Assignement_2

Anshupriya Srivastava (netID: as996)

9/15/2019

Question_1

```
## Loading required package: carData
```

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).

Fitting the Regression Line

```
model_1 <- lm(data = OldFaithful, Interval ~ Duration + as.factor(Date))
summary(model_1)
```

```
##
## 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
```

When we use categorical variable in a regression model we need a baseline for comparison. The baseline group here is Day = 1 (value = 32.8770). Date6 seems to have a significance difference of mean intervals compared to the first day. The slope offset for Date6 is 1.9918 and absolute t-value offset is also the highest 0.740. Date2 also seems to have a significant difference since the slope offset for Date2 is 1.3275 and absolute t-value offset is 0.489. T-value measures the difference in means of population distributions. The regression model shows that Date2 and Date6 have the highest values of difference in means. So, we can say that these dates have a significant difference of mean as compared to the Date1.

Controlling for Day shows that there is a significance differnce at Date2 and Date6. These two dates have the higher impact on the final output as compared to the other dates.

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 use ANOVA to capture which model is better out of the two.

```
model_2 <- lm(data = OldFaithful, Interval ~ Duration)
anova(model_1, model_2)
```

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

We can see from the results that the P-Value is significantly large. This implies that adding an additional parameter does not improve the model.

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?

```
## [1] "RMSE (new model): 6.45932682262124"
```

```
## [1] "RMSE (old model): 6.49555986265575"
```

The new model has a lower RMSE. But the difference in the RMSE between the two models is not significant. We cannot say one model is better than the other.

Question 2

Introduction

The data for this exercise has been obtained from an Obsevational Study conducted by Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. The original dataset contains details of about 15,000 families but we have taken into account 869 male single births where the baby lived for at least 28 days. The mothers were interviewed quite early in their pregnancy to collect information on socioeconomic and demographic characteristics, with an indicator of whether the mother smoked during pregnancy. Our aim is to analyze this data and identfy any associations between smoking and birth weight. By the doing so, we aim to answer the following questions:

```
* Do mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke?
* What is a likely range for the difference in birth weights for smokers and non-smokers?
* Is there any evidence that the association between smoking and birth weight differs by mother’s race?
* Are there other interesting associations with birth weight that are worth mentioning?
```

Data

The data for this exercise was made available Sakai - filename “smoking.csv” has been used to answer the questions mentioned above. The list of variables present in the dataframe:

```
    *id: (numeric): id number
    *date: birth date where 1096 = January1, 1961
```

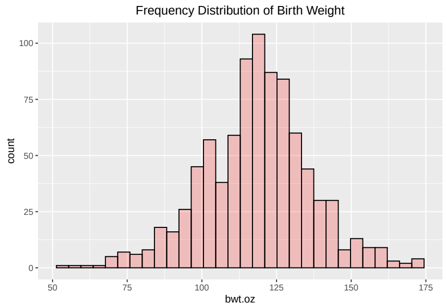
```
*gestation: length of gestation in days
*bwt.oz: birth weight in ounces
*parity: total number of previous pregnancies, including fetal deaths and still births
*mrace:mother's race or ethnicity
*mage: mother's age in years at termination of pregnancy
*med: mother's education
*mht: mother's height in inches
*mpregwt: mother's pre-pregnancy weight in pounds
*inc: family yearly income in 2500 increments
*smoke: does mother smoke?
```

Transformations:

```
*Race category 0 - 5 has been collapsed into one category for race = white.
*Mother's education has been renamed and categories 6-7 have been collapsed into 6,7 = trade school.
```

Our **response variable is birth weight or bwt.oz**. I have used a histogram to view the distribution of this variable.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



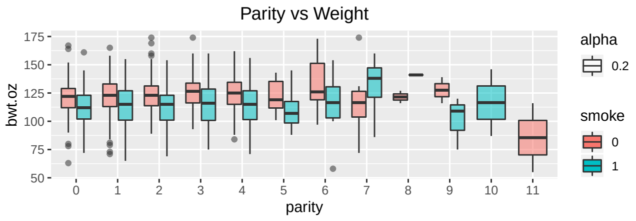
The birthweight data looks **normally distributed** with a slight left skew.

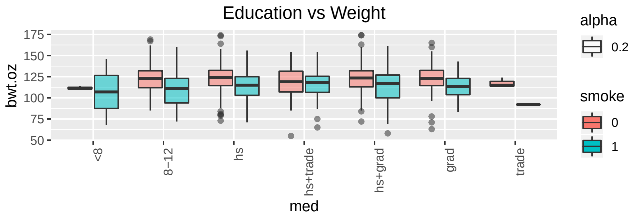
Hypothesis:

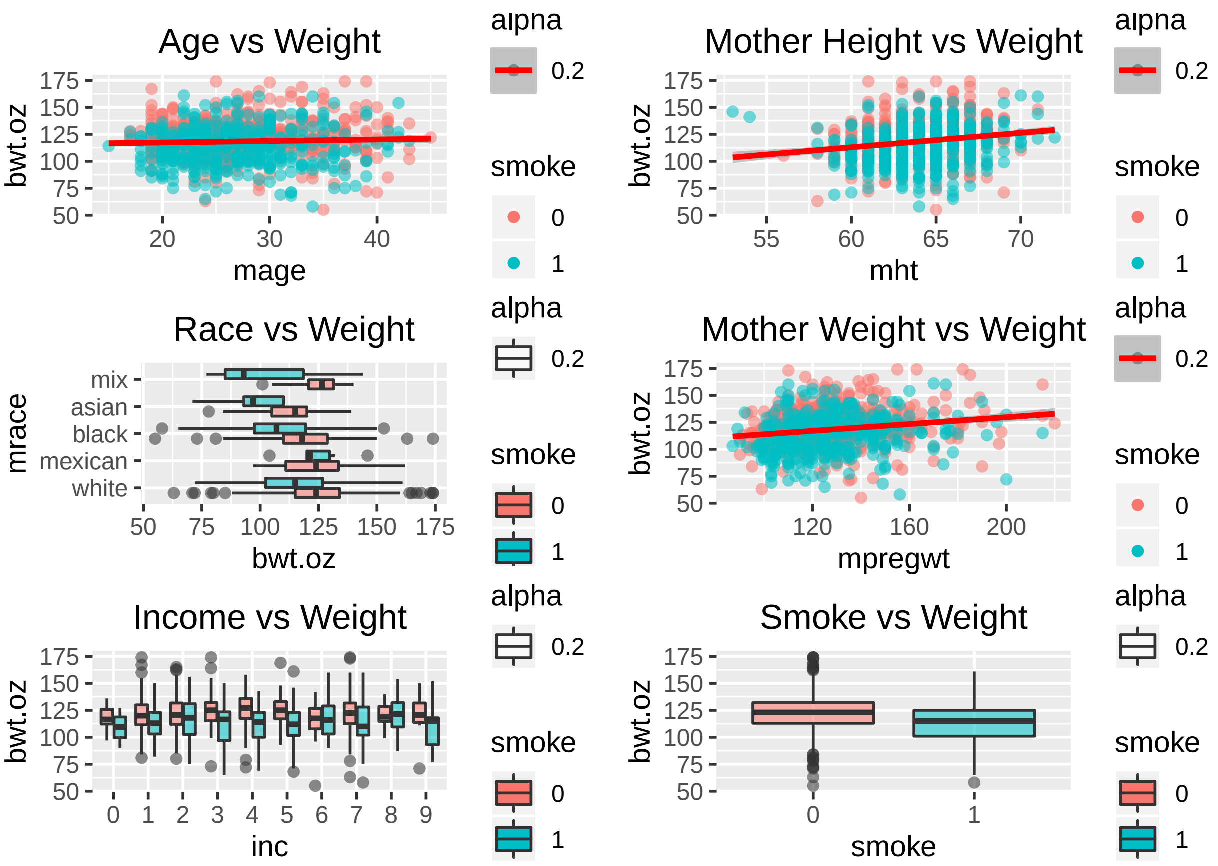
$H_0 :$ *The difference in the mean weight of the babies whose mothers smoke versus those who do not is* 0.

$H_1 :$ *The difference in the mean weight of the babies whose mothers smoke versus those who do not is greater than* 0.

Relationships:







By studying the graphs above we can can determine the following about our dataset:

i. Parity vs Weight: All categories show significant difference in mean weights. Women across all parity level who smoke seem have babies with weight lesser than the average. Based on this observation I will be using this in my model.

ii. Race vs Weight: Women of mixed races seem to show a very significant difference of weights. Apart from mexicans all other races show variations in their median weight. Race will be a part of the model as well.

iii. Age vs Weight: Age does not seem to be linearly related to the weight of the baby. I will be not be using Age in my model.

iv. Mother Height vs Height: There seems to be a linear relationship between the two categories even though it seems to be heavily infuenced by the dispersed weights around it. Since this shows a linear relationship this will be a part of the model as well.

v. Education vs Weight: Education does not show a very largr difference in weights. So, it will not be a part of my model.

vi. Mother Weight vs Weight: The seems to be a linear relationship between the two categories even though it seems to be heavily infuenced by the dispersed weights around it specially towards one side. Since this shows a linear relationship this will be a part of the model as well.

vii. Income v/s Weight: Doesn't seem to show any significant difference. I will not be using this in my model.

viii. Smoke v/s Weight: The weights of babies whose mothers smoke seems to be lesser than those of don't. This variable will be a part of the model.

Model

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

Call: lm(formula = bwt.oz ~ smoke, data = smoking)

At 95% confidence level, there is a significant difference (**p-value = 9.39e-14**) of the two weights. Here, the p-value is less than 0.05 so we can **reject the null hypothesis**. This gives evidence to support our assumption that mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke.

```
##                 2.5 %     97.5 %
## (Intercept) 120.94824 124.129009
## smoke1      -11.34548  -6.674704
```

The maximum difference of the mean at 95% confidence interval can be as low as -11.34548 and as high as -6.674704.

```
## [1] "RMSE: 17.4718654891152"
```

Impact of other variables on birth weight

Is there any evidence that the association between smoking and birth weight differs by mother's race? If so, characterize those differences.

The "Race vs Weight" plot shows that women of mixed races seem to show a very significant difference of weights. Apart from mexicans all other races show variations in their median weight. Using a regression model with smoke and race as intercation variables we can quantify the relationship.

```
##
## Call:
## lm(formula = bwt.oz ~ smoke + mrace + smoke:mrace, data = smoking)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62.37 -10.71   0.04  10.63  56.63
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         124.70886    0.96527 129.196  < 2e-16 ***
## smoke1               -9.49918    1.37169  -6.925 8.53e-12 ***
## mracemexican         -2.18254    4.05317  -0.538  0.59039
## mraceblack           -7.33652    2.01594  -3.639  0.00029 ***
## mraceasian          -11.10886    3.56498  -3.116  0.00189 **
## mracemix             -0.20886    5.04656  -0.041  0.96700
## smoke1:mracemexican  11.13953    8.15169   1.367  0.17213
## smoke1:mraceblack     0.08684    2.98991   0.029  0.97684
## smoke1:mraceasian    -6.21193    6.80981  -0.912  0.36192
## smoke1:mracemix     -10.33415   11.16072  -0.926  0.35474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.16 on 859 degrees of freedom
## Multiple R-squared:  0.1057, Adjusted R-squared:  0.09636
## F-statistic: 11.28 on 9 and 859 DF,  p-value: < 2.2e-16
```

A **p-value (2.2e-16)** < 0.05 implies that we can reject the null hypothesis. The interation between smoke and race has increased R-squared value from 0.06095 to 0.09636 suggesting that this model fits the data better.

```
## [1] "RMSE: 17.0600226515754"
```

The RMSE value for this model is lesser than the previous model which agaian implies a better fit.

Are there other interesting associations with birth weight that are worth mentioning?

The variables smoke, mrace, parity, mht and mpregwt are *variables of interest*. Explanations for the same has been given above.

```
## 
## Call:
## lm(formula = bwt.oz ~ smoke + mrace + parity + mht + mpregwt, 
##     data = smoking)
## 
## 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 ** 
## smoke1        -9.51225    1.15340  -8.247 6.14e-16 ***
## mracemexican   3.84011    3.48647   1.101 0.271020    
## mraceblack    -8.68070    1.52411  -5.696 1.69e-08 ***
## mraceasian    -7.97365    3.03160  -2.630 0.008688 ** 
## mracemix      -1.91987    4.39028  -0.437 0.662004    
## 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 *  
## mht            0.98204    0.26077   3.766 0.000177 ***
## mpregwt        0.09906    0.03215   3.081 0.002126 ** 
## ---
## 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
```

```
## [1] "RMSE: 16.4316140121331"
```

Results

Using **R-Squared** and **RMSE** values I have decided to use the **third model**. It is a multiple linear regression that uses smoke, mrace, parity, mht and mpregwt to predict the birthweight of newborn babies. I am using this model because it has the **highest R-squared** valiue of **0.1528**. Also, this model has the **lowest RMSE** which is **16.43**. The regression output for the model is available above and the RMSE calculation.

Conclusion - Limitaion, takeaway, future work

None of the variables show a strong linear relationship with the response variable. The linearity is heavily influenced by the datapoints that do not lie on the regression line. Hence using a simple linear regression model is not sufficient in this case. The model with the highest R-squared value (0.1528) is also not the best fit for this data because it only justifies 15% of the relationship which is quite low. This dataset requires a more complex model to explain the dependency of birthweight on the variables of interest.