Stats 110 Homework 3

Owen Lin

1. a. Yes, Model 1 is nested under Model 2
   b. $\beta_1 = 0$ implies X1 does not affects the response Y given X2 is in the model
   c. If $\beta_3 \neq 0$, it implies that the impact of X1 on Y depends on the value of X2
   d. H0: $\beta_2 = \beta_3 = 0$
      Ha: At least one of the β_2 or β_3 is not 0
      β_2 is direct effect and β_3 is the indirect effect.
   e. H0: $\beta_1 = \beta_2 = 0$
      Ha: H0 is not true
   f. When X2 = 0, model 2 becomes $y_i = \beta_0 + \beta_1 * x_{i1}$, so one unit change in X1 results in β_1 unit change in Y
      When X2 = 1, model 2 becomes $y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) * x_{i1}$, one unit change in X1 results in $\beta_1 + \beta_3$ unit change in Y
2. a. Need more information, depend on how effective is the new regressor given the existing p regressors.
   b. Model 1 and 2 have the same SSTO, because the dataset didn't change based on the model.
   c. Model 1 likely has a lower sum of square error(SSE) than model 2. The additionally regressor only helps the prediction or it doesn't help at all at worse.
   d. β_j Very likely to be different. The new regressor will likely affecting the previous regressors. When the new regressor is categorical, the slope coefficience may not change
3. a. $rest_i = \beta_0 + \beta_1 * Hgt_i + \beta_2 * Wgt_i + \beta_3 * Smoke_i + \beta_4 * Hgt_i * Wgt_i + \epsilon_i$
   b. For example, a weight increase of 20 pounds might not be as big of a deal to a 190cm tall person, but it will impact someone who's 150cm greatly.
   c. $rest_i = 181.48 - 1.61 * Hgt_i - 0.50 * Wgt_i + 5.75 * Smoke_i + 0.01 * Hgt_i * Wgt_i$, The adjusted r square is 0.08

```r
pulse <- read.table("D:\\Coding\\R\\Stats 110\\Data\\Pulse.txt", fill = TRUE, header = TRUE) #nolint
model_pulse <- lm(Rest ~ Hgt + Wgt + Smoke + Hgt*Wgt, data = pulse)
summary(model_pulse)
```

```
##
## Call:
## lm(formula = Rest ~ Hgt + Wgt + Smoke + Hgt * Wgt, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.405  -6.300  -0.815   5.667  34.342
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 181.484803  55.771278   3.254  0.00131 **
## Hgt          -1.611175   0.811581  -1.985  0.04832 *
## Wgt          -0.496353   0.371181  -1.337  0.18249
## Smoke         5.751786   2.011254   2.860  0.00463 **
## Hgt:Wgt       0.006861   0.005251   1.307  0.19264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.528 on 227 degrees of freedom
## Multiple R-squared:  0.09871,    Adjusted R-squared:  0.08283
## F-statistic: 6.215 on 4 and 227 DF,  p-value: 9.194e-05
```

d. Estimated SSE: 9.528^2*(227) = 20607.69
e. H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
   Ha: At least one of the β is not 0
   Test-statistic: 6.215 ~ F(4, 227)
   p-value: 9.149*10^-5
   Conclusion: reject the null at 95% significant level and conclude that at least one variable is a significant indicator of y
f. H0: $\beta_4 = 0$
   Ha: $\beta_4 \neq 0$
   Test-statistic: 1.307 ~ t(227)
   p-value: 0.19264
   Conclusion: fail to reject the null at 95% significant level, inconclusive.
g. H0: $\beta_2 = \beta_4 = 0$
   Ha: At least one of the β is not 0
h. Conclusion: fail to reject the null at 95% significant level with a p-value of 0.3903. In other word, we don't have evident that the model with weight as a regressor is better than the reduced model.

```
model_full <- lm(Rest ~ Hgt + Wgt + Smoke + Hgt*Wgt, data = pulse)
model_reduced <- lm(Rest ~ Hgt + Smoke, data = pulse)
anova(model_reduced, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: Rest ~ Hgt + Smoke
## Model 2: Rest ~ Hgt + Wgt + Smoke + Hgt * Wgt
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    229 20781
## 2    227 20610  2    171.57 0.9449 0.3903
```

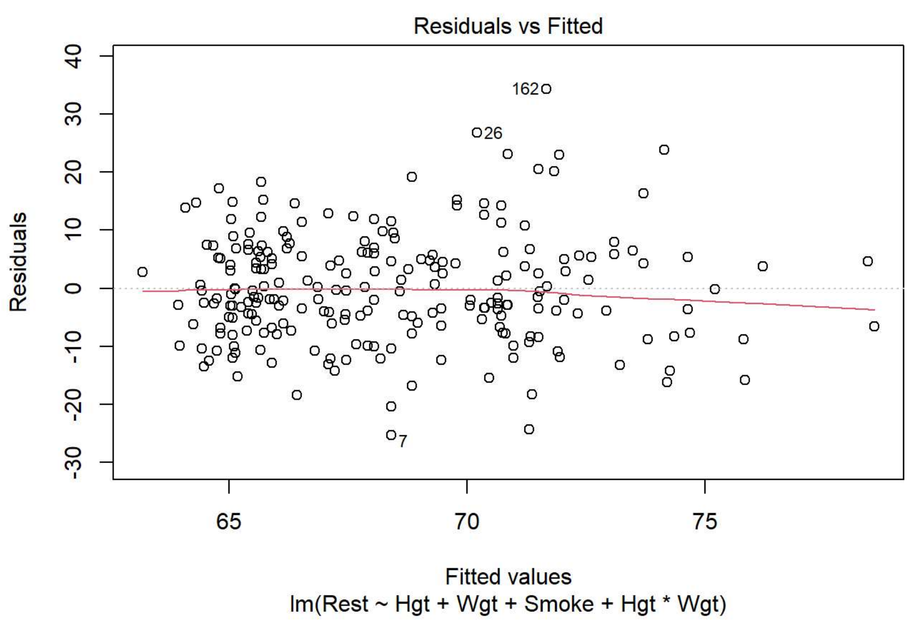
i. Adding weight when height is already in the model didn’t add any explanatory strength to the model, SSR only increases by less than 0.1

```
anova(model_full)
```

```
## Analysis of Variance Table
##
## Response: Rest
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Hgt         1  1346.2 1346.18 14.8273 0.0001533 ***
## Wgt         1     0.0    0.03  0.0003 0.9857152
## Smoke       1   756.0  755.99  8.3267 0.0042833 **
## Hgt:Wgt     1   155.0  155.02  1.7075 0.1926369
## Residuals 227 20609.5   90.79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

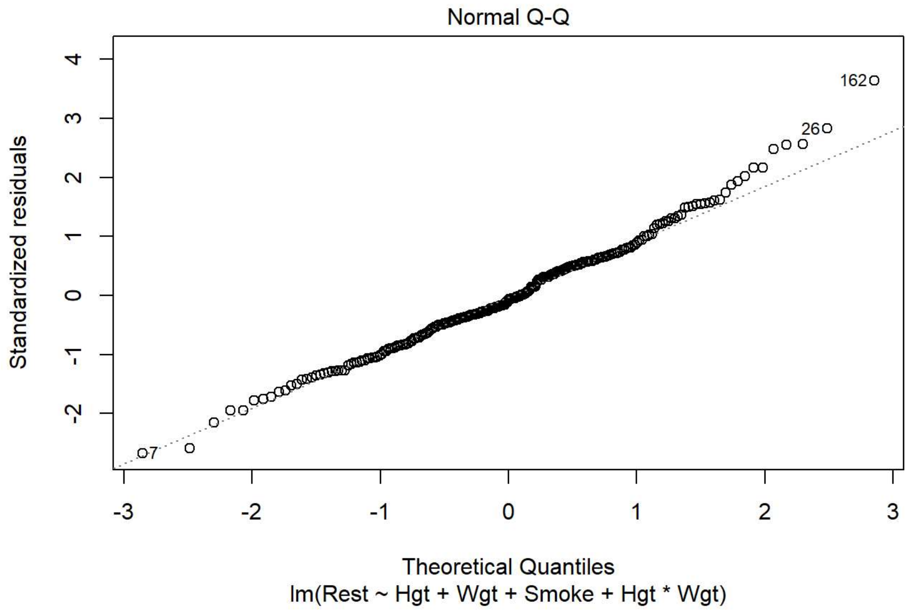
j. SSTO = 1346.2 + 0 + 756 + 155 + 20609.5 = 22866.7
k. SSE: 22866.7 - 1346.2 = 21520.5
SSR: 1346.2
SSTO: 22866.7
l. People who have a low weight might get that from exercise frequently, which could be the actual cause that cause the decrease in rest heart rate.
m. It seems that fitted values around 72 varies more than other values. So a single constant σ^2 for the entire model is invalid.

```
plot(model_pulse, 1)
```



n. The line is rather flat, so the linearity assumption is not violated.
o. Most of the points still fit the line, other than one of the tails. The normality assumption does not have a big problem.

```
plot(model_pulse, 2)
```



p. No, it doesn't make sense to predict the resting heart rate for someone who weight 350 pound, because it is outside of our sample range and we have no way to gurantee the extrapolation follows the same trend.

```
summary(pulse)
```

```
##      Active          Rest            Smoke            Gender      
##  Min.   : 51.0   Min.   : 43.00   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.: 79.0   1st Qu.: 62.00   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median : 88.5   Median : 68.00   Median :0.0000   Median :0.0000  
##  Mean   : 91.3   Mean   : 68.35   Mean   :0.1121   Mean   :0.4741  
##  3rd Qu.:102.0   3rd Qu.: 74.00   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :154.0   Max.   :106.00   Max.   :1.0000   Max.   :1.0000  
##     Exercise          Hgt             Wgt       
##  Min.   :1.000   Min.   :60.00   Min.   :102.0  
##  1st Qu.:2.000   1st Qu.:65.00   1st Qu.:135.0  
##  Median :2.000   Median :68.00   Median :150.0  
##  Mean   :2.254   Mean   :68.25   Mean   :157.9  
##  3rd Qu.:3.000   3rd Qu.:71.00   3rd Qu.:175.0  
##  Max.   :3.000   Max.   :78.00   Max.   :260.0
```

4. a. $Arsenic_i = \beta_0 + \beta_1 * Year_i + \beta_2 * Miles_i + \beta_3 * Year_i * Miles_i + \epsilon_i$
   b. $Lead_i = \beta_0 + \beta_1 * Year_i + \beta_2 * Iclean + \epsilon_i$
   Iclean = 0: $Lead = \beta_0 + \beta_1 * Year_i$
   Iclean = 1: $Lead = \beta_0 + \beta_2 + \beta_1 * Year_i$
   c. $Titanium_i = \beta_0 + \beta_1^2 * Miles_i$
   d.
   $Sulfide_i = \beta_0 + \beta_1 * Year_i + \beta_2 * Miles_i + \beta_3 * Depth_i + \beta_4 * Year_i * Miles_i + \beta_5 * Year_i * Depth_i + \beta_6 * Miles_i * Dept$
5. a. As Mileage increase, the average price decreases.

```
car <- read.table("D:\\Coding\\R\\Stats 110\\Data\\ThreeCars.txt", fill = TRUE, header = TRUE) #nolint
plot(x = car$Mileage, y = car$Price, data = car)
```

```
## Warning in plot.window(...): "data" is not a graphical parameter
```

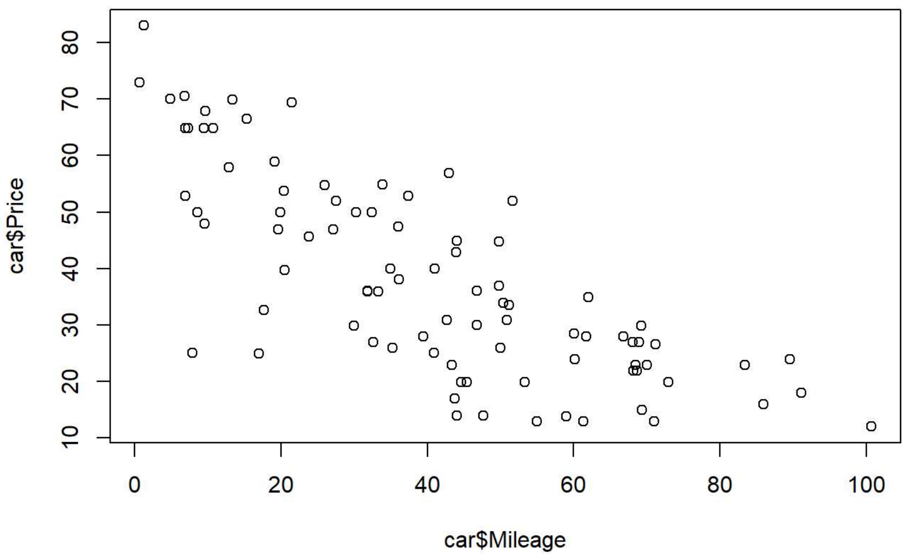
```
## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter
```

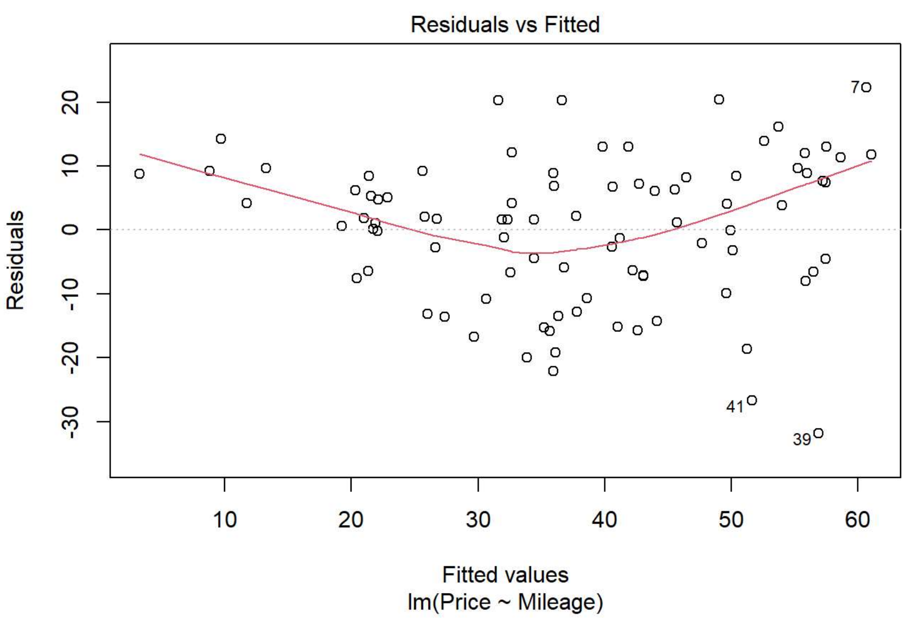
```
## Warning in box(...): "data" is not a graphical parameter
```

```
## Warning in title(...): "data" is not a graphical parameter
```



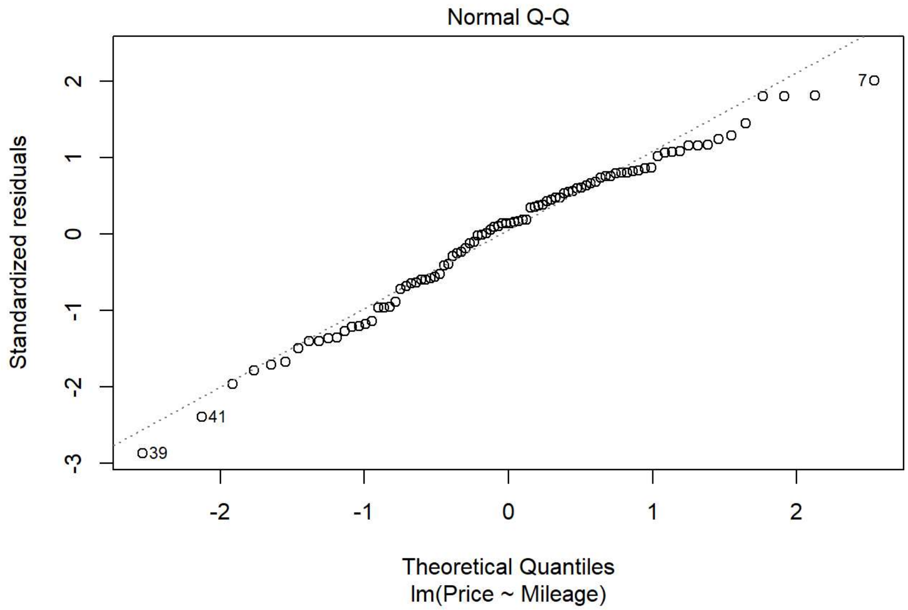
b. See Code below

```
model_car <- lm(Price ~ Mileage, data = car)
plot(model_car, 1)
```



c. It seems that the relation is not linear but rather quadratic
d. The variance is not constant. Generally speaking, cars with large Mileage also has a large variance
e. Since the observations are all located closely around the line, the error is rather normal and is consistent with the assumption

```
plot(model_car, 2)
```


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Mileage)
-3
-2
-1
0
1
2
39
41
7