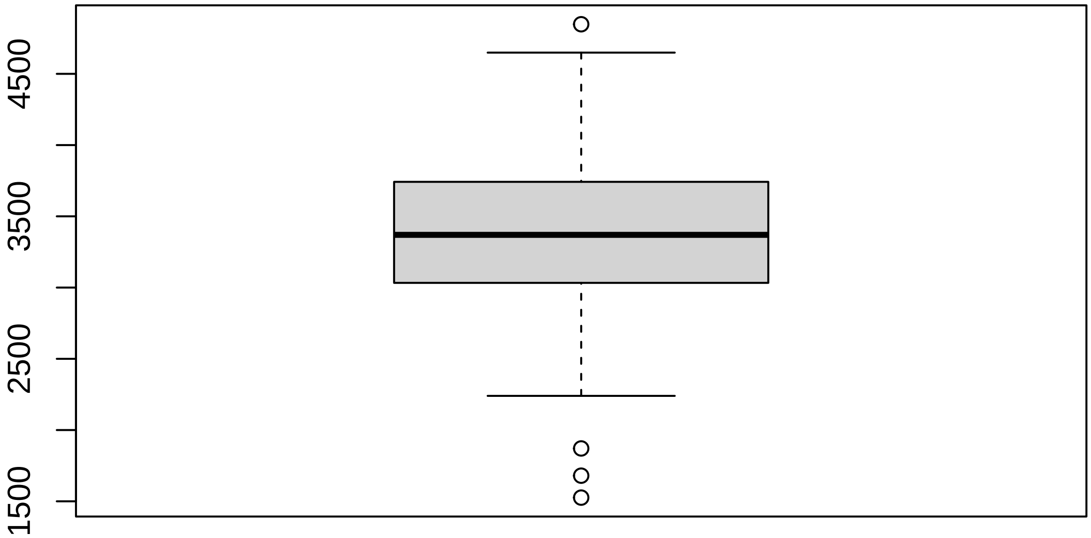
Midterm Exam_StatisticalModeling_Fall2021

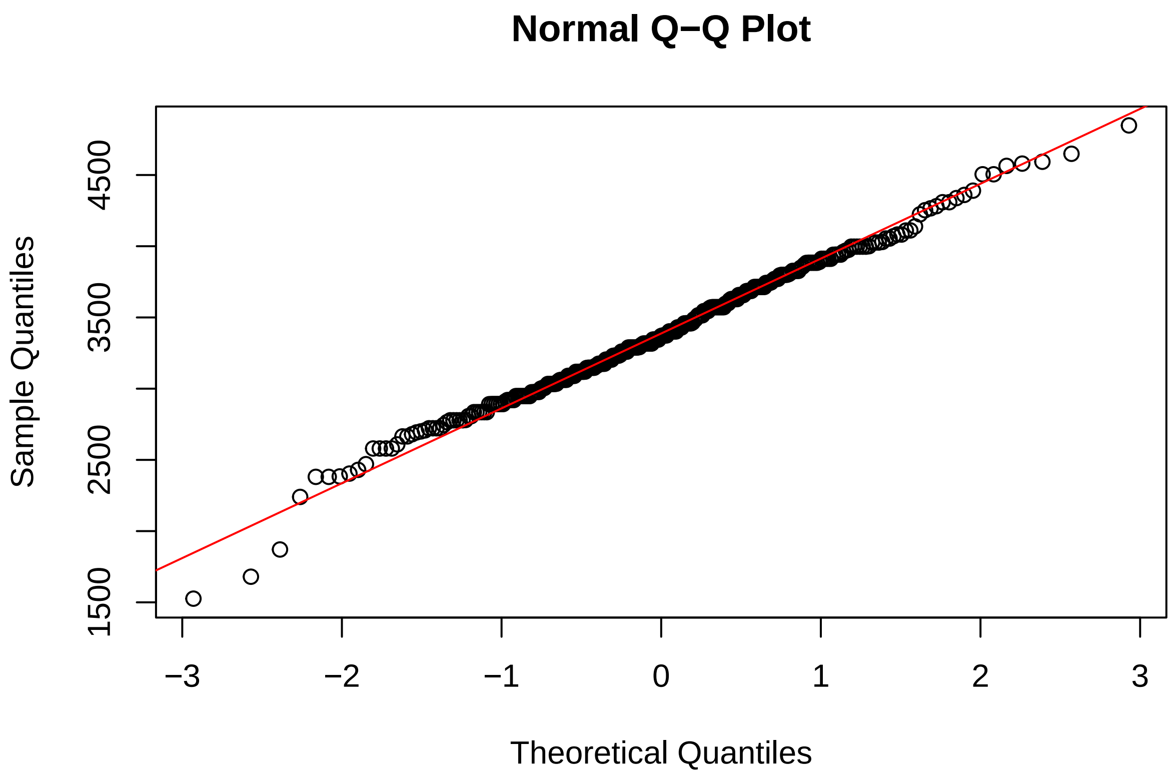
zho125 - Diego Aldo Pettorossi

10/5/2021

```r
boxplot(bweight$Weight)
```



```r
qqnorm(bweight$Weight)
qqline(bweight$Weight, col = "red")
```

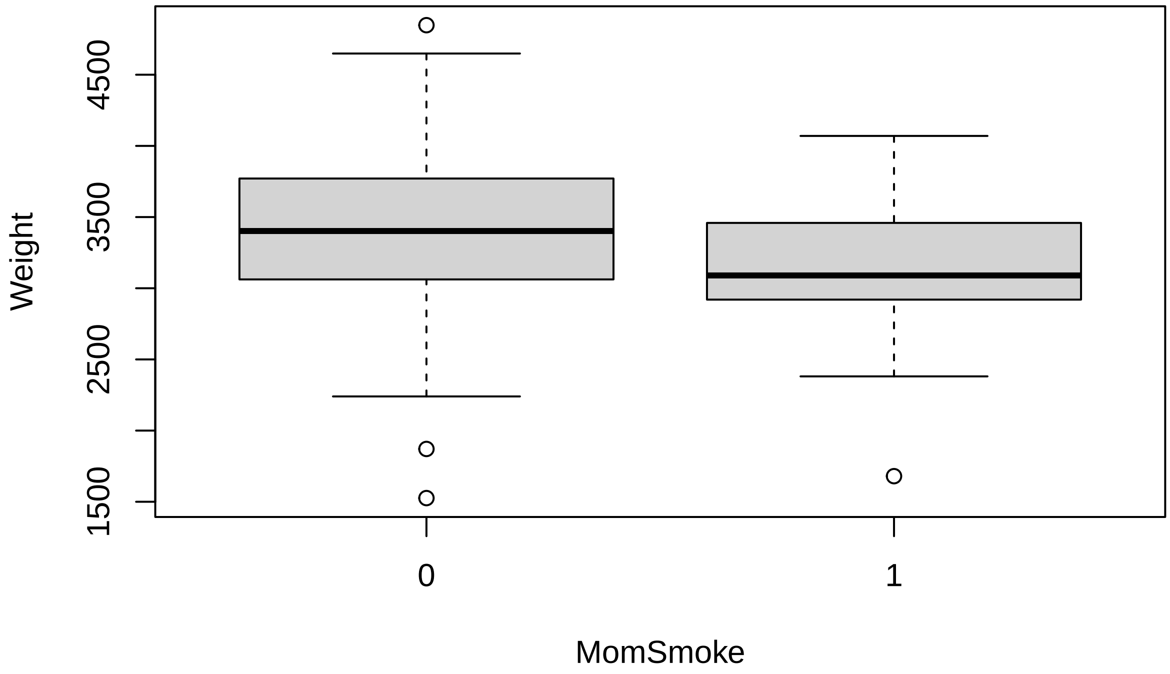


```
shapiro.test(bweight$Weight)
```

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

a) The boxplot and the QQplot of the variable Weight shows that the central values follow a normal distribution. However, the presence of outliers indicates that the distribution as a whole is not normal . In fact, the p-value of the Shapiro-Wilk test is greater the significance level, so we should reject the null hypothesis: the data doesn't follow a normal distribution.

```
boxplot(Weight ~ MomSmoke, data= bweight)
```



b) The within group variation and median of smoking mom's infant birth weight is smaller than the infant from smoking mom group.

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

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

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

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

c) The Shapiro-Wilk test results indicates that infants weights from smoking and non-smoking moms do not follow a normal distribution.

#Exercise 2

```
var.test(Weight ~ MomSmoke, data = 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
```

a) Since the data don't follow the normal distribution and they don't have equal variance we should perform the Wilcoxon Signed Rank Test

- Null hypothesis (H0) : Infants from smoking and non-smoking moms weights have the same median.

- Alternative hypothesis (H1) : Infants from smoking and non-smoking moms weights have different medians.

```
wilcox.test(Weight ~ MomSmoke, data = bweight)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Weight by MomSmoke
## W = 6717.5, p-value = 0.002886
## alternative hypothesis: true location shift is not equal to 0
```

b) The test indicates that the two populations have different medians.

#Exercise 3

```
aov.bweight1 = aov(Weight ~ MomSmoke, data = bweight)
summary(aov.bweight1)
```

```
##              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
```

```
LeveneTest(aov.bweight1)
```

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

```
oneway.test(Weight ~ MomSmoke, data = bweight, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Weight and MomSmoke
## F = 9.9617, num df = 1.000, denom df = 54.877, p-value = 0.002595
```

```
TukeyHSD(aov.bweight1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Weight ~ MomSmoke, data = bweight)
##
## $MomSmoke
##          diff       lwr       upr     p adj
## 1-0 -260.0171 -426.6549 -93.37931 0.0023339
```

```
ScheffeTest(aov.bweight1)
```

```
##
##   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
```

Since the two groups have different variances we should perform the Welch's ANOVA test.
Based on the test result, the conclusion is that infants from smoking and non-smoking moms have different mean weights.
Specifically, the post-hoc tests indicate the weight of second group is higher compared to the first one by about 260 g.

Accordingly with Exercise 2 and Exercise 3 the variable MomSmoke has a significant impact on Weight.

#Exercise 4

```
model = Anova(aov(Weight ~ Black + MomSmoke, data = bweight), type = 3) #STOP REMOVING
model
```

```
## 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
```

```
model = Anova(aov(Weight ~ Black*MomSmoke, data = bweight), type = 3) # NOT SIGNIFICANT
model
```

```
## Anova Table (Type III tests)
##
## Response: Weight
##                    Sum Sq  Df    F value    Pr(>F)
## (Intercept)    2546671287   1 10513.4642 < 2.2e-16 ***
## Black             3292713   1    13.5934 0.0002707 ***
## MomSmoke          2222570   1     9.1755 0.0026729 **
## Black:MomSmoke       5461   1     0.0225 0.8807474
## Residuals        70488788 291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) I performed a backward selection based on type3 SS results 0.05 criteria on p-value.

- STEP 1: I removed education level from the full model because it was the least significant variable (p-value = 0.86).

- STEP 2: I removed marriage from the model for the same reason (p-value = 0.62).

- STEP 3: I removed Boy (gender) for the same reason (p-value = 0.39).

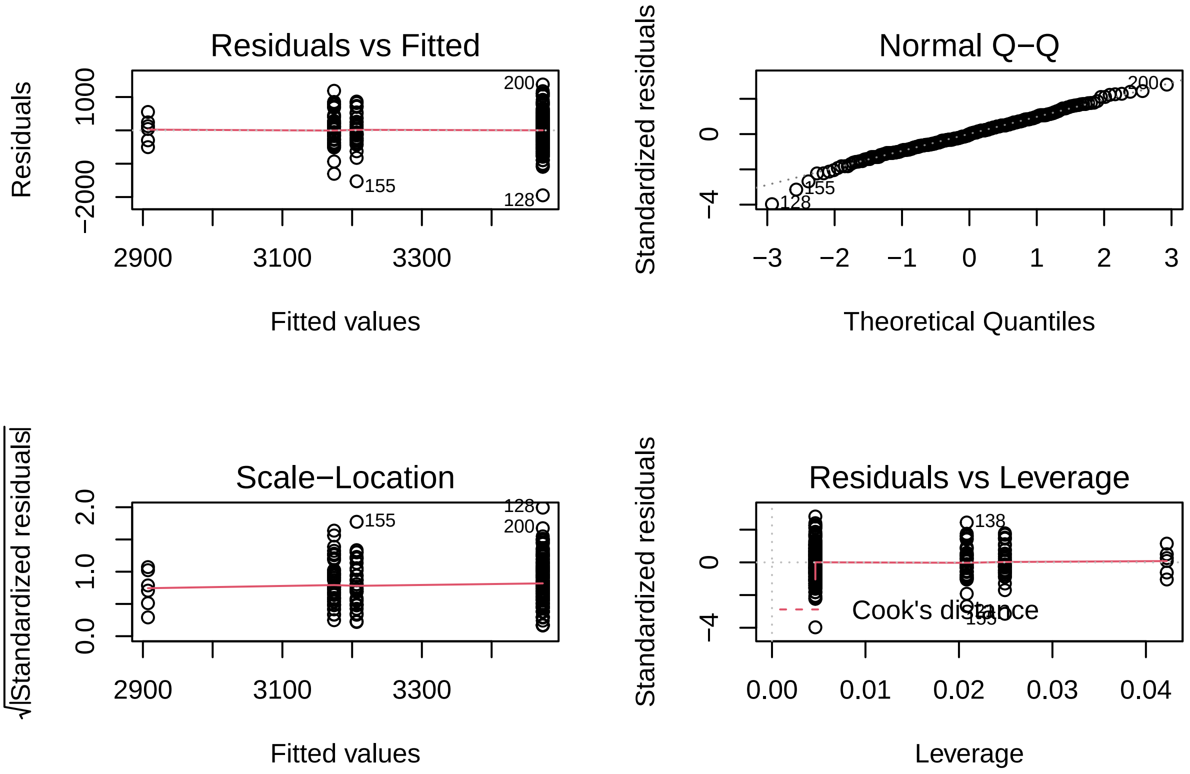
- STEP 4: I kept MomSmoke and Black (race) variables because they are both significant (p-value < 0.05).

- STEP 5: I didn't add the interaction effects in the model because it's not signifcant (p-value = 0.88).

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

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

```
par(mfrow= c(2,2))
plot(lm(Weight ~ Black + MomSmoke, data = bweight))
```



b) The final model, which includes the variables MomSmoke and Black, explains about 78.96% of the total variation. The normality assumption is validated by the diagnostic plots.

```
TukeyHSD(aov(Weight ~ Black + MomSmoke, data = bweight))
```

```
##   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
```

```r
ScheffeTest(aov(Weight ~ Black + MomSmoke, data = bweight))
```

```
##
##   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
```

c) The post-hoc tests gave the following results:

- White infants have an higher mean weight compared to the black ones. The difference is about 293.94g.

- Infants from non-smoking moms have an higher mean weight compared to the ones whose mum smokes. The difference is about 266.76g