Methods and Data Analysis 2

Srishti Saha (ss1078)

18 September, 2019

Question 1

After importing the OldFaithful dataset, let us now fit a model on Interval using Duration and Date.

Q1A: Fitting Regression Model

```
##
## Call:
## lm(formula = Interval ~ Duration + as.factor(Date), data = oldfaithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3886  -4.7332  -0.5622   3.9759  15.9639
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       32.8770     3.0672  10.719   <2e-16 ***
## Duration          10.8813     0.6622  16.431   <2e-16 ***
## as.factor(Date)2   1.3275     2.7173   0.489    0.626
## as.factor(Date)3   0.7825     2.6994   0.290    0.773
## as.factor(Date)4   0.1625     2.6461   0.061    0.951
## as.factor(Date)5   0.2463     2.6459   0.093    0.926
## as.factor(Date)6   1.9918     2.6580   0.749    0.455
## as.factor(Date)7  -0.1700     2.7020  -0.063    0.950
## as.factor(Date)8  -0.6944     2.6957  -0.258    0.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.866 on 98 degrees of freedom
## Multiple R-squared:  0.7408, Adjusted R-squared:  0.7196
## F-statistic:    35 on 8 and 98 DF,  p-value: < 2.2e-16
```

From the model summary above, we see that the intercept comes out to be significant as the p-value is low (in the range of e^{-16}). The predictor variable Duration also seems to be significant. The most significant variable on the basis of the t-statistic (highest t-value) would be Duration. However, no other day seems to have a significant impact on the y-variable (i.e. Interval), purely on the basis of the p-values (all p-values are > 0.4). The estimate of Duration is 10.88 which means the coefficient of Duration would be 10.88. Thus, just on the basis of Duration (duration of previous eruption) and Date=1, having any and all other variables constant, an eruption will have an Interval of 10.88 greater than the baseline. The intercept of 32.87 includes the baseline evaluation of Date=1 (level of Date variable). This means that for an eruption instance with the duration of the previous eruption equal to 0 and Date=1 (all other levels of Date have a 0 flag), the interval will be 32.87.

If we look at the levels of the Date variable, Date=1 has been absorbed in the intercept as baseline (value=32.88). **Although none of the levels of Date come out to be statistically significant with respect to their p-values (as the p-values are significantly high), we can compare the**

difference of the mean intervals between the various levels of Date. From the above result, Date=6 and Date=2 have a high value of t-statistic and the estimate. The estimate of 1.99 for Date=6 implies that with Duration kept as constant and considering Date=1 as baseline, the Interval of Day 6 is going to be higher by 1.99 units than on Day1. Similarly, for Day2, it is going to be higher by 1.33 than baseline of Day1.

If we look at the **adjusted R-squared value**, it is 0.72 which indicates 72% of the variation in Interval has been explained by this model. On comparing it with the multiple R-squared value of 0.74, we can conclude that addition of the Date variable has led to some penalty which shows as a difference between R-squared and adjusted R-squared.

Q1B: F-test to compare models

- From the above question, model 1 can be defined as: Model 2: prediction of Interval using Duration and Date converted as factor
- Let us define model_2 as: Model 2: prediction of Interval using Duration as the only predictor

On comparing the 2 models using ANOVA.

```
anova(lm_model2,lm_model1)
```

```
## Analysis of Variance Table
##
## Model 1: Interval ~ Duration
## Model 2: Interval ~ Duration + as.factor(Date)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    105 4689.0
## 2     98 4620.2  7    68.853 0.2086 0.9828
```

In the above F-test, our null and alternate hypothesis is as follows:

- H_0 : The coefficients for Date variable (all levels) are 0.
- H_A : The coefficients for Date variable (all levels) are not 0.

On evaluating the results of the ANOVA test above, we see that the p-value is 0.98 which not significant statistically. Thus, we cannot reject the null hypothesis. This implies that the coefficients of the Date variable (all levels) are not very different from 0. **In other words, the Date variable is not statistically significant for our model and does not contribute anything to it. It will thus perform the same with or without the Date variable.**

Q1C: K-fold validation to compare average RMSE

```
## [1] "RMSE of model 1 (with Duration and Date) after K-fold is= 7.05298869896774"
```

```
## [1] "RMSE of model 2 (with Duration) after K-fold is= 6.61109992685862"
```

We see above that the RMSE of the model with only Duration as a predictor (model 2) is lower than the RMSE of model 1 (with both Duration and Date). This means that the old model (with only Duration) is more accurate than the new model (including Date). **The old model has higher predctive accuracy.**

Question 2

Summary

This report covers analysis on data derived from the Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital. While the original study covered 15000 families, we consider a subset of the data. After removing records with missing data, we deal with 869 family records. A linear regression model was built to study the association of various parameters with the weight of the baby at birth.

Introduction

This report covers the different data transformation steps, the process of exploratory data analysis and the exercise of determining the relationship of different variables with the babies' weight at the time of its birth. The objective of the study is based on the Surgeon Generals' claim that mothers who smoke have increased rates of premature delivery (before 270 days) and low birth weights. The analysis will **check and determine if there is an association between smoking (along with other variables like age, education, income, race etc.) and birth weight**. We compared babies whose mothers smoke to babies whose mothers have never smoked. Another exercise was to **model a linear relationship between different predictor variables and birth weight of the babies.**

Data

Data Dictionary and Statistics

This section will give an overview of the dataset.

```
## 'data.frame':    869 obs. of  12 variables:
##  $ id       : int  4604 7435 7722 2026 3553 3491 6757 6153 8187 8403 ...
##  $ date     : int  1598 1527 1563 1503 1638 1705 1444 1405 1669 1669 ...
##  $ gestation: int  148 181 204 225 233 234 234 235 236 241 ...
##  $ bwt.oz   : int  116 110 55 132 105 85 97 129 63 128 ...
##  $ parity   : Factor w/ 12 levels "0","1","2","3",..: 8 8 12 5 5 8 1 4 1 1 ...
##  $ mrace    : Factor w/ 5 levels "5","6","7","8",..: 3 3 3 3 3 3 2 3 1 3 ...
##  $ mage     : int  28 27 35 28 34 33 26 24 24 17 ...
##  $ med      : Factor w/ 7 levels "0","1","2","3",..: 2 2 4 3 4 2 6 5 6 2 ...
##  $ mht      : int  66 64 65 67 61 67 65 66 58 64 ...
##  $ mpregwt  : int  135 133 140 148 130 130 112 135 99 126 ...
##  $ inc      : Factor w/ 10 levels "0","1","2","3",..: 3 2 7 4 4 3 7 2 8 3 ...
##  $ smoke    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##           vars   n   mean    sd median trimmed   mad min max range  skew kurtosis   se
## gestation    1 869 278.51 15.70    279  279.26 10.38 148 338   190 -1.25     8.61 0.53
## bwt.oz       2 869 118.36 18.05    119  118.51 16.31  55 174   119 -0.11     0.56 0.61
## mage         3 869  27.29  5.71     26   26.90  5.93  15  45    30  0.58    -0.28 0.19
## mht          4 869  64.07  2.53     64   64.09  2.97  53  72    19 -0.15     0.25 0.09
## mpregwt      5 869 128.48 20.78    125  126.63 19.27  87 220   133  1.02     1.66 0.70
```

We see that we 869 observations across 10 variables. Following are the descriptions of these columns: * **id-** *Unique ID number for each observation* * **date-** *birth date where 1096 = January1, 1961* * **gestation-** *length of Gestation period of the child-OUTCOME* * **bwt.oz-** *birth* *weight in ounces (999 = unknown)-OUTCOME* * **parity-** *total number of previous pregnancies, including fetal deaths and still births. (99=unknown* * **mrace-** *mother's race: 5- White, 6=mexican, 7=black, 8=asian, 9=mix, 99=unknown* * **mage-** *mother's age in years at termination of pregnancy* * **med-** *mother's education: 0 = less than 8th grade, 1 = 8th to 12th grade, did not graduate high school, 2 = high school graduate, no other schooling, 3 = high school graduate + trade school, 4 = high school graduate + some college, 5 = college graduate, 7 = trade school but unclear if graduated from high school, 9 = unknown* * **mht-** *mother's height in inches* * **mpregwt-** *mother's pre-pregnancy weight in pounds* * **inc-** *family yearly income in 2500 increments. 0 = under 2500, 1 = 2500-4999, . . . , 9 = 15000+. 98=unknown, 99=not asked* * **smoke-** *does mother smoke?:0 = never,1 = smokes now,2 = until preg,3 = once did, not now* The categorical variables have been shown as factors in the above summary. A few summary statistics of the numerical columns have also been given here.

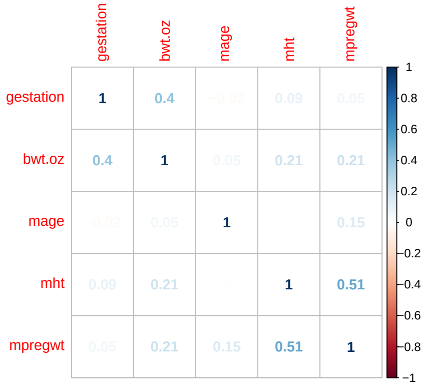
Data Transformations

We have already performed a few data transformations, namely: >1. converted education, race, income and smoke to factor variables 2. In race: collapsed levels 0-5 to '5' indicating 'White' as per instructions 3. In education: collapsed levels 6-7 to '7' indicating 'trade school' as per instructions (however, there were no records for med=6) We will now drop the ID and date columns owing to the nature of the variables. These variables cannot be used for predictions or establishing associations.

```
## [1] "Resultant dimensions of the dataset are: 869 rows and 10 columns"
```

EDA

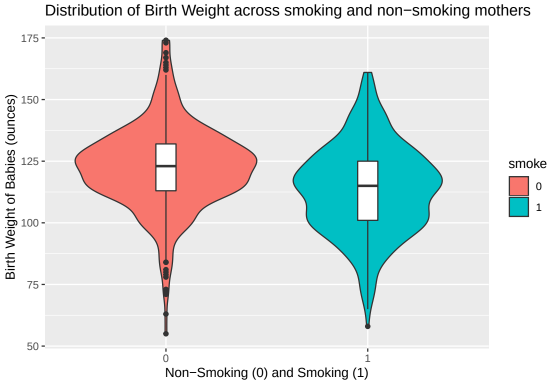
Let us first plot the correlation of these variables.



From the correlation plot, we see that:

1. There is a *positive* correlation between Mother's height and weight of about 0.51.
2. There is a *relatively high positive* correlation (0.4) between birth weight and gestation. As explained in the question, both of these are outcome variables. One should not be considered as the predictor for the other based on scientific reasons.
3. There is a *low positive* correlation of birth weight with mothers' age (0.05).
4. There is a *positive* correlation of both mothers' height and weight with the birth weight (0.21 each).

Let us look at the relation with the categorical variables



On doing a preliminary analysis with the 'smoke' variable, we see that mothers who never smoked have shown a higher mean birth weight than the mothers who smoke. This might establish a relationship between the two variables

We also did a primary check of correlation between the categorical variables (Parity and mothers' race) with birth weight across the population of smoking and non-smoking mothers (Plots in Appendix). We observe that: 1. **Parity** (number of previous births) shows a linear relationship with birth weight. It also has a high-variance effect on birth weight. 2. **Mother's race** shows an effect on birth weight but in the smoking mothers belonging to the races: Black, Asian or mix.

On doing a similar check with numerical variables (mothers' height and weight), we observe that both these variables have a linear relationship with the weight of the babies at birth (plots in appendix).

Model

Let us fit a model with the variables: parity, mrace, mage, med, mht, mpregwt, in and smoke.

Rationale of final model building:

On evaluating the model, adjusted R-squared (after penalizing for a lot of predictors) is 14.8%. Only 14.8% variance is explained. Removing the following variables as they seem to be insignificant on the basis of their high p-values and low t-values: **mage, med and inc**. Retaining Parity because some levels (2,3,4,6,8,11) showed moderate impact on the basis of their p-value. Of these, parity=3 level showed a t-value of 2.9. Thus, it is moderately significant when compared to other t-statistics of the range (2-7.8). Retaining mht and mpregwt due to their high significance based on low p-values and a moderate influence with a t-value of 3.63 and 3.11 respectively. smoke variable has a high t-statistic with a n absolute value of 7.98. This has a high significance. In mrace, 2 levels: 7 and 8 came out to be significant on the basis of their low p-values. mrace=7 has a relatively high absolute t-statistic (5.63) implying significance. Thus, race is partially significant.

Let us try creating a model with: *parity*, *mht*, *mpregwt*, *smoke* and *mrace*

The adjusted R-squared is now 15.3%. The residual standard error is 16.61 which is marginally lower than the error of 16.66 of the first model. Smoke (=1) still seems to be the mst significant variable with absolute t-stat = 8.25.

In our EDA process, we saw that race and smoke indicator had an association. Let us create an interaction variable on the same: (smoke:race). Let us create a model with these selected variables: *parity*, *mht*, *mpregwt*, *smoke*, *mrace* and *smoke:mrace*

The adjusted R-squared of this model is 15.4% i.e. this model explains 15.4% of the variation in birth weight of babies. The most significant variable is smoke (level=1) with a low p-value (e^{-12}) and an abs t-value of 7.192.

```
## [1] "RMSE of final model after K-fold is= 16.328465778908"
```

The root mean squared error of this model is 16.32, which has also decreased from the first model which has increased its predictive accuracy.

Assessing assumptions for the model

1. If we look at the residuals versus fitted plot (Appendix), we see that the residuals have a linear trend with no apparent pattern in the plot. This satisfies the assumption of Linearity
2. The Q-Q plot shows a faily linear line (except on the extremely high and low quantiles where there is a slight distortion from the straight line). Thus the assumption of Normality is fairly met.
3. The residual plot shows no conical pattern, thus implying no heteroskedasticity. It is linear and thus proves the assumption of Independence and Equal Variance.

Potential Limitations

1. The adjusted R-squared is only 15.4% which is very low. We have only been able to explain 15.4% variation in our target variable.
2. We have lost a lot of data (169 rows) while removing the missing values. This might have affected the model since it did not have the complete information
3. The leverage plot shows that there are a few high leverage-points. They have not been investigated or justified by this model yet. This model is susceptible to outliers. If we look at the Leverage plot, there are a few points with high leverage. They lie near the 0.5 cook's distance line- both above and below the 0-line. The 487th record seems to have a high leverage.

Results

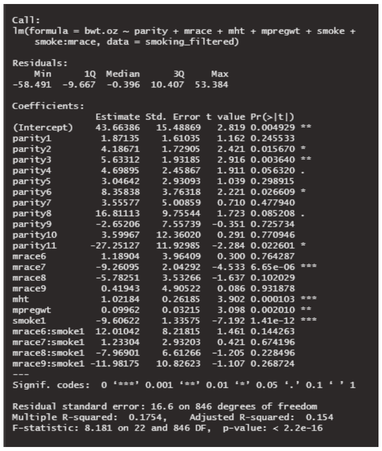
Final Model: Summary of final model with parity, mht, mpregwt, smoke, mrace, smoke:mrace

(Please find snapshot of summary of final model below (Figure 1). All model summaries are also in appendix)

Let us also look at the confidence intervals.

```
##          2.5 %    97.5 %
## smoke1 -12.228 -6.984439
```

- The final model (with the interaction variable smoke:mrace) explains 15.4 % variation in the y-variable



```
Call:
lm(formula = bwt.oz ~ parity + mrace + mht + mpregwt + smoke +
    smoke:mrace, data = smoking_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-58.491  -9.667  -0.396  10.407  53.384

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    43.66386   15.48869   2.819 0.004929 **
parity1         1.87135    1.61035   1.162 0.245533
parity2         4.18671    1.72905   2.421 0.015670 *
parity3         5.63312    1.93185   2.916 0.003640 **
parity4         4.69895    2.45867   1.911 0.056320 .
parity5         3.04642    2.93093   1.039 0.298915
parity6         8.35838    3.76318   2.221 0.026609 *
parity7         3.55577    5.00859   0.710 0.477940
parity8        16.81113    9.75544   1.723 0.085208 .
parity9        -2.65206    7.55739  -0.351 0.725734
parity10        3.59967   12.36020   0.291 0.770946
parity11      -27.25127   11.92985  -2.284 0.022601 *
mrace6          1.18904    3.96409   0.300 0.764287
mrace7         -9.26095    2.04292  -4.533 6.65e-06 ***
mrace8         -5.78251    3.53266  -1.637 0.102029
mrace9          0.41943    4.90522   0.086 0.931878
mht             1.02184    0.26185   3.902 0.000103 ***
mpregwt         0.09962    0.03215   3.098 0.002010 **
smoke1         -9.60622    1.33575  -7.192 1.41e-12 ***
mrace6:smoke1  12.01042    8.21815   1.461 0.144263
mrace7:smoke1   1.23304    2.93203   0.421 0.674196
mrace8:smoke1  -7.96901    6.61266  -1.205 0.228496
mrace9:smoke1 -11.98175   10.82623  -1.107 0.268724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.6 on 846 degrees of freedom
Multiple R-squared:  0.1754,    Adjusted R-squared:  0.154
F-statistic: 8.181 on 22 and 846 DF,  p-value: < 2.2e-16
```

Figure 1: Snapshot of model summary

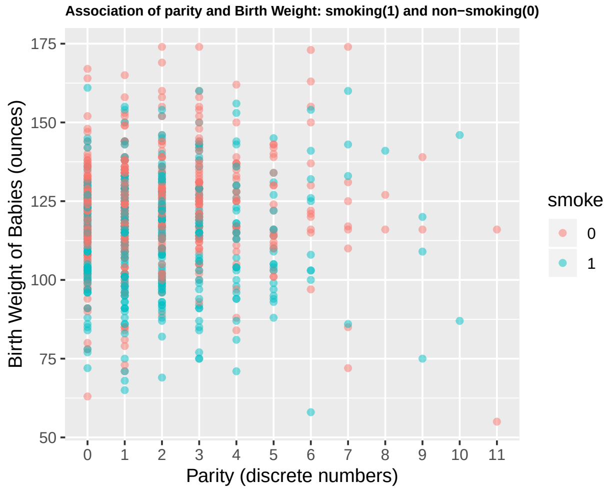
- **The estimate of smoke(=1) is -9.61 which means for all variables kept as constant, mothers who smoke have babies with average weight 9.61 ounces lower than those who do not smoke (=0).** The intercept estimate is 43.66 which means that **the baseline weight of a baby from a mother who does not smoke will be 43.66 ounces.**
- The most significant variable is smoke (=level 1 i.e. mothers who smoke).
- The interaction terms (for race levels 6,7,8 and 9) are not statistically significant on the basis of the high p-value.
- We can say with 95% confidence that the true difference of babies' mean bwt.oz across smoke=0 and smoke=1 lies between **-12.23 and -6.98 ounces**.

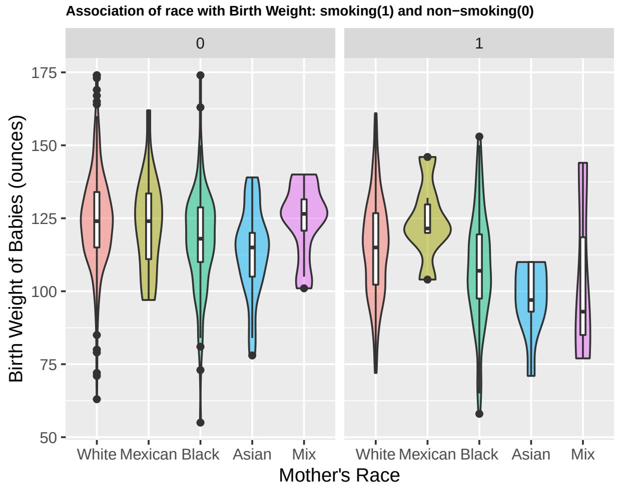
Conclusion

1. As shown in model results and from the EDA steps, we have concluded that **mothers who smoke have shown a lower average baby weight than the mothers who do not smoke**. Our model shown that this difference is almost 9.6 ounces given all other factors are constant.
2. Based on the 95% Confidence Intervals, we can say with 95% confidence that the difference of babies' mean birth weights across the population of smoking and non-smoking mothers lies between **-12.23 and -6.98 ounces**.
3. If so, characterize those differences. The EDA steps show that the **effect of race on birth weight can be seen on smoking mothers, but predominantly on the ones belonging to the races: Black, Asian or mix.** (plots in appendix)
4. The impact of all variables have been explained and taken into consideration to build the final model above. All supporting plots are in the appendix

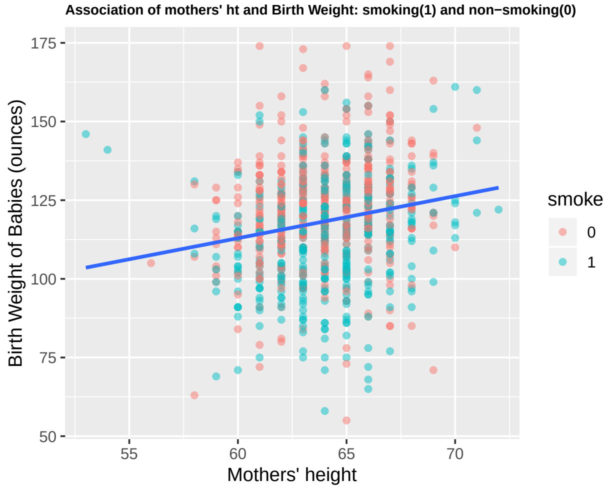
Appendix

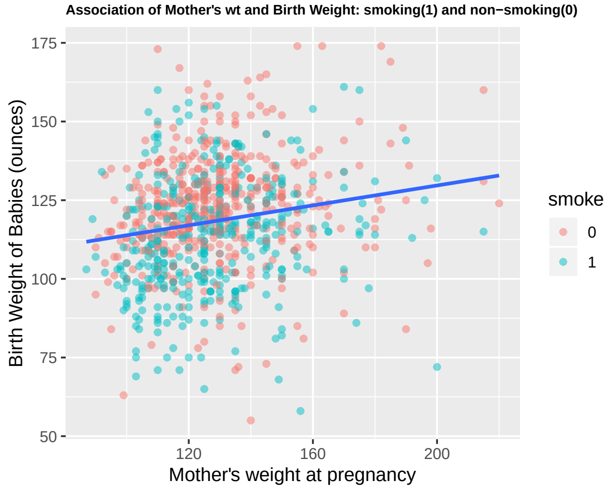
Plots to show association of categorical variables with babies' birth weight across smoking and non-smoking mothers


Association of parity and Birth Weight: smoking(1) and non−smoking(0)


Association of race with Birth Weight: smoking(1) and non−smoking(0)

Plots to show association of numerical variables with babies' birth weight across smoking and non-smoking mothers


Association of mothers' ht and Birth Weight: smoking(1) and non−smoking(0)


Association of Mother's wt and Birth Weight: smoking(1) and non−smoking(0)

Model 1: Summary of model 1 with all predictor variables

```
##
## Call:
## lm(formula = bwt.oz ~ parity + mrace + mage + med + mht + mpregwt +
##     inc + smoke, data = smoking_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.170  -9.736  -0.169  10.316  52.132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.31032   18.21406   2.048 0.040830 *
## parity1       1.70112    1.66513   1.022 0.307261
## parity2       4.15387    1.85350   2.241 0.025283 *
## parity3       6.18044    2.12399   2.910 0.003712 **
## parity4       5.40998    2.71975   1.989 0.047011 *
## parity5       3.10111    3.24498   0.956 0.339519
## parity6       9.69695    4.06658   2.385 0.017323 *
## parity7       4.71120    5.30839   0.887 0.375066
## parity8      18.96461   10.16797   1.865 0.062514 .
## parity9      -1.10612    7.86996  -0.141 0.888260
## parity10     13.57128   12.99740   1.044 0.296717
## parity11    -26.36298   12.20550  -2.160 0.031063 *
## mrace6        4.57849    3.54474   1.292 0.196843
## mrace7       -9.01947    1.60316  -5.626 2.52e-08 ***
## mrace8       -7.56575    3.12120  -2.424 0.015563 *
## mrace9       -2.67407    4.44837  -0.601 0.547914
## mage         -0.07764    0.13507  -0.575 0.565572
## med1          6.63048    8.18489   0.810 0.418121
## med2          8.76426    8.08934   1.083 0.278930
## med3          7.60741    8.37365   0.908 0.363880
## med4          9.20402    8.11906   1.134 0.257275
## med5          8.62000    8.15867   1.057 0.291025
## med7         -4.02772   11.57981  -0.348 0.728061
## mht           0.97716    0.26948   3.626 0.000305 ***
## mpregwt       0.10240    0.03293   3.110 0.001937 **
## inc1          2.92325    3.60193   0.812 0.417264
## inc2          4.85614    3.60797   1.346 0.178685
## inc3          1.77858    3.64265   0.488 0.625490
## inc4          2.17105    3.72610   0.583 0.560280
## inc5          1.95680    3.75629   0.521 0.602547
## inc6          1.86191    4.02759   0.462 0.643995
## inc7          1.54673    3.72974   0.415 0.678467
## inc8          3.15145    5.48435   0.575 0.565699
## inc9         -1.77943    5.07431  -0.351 0.725922
## smoke1       -9.44136    1.18291  -7.981 4.77e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.66 on 834 degrees of freedom
## Multiple R-squared:  0.1812, Adjusted R-squared:  0.1478
## F-statistic: 5.427 on 34 and 834 DF,  p-value: < 2.2e-16
```

Model 2: Summary of model 1 with parity, mht, mpregwt, smoke, mrace

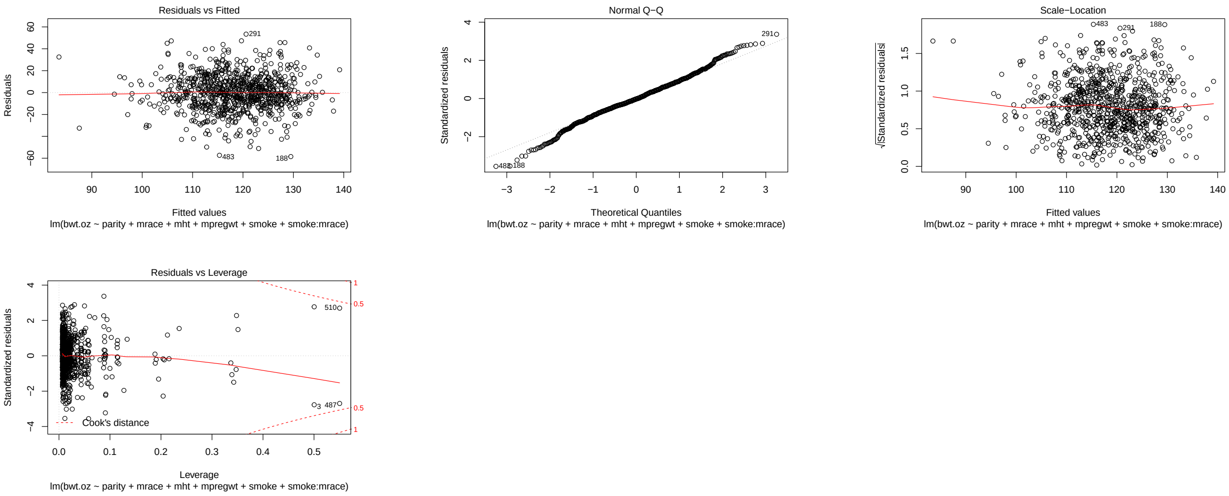
```
##
## Call:
## lm(formula = bwt.oz ~ parity + mrace + mht + mpregwt + smoke,
##     data = smoking_filtered)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.207  -9.697  -0.227  10.328  53.020
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.27123   15.43348   2.998 0.002796 **
## parity1       1.80204    1.60884   1.120 0.262994
## parity2       4.18477    1.72575   2.425 0.015520 *
## parity3       5.63491    1.93109   2.918 0.003616 **
## parity4       4.61482    2.45065   1.883 0.060028 .
## parity5       2.69204    2.92519   0.920 0.357679
## parity6       8.75790    3.75189   2.334 0.019814 *
## parity7       3.41048    5.00075   0.682 0.495429
## parity8      16.63051    9.75715   1.704 0.088664 .
## parity9      -3.11114    7.55007  -0.412 0.680394
## parity10      7.99373   11.99375   0.666 0.505278
## parity11    -27.81853   11.88855  -2.340 0.019517 *
## mrace6        3.84011    3.48647   1.101 0.271020
## mrace7       -8.68070    1.52411  -5.696 1.69e-08 ***
## mrace8       -7.97365    3.03160  -2.630 0.008688 **
## mrace9       -1.91987    4.39028  -0.437 0.662004
## mht           0.98204    0.26077   3.766 0.000177 ***
## mpregwt       0.09906    0.03215   3.081 0.002126 **
## smoke1       -9.51225    1.15340  -8.247 6.14e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.61 on 850 degrees of freedom
## Multiple R-squared:  0.1704, Adjusted R-squared:  0.1528
## F-statistic: 9.699 on 18 and 850 DF,  p-value: < 2.2e-16
```

Final Model with parity, mht, mpregwt, smoke, mrace, smoke:mrace

```
##
## Call:
## lm(formula = bwt.oz ~ parity + mrace + mht + mpregwt + smoke +
##     smoke:mrace, data = smoking_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.491  -9.667  -0.396  10.407  53.384
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    43.66386   15.48869   2.819 0.004929 **
## parity1         1.87135    1.61035   1.162 0.245533
## parity2         4.18671    1.72905   2.421 0.015670 *
## parity3         5.63312    1.93185   2.916 0.003640 **
## parity4         4.69895    2.45867   1.911 0.056320 .
## parity5         3.04642    2.93093   1.039 0.298915
## parity6         8.35838    3.76318   2.221 0.026609 *
## parity7         3.55577    5.00859   0.710 0.477940
## parity8        16.81113    9.75544   1.723 0.085208 .
## parity9        -2.65206    7.55739  -0.351 0.725734
## parity10        3.59967   12.36020   0.291 0.770946
## parity11      -27.25127   11.92985  -2.284 0.022601 *
## mrace6          1.18904    3.96409   0.300 0.764287
## mrace7         -9.26095    2.04292  -4.533 6.65e-06 ***
## mrace8         -5.78251    3.53266  -1.637 0.102029
## mrace9          0.41943    4.90522   0.086 0.931878
## mht             1.02184    0.26185   3.902 0.000103 ***
## mpregwt         0.09962    0.03215   3.098 0.002010 **
## smoke1         -9.60622    1.33575  -7.192 1.41e-12 ***
## mrace6:smoke1  12.01042    8.21815   1.461 0.144263
## mrace7:smoke1   1.23304    2.93203   0.421 0.674196
## mrace8:smoke1  -7.96901    6.61266  -1.205 0.228496
## mrace9:smoke1 -11.98175   10.82623  -1.107 0.268724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.6 on 846 degrees of freedom
## Multiple R-squared:  0.1754, Adjusted R-squared:  0.154
## F-statistic: 8.181 on 22 and 846 DF,  p-value: < 2.2e-16
```

Residual Plots of final Model


Residuals vs Fitted
Residuals
Fitted values
lm(bwt.oz ~ parity + mrace + mht + mpregwt + smoke + smoke:mrace)
291
483
188
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(bwt.oz ~ parity + mrace + mht + mpregwt + smoke + smoke:mrace)
291
483
188
Scale−Location
√|Standardized residuals|
Fitted values
lm(bwt.oz ~ parity + mrace + mht + mpregwt + smoke + smoke:mrace)
483
291
188
Residuals vs Leverage
Standardized residuals
Leverage
lm(bwt.oz ~ parity + mrace + mht + mpregwt + smoke + smoke:mrace)
510
3
487
Cook's distance
1
0.5
0.5
1