# Midterm Exam

Rudy Martinez

10/7/2020

### Set Working Directory

```
setwd("/Users/rudymartinez/Desktop/MSDA/Fall 2020/STA 6443_Algorithms I/STAT-Algorithms-1/Week 7/Midterm Exam"
```

### Read File

```
bweight = read.csv("birthweight.csv", header=TRUE);
bweight$Black = as.factor(bweight$Black);
bweight$Married = as.factor(bweight$Married);
bweight$Boy = as.factor(bweight$Boy);
bweight$MomSmoke = as.factor(bweight$MomSmoke);
bweight$Ed = as.factor(bweight$Ed);
bweight$Weight = as.numeric(bweight$Weight)
```

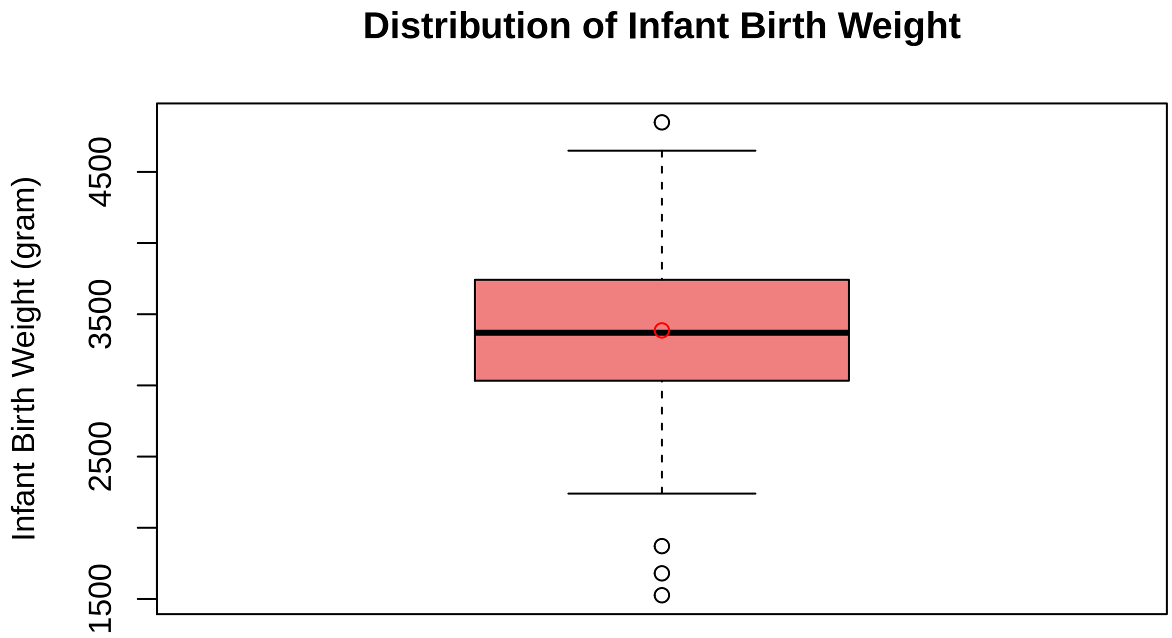
### Libraries

```
library(DescTools)
library(MASS)
library(car)
library(tidyverse)
```

### Exercise 1.A

Generate Boxplot for infant birth weight (`Weight`) and comment on the general features of the distribution. Generate a normal QQ-plot and perform Shapiro-Wilk test to check whether normality is a reasonable assumption for `Weight`. Make a conclusion.

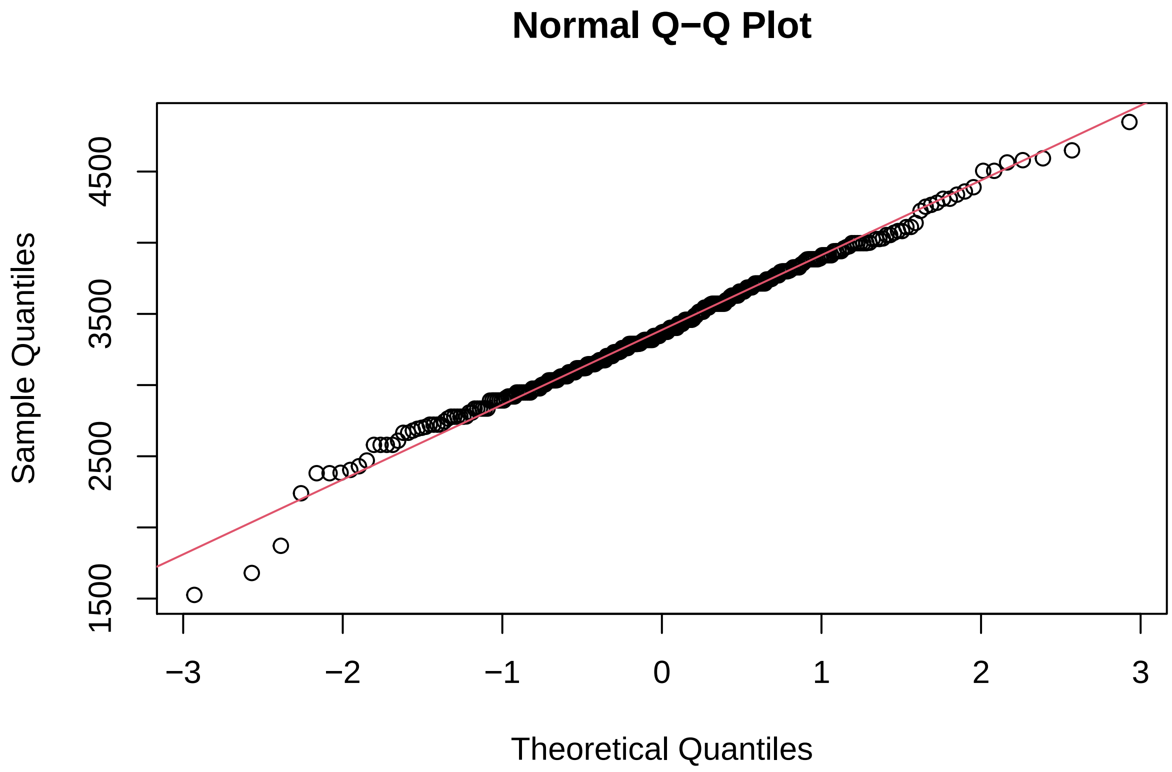
```
boxplot(bweight$Weight,
        main = "Distribution of Infant Birth Weight",
        ylab = "Infant Birth Weight (gram)",
        col = "lightcoral",
        border = "black",
        horizontal = FALSE
)

points(mean(bweight$Weight, na.rm=TRUE), col="red")
```



**Comments:** The Boxplot above represents the distribution of Infant Birth Weight (grams):

- Visually, the Boxplot appears to be symmetrical
- The mean and median nearly overlap, indicating that the distribution is likely to be symmetric
- In sum, the mean and median Infant Birth Weight are positioned within the `Weight` range of $3000 - 3500$ grams

```r
qqnorm(bweight$Weight); qqline(bweight$Weight, col = 2)
```



```
shapiro.test(bweight$Weight)
```

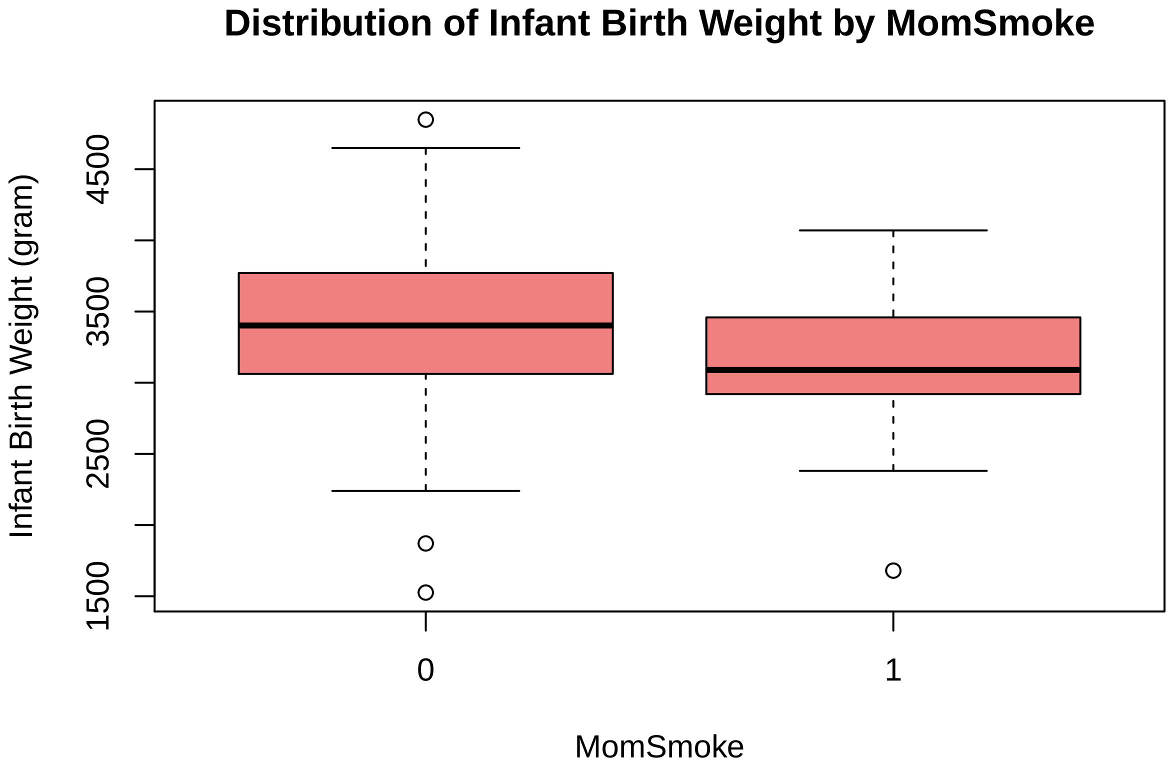
```
##
##  Shapiro-Wilk normality test
##
## data:  bweight$Weight
## W = 0.99206, p-value = 0.1153
```

**Conclusion:** Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable. Moreover, the Shapiro-Wilk normality test produces a p-value **above** the significance level of 0.05, meaning that there is not enough evidence to reject the null hypothesis. Thus, `Weight` follows a normal distribution.

### Exercise 1.B

Generate a boxplot of `Weight` by `MomSmoke` and compare infant birth weights between smoking levels.

```
boxplot(Weight ~ MomSmoke, data=bweight,
        main = "Distribution of Infant Birth Weight by MomSmoke",
        xlab = "MomSmoke",
        ylab = "Infant Birth Weight (gram)",
        col = "lightcoral",
        border = "black",
        horizontal = FALSE
)
```



**Note:** `MomSmoke` is a categorical variable where 0 is **non-smoking mom**, 1 is **smoking mom**

- Visually, it appears that Categorical Variables **non-smoking mom** and **smoking mom** do not have the same average Infant Birth Weight
- The Group 0 (**non-smoking mom**) Boxplot appears to be symmetrical, while the Group 1 (**smoking mom**) Boxplot appears to be slightly right skewed.

## Exercise 1.C

For each level in `MomSmoke`, perform Shapiro-Wilk test for checking the Normality of `Weight.` Make a conclusion.

```
shapiro.test(bweight[bweight$MomSmoke=="0", "Weight"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bweight[bweight$MomSmoke == "0", "Weight"]
## W = 0.99362, p-value = 0.3549
```

```
shapiro.test(bweight[bweight$MomSmoke=="1", "Weight"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bweight[bweight$MomSmoke == "1", "Weight"]
## W = 0.96299, p-value = 0.2
```

**Conclusion:**

- The Shapiro-Wilk normality test results produces a p-value that is **above** the significance level of 0.05 for each `MomSmoke` group (**non-smoking mom** and **smoking mom**), meaning that there is not enough evidence to reject the null hypothesis. Thus, an assumption of Normality is reasonable for the `Weight` variable by each `MomSmoke` group.

## Exercise 2

We want to test if there is a significant difference in birth weights between infants from smoking mom and non-smoking mom. Perform a hypothesis test of whether infants from smoking moms have different weights than infants from non-smoking moms. Which test do you choose? Use the answer in Exercise 1 for choosing the proper test. Specify null and alternative hypotheses and state your conclusion.

```r
var.test(Weight ~ MomSmoke, bweight, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  Weight by MomSmoke
## F = 1.0786, num df = 253, denom df = 40, p-value = 0.8009
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6421109 1.6671729
## sample estimates:
## ratio of variances
##           1.078555
```

**Comments:** Although **non-smoking mom** and **smoking mom** groups both follow a normal distribution, we must check for equal variance to determine the proper test.

- After conducting an F test (var.test), the results indicated a p-value of **0.8009**. This high p-value means we can't reject the null hypothesis, indicating that the two groups have an **equal variance**.
- Based on these findings, the next step is to conduct a **Pooled T-test**

### Specify the Null and Alternative Hypothesis

Null: mean `Weight` of **non-smoking mom** group is equivalent to the mean `Weight` of **smoking mom** group.

Alternative: mean `Weight` of **non-smoking mom** group is not equivalent to the mean `Weight` of **smoking mom** group.

### Pooled T-test

```r
t.test(Weight ~ MomSmoke, bweight ,alternative ="two.sided", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Weight by MomSmoke
## t = 3.071, df = 293, p-value = 0.002334
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   93.37931 426.65488
## sample estimates:
## mean in group 0 mean in group 1
##        3422.724        3162.707
```

**Conclusion:** The Pooled t-test produced a p-value **below** the significance level of 0.05, meaning we reject the null hypothesis. This indicates that the Mean `Weight` of **non-smoking mom** group and **smoking mom** group are NOT equivalent.

## Exercise 3.A

Now perform one-way ANOVA on `Weight` with `MomSmoke`. Check homogeneity of variance assumption. Does it hold, and is it okay to perform ANOVA?

```r
aov.res_weight = aov(Weight ~ MomSmoke, data = bweight)

summary(aov.res_weight)
```

```
##              Df   Sum Sq Mean Sq F value  Pr(>F)   
## MomSmoke      1  2386708 2386708   9.431 0.00233 **
## Residuals   293 74151291  253076                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

**Comments:** The p-value of 0.00233 is below the significance level of 0.05, meaning that we **reject** the null hypothesis. Therefore, `MomSmoke` has a significant effect on `Weight` (at least one group in `MomSmoke` has a different mean of `Weight`).

### Check Homogeneity of Variance Assumption

```r
LeveneTest(aov.res_weight)
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.6767 0.4114
##       293
```

**Comments:** The p-value is above the significance level of 0.05, meaning that we **can’t reject** the null. Therefore, all groups in `MomSmoke` have the same variance.The homogeneity of variance assumption **holds**. It is valid to perform ANOVA.

## Exercise 3.B

```r
ScheffeTest(aov.res_weight)
```

```
## 
##   Posthoc multiple comparisons of means: Scheffe Test 
##     95% family-wise confidence level
## 
## $MomSmoke
##          diff    lwr.ci    upr.ci   pval    
## 1-0 -260.0171 -426.6549 -93.37931 0.0023 ** 
## 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

**Conclusion:** Due to a p-value (0.0023) below the significance level of 0.05, `MomSmoke` pair 1-0 has a significant effect on `Weight`.

- Specifically, `MomSmoke` group 0 (**non-smoking mom**) infants have *greater average weights* than infants from `MomSmoke` group 1 (**smoking moms**).
- Simply put, the mean `Weight` of group 0 (**non-smoking mom**) is greater than the mean `Weight` of group 1 (**smoking moms**).

## Exercise 4

Using `Black`, `Married`, `Boy`, and `MomSmoke`, and `Ed` variables as possible effects, find the best ANOVA model for Weight. Manually perform backward selection based on **type3 SS** result with **0.05** criteria on p-value. Perform backward selection only with main effects and then check the interaction effects only based on significant main effect terms.

## Exercise 4.A

Write down step by step how you perform backward selection and how you find the final model. Please do NOT include all intermediate tables and graphs in the report. Just describe each step which variable you delete and why.

### Start With Full Model (5 Variables)

```
aov.res_weight_2 = aov(Weight ~ Black + Married + Boy + MomSmoke + Ed, data = bweight)
Anova(aov.res_weight_2, type = 3)
```

**Results:** Based on the results of the Type 3 ANOVA test, `Black` and `MomSmoke` have a p-value below the significance level of 0.05 (0.0008217 and 0.0027017 respectively), meaning we reject the null. Both of these categorical predictors have a significant effect on `Weight`.

On the contrary, `Married`, `Boy`, and `Ed` have a p-value above the significance level of 0.05 (0.6394546, 0.3763046, and 0.8625846 respectively), meaning they **do not** have a significant effect on `Weight`. **`Ed` is removed from the model as it has the highest p-value.**

### Model With 4 Variables

```
aov.res_weight_3 = aov(Weight ~ Black + Married + Boy + MomSmoke, data = bweight)
Anova(aov.res_weight_3, type = 3)
```

**Results:** Based on the results of the Type 3 ANOVA test, `Black` and `MomSmoke` have a p-value below the significance level of 0.05 (0.0007778 and 0.0026466 respectively), meaning we reject the null. Both of these categorical predictors have a significant effect on `Weight`.

On the contrary, `Married` and `Boy` have a p-value above the significance level of 0.05 (0.6157671 and 0.3807876 respectively), meaning they **do not** have a significant effect on `Weight`. **`Married` is removed from the model as it has the highest p-value.**

### Model With 3 Variables

```
aov.res_weight_4 = aov(Weight ~ Black + Boy + MomSmoke, data = bweight)
Anova(aov.res_weight_4, type = 3)
```

**Results:** Based on the results of the Type 3 ANOVA test, `Black` and `MomSmoke` have a p-value below the significance level of 0.05 (0.0001223 and 0.0014471 respectively), meaning we reject the null. Both of these categorical predictors have a significant effect on `Weight`.

On the contrary, `Boy` has a p-value above the significance level of 0.05 (0.3888071), meaning it **does not** have a significant effect on `Weight`. **`Boy` is removed from the model as it has the highest p-value.**

### Model With 2 Variables

```r
aov.res_weight_5 = aov(Weight ~ Black + MomSmoke, data = bweight)
Anova(aov.res_weight_5, type = 3)
```

**Results:** Based on the results of the Type 3 ANOVA test, `Black` and `MomSmoke` have a p-value below the significance level of 0.05 (0.0001232 and 0.0013954 respectively), meaning we reject the null. Both of these categorical predictors have a significant effect on `Weight`.

#### Model With 2 Variables and Interaction

```r
aov.res_weight_6 = aov(Weight ~ Black * MomSmoke, data = bweight)
Anova(aov.res_weight_6, type = 3)
```

**Results:** Based on the results of the Type 3 ANOVA test, `Black` and `MomSmoke` have a p-value below the significance level of 0.05 (0.0002707 and 0.0026729 respectively), meaning we reject the null. Both of these categorical predictors have a significant effect on `Weight`.

On the contrary, the Interaction between `Black` and `MomSmoke` has a p-value above the significance level of 0.05 (0.8807474), meaning it **does not** have a significant effect on `Weight`. **The Interaction is removed from the model as it has the highest p-value.**

## Exercise 4.B

Specify the final model and report the amount of variation explained by the model. Also, check the Normality assumption through diagnostics plots.

#### Final Model

Model With **2 Variables and No Interaction**

#### Variation Explained by the Model (Predictors = `Black` and `MomSmoke`)

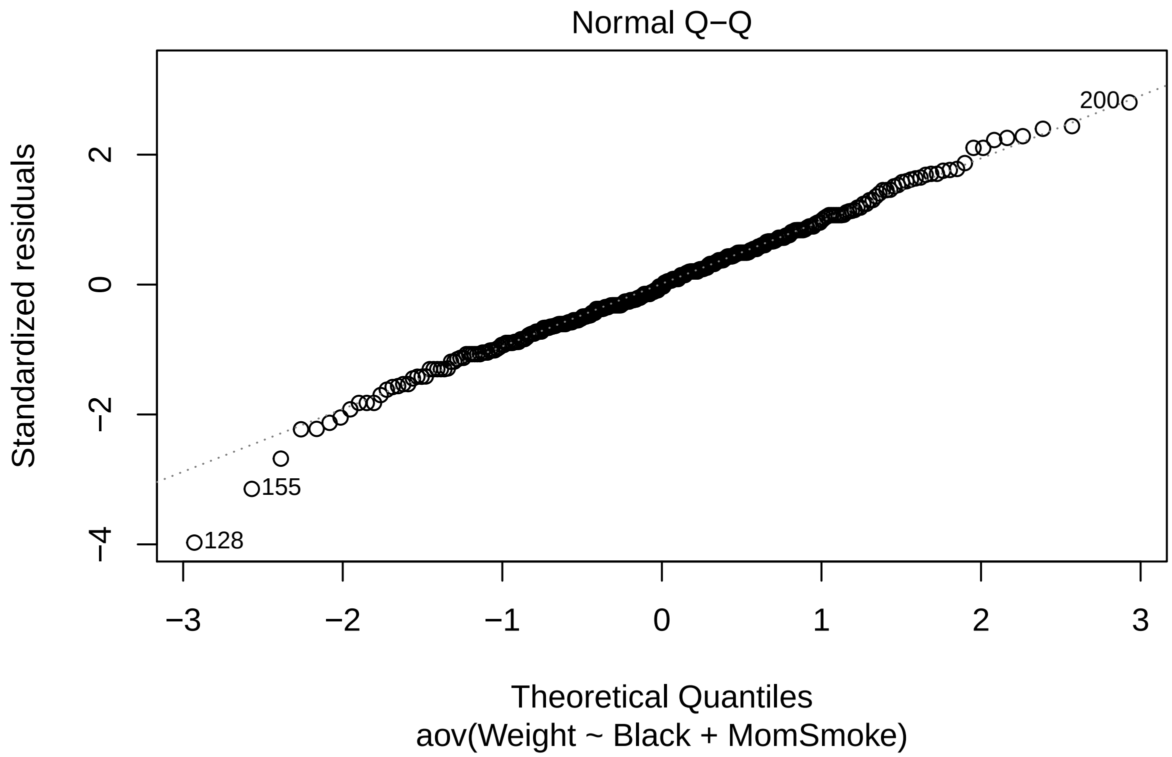
```r
lm.res_weight_7 = lm(Weight ~ Black + MomSmoke , data = bweight)
summary(lm.res_weight_7)$r.squared
```

```
## [1] 0.07896405
```

**Results:** 7.9% of the variation of `Weight` can be explained by the model (`Black` and `MomSmoke`).

#### Normality Check

```r
par(mfrow=c(1,1))
plot(aov.res_weight_5, 2)
```



**Diagnostics Plot:** Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable.

## Exercise 4.C

State conclusions about significant differences in `Weight` across groups. For each significant variable, state specifically which level has a larger or smaller mean value of `Weight`.

### Post-hoc Test

```r
TukeyHSD(aov.res_weight_5)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Weight ~ Black + MomSmoke, data = bweight)
##
## $Black
##          diff       lwr       upr     p adj
## 1-0 -293.9412 -445.2216 -142.6608 0.0001605
##
## $MomSmoke
##         diff       lwr       upr     p adj
## 1-0 -266.763 -429.5199 -104.0061 0.0013989
```

`Black`

- Due to a p-value below the significance level of 0.05 (0.0001605), `Black` pair **1-0** has a significant effect on `Weight`.
- Specifically, the following effects of `Black` on `Weight` can be seen:

    - 1 $<$ 0 (The mean `Weight` of 0 (White) is greater than the mean `Weight` of 1 (Black))

`MomSmoke`

- Due to a p-value below the significance level of 0.05 (0.0013989), `MomSmoke` pair **1-0** has a significant effect on `Weight`.
- Specifically, the following effects of `MomSmoke` on `Weight` can be seen:
    - 1 $<$ 0 (The mean `Weight` of 0 (**Non-Smoking Mom**) is greater than the mean `Weight` of 1 (**Smoking Mom**))

In summary, the analysis above indicates the following:

- Mothers classified as Black have a **lower** mean value of Infant Birth Weight compared to Mothers classified as White
- Mothers who smoke have a **lower** mean value of Infant Birth Weight compared to Mothers that **do not** smoke