmoke now	-28.97	-3.45
mother_raceblack	-10.23	4.99
mother_racemexican	-3.52	16.51
mother_racemix	-4.50	18.65
mother_racewhite	-0.65	13.26
mpregwtc	0.06	0.18
mht	0.42	1.44
mother_smokesmoke now:mother_raceblack	-5.47	22.03
mother_smokesmoke now:mother_racemexican	1.17	41.25
mother_smokesmoke now:mother_racemix	-30.48	19.01
mother_smokesmoke now:mother_racewhite	-6.38	19.67

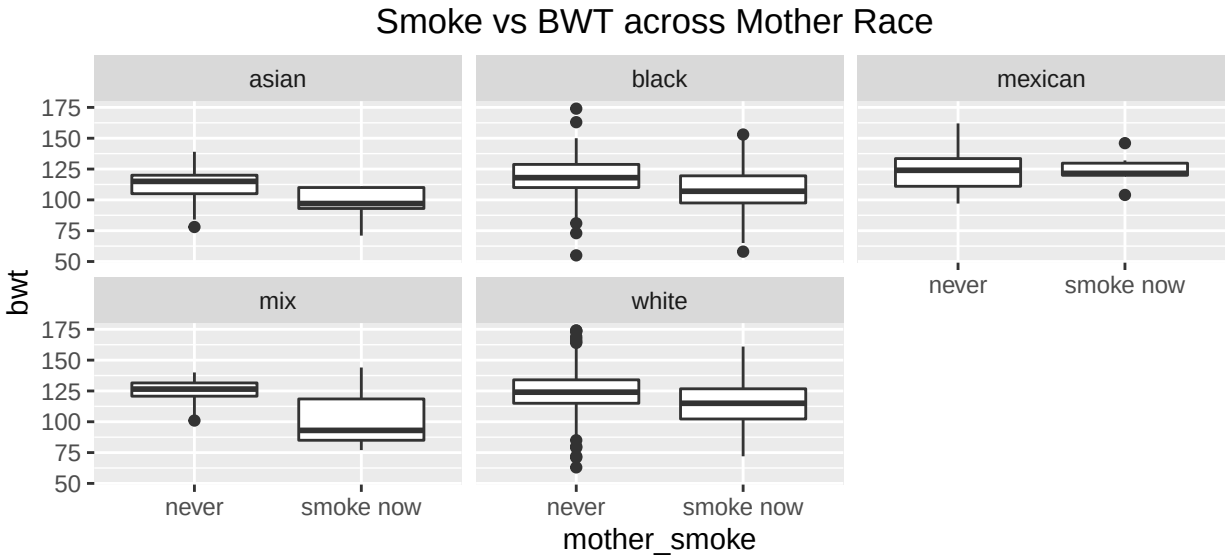
Exploratory Data Analysis

We used the smoking.csv data set provided to us. This is a modified version of the babies.csv data set. All the data entries that had any column with a missing value were removed. Therefore, the smoking.csv data set had no missing values.

During the EDA we found that our dependent variable Birth Weight followed a normal distribution. We also notice that majority of variable have an entirely random relationship with our dependent variable (bwt.oz). These variables include mother's education and mother's age. The variables parity, mother's height and mother's weight at pregnancy show a somewhat linear trend although there is some randomness included. Therefore, we decided to include them in our model.

Given that we notice randomness when plotting scatter plots for most predictor variables against a response variable, we are unable to decipher any appropriate transformations for our x-variables to fit our model well.

To take into account the interaction effects we looked at several plots. Given our variable of interest is the mother_race and her smoking habits, we explore the trend of the affect of smoking on the birth weight across all races. For all the races, the smoking mothers appear to given birth to lower weight babies.For Mexican mother's our box plot indicated that the median value for birth weight for smoking mothers and non-smoking mothers is nearly the same. Given that we notice something worth exploring here and as we specifically want to capture whether the affect of smoking on birth weight differs by race, we deliberately include the interaction term of race and smokers in our model. The trend is shown below:



We also explore other interaction terms. Given that our smoking variable is our variable of interest we also explore its interactions with the other variables as well apart from race. We notice that affect of all other variables remains same for both smokers and non smokers. We also explored the interaction of some other variables which made intuitively made sense . This included mother's education and income, mother's education and mother's weight at pregnancy. However, none of these yielded any interesting results.

The Model

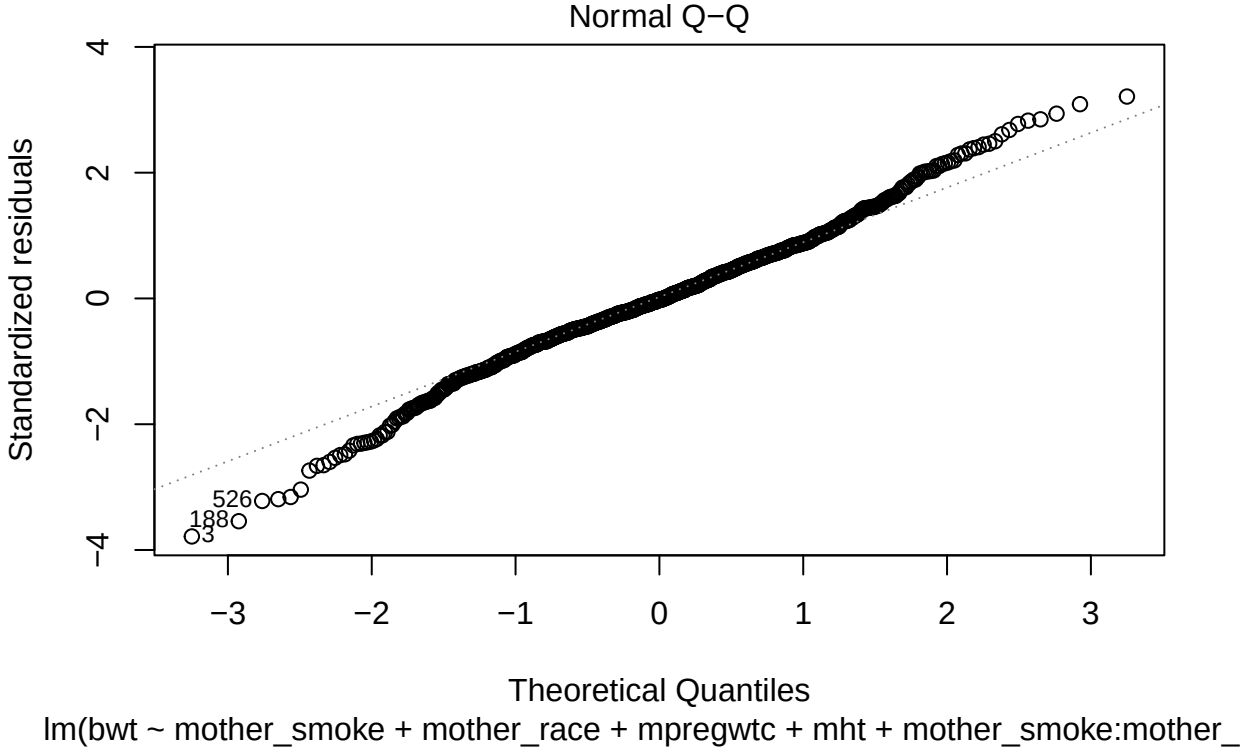
Based on our results from EDA, we preceded to build a linear regression model where our independent variable was bwt (birth weight) and smoke, mother's height, mother's weight at pregnancy, mother_race, income, interaction for smoker and mother's race were included as predictor variables. Used backward selection in R with BIC (Baysean Information Criteria), we were able to remove variables like race and interaction of smokers and race. Again as we are interested in the race variable and its combined affect with smoking on the birth weight we add those variables back to our model. After backward selection our final model looks like this:

$$BirthWeight_i = \beta_0 + \beta_1 * smoker_i + \beta_2 * race_i + \beta_3 * gestation_i + \beta_4 * motherheight_i + \beta_5 * mpregwt_i + \beta_6 * smoker_i * race_i \quad (1)$$

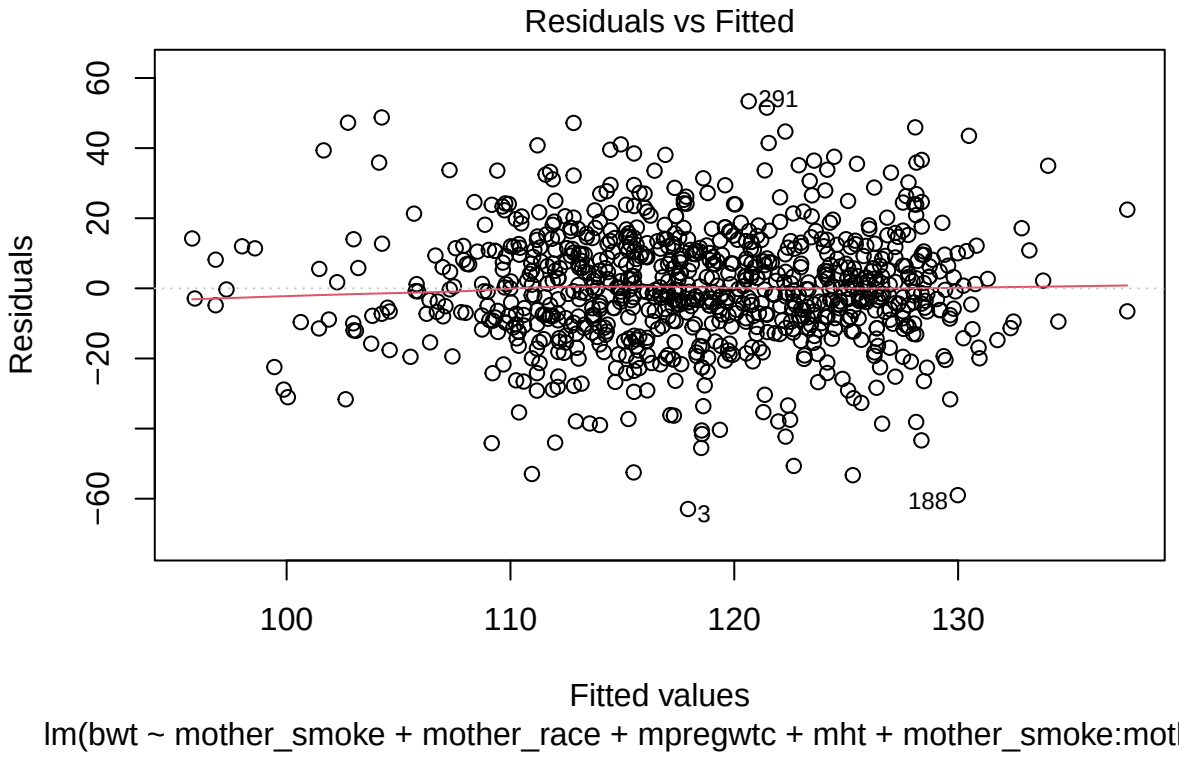
Given that the continuous variables in our model are not centered, the intercept cannot be interpreted as it does not make intuitive sense. We do not that smoking mothers tend to give birth to children whose birth weight on average is 16.21 ounces lower as compared to non smoking mothers. This variable is also significant as indicated by the small p-value. Holding all else constant, one unit increase in mother's weight at pregnancy can lead to a 0.12 ounces increase in the birth of the new born child. According to our model, height also has a positive affect on birth weight increasing it by 0.93 ounces on average. Both the mother's height and her weight at pregnancy had standard estimates that are statistically significant. All of the standard estimates of ethnicity and their interaction term with smoking are statistically insignificant except Mexican as we have discussed before. We do note that although the R-squared is low for our model,the F test results in a lower p-value.This means that our model is good fit for the data as compared to a model with no independent variable. The F-test would make more sense when we compare our model to a model without the interaction term. The F-test comparing the two models results in a large p-value indicating that as compared to a model with interactions does not fit the data significantly better than a model without interactions. The detailed results of our model are provided on the next page.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	58.70	16.65	3.53	0.00
mother_smokesmoke now	-16.21	6.50	-2.49	0.01
mother_raceblack	-2.62	3.88	-0.67	0.50
mother_racemexican	6.50	5.10	1.27	0.20
mother_racemix	7.07	5.90	1.20	0.23
mother_racewhite	6.30	3.54	1.78	0.08
mpregwtc	0.12	0.03	3.70	0.00
mht	0.93	0.26	3.56	0.00
mother_smokesmoke now:mother_raceblack	8.28	7.00	1.18	0.24
mother_smokesmoke now:mother_racemexican	21.21	10.21	2.08	0.04
mother_smokesmoke now:mother_racemix	-5.74	12.61	-0.46	0.65
mother_smokesmoke now:mother_racewhite	6.65	6.64	1.00	0.32

We also notice by looking at the QQ Plot and the "Residuals vs Fitted Plots" that the model satisfies the assumption of normality, constant variance in residuals and independence of residuals. However, the normality assumption is not entirely satisfied, but we can allow this much to happen.



We see an entirely random pattern for the residuals in the fitted vs residuals graph providing us proof for independence of residuals. Moreover, the residuals are distributed equally around the zero line vertically. Its seems that the model has done well on the equal variance assumptions as well.



The linearity assumption also seems to satisfied when we look at the scatter plots of residuals against our continuous predictor variables, as we note a random distribution of points around the origin. The relevant plots have been included in the Appendix section of the report.

In terms of outliers, leverage points and influential points the models seems to do a really good job. As the Cooks Distance for all points is significantly below 0.5, we can safely assume that the model has no influential points. We also notice that the model has a few outliers which are not influential points. However, we chose not to exclude them from our model as they are a part of the real data set and not a result of a human error during the data entry process. The points with leverage greater than 0.014 have been identified as the leverage points. This threshold has been calculated by using the formula below where p is the number of predictors and n is the number of data observations:

$$Thresh - hold = \frac{2 * (p + 1)}{n} \tag{2}$$

We do note that there are few leverage points in the data set but none of these points seem to be influential points. The "Residuals vs Leverage" plot has also been included in our analysis.

To check if our model is faced with the issue of multicollinearity we see calculate the VIF (Variance Inflation Factor) for each of our predictor variables. We do notice a high VIF for our interaction term and our smoking variable. However, this has to happen by default whenever we include interaction terms of dummy variables. As we included two interaction terms for our smoking variable, we do see a high VIF for it. For all the other variables we not that VIF is below 10.

Conclusion and Limitations

From our model we can easily note that smoking does have a statistically significant effect on the birth weight of a new born baby regardless of the ethnicity of the mother except "Mexican". For Mexican we note that the standard estimate was significantly significant for the Mexican and smoker variable. Mother's

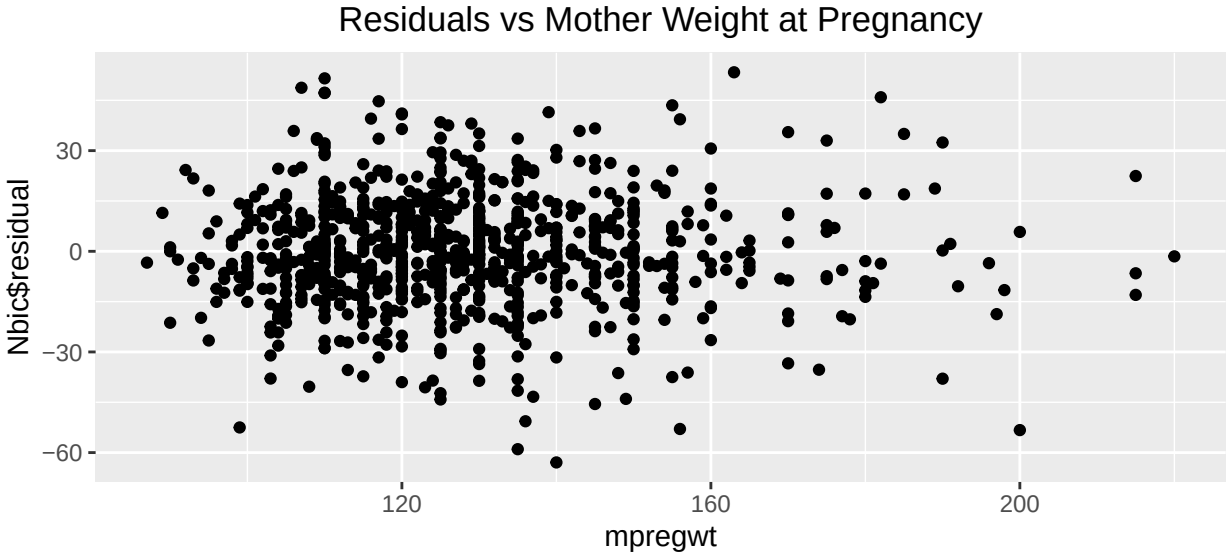
weight at pregnancy and her height also tend to have a statistically significant affect on the child's birth weight. These variable tend to have a positive affect on the birth weight.

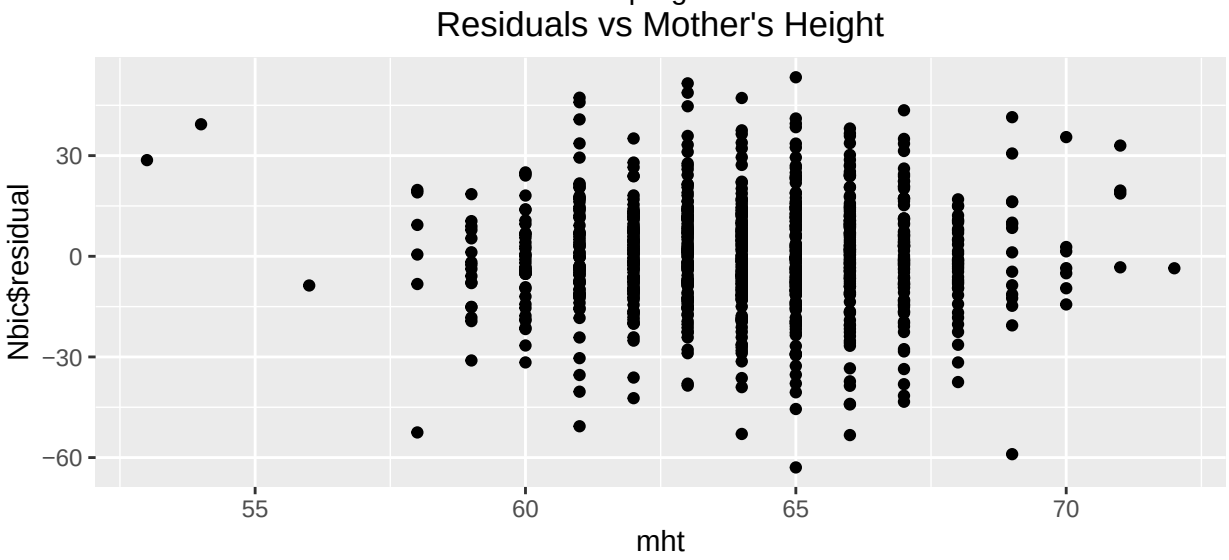
However, there are a few limitations in our model.

1. We note that we seem to lack data for our one of our major variable variable of interest and that is race. Most of the data that have is for white and black mothers, but lack numbers for mix (15 data entries), Asian(34 data entries) and Mexican mothers (25 data entries).
2. Another limitation is that the data set used is just a small subset of the actual data. Many relevant variables like father's biological data was not used. Many data entries were deleted as well to remove missing values.
3. Lastly, the adjusted R-squared for the model is really low.

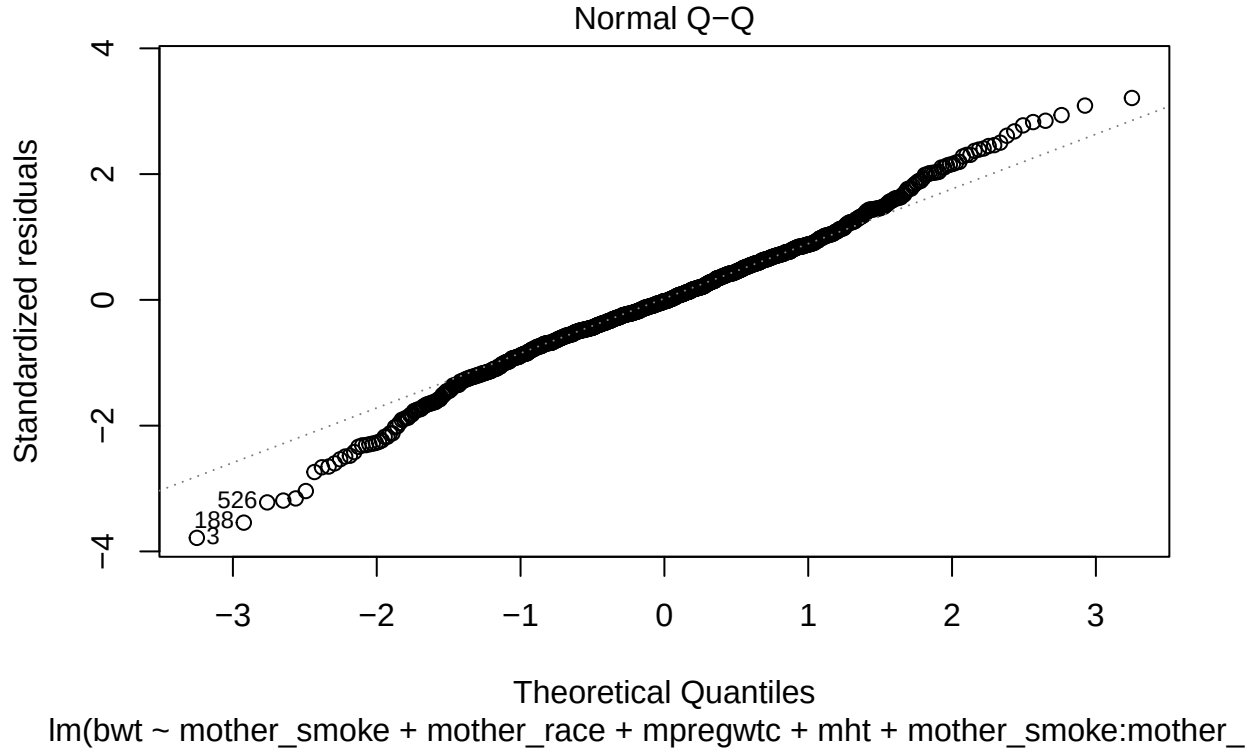
Appendix

To check for the linearity assumption we plot the residuals of our model against the predictor variables. For most predictors, we note that the data points are scattered randomly around the origin.

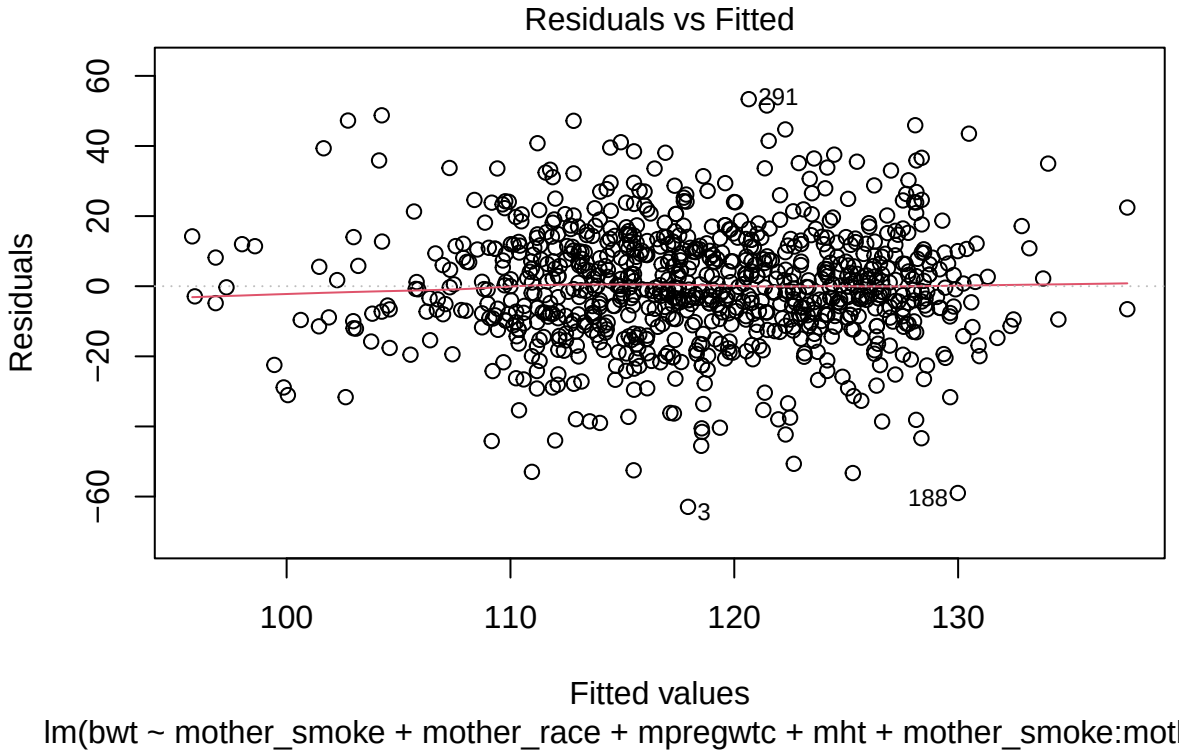




Here is the QQ plot of out model to check for normality assumption.

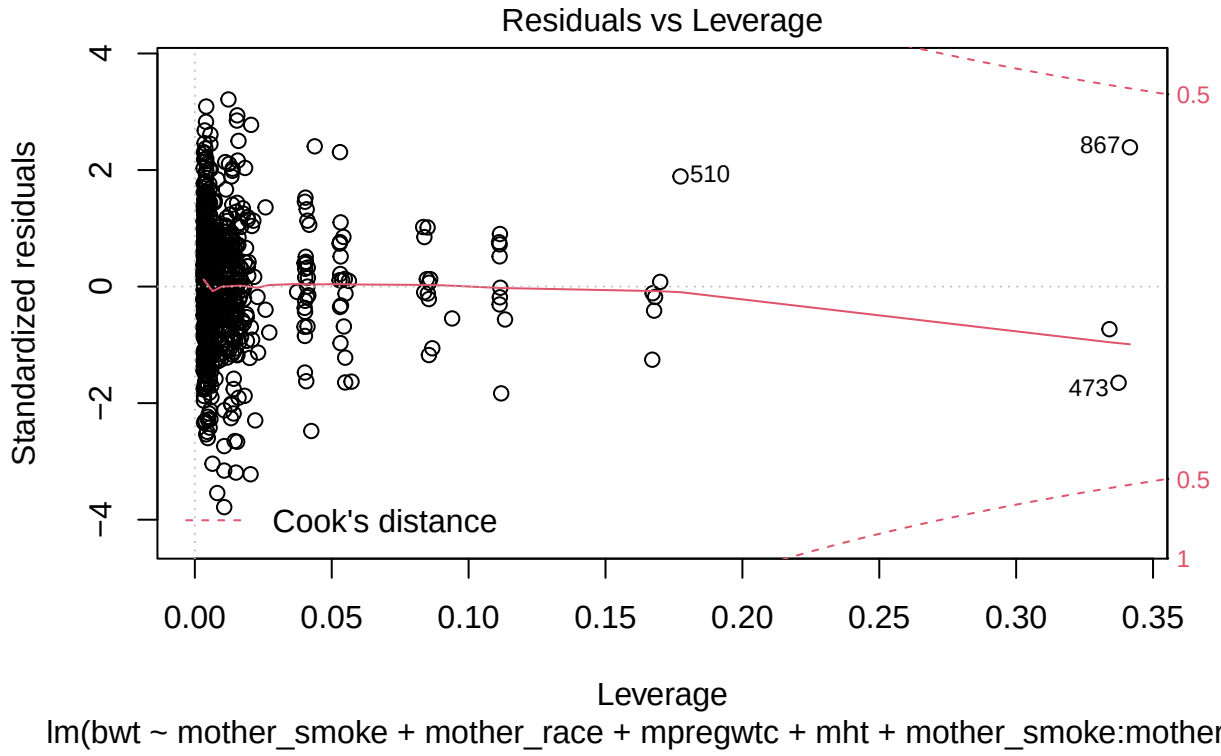


The residuals allow us to check for the linearity, independence and the constant variance assumption. Provided below is the "Residuals vs Fitted" plot which indicates that the independence and the constant variance assumptions are satisfied.



Lastly we used "Leverage vs Standardized Residuals plot" to dedect influential points, leverage points and

outliers.


Residuals vs Leverage
Standardized residuals
Leverage
lm(bwt ~ mother_smoke + mother_race + mpregwtc + mht + mother_smoke:mother_
Cook's distance
0.5
0.5
1
867
510
473
4
2
0
−2
−4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35