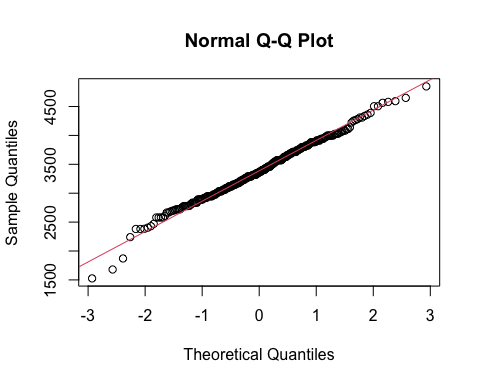
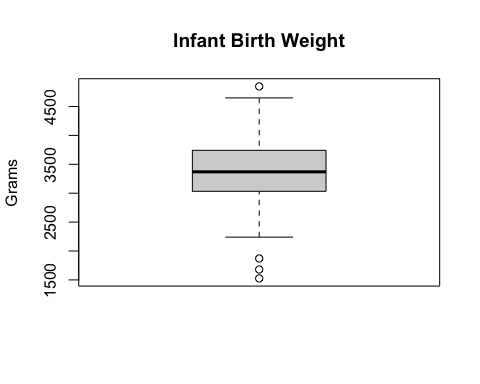
Midterm

Ashley Cortez

10/9/2021

Exercise 1

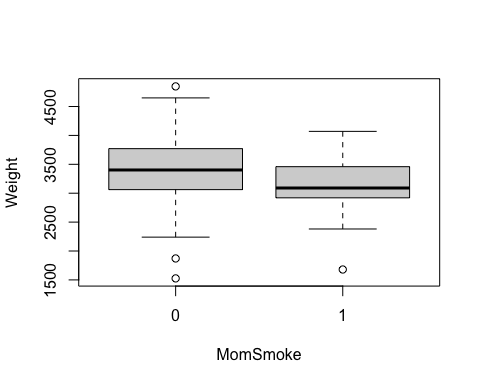
1. **Generate Boxplot for infant birth weight (Weight) and comment on the general features of the distribution. Generate a normal QQ-plot and perform Shapiro-wilk test to check whether normality is a reasonable assumption for Weight. Make a conclusion.**



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

From the shapiro test the p-value is 0.1153 > 0.05, therefore, we do not have enough evidence to reject the null hypothesis - conclusion: data follows normal distribution

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



The mean of infant weights with smoking moms is lower than the mean of infant weights with non-smoking moms. The mean = median for the non-smoking moms, meaning skewness is zero. The median > mean for smoking moms, meaning the data is slightly right skewed.

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

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

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

H0:Data follows normal distribution

H1: Data does not follow normal distribution

non-smoking mom p-value 0.3549

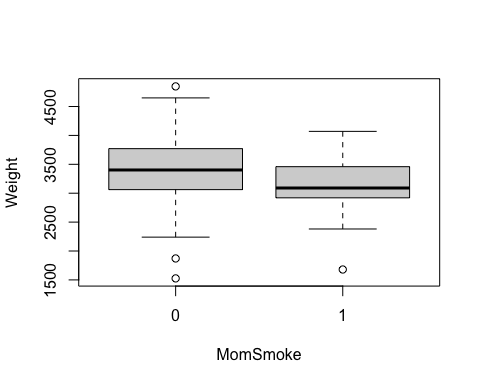
smoking mom p-value 0.2

Both p-values are larger than significance level, so we do not have enough evidence to reject the null hypothesis - our conclusion is that the data follows a normal distribution

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**

**NOTE: If you decide to use the parametric test, perform two-sample t-test rather than ANOVA.**



##   
## 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

##   
## Welch Two Sample t-test  
##   
## data: Weight by MomSmoke  
## t = 3.1562, df = 54.877, p-value = 0.002595  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 94.91074 425.12344  
## sample estimates:  
## mean in group 0 mean in group 1   
## 3422.724 3162.707

H0: mean of Weight = mean of MomSmoke

Ha:mean of Weight ≠ mean of MomSmoke

P-value - 0.002595 - we reject the null hypothesis and conclude that infants from smoking moms have different weights than infants from non-smoking moms.

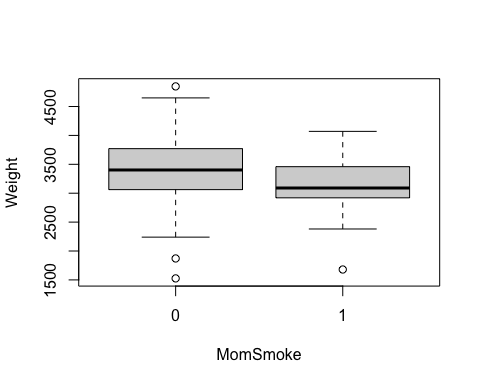
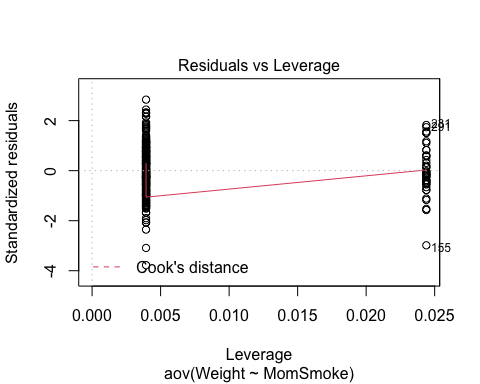
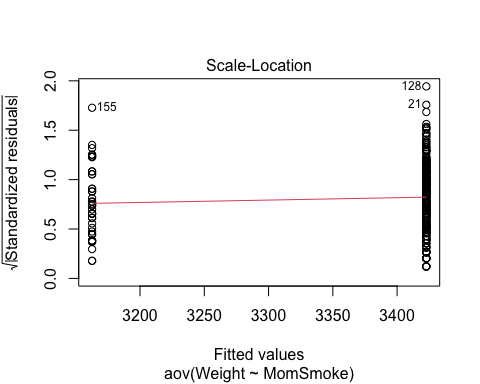
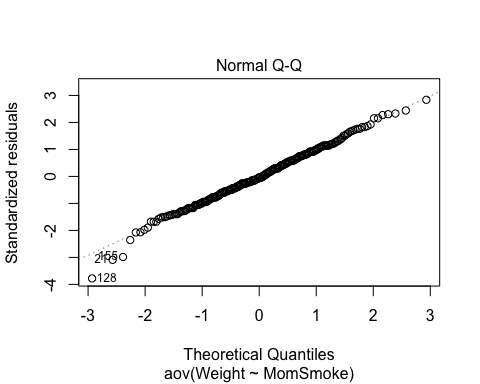
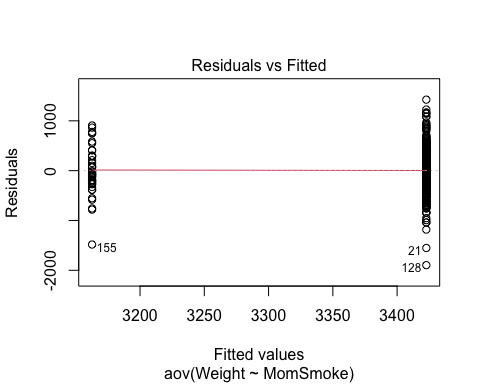
Exercise 3

Now perform one-way ANOVA on Weight with MomSmoke.

1. **Check homogeneity of variance assumption. Does it hold and okay to perform ANOVA?**

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

## 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



## Analysis of Variance Table  
##   
## Response: Weight  
## Df Sum Sq Mean Sq F value Pr(>F)   
## MomSmoke 1 2386708 2386708 9.4308 0.002334 \*\*  
## Residuals 293 74151291 253076   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## [1] 0.03118331

With a large p-value from the LeveneTest, we fail to reject the null concluding that the groups have the same variance. Following the homoscedasticity assumption. The plots match the normality assumption, specifically the Normal QQ plot. 31% of the variation of weight can be explained by the MomSmoke. Okay to perform ANOVA.

1. **Make a conclusion on the effect of MomSmoke. Compare your result with the conclusion of Exercise 2.**

## Anova Table (Type III tests)  
##   
## Response: Weight  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 2975620765 1 11757.8112 < 2.2e-16 \*\*\*  
## MomSmoke 2386708 1 9.4308 0.002334 \*\*   
## Residuals 74151291 293   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

H0: MomSmoke has no effect on infant weight H1: MomSmoke has an effect on infant weight

P-value 0.002334 < 0.05. We reject the null hypothesis, and our conclusion is that there exists a significant effect on MomSmoke on infant weight. Our conclusion is the same in exercise 2.

Exercise 4

1. **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.**

Continuous response: Weight (y)

Categorical predictors (Variables): Black, Married, Boy, MomSmoke, Ed

Start with a model that contains all the variables (Full Model) to find significant effect

Black - 0.0008217

Married - 0.6394546

Boy - 0.3763046

MomSmoke - 0.0027017

Ed - 0.8625846

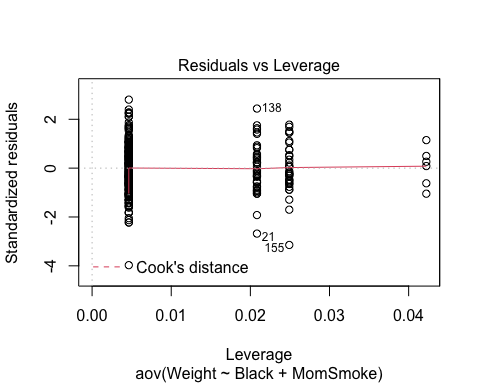
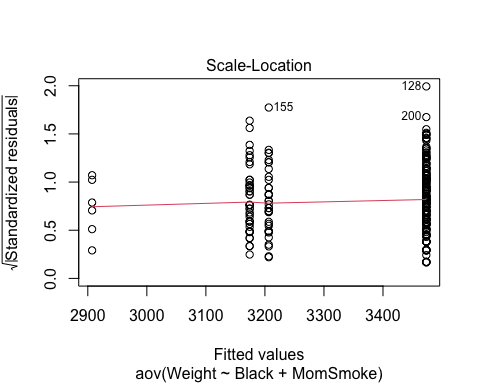
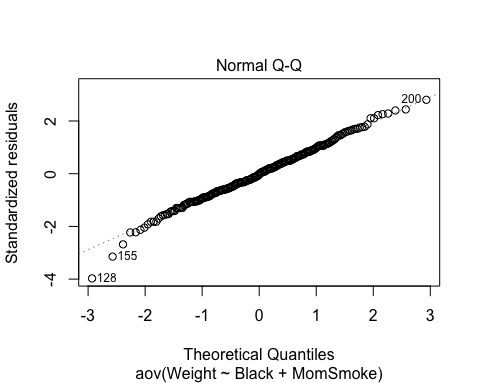
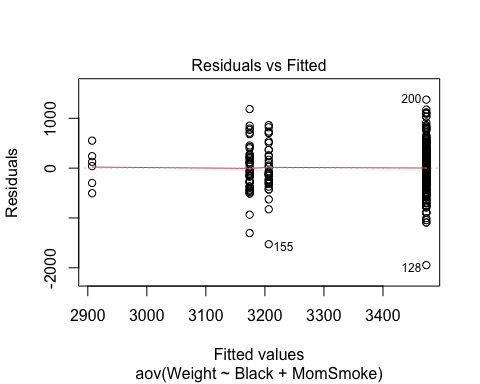
Remove the least significant variable Variables: Married, Boy, and Ed all have p-values larger than our significance level 0.05, therefore they do not have an effect on weight and should not be kept.

Keep removing the least significant variable until reaching the stopping rule (no more variables with a p-value greater than significance level of 0.05) or running out of variables

The final model should be: aov(Weight ~ Black + MomSmoke, data = bweight) because variables Black and MomSmoke have p-values less than the signifiance level of 0.05, therefore, they do have an effect on weight

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

## Anova Table (Type III tests)  
##   
## Response: Weight  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 2600800716 1 10772.989 < 2.2e-16 \*\*\*  
## Black 3657042 1 15.148 0.0001232 \*\*\*  
## MomSmoke 2513301 1 10.411 0.0013954 \*\*   
## Residuals 70494249 292   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



## Analysis of Variance Table  
##   
## Response: Weight  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Black 1 3530450 3530450 14.624 0.0001605 \*\*\*  
## MomSmoke 1 2513301 2513301 10.411 0.0013954 \*\*   
## Residuals 292 70494249 241419   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## [1] 0.07896405

The plots match the normality assumption, specifically the Normal QQ plot. 79% of the variation of weight can be explained by Black and MomSmoke

1. **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.**

ScheffeTest(aov.weight1)

##   
## Posthoc multiple comparisons of means: Scheffe Test   
## 95% family-wise confidence level  
##   
## $Black  
## diff lwr.ci upr.ci pval   
## 1-0 -293.9412 -483.0575 -104.8249 8e-04 \*\*\*  
##   
## $MomSmoke  
## diff lwr.ci upr.ci pval   
## 1-0 -266.763 -470.2261 -63.29987 0.0060 \*\*   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Black: Mean(white)>Mean(black)

MomSmoke: Mean(non-smoking mom)>Mean(smoking mom)