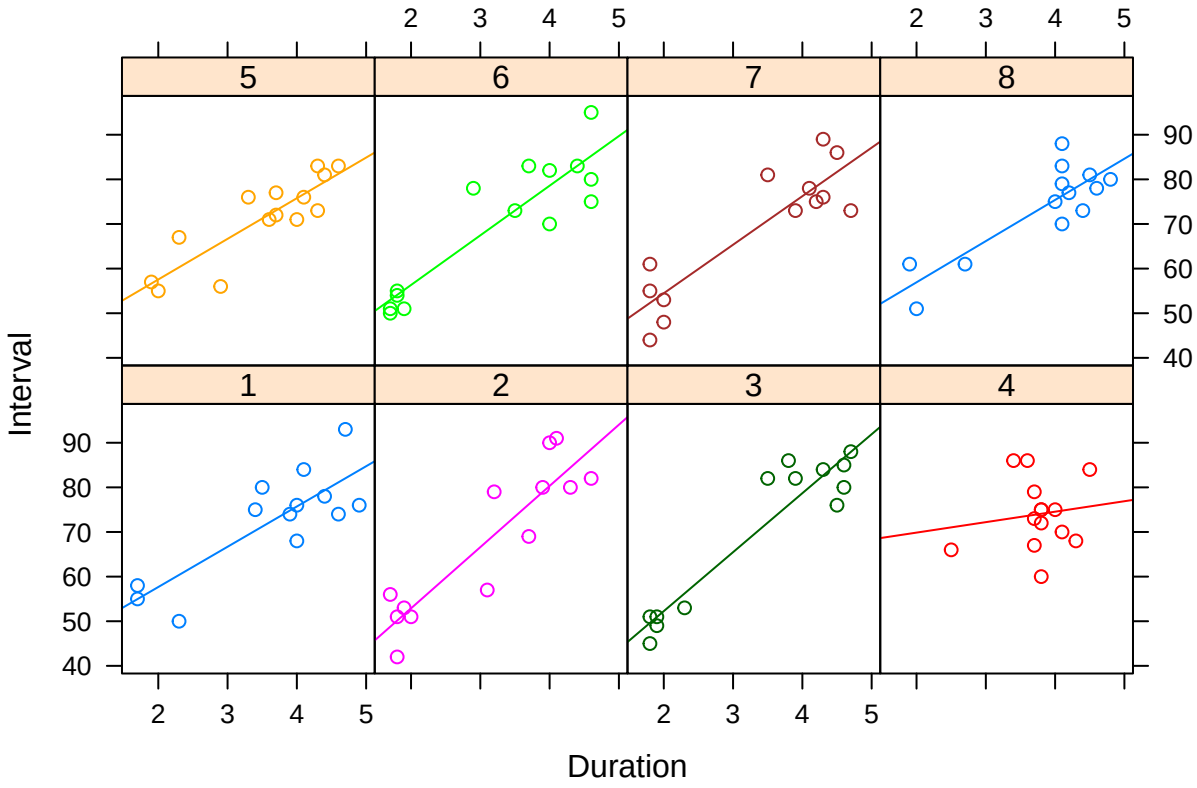
Methods and Data Analysis 2

Ashwini Marathe

Question 1a:

Fit a regression of interval on duration and day (treated 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).



The date column is numeric in the original data. Converting date column to factor variable, there are 8 levels/factors in the column.Above figure show the plot of duration vs. interval for different date values. From the plot we observe that day 4 slope is different from other days. But we also observe that there are less data points for day 4. We can confirm if there is significant difference between different days by including it in linear regression. Using linear regression of interval on duration and day, the p-values for all days come out to be very high and hence we cannot reject the null hypothesis that slopes corresponding to each day is equal to zero. Also, compared to the previous model without using day variable has a better R^2-value. Though statistically insignificant, day 6 has the highest absolute t-value. The mean interval difference for day 6 as compared to day 1 is approximately 2 minutes. From the EDA it seemed day 4 is different, but the regression analysis states otherwise.

Question 1b:

Perform an F-test to compare this model to your model for this data from the last homework. In context of the question, what can you conclude from the results of the F-test?

We can test if the addition of new variable 'Date' to the model improves the analysis or not. We can use ANOVA method to check this. The ANOVA method uses the following hypothesis:

H_0: The extra estimator does not make the model better significantly.

H_1: The extra estimator does make the model better.

```
## Analysis of Variance Table
##
## Model 1: Interval ~ Duration
## Model 2: Interval ~ Duration + 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
```

Comparing the two models with and without day variable, the degree of freedom goes down from 105 to 98 due to addition of 8 levels from date variable. The f-value for two models can be calculated by as,

$$F = \frac{(SS_1 - SS_2)(DF_2)}{(SS_2)(DF_1 - DF_2)}$$

The f-value is calculated as 0.2086 and the p-value for this f-value is 0.98. As the p-value is much higher, we can conclude that the addition of date variable does not improve the model.

Question 1c:

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

Using k-fold cross validation (with k=10), the average RMSE for old and new model is 6.45 and 6.44 respectively. Thus there is no significant improvement with the addition of new variable date in regression analysis. From these RMSE values the model without date variable seems to have higher predictive accuracy.

Question 2:

Summary:

In this problem we analyzed data from around 15,000 babies with parents from diverse races, educational and financial backgrounds and tried to analyze if a mother smokes does this habit lead to lesser average birth weight than babies whose mothers do not smoke. We also included other predictors to predict birth weight and found that statistically only a few of the variables are significant contributors in the analysis.

Introduction:

It is known fact today that smoking by mothers can lead to a variety of health issues in babies. One of the first studies conducted to analyze this claim was done fifty years ago at Kaiser Foundation Hospital in Oakland, CA. This study collected data over a span of seven years from familes with different socio economic backgrounds. The goal of this assignment is to analyze this data and come up with paramters that affect the health of new born babies. We also analyze the effect of smoking by mothers on birth weight.

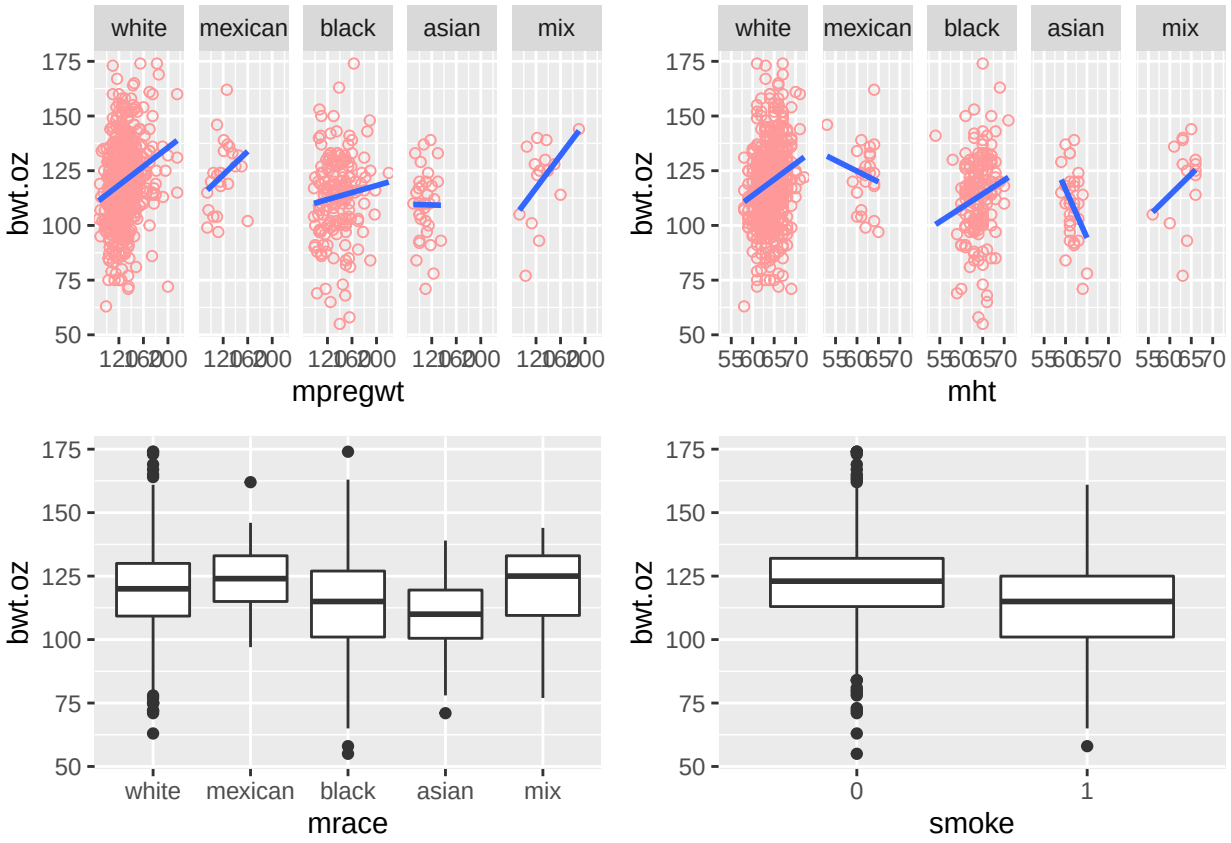
Data:

The original data provided has information columns about mother as well father. But the father information columns are empty for many data points. For the purpose of this study we have ignored those columns. The data used in this assignment has 12 fields. The data dictionary is as follows:

1. **id:** Index Number.
2. **date:** Birth date (number of days from January1, 1961) *(discrete variable)*
3. **gestation:** Length of gestation in days. *(discrete variable)*
4. **bwt.oz:** Birth weight in ounces. *(continuous variable)*
5. **parity:** Total number of previous pregnancies, including fetal deaths and still births. *(discrete variable)*
6. **mrace:** Mother's race *(categorival variable)*
7. **mage:** Mother's age *(discrete variable)*
8. **med:** Mother's education *(discrete variable)*
9. **mht:** Mother's height *(discrete variable)*
10. **mpregwt:** Mother's pre-pregnancy weight in pounds *(discrete variable)*
11. **inc:** Family income *(categorical variable)*
12. **smoke:** Mother's smoking habit *(categorical variable)*

The variables *mrace*, *med*, *income* and *smoke* in the data are discrete, which have been converted to categorical variables as they do not have a relative order.

From preliminary analysis of data the following variables seem to affect the birth weight,



From exploratory data analysis, the variables *parity*, *mrace*, *smoke*, *mht*, *pregwt*, *mage* seem to have a relationship with the response variable *bwt.oz*. Though these relationships are not very 'linear' in nature, no transformation seemed to fit the variables well. So we use them untransformed. The interaction between *mpregwt* and *mht* was tested as intuitively the variables seem related. But the interaction did not seem signifcant. Other interactions tested included *mrace*:*mpregwt*, *med*:*mht*, but none of the interaction terms were significant enough to be added in the regression analysis.

Model:

1. We fit the model using the variables *parity*, *mrace*, *smoke*, *mht*, *pregwt*, *mage*. The summary of this model indicates that *mage* is not significant. To check if *mage* has significant effect on the analysis we create another model without the *mage* variable and perform anova test. The p-value for anova test is 0.29 and thus we can conclude that *mage* does not make any significant improvement in the model and hence can be omitted.

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ parity + mrace + smoke + mpregwt + mht + mage
## Model 2: bwt.oz ~ parity + mrace + smoke + mpregwt + mht
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    859 239871
## 2    860 240014 -1   -142.31 0.5096 0.4755
```

2. To check if the interaction term *mpregwt*:*mht* is significant or not, we build two models with and without the intercation term.

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ parity + mrace + smoke + mpregwt + mht + mpregwt:mht
## Model 2: bwt.oz ~ parity + mrace + smoke + mpregwt + mht
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    859 240002
## 2    860 240014 -1   -11.261 0.0403 0.8409
```

The p-value of anova test is 0.91 and hence we conclude that interaction is insignificant and need not be included.

3. In the model with variables *parity*, *mrace*, *smoke*, *mht*, *pregwt*, the variable *parity* has realtively higher p-values(less than 0.05 but higher than others). Importance of this variable is checked by ANOVA test.

```
## Analysis of Variance Table
##
## Model 1: bwt.oz ~ mrace + smoke + mpregwt + mht
## Model 2: bwt.oz ~ parity + mrace + smoke + mpregwt + mht
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    861 241264
## 2    860 240014  1    1250.3 4.4798 0.03458 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the anova test is 0.01 and hence we reject the null hypothesis and conclude that parity is in fact an important variable.

4. **Final model** After trying out different variable combinations and interactions the variables included in the final model are *parity*, *mrace*, *smoke*, *mht* and *mpregwt*. The summary of the final model is shown below:

```
##
## Call:
## lm(formula = bwt.oz ~ parity + mrace + smoke + mht + mpregwt,
##     data = smoking)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.969  -9.525  -0.336  10.131  50.206
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.64712   15.38492   3.227 0.001298 **
## parity        0.66507    0.31422   2.117 0.034584 *
## mracemexican  3.29715    3.46725   0.951 0.341902
## mraceblack   -8.82690    1.51623  -5.822 8.22e-09 ***
## mraceasian   -7.93888    3.03506  -2.616 0.009060 **
## mracemix     -1.98421    4.38639  -0.452 0.651126
## smoke1       -9.35194    1.15218  -8.117 1.65e-15 ***
## mht           0.93387    0.26070   3.582 0.000360 ***
## mpregwt       0.10808    0.03217   3.360 0.000814 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.71 on 860 degrees of freedom
## Multiple R-squared:  0.1514, Adjusted R-squared:  0.1435
## F-statistic: 19.17 on 8 and 860 DF,  p-value: < 2.2e-16
```

From the summary above the R^2- value for the fnal model is 15.28%. The variable *smoke* has the highest contribution as it has the highest absolte t-value of 8.24.

From the residual plots the error seems to approximately normal and equal variance. The *qqplot* is straight line but diverges towards the ends. The residual vs. leverage plot has a couple of points with high Cook's distance and could be potential outliers.

Result

1. *Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke?*

From the summary of the final model it is evident that on an average the weight of babies whose mothers smoke is 9.5 ounces lesser than the average weight of babies whose mothers do not smoke.

2. *What is a likely range for the difference in birth weights for smokers and non-smokers?*

The 95% confidence range for the difference in birth weights for smokers and non-smokers is [120.9482446, -11.3454774]

3. *Is there any evidence that the association between smoking and birth weight differs by mother's race?*

Visually the plot for *mrace* and *bwt.oz* factored by *smoke* variable seems to have some pattern. Specillay for white and black races the birth weight is seen to be lower for non-smoker mothers. But using *mrace*:*smoke* does not yeild any significant results.

Shorcoming

The R^2-value for the model is quite low and hence the given model does not do a good job. The variables do not seem to have strong linear relationship with the birth weight. Probably techniques other than linear regression can yield better results for this data.