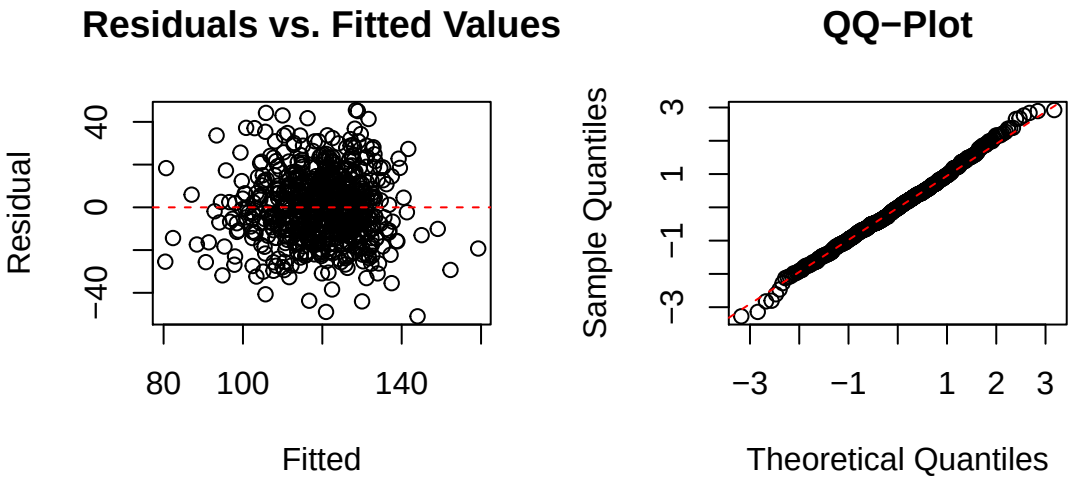
Model Diagnostics

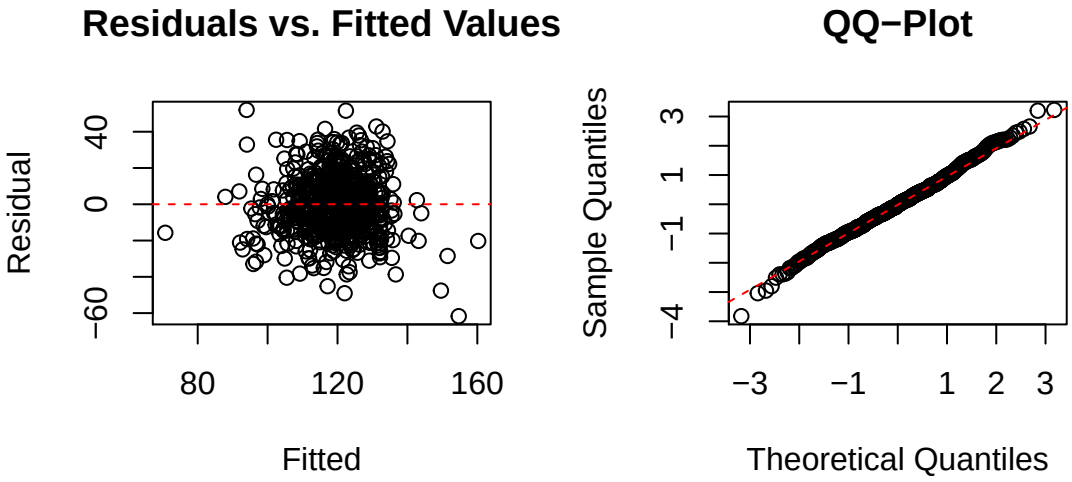
Residual Plots

Model 1 (Selected via Automated Model Selection)



From the residuals vs. fitted values plot, we can see that the conditional mean of the response ($weight_i$) is a linear function. We can also see that the conditional variance of $weight_i$ is constant. From the QQ-plot, we can see that the errors are iid normals. Hence, from these two plots we can conclude that Model 1 satisfies the linear regression model assumptions.

Model 2 (Selected via Manual Model Selection)



From the residuals vs. fitted values plot, we can see that the conditional mean of the response ($weight_i$) is a linear function. We can also see that the conditional variance of $weight_i$ is constant. From the QQ-plot, we can see that the errors are iid normals. Hence, from these two plots we can conclude that Model 2 satisfies the linear regression model assumptions.

K-fold cross-validation

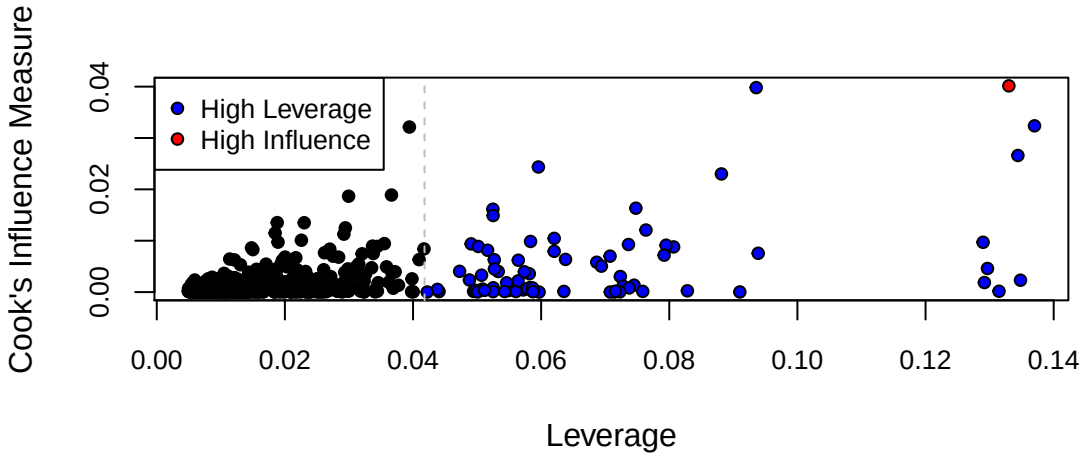
We used the k-fold cross validation to compare the two models(Automatic and Manual). We partitioned the original data into k (10) equal subsets. Then, we trainied each of our models on k-1 folds of the data, which formed the training sets.The accuracy of our models was calculated by validating the predicted results for the kth fold, which formed the test set. The model with a lower value of **MSPE** (Mean Squared Predictive Error) was chosen due to higher predictive accuracy.

```
##   Auto_model Manual_model
##      88067.9      91890.7
```

Looking at the two MSPE values we can say that the model selected through automatic model selection is better since it has a lower value of MSPE on the testing data.

Leverege and Influence measures

Model 1

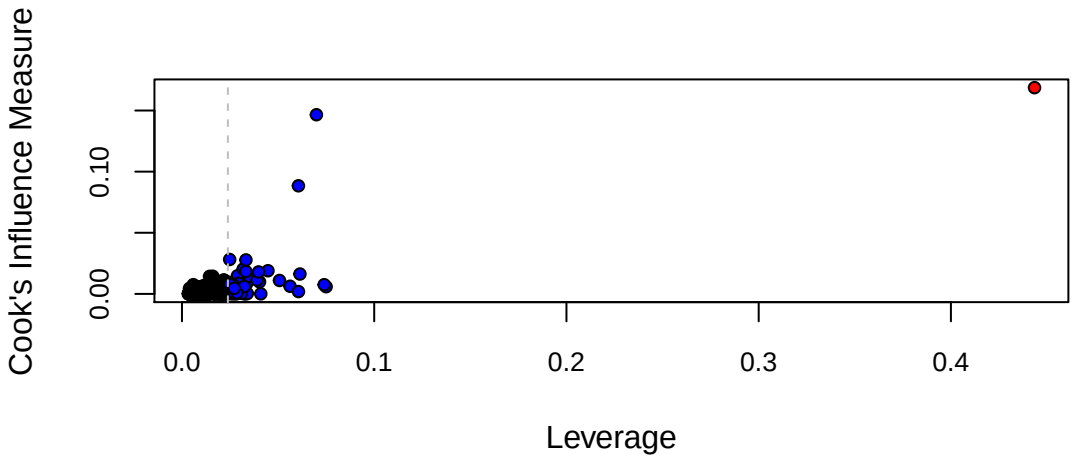


In this graph, we can see that there is one point which has high influence according to Cook's Influence Measure (in red). There are several blue points, which are points that have high leverage. The point with the highest influence is:

```
##      wt gestation parity  time number smoke mage mwt mht             meth
## 150 126       282      8 never  never never   38 250  66 African-American
##          income
## 150 12500-14999
```

So the parity, mother's age and mother's weight are greater than the mean values (1.9,27, 130 respectively) which would explain why this is a high-influence point.

Model 2



In this graph, we can see that there is one point which has high influence according to Cook's Influence Measure (in red). There are several blue points, which are points that have high leverage, however these points are less spread out than in Model 1. The point with the highest influence is:

```
##      wt gestation parity  time number smoke mage mwt mht      meth
## 413 138       288      0 never  never never   19 124  66 Caucasian
##          income
## 413 12500-14999
```

Here, the mother's age is significantly lower than the mean age, which would explain why the point is of greatest influence.

AIC

The AIC estimates the quality of each model, relative to the other model. The larger the difference in AIC, indicates stronger evidence for one model over the other. In this case, we want to compare two models. If $AIC_1 < AIC_2$ then model 1 would be preferred over model 2.

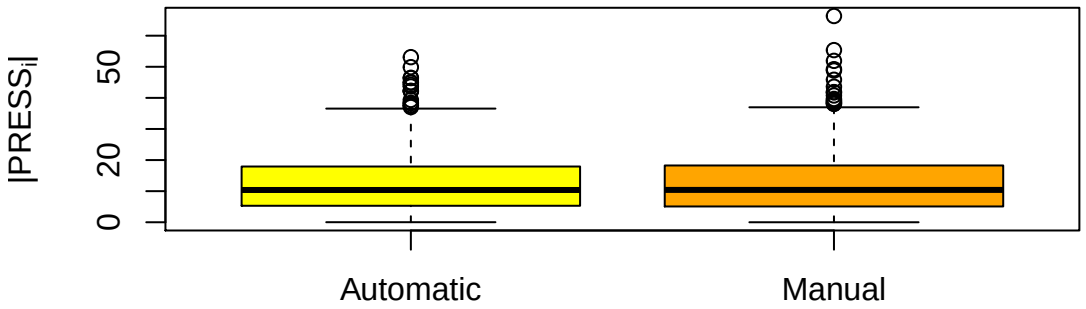
```
## [1] 5608.736
```

```
## [1] 5643.118
```

Since model 1(automatic model) has lower AIC than model 2 (Manually selected), we can say that using model 1 has a greater predictive advantage over model 2.

PRESS Statistic

Press statistic is a form of cross validation to provide a summary measure of the fit of a model to our training set. It should be noted that the training and test set are disjoint. In general, the smaller the PRESS value, the better the model's predictive ability.



We can see that the results for the sum of squared PRESS residuals for the two models are consistent with the cross-validation results, giving a bit of a predictive advantage to the automatic model (Model 1).

Discussion

The final model used following covariates: mother's age, mother's body-mass index, smoke, mother's height, mother's ethnicity, interaction between mother's age and gestation, and interaction between parity and gestation. As determined in the previous section, the model does not seem to violate any of the regression assumptions. There was a point with great influence corresponding to an overweight 38 year old woman. However it would not be wise to exclude it as an "outlier" since there is no justification for why the model would not be able to predict the birth weight of 38-year old mothers.

As discussed in the model selection section, the covariate with the smallest p-value, and hence the most significant (in the presence of the other covariates) was the mother's age. Also, as mentioned in the previous section, for model 1, the point with the highest influence was that with the highest age. It is also very interesting to note that both the gestation-age and the gestation-parity interactions were significant in this model. Both interaction covariates had positive coefficients. Nonetheless, further study would be required to determine the exact extent and nature of these interactions.

Consider two women of the same age, ethnicity, height, weight, parity, and gestation, but one of them smokes now while the other stopped smoking. Then, according to this model, the weight of the baby born to the woman who currently smokes is expected to be lower than the weight of the baby born to the woman who has stopped smoking (since all other factors have been held constant). However, most interestingly, this is true irrespective of when the mother stopped smoking (the coefficient + intercept is positive for both "used to smoke" and "smoked until pregnancy").

It should also be noted that the BMI is a significant covariate and the coefficient is positive. Thus, advice for prospective parents to reduce the probability of having an underweight baby would be to ensure that the mother has a healthy body-mass index, and to also stop smoking. The model suggests that it is better to quit smoking before pregnancy, however quitting at any point is still helpful.

Future Direction

As noted earlier, a significant portion of the father's data was missing. Due to this, the models did not use information from the father, which is a serious shortcoming since a child's genetic information is determined by both parents. Should this investigation be continued, data which is not missing the height/weight of fathers should be collected. Strategies used to reduce non-response bias could be employed as well (e.g. re-surveying a subset of the non-response group). A limitation was that there were only two socio-economic factors in the variable set (ethnicity and income) and one of them, income, did not have great predictive power. This could be due to the tendency that an individual has to lie about their income if they either make significantly less or significantly more than the average person.