Problem Set 2

Elana Nelson

March 4, 2019

Model Building

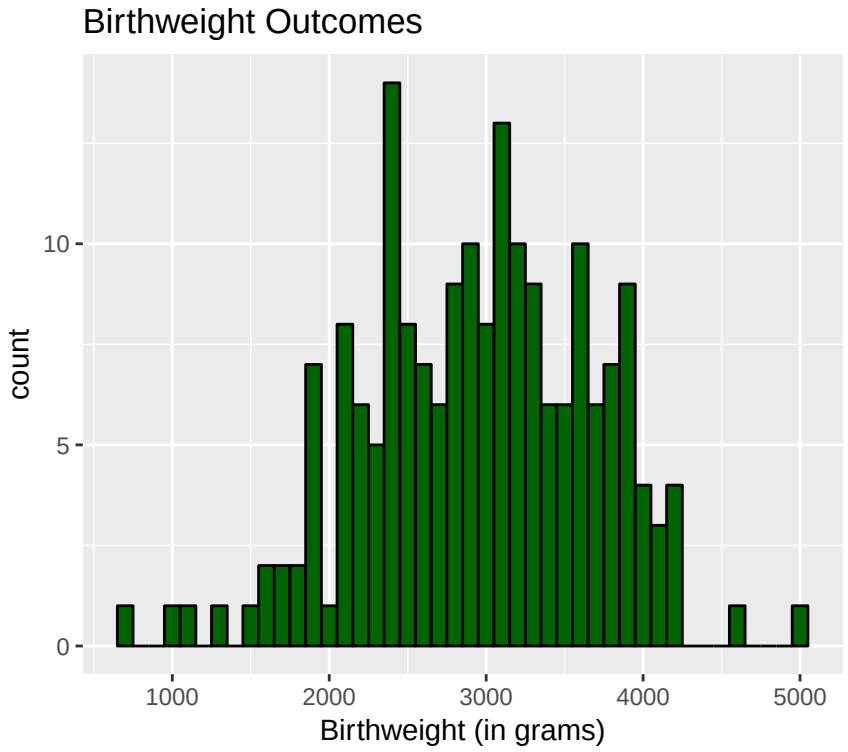
The data considers different factors of low birth weight. The goal is to build a model to predict birth eight.

1. We begin by using number summaries and graphs to explore relationships of variables in the data set and birthweight, bwt.

Table 1:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age of Mother (lbs)	189	23.238	5.299	14	19	26	45
Weight at Last Menstraul Period	189	129.815	30.579	80	110	140	250
Physician Visits in First Trimester	189	0.794	1.059	0	0	1	6
Birthweight of Baby (grams)	189	2,944.656	729.022	709	2,414	3,475	4,990

We can also visualize the birthweight outcomes of the babies with a histogram.



We also have different indicator variables including whether or not the birthweight is considered low, race, smoking habits, premature labors, history of hypertension, and presence of uterine irritability. We can get a general overview of these variables with a frequency table.

And now we visualize relationships between different variables and birthweight.

x		Frequency	Percent	Cum. percent
low.weight$low :	0	130	68.8	68.8
	1	59	31.2	100.0
	Total	189	100.0	100.0

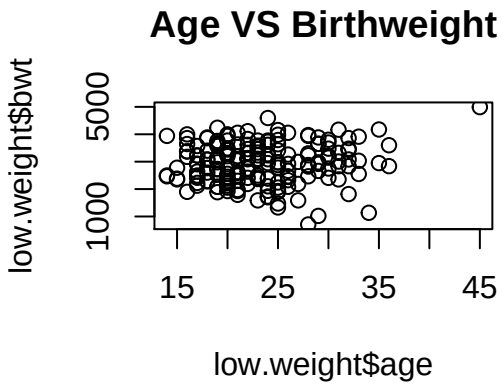
x		Frequency	Percent	Cum. percent
low.weight$race :	1	96	50.8	50.8
	2	26	13.8	64.6
	3	67	35.4	100.0
	Total	189	100.0	100.0

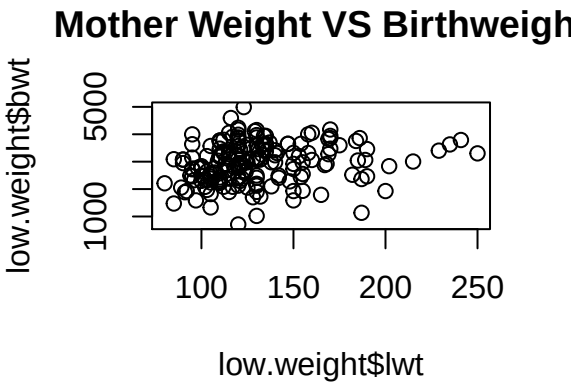
x		Frequency	Percent	Cum. percent
low.weight$smoke :	0	115	60.8	60.8
	1	74	39.2	100.0
	Total	189	100.0	100.0

x		Frequency	Percent	Cum. percent
low.weight$ptl :	0	159	84.1	84.1
	1	24	12.7	96.8
	2	5	2.6	99.5
	3	1	0.5	100.0
	Total	189	100.0	100.0

x		Frequency	Percent	Cum. percent
low.weight$ht :	0	177	93.7	93.7
	1	12	6.3	100.0
	Total	189	100.0	100.0

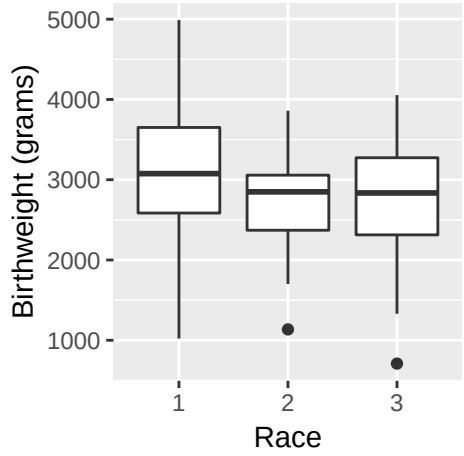
x		Frequency	Percent	Cum. percent
low.weight$ui :	0	161	85.2	85.2
	1	28	14.8	100.0
	Total	189	100.0	100.0

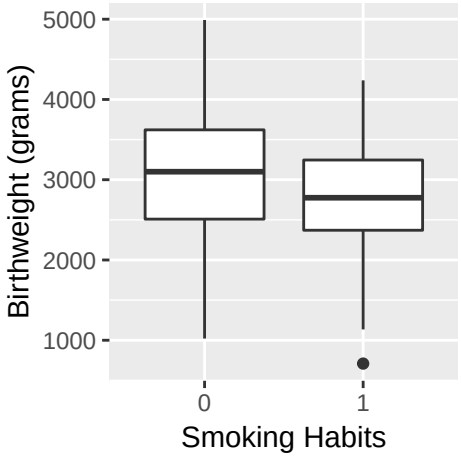


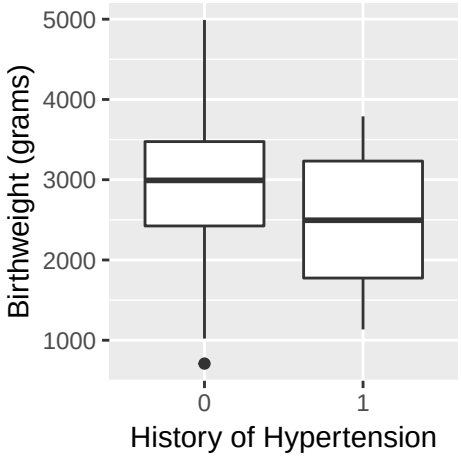


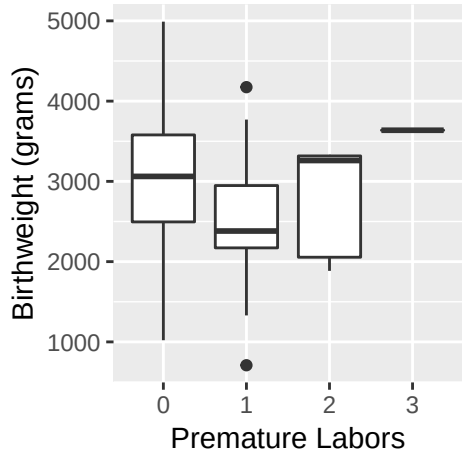
The plots of age of mother vs birthweight do not initially show a clear relationship. The same is true for mother weight vs birthweight.

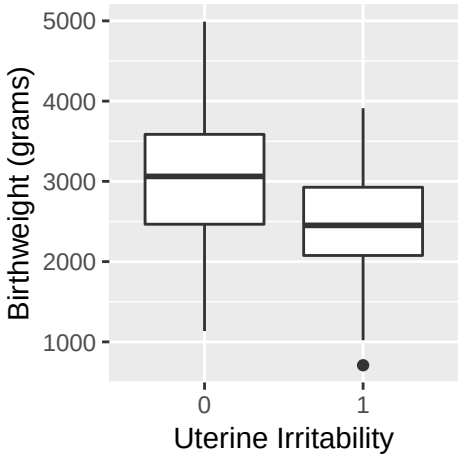
We now look at the relationship between birthweight and the different indicator variables including race, premature labors, smoking habits, history of hypertension, and history of uterine irritability.











The above density graphs show the relationship between birthweight and various characteristics of the mother: * It appears that Black women tend to have lower birthweight babies compared to both the White and other category * Smokers tend to have lower birthweight babies than non smokers * Those with a history of

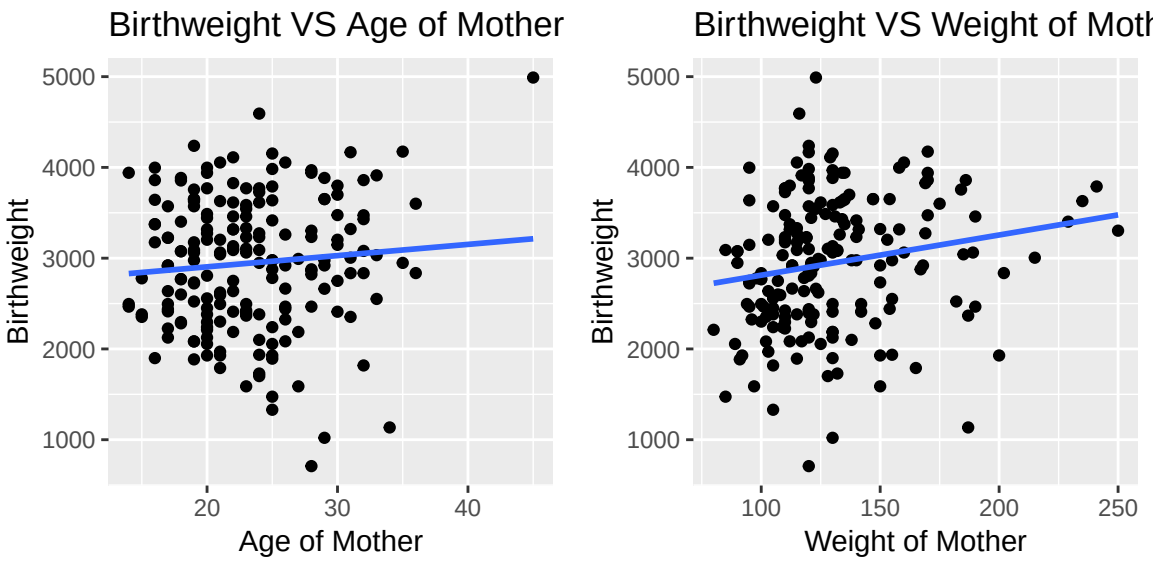
hypertension tend to have lower birthweight babies than those without.

2. We now need to make sure that our low, race, and ui variables are factors with correct names.

3. We now begin model building by looking at simple linear regressions for each of the 8 predictor variables. We will also examine relevant plots and create a nice combined table of summary stats.

Variable	Estimate	p-Value	Conf.Low	Conf.High
age	12.3643	0.2188	-7.4035	32.1322
lwt	4.4293	0.0105	1.0499	7.8086
factor(race)Black	-384.0473	0.0159	-695.5019	-72.5927
factor(race)Other	-299.7247	0.0091	-523.9878	-75.4615
smoke	-281.7133	0.0092	-492.7338	-70.6927
ptl	-228.6506	0.0335	-439.2600	-18.0411
ht	-435.5607	0.0449	-861.0973	-10.0241
uiYes	-580.1801	0.0001	-863.3298	-297.0304
ftv	40.0971	0.4258	-59.0172	139.2114

Graphically



- The plot of age vs birthweight shows a weak relationship, as is confirmed by the summary table showing that age is insignificant to the outcome.
- The plot of weight of mother vs birthweight shows a slightly positive relationship. The p value of this relationship shows that this relationship is significant

4. Commenting on significance of variables

Considering the pvalues of each simple relationship, it can be determined that the following variables may be significant to the outcome:

- Weight of Mother (lwt)
- Race, Black
- Race, Other
- Smoking Habits (smoke)

- Number of Premature Labors (ptl)
- History of Hypertension (ht)
- Uterine Irritability (ui)

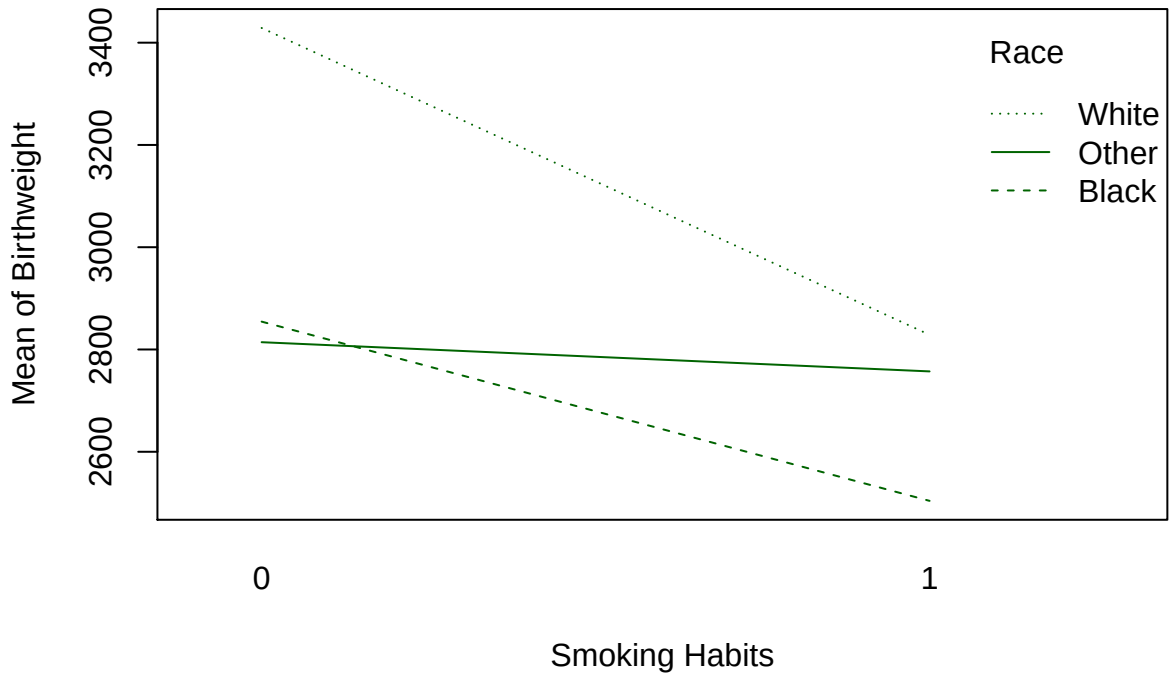
It is worth noting however that the intercept of age of mother and weight of mother is at zero, and it may make more sense to mean-center these to better understand their influence on the outcome.

5. Explore the possibility of interaction between smoking and race. Display a graph that would allow you to explore this and then run a regression with the interaction term. Interpret the results of this model.

There is a helpful interaction function that can help us generate an interaction term between two categorical variables. There is also an interaction plot built into r that will be helpful for visualizing relationships.

	Nonsmoker	Smoker
White	3428.750	2828.731
Black	2854.500	2504.000
Other	2814.236	2757.167

We can see with this cross tabulation table that there is an interaction between race and smoking habits. The mean of the outcome, birthweight, decreases for each race from white in different amounts depending on smoking habits. We can also visualize these interactions with an interaction plot.



This interaction plot shows us that there is evidence of an interaction based on the crossing of the lines. We at this point should run a regression with the interation term. We can create an interaction term with the interaction function.

term	estimate	std.error	statistic	p.value
(Intercept)	3428.7500	103.0326	33.278291	0.0000000
interactionBlack.0	-574.2500	199.5218	-2.878131	0.0044766
interactionOther.0	-614.5136	138.2328	-4.445498	0.0000152
interactionWhite.1	-600.0192	139.9938	-4.286042	0.0000294
interactionBlack.1	-924.7500	239.4262	-3.862359	0.0001557
interactionOther.1	-671.5833	222.5759	-3.017322	0.0029135

According to the model, every interaction term is statistically significant and results in a large decrease in birthweight from the intercept (white, nonsmokers). There is a very large decrease in birthweight for Black nonsmokers.

6. Build a multiple regression model with what you have found in problems 4 and 5. Do the coefficients change from the simple regressions? Comment on both direction and magnitude changes.

Considering the significance of the simple linear regression, and the significance of the interaction term, we can build a multiple regression model that includes the variables that remained significant to the outcome, including: * Weight of Mother (lwt) * Race, Black * Race, Other * Smoking Habits (smoke) * Number of Premature Labors (ptl) * History of Hypertension (ht) * Uterine Irritability (ui) * Interaction

term	estimate	std.error	statistic	p.value
(Intercept)	2983.531079	255.991814	11.6547909	0.0000000
lwt	3.670659	1.697372	2.1625547	0.0319022
ptl	-56.292699	100.116704	-0.5622708	0.5746350
ht	-537.207208	200.945252	-2.6734009	0.0082031
uiYes	-532.084076	138.717882	-3.8357281	0.0001733
interactionBlack.0	-503.611889	190.304939	-2.6463417	0.0088616
interactionOther.0	-480.961822	135.535852	-3.5485948	0.0004947
interactionWhite.1	-467.696806	136.198467	-3.4339359	0.0007392
interactionBlack.1	-937.987752	226.957992	-4.1328694	0.0000550
interactionOther.1	-480.709444	214.813832	-2.2377956	0.0264675

Let's now compare this multiple regression model to the simple linear regressions found earlier.

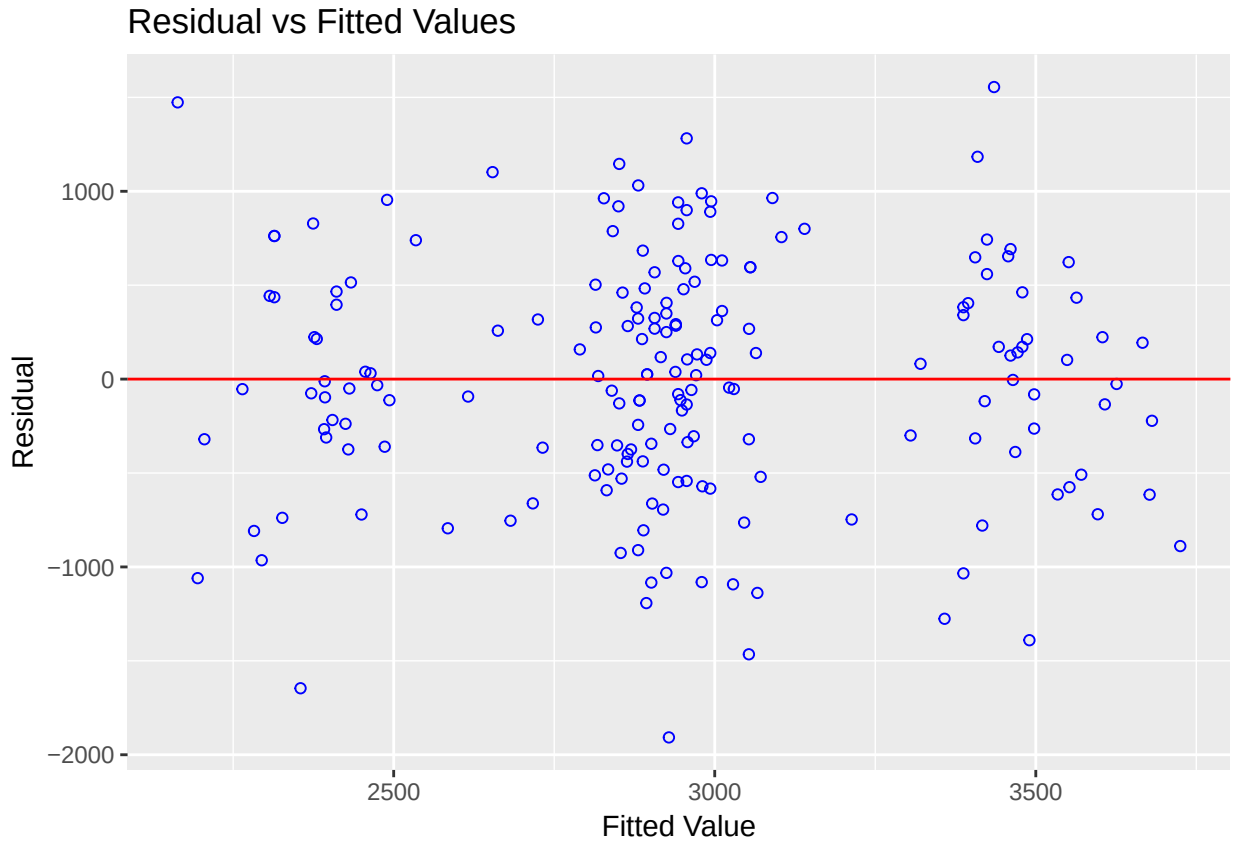
term	estimate	p.value	conf.low	conf.high
age	12.364332	0.2187898	-7.403527	32.13219
lwt	4.429264	0.0104807	1.049927	7.80860
factor(race)Black	-384.047276	0.0159369	-695.501887	-72.59266
factor(race)Other	-299.724658	0.0090814	-523.987812	-75.46150
smoke	-281.713279	0.0091557	-492.733821	-70.69274
ptl	-228.650555	0.0335101	-439.259967	-18.04114
ht	-435.560735	0.0448941	-861.097339	-10.02413
uiYes	-580.180124	0.0000773	-863.329848	-297.03040
ftv	40.097141	0.4258379	-59.017158	139.21144

The most interesting difference between these two models is the drastic change in the effect of the count of premature labors. In the simple linear model it had a much greater negative effect on the outcome than in the multiple regression model. This may be due to interactions with the other variables, or that other variables are able to explain the outcome better in the multiple regression model. The count of premature

labors also becomes insignificant in the multiple regression model. The inclusion of interaction terms may explain the outcome including premature labors, thus making ptl insignificant.

7. Use the plots we have identified to check the model fit. a. Are the assumptions of linear regression met by this? b. How does this model fit? c. Comment on if you see any possible outliers or collinearity

To begin checking model fit, we must first address the underlying assumptions of linear regression - the residuals are random, and normally distributed about zero with homogeneous variance. We can test this with a standardized residual plot and some score and f tests for the variance.



Score Test for Heteroskedasticity ——————————— Ho: Variance is homogenous Ha: Variance is not homogenous

Variables: fitted values of bwt

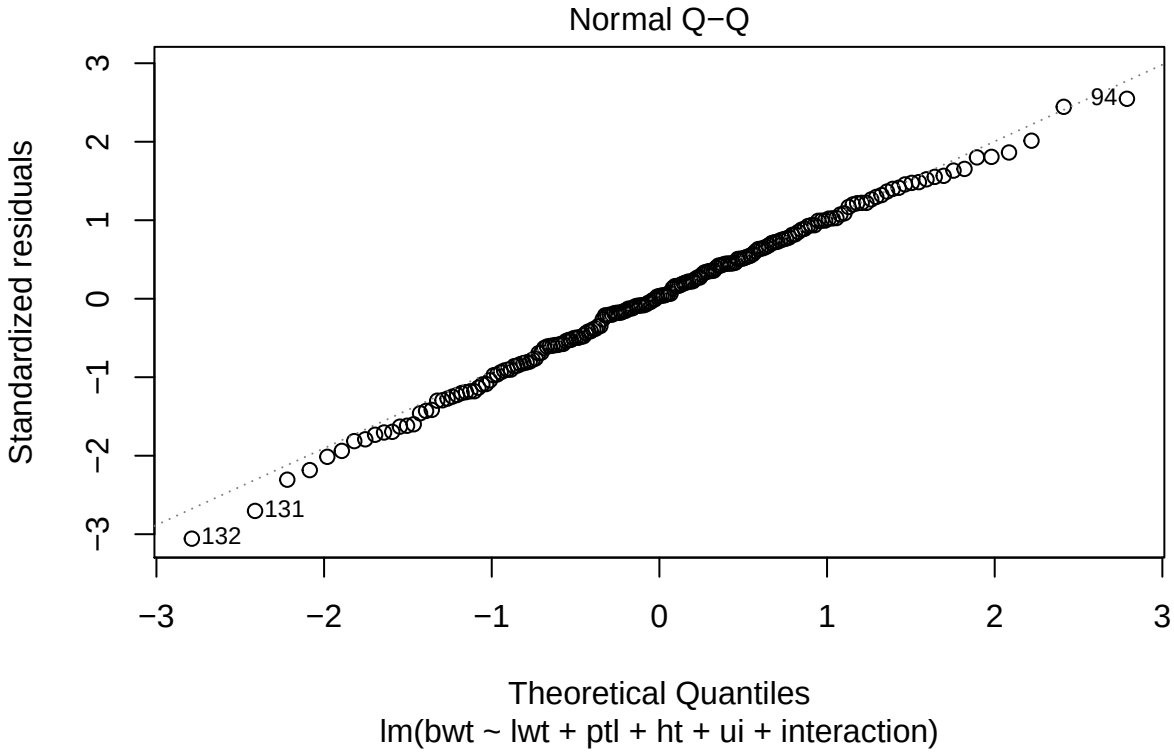
Test Summary
DF = 1
Chi2 = 0.1117979
Prob > Chi2 = 0.7381064

F Test for Heteroskedasticity
Ho: Variance is homogenous
Ha: Variance is not homogenous

Variables: fitted values of bwt

Test Summary
Num DF = 1
Den DF = 187
F = 0.1106804
Prob > F = 0.7397435

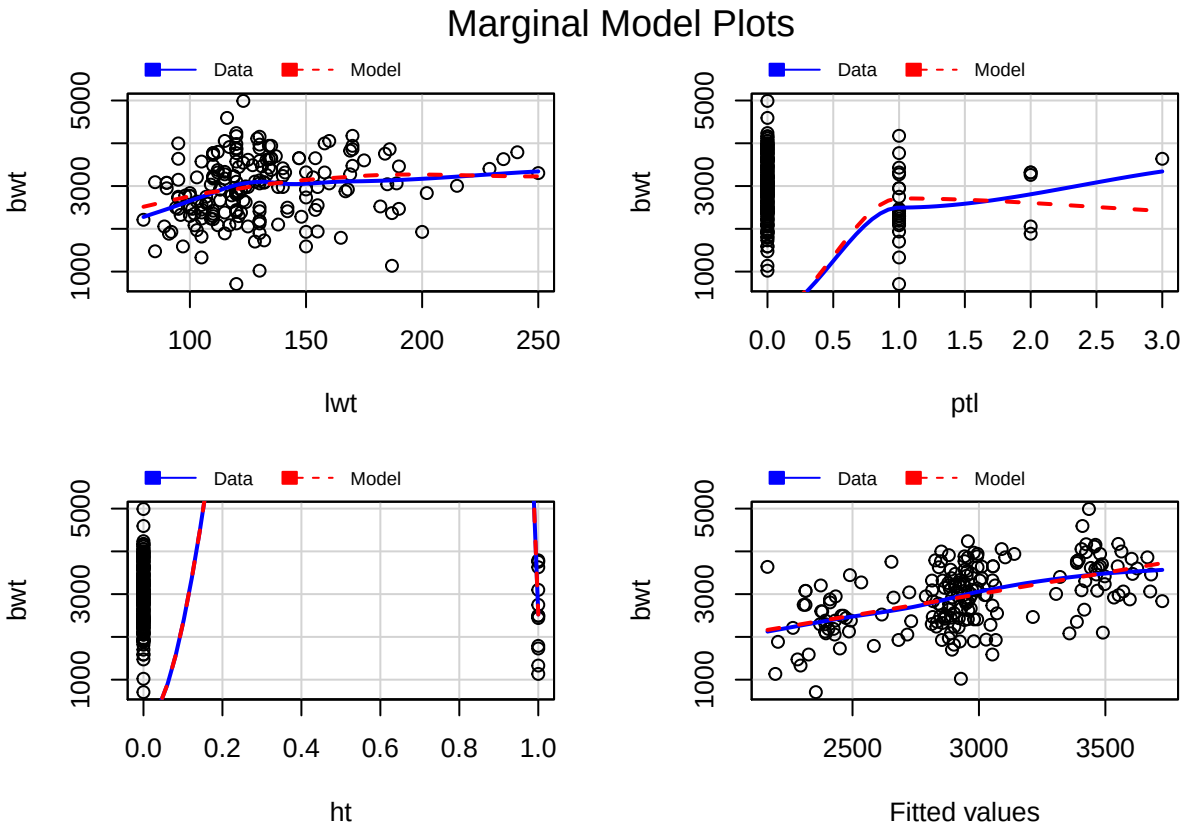
We see with the residual plot that the residuals are quite nicely randomly distributed about zero, which is a good sign. There do however seem to be clusters in the distribution of the residuals. Furthermore, both the score and F test prove the null hypothesis of homogeneous variance. So, thus far, we can say that our assumptions of linear regression hold. For further confirmation we can check the normality of our residuals with a QQ plot, and check the mean to ensure it is zero.



[1] -1.287183e-16

Above we see further confirmation that our assumptions of normally distributed residuals hold, as the residuals fit the theoretical values of normally distiributed residuals pretty well and the mean is quite small, nearly zero. However, there are a few values that do not lie within the normal distribution. These could be outliers in the data.

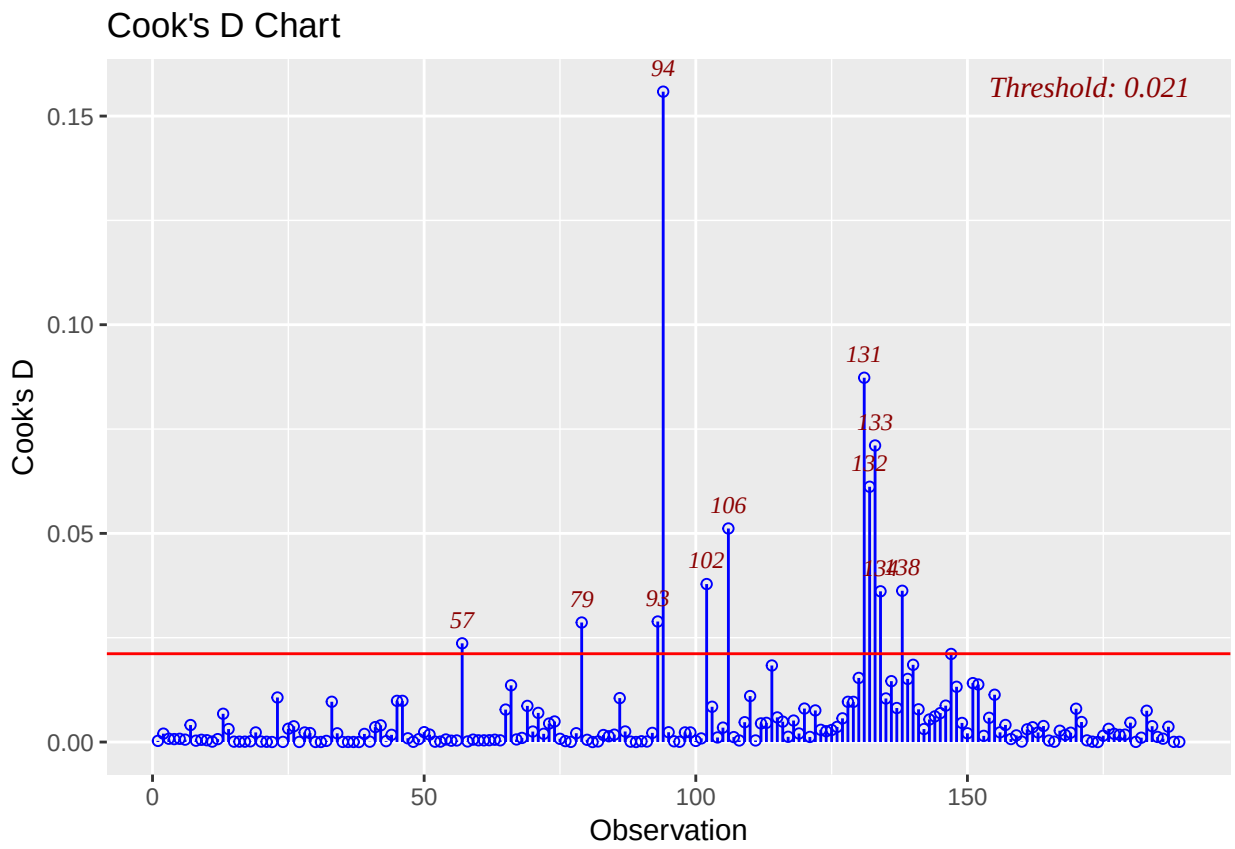
We should then see how well this model fits with a marginal model plot.

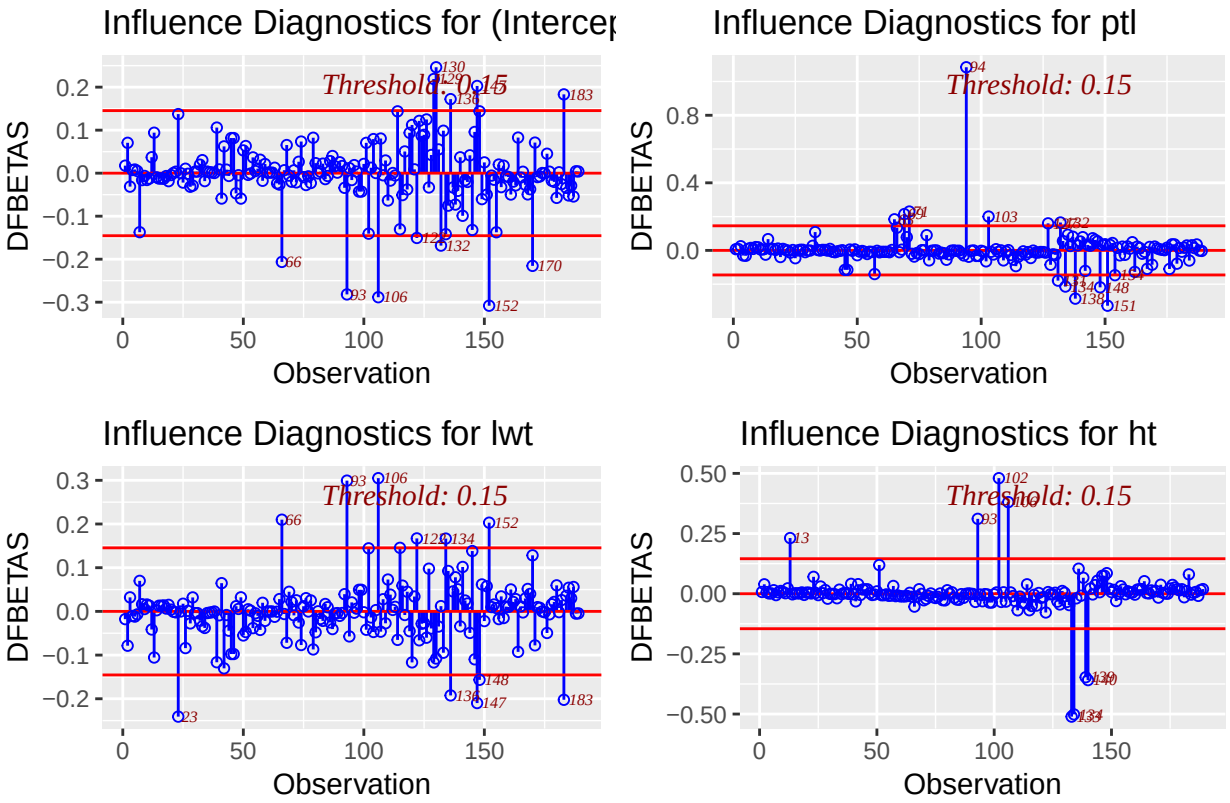


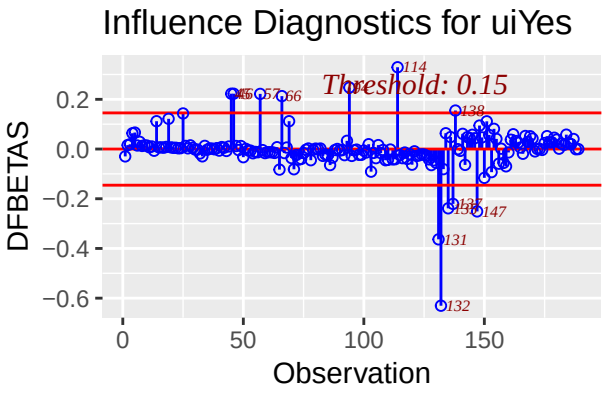
r.squared	adj.r.squared	sigma	statistic
0.2567177	0.219346	644.125	6.869302

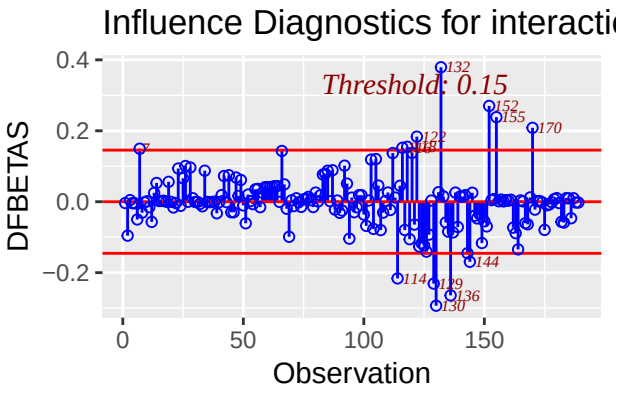
All of the marginal model plots show that our model fits the data pretty well, except for the plot of birthweit against premature labors, ptl. This is not surprising as earlier it was shown that in the multiple regression model, ptl becomes insignificant, indicating that something off is happening with this variable, or that it is not a good predictor. Furthermore, the multiple R squared value is only about 0.26. So, although the model fits well and our assumptions of linear regression hold, the model is not doing a good job of explaining all of the variance.

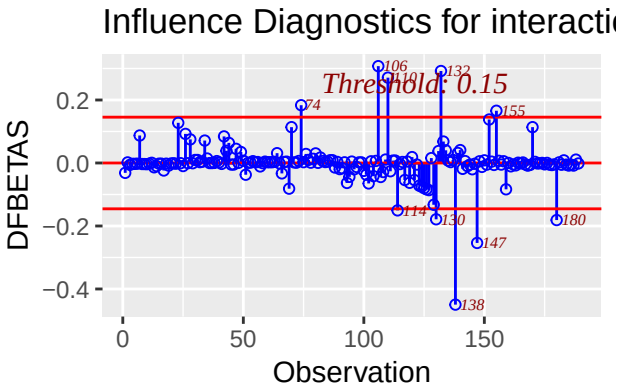
At this point we can check for any possible outliers or chances of multicollinearity with the combination of Cook's Distance, DFBETAS, and DFFITS.

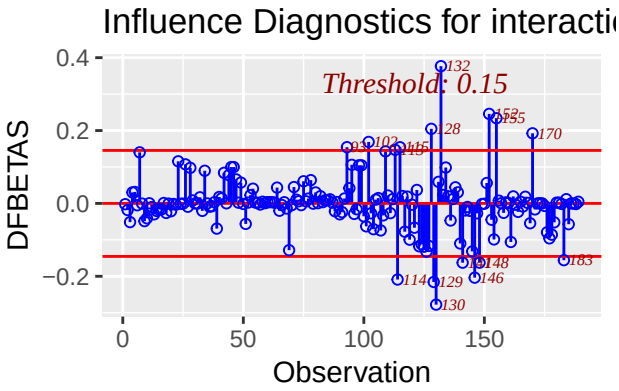

Cook's D Chart
Threshold: 0.021
94
131
133
132
106
102
57
79
93
Cook's D
0.15
0.10
0.05
0.00
0
50
100
150
Observation


Influence Diagnostics for (Intercep
Threshold: 0.15
DFBETAS
Observation
Influence Diagnostics for ptl
Threshold: 0.15
DFBETAS
Observation
Influence Diagnostics for lwt
Threshold: 0.15
DFBETAS
Observation
Influence Diagnostics for ht
Threshold: 0.15
DFBETAS
Observation

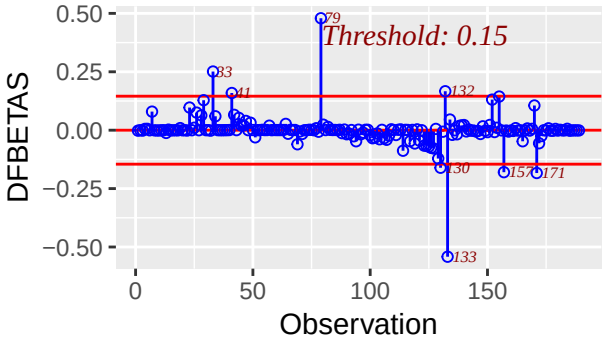

Influence Diagnostics for uiYes
Threshold: 0.15
DFBETAS
Observation


Influence Diagnostics for interacti
Threshold: 0.15
DFBETAS
Observation

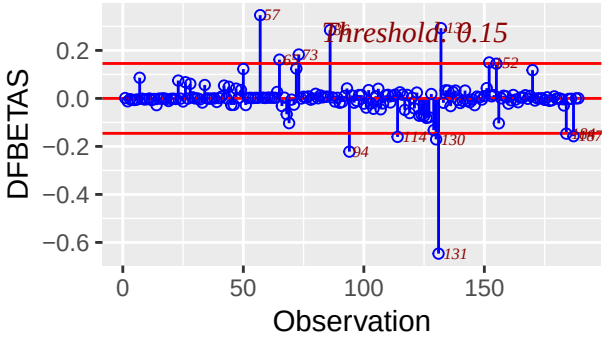

Influence Diagnostics for interacti
Threshold: 0.15
DFBETAS
Observation

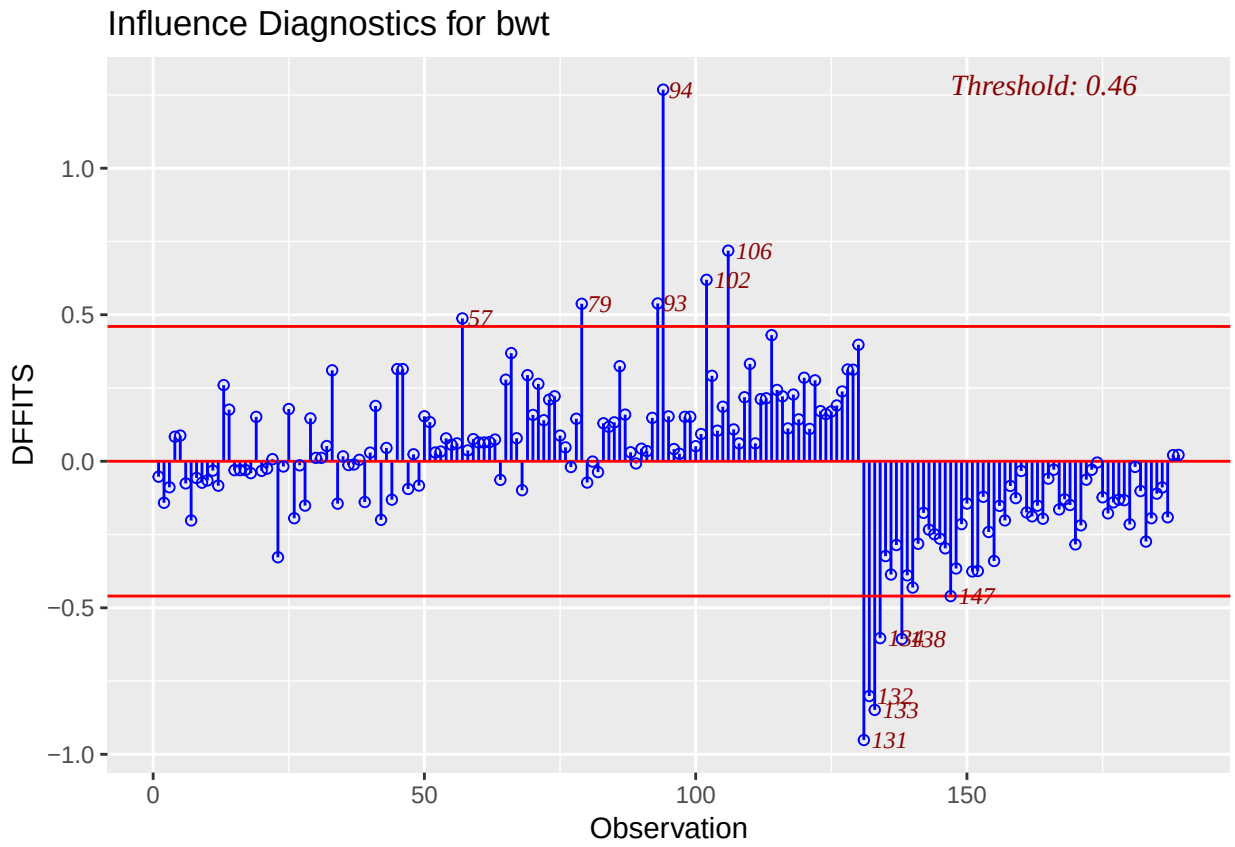

Influence Diagnostics for interacti
Threshold: 0.15
DFBETAS
Observation

Influence Diagnostics for interactionBlack.1



Influence Diagnostics for interactionOther.1





What immediately stands out from all of these different kinds of influence plots is the significance of the 94th observation, which comes from the ptl variable. It is a major outlier, which is also interesting for the changes in the significance of the variable to the outcome when switching from a simple to a multiple linear regression model. Furthermore, some of the outliers for history of hypertension are the same for our interaction terms. This may indicate some multicollinearity happening between our interaction terms. Perhaps history of hypertension and race/smoking habits are all related in some way.