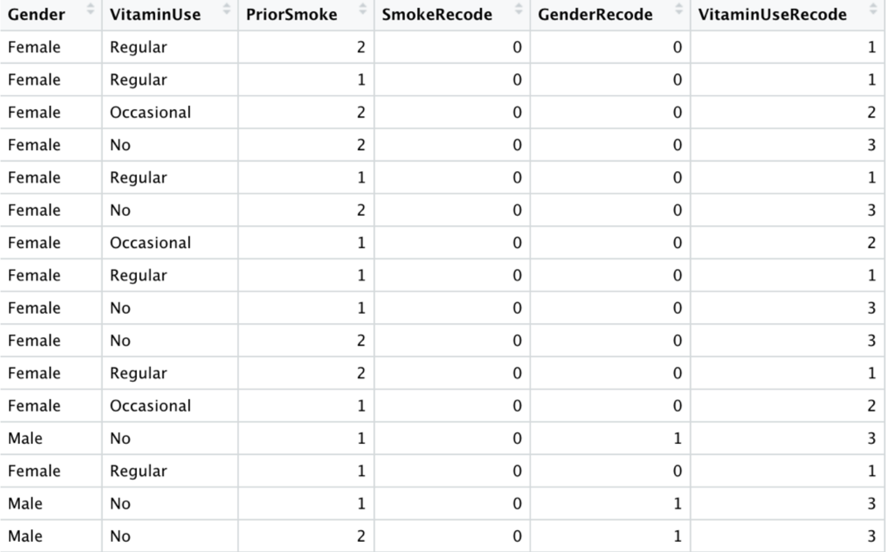
Michael Venit

MSDS 410

Computational Assignment #3

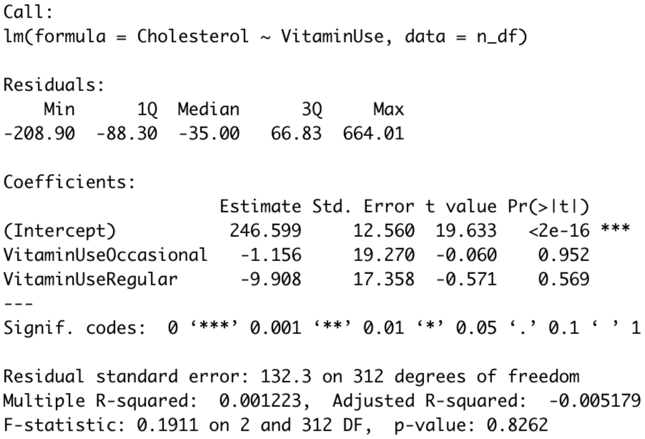
1. All of the categorical variables in the dataset were recoded from text-based categories into numerical values that indicate group. The VitaminUseRecode variable was coded so that code it so that: 1=regular, 2=occasional, 3=never. Meanwhile, GenderRecode was recoded in binary fashion so that 1=male and 0=female. Similarly, the SmokerRecode feature was recoded so that 1=does smoke and 0=does not smoke.



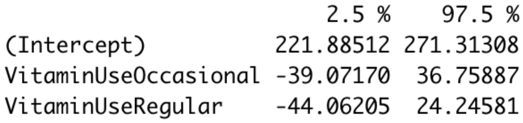
Gender	VitaminUse	PriorSmoke	SmokeRecode	GenderRecode	VitaminUseRecode
Female	Regular	2	0	0	1
Female	Regular	1	0	0	1
Female	Occasional	2	0	0	2
Female	No	2	0	0	3
Female	Regular	1	0	0	1
Female	No	2	0	0	3
Female	Occasional	1	0	0	2
Female	Regular	1	0	0	1
Female	No	1	0	0	3
Female	No	2	0	0	3
Female	Regular	2	0	0	1
Female	Occasional	1	0	0	2
Male	No	1	0	1	3
Female	Regular	1	0	0	1
Male	No	1	0	1	3
Male	No	2	0	1	3

2. We can see from the Model 1 summary that the intercept term is the only coefficient with which we can reject the null hypothesis at alpha= 0.05. In this particular model, the intercept represents those who "never" use vitamins. This means that the baseline cholesterol for someone with no vitamin use is 246.599. The regression coefficients for the other two vitamin use categories can be interpreted as subtracting cholesterol from the baseline value. If a person is a user of vitamins, then cholesterol is expected to decrease. However, neither regression coefficient for the two usage categories is able to reject the null hypothesis. This can be seen

further by examining their confidence intervals, which can be seen below (both include 0). Additionally, the R-squared value of the model is nearly 0 which indicates that our dependent variable is accounting for virtually none of the variance in Cholesterol. There is also an observed F-statistic of 0.1911 which is less than the calculated critical F-value of 3.0247, hence we fail to reject the null hypothesis.

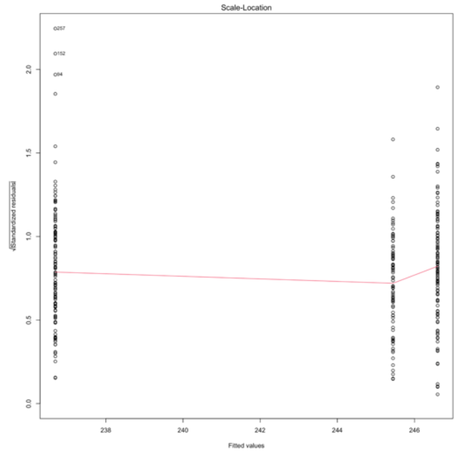


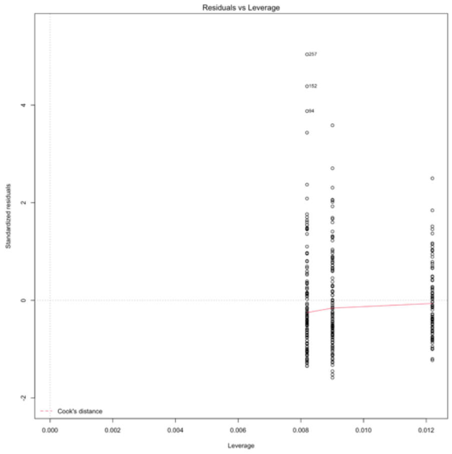
```
Call:
lm(formula = Cholesterol ~ VitaminUse, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-208.90  -88.30  -35.00   66.83  664.01 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           246.599     12.560  19.633   <2e-16 ***
VitaminUseOccasional   -1.156     19.270  -0.060    0.952    
VitaminUseRegular      -9.908     17.358  -0.571    0.569    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,	Adjusted R-squared:  -0.005179 
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

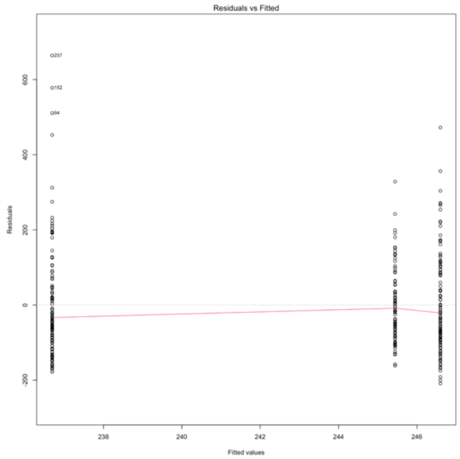


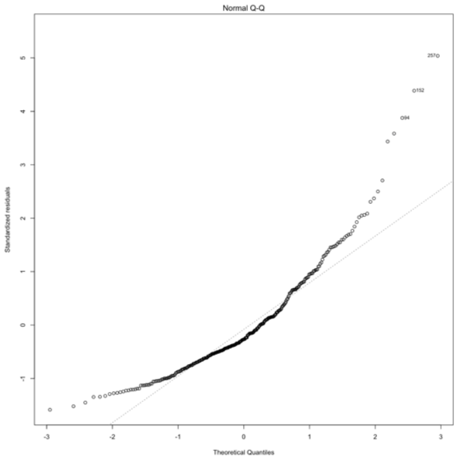
```
                         2.5 %    97.5 %
(Intercept)          221.88512 271.31308
VitaminUseOccasional -39.07170  36.75887
VitaminUseRegular    -44.06205  24.24581
```

$$\hat{Y} = 246.599 - 1.156X_1 - 9.908X_2$$



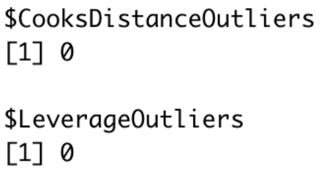






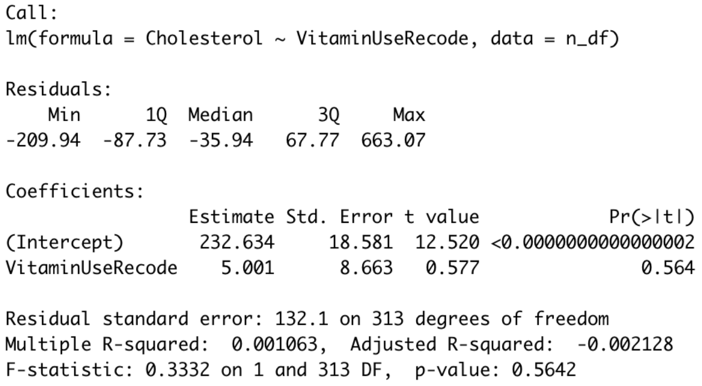
Below the results from Cooks Distance and leverage calculations can be seen. For reference, will investigate values with Cooks Distance greater than 1 and leverage values greater than:

$$\frac{2 * (k+1)}{n}$$



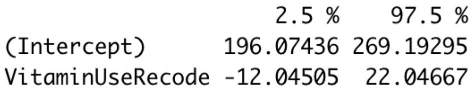
```
$CooksDistanceOutliers
[1] 0

$LeverageOutliers
[1] 0
```

It appears that there are several potential outlier values based on their extremely large variance in the Residuals vs Fitted graph (top left). The residuals also follow a strict pattern since we only have 3 values for the categorical variable. However, results from Cooks Distance and leverage calculations reveal that there are no values that surpass either threshold.



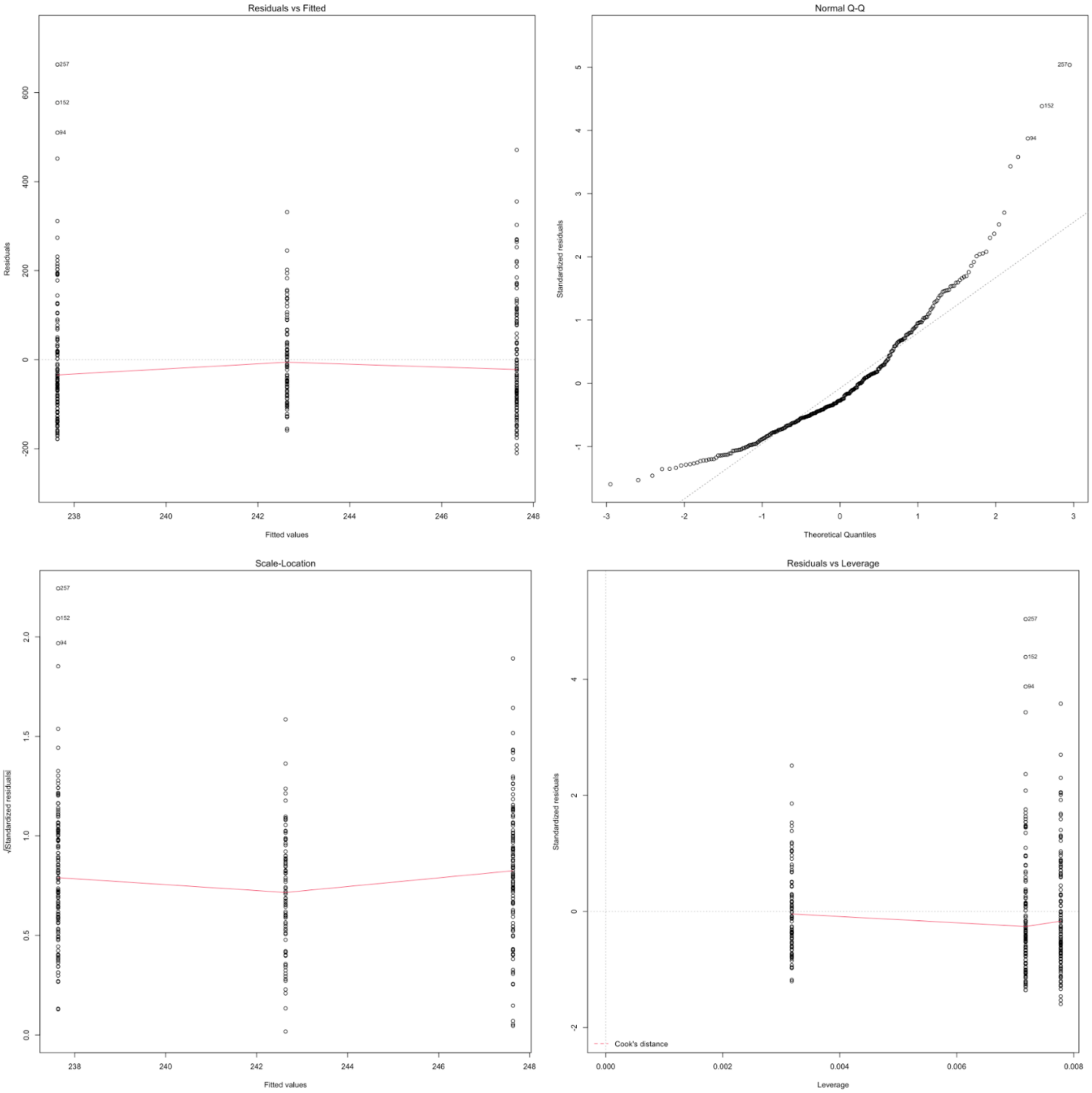
```
Call:
lm(formula = Cholesterol ~ VitaminUseRecode, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-209.94  -87.73  -35.94   67.77  663.07 

Coefficients:
                 Estimate Std. Error t value            Pr(>|t|)    
(Intercept)       232.634     18.581  12.520 <0.0000000000000002
VitaminUseRecode    5.001      8.663   0.577               0.564    

Residual standard error: 132.1 on 313 degrees of freedom
Multiple R-squared:  0.001063,	Adjusted R-squared:  -0.002128 
F-statistic: 0.3332 on 1 and 313 DF,  p-value: 0.5642
```

$$\hat{Y} = 232.634 + 5.001$$



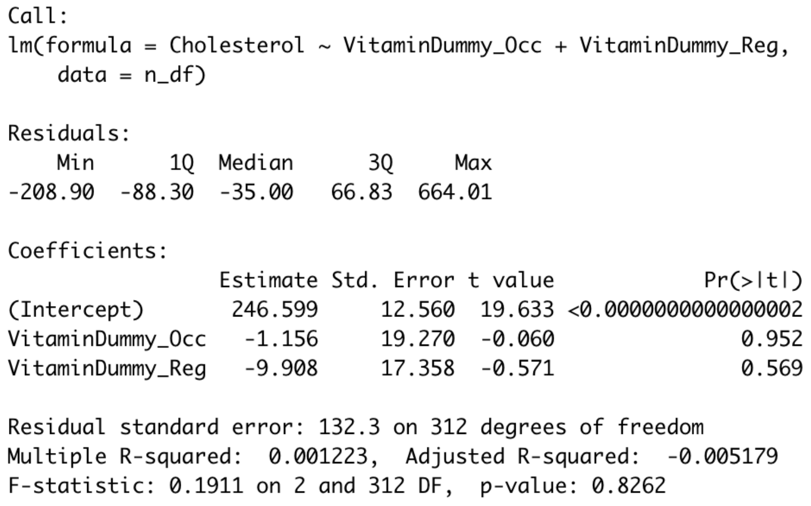
```
                       2.5 %    97.5 %
(Intercept)      196.07436 269.19295
VitaminUseRecode -12.04505  22.04667
```

As it pertains to Model 2, the Vitamin variable is now being treated as a numeric variable in the model. The coefficient can be interpreted as a 1 unit increase in vitamin use will increase Cholesterol by 5.001 units, though it is not statistically significant as seen by the p-value (0.564). The confidence interval of the coefficient can be seen above which includes zero. The R-squared value is once again nearly zero (0.001063) which tells us that the VitaminUseRecode variable accounts for virtually no variance in Cholesterol. As it pertains to our F-statistic of 0.332, the critical F-value was calculated to be 3.8713, indicating that we fail to reject the null hypothesis.



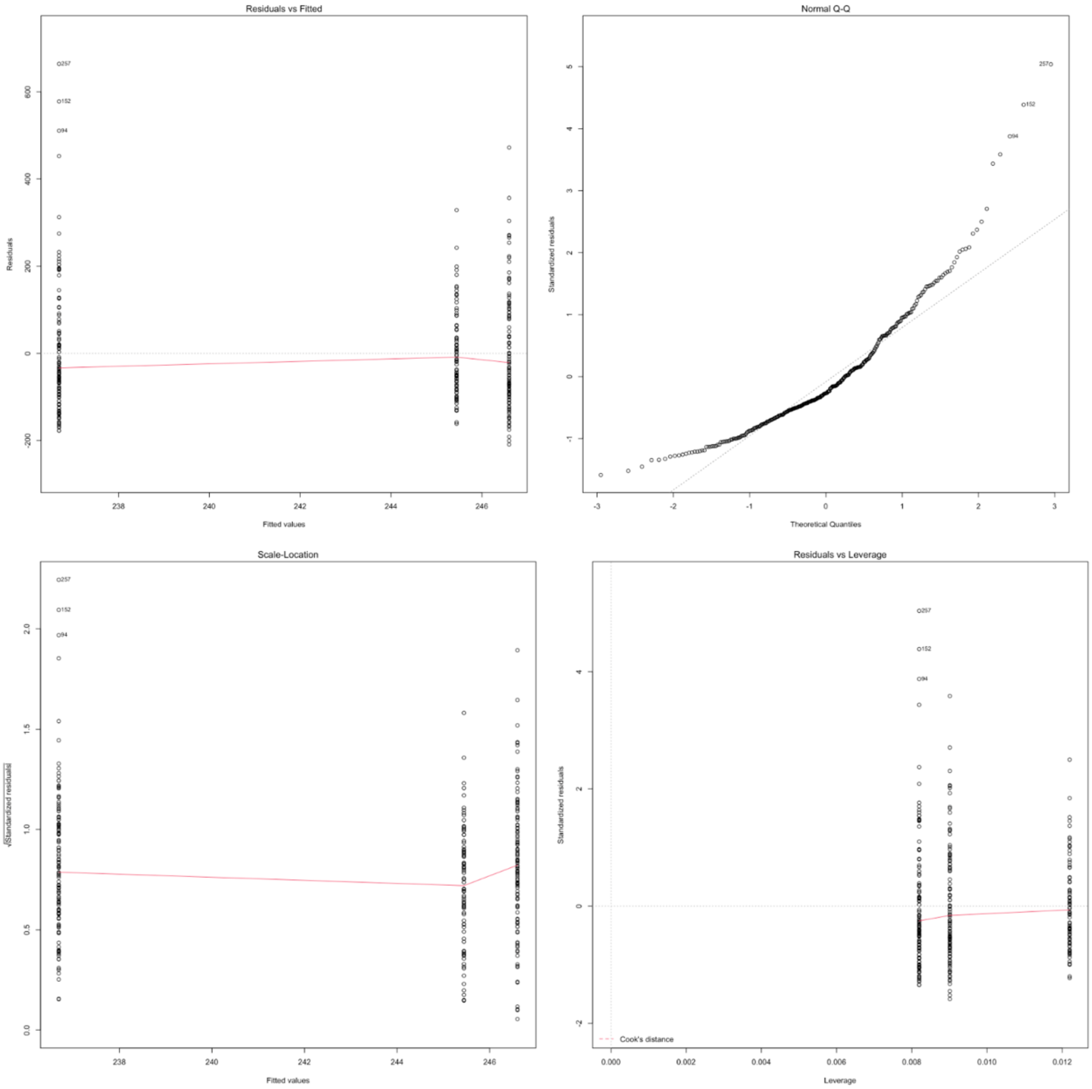
We can also see that once again the residual graph shows 3 extreme values indicating that we potential outliers in our dataset. This was also observed in the histogram of cholesterol as it showed a significant right skew. The regression coefficients have flipped in the second model and there is now only one. Despite the recoding, the variable VitaminUse doesn't appear to explain any of the variance in Cholesterol. The same calculations for Cooks Distance and leverage were conducted on Model 2 for the purpose of outlier detection. However, similar to Model 1, there appear to be no values meeting either threshold for these diagnostic measures.

3. We can see from the model summary below that the exact same model as Model 1 was obtained. This is not surprising as R coerces character columns to factors which, essentially dummy codes them under the hood. By leaving out the dummy coded variable, VitamineUse == "No", we can interpret the coefficients as the following. The intercept (Beta_0) can be seen as the average cholesterol of someone who does not take vitamins. The first regression coefficient, VitaminDummy_Occ (Beta_1) is the difference in cholesterol between those that don't use vitamins and those that use them occasionally. Beta_1 + Beta_0 equals the average cholesterol for someone with occasional vitamin use. Meanwhile, VitaminDummy_Reg (Beta_2) represents the difference in cholesterol between those that use vitamins regularly and those that don't use them. Beta_0+ Beta_2 is equal to the average cholesterol for someone who uses vitamins regularly.

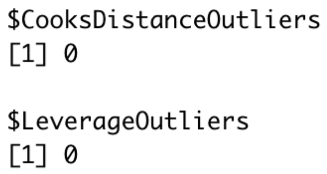


```
Call:
lm(formula = Cholesterol ~ VitaminDummy_Occ + VitaminDummy_Reg,
    data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-208.90  -88.30  -35.00   66.83  664.01

Coefficients:
                 Estimate Std. Error t value            Pr(>|t|)
(Intercept)       246.599     12.560  19.633 <0.0000000000000002
VitaminDummy_Occ   -1.156     19.270  -0.060               0.952
VitaminDummy_Reg   -9.908     17.358  -0.571               0.569

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,  Adjusted R-squared:  -0.005179
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

As can be seen in Model 1, the coefficients don't hold any statistical significance, and we are unable to reject the null hypothesis that they are 0 due to their p-values, while the R-squared value is nearly zero. The model seems to indicate that these dummy coded variables do not account for any variance in cholesterol. The F-statistic as can be seen in the model summary was calculated to be 0.1911, while the critical F-value was found to be 3.0247, through an Omnibus F-test. The critical F-value is larger than that of the F-statistic, confirming that we fail to reject the null hypothesis.

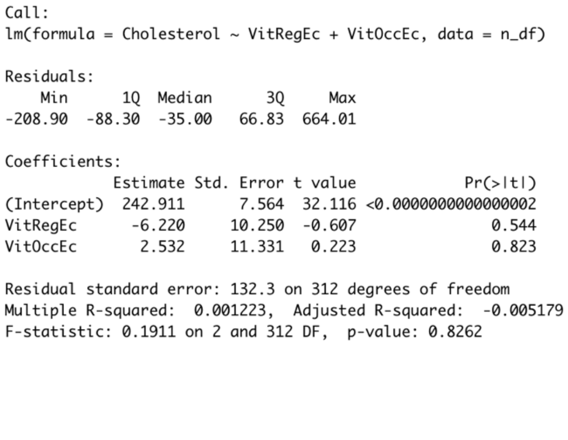


Upon observing the graphical output, there appears to be no change between Model 2 and Model 1, as all of the coefficients are exactly the same. There appears to be several potential outlier values based on their extremely large variance in the Residuals vs Fitted graph (top left). The residuals also follow a strict pattern since we only have 3 values for the categorical variable. However, results from Cooks Distance and leverage calculations reveal that there are no values that surpass either threshold, similarly to Model 1.

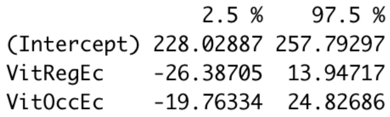


```
$CooksDistanceOutliers
[1] 0

$LeverageOutliers
[1] 0
```

4. For Model 4, the observed F-statistic and R-squared values are exactly the same as Models 1 and 3. The coefficients have changed slightly as, though we are still only accounting for an extremely small amount of variation in cholesterol as can be seen in the model summary below. The sign for the regular usage coefficient has become negative indicating that the average cholesterol for those that consume vitamins regularly is less than those that don't. Additionally, the occasional vitamin user has a higher average cholesterol than a person who does not consume vitamins. Despite these interpretations, the coefficients are not statistically significant and we fail to reject the null hypothesis that they are equal to zero. Their confidence intervals can be seen below (they include zero).

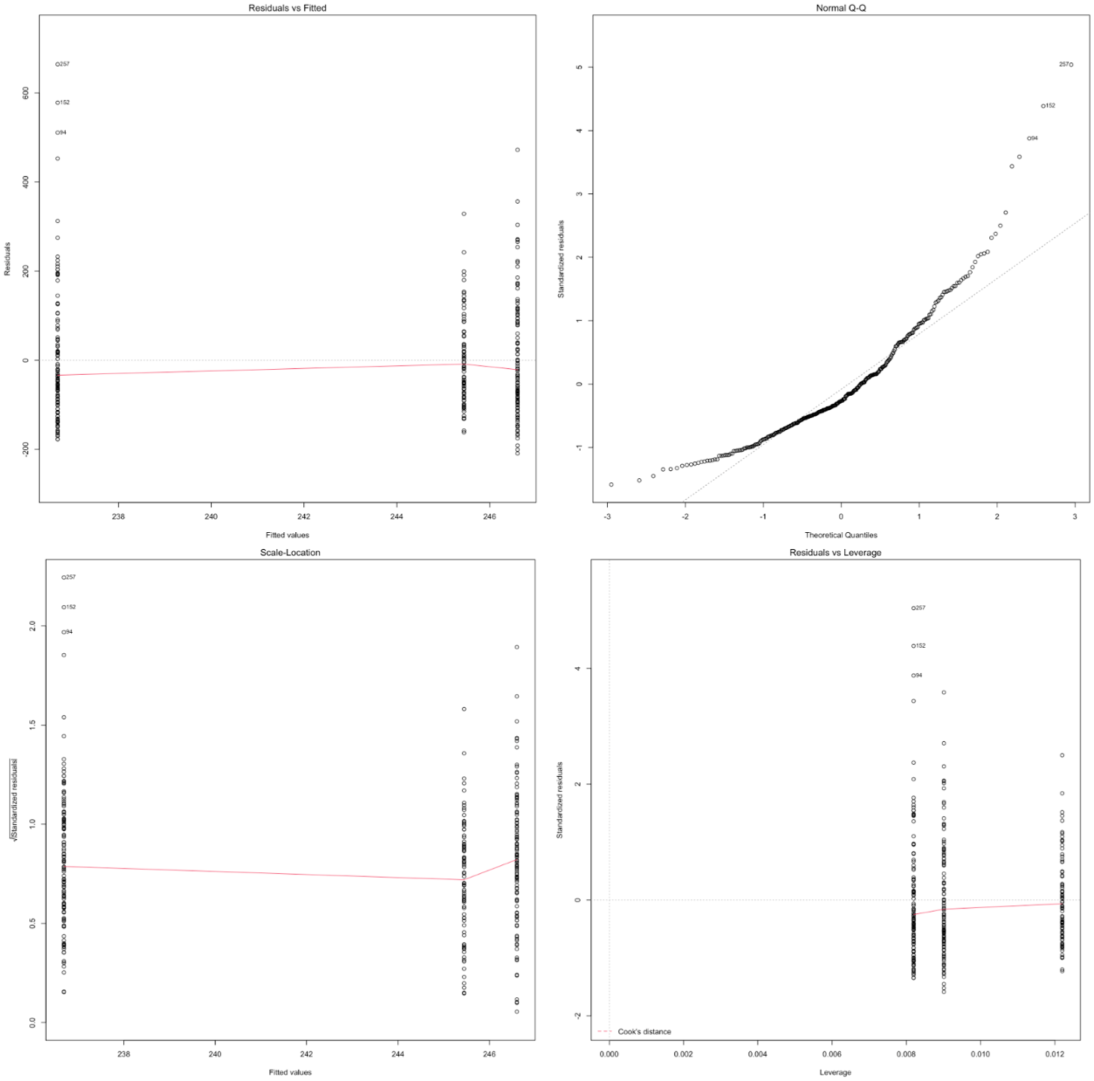


```
Call:
lm(formula = Cholesterol ~ VitRegEc + VitOccEc, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-208.90  -88.30  -35.00   66.83  664.01 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  242.911      7.564  32.116 <0.0000000000000002
VitRegEc      -6.220     10.250  -0.607               0.544
VitOccEc       2.532     11.331   0.223               0.823

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,	Adjusted R-squared:  -0.005179 
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```



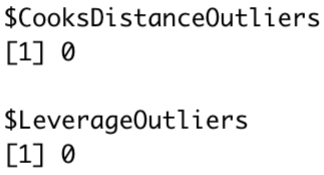
```
                 2.5 %    97.5 %
(Intercept) 228.02887 257.79297
VitRegEc    -26.38705  13.94717
VitOccEc    -19.76334  24.82686
```

The F-statistic as can be seen in the model summary was calculated to be 0.1911, while the critical F-value was found to be 3.0247, through an Omnibus F-test. The critical F-value is larger than that of the F-statistic, confirming that we fail to reject the null hypothesis.



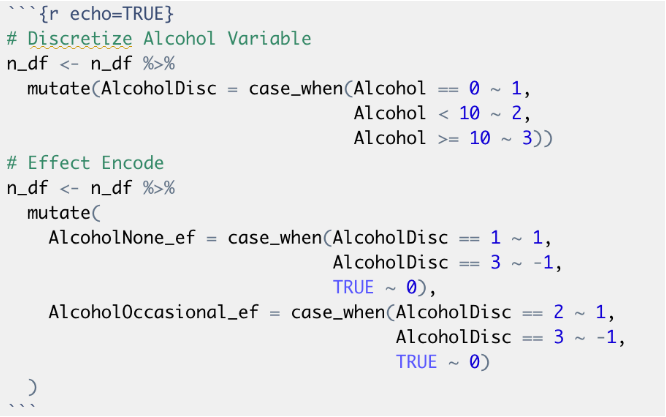
When observing the graphical output, the only noticeable change detected is the interpretation of the coefficients. However, the coefficient interpretation can/will be different if a different value is held out as the baseline. The dummy-coded variables (or one-hot-encoded) are preferred as I normally describe them as it is easier to remember a 1 or 0 combination for each value of categorical variable.

There appears to be several potential outlier values based on their extremely large variance in the Residuals vs Fitted graph (top left). The residuals also follow a strict pattern since we only have 3 values for the categorical variable. However, results from Cooks Distance and leverage calculations reveal that there are no values that surpass either threshold, similarly to Models 1 and 3.

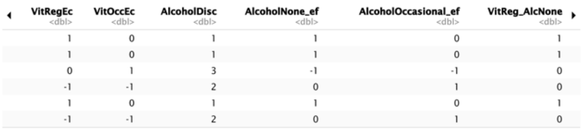


```
$CooksDistanceOutliers
[1] 0

$LeverageOutliers
[1] 0
```

5. The following code was used to discretize the Alcohol variable to form a new categorical variable with 3 levels. The code and output can be seen below

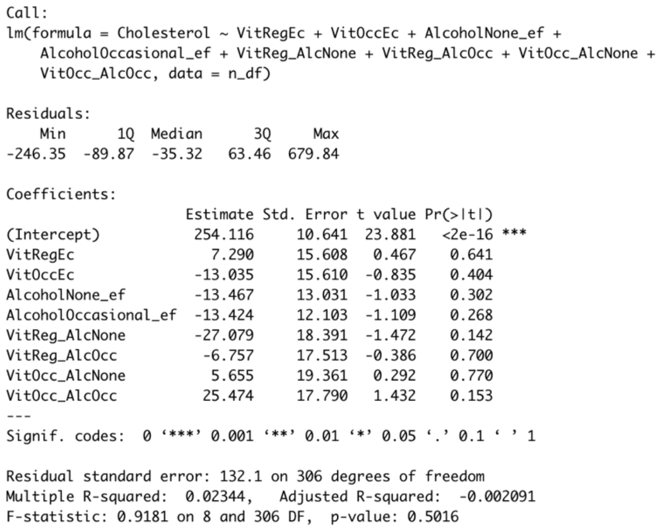


~~~
```{r echo=TRUE}
# Discretize Alcohol Variable
n_df <- n_df %>%
  mutate(AlcoholDisc = case_when(Alcohol == 0 ~ 1,
                                 Alcohol < 10 ~ 2,
                                 Alcohol >= 10 ~ 3))
# Effect Encode
n_df <- n_df %>%
  mutate(
    AlcoholNone_ef = case_when(AlcoholDisc == 1 ~ 1,
                               AlcoholDisc == 3 ~ -1,
                               TRUE ~ 0),
    AlcoholOccasional_ef = case_when(AlcoholDisc == 2 ~ 1,
                                     AlcoholDisc == 3 ~ -1,
                                     TRUE ~ 0)
  )
```
~~~

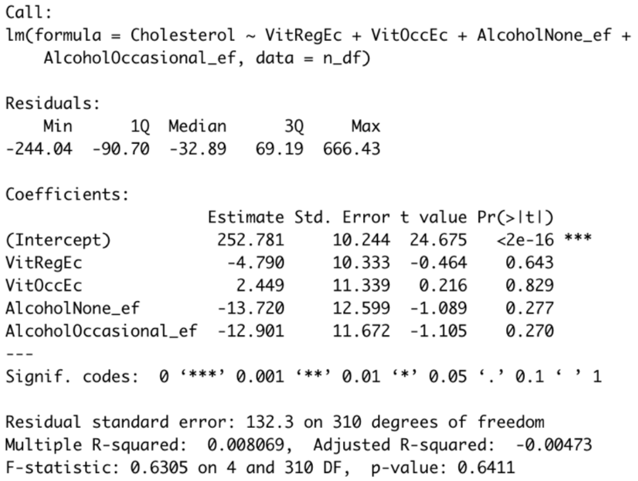


VitRegEc <dbl>	VitOccEc <dbl>	AlcoholDisc <dbl>	AlcoholNone_ef <dbl>	AlcoholOccasional_ef <dbl>	VitReg_AlcNone <dbl>
1	0	1	1	0	1
1	0	1	1	0	1
0	1	3	-1	-1	0
-1	-1	2	0	1	0
1	0	1	1	0	1
-1	-1	2	0	1	0

6. Below is the summary of our full model (left) with the interaction terms as well as the reduced model (right) without interaction terms.



```
Call:
lm(formula = Cholesterol ~ VitRegEc + VitOccEc + AlcoholNone_ef +
    AlcoholOccasional_ef + VitReg_AlcNone + VitReg_AlcOcc + VitOcc_AlcNone +
    VitOcc_AlcOcc, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-246.35  -89.87  -35.32   63.46  679.84

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           254.116     10.641  23.881   <2e-16 ***
VitRegEc                7.290     15.608   0.467    0.641
VitOccEc              -13.035     15.610  -0.835    0.404
AlcoholNone_ef        -13.467     13.031  -1.033    0.302
AlcoholOccasional_ef  -13.424     12.103  -1.109    0.268
VitReg_AlcNone        -27.079     18.391  -1.472    0.142
VitReg_AlcOcc          -6.757     17.513  -0.386    0.700
VitOcc_AlcNone          5.655     19.361   0.292    0.770
VitOcc_AlcOcc          25.474     17.790   1.432    0.153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.1 on 306 degrees of freedom
Multiple R-squared:  0.02344,	Adjusted R-squared:  -0.002091
F-statistic: 0.9181 on 8 and 306 DF,  p-value: 0.5016
```



```
Call:
lm(formula = Cholesterol ~ VitRegEc + VitOccEc + AlcoholNone_ef +
    AlcoholOccasional_ef, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-244.04  -90.70  -32.89   69.19  666.43

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           252.781     10.244  24.675   <2e-16 ***
VitRegEc               -4.790     10.333  -0.464    0.643
VitOccEc                2.449     11.339   0.216    0.829
AlcoholNone_ef        -13.720     12.599  -1.089    0.277
AlcoholOccasional_ef  -12.901     11.672  -1.105    0.270
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.3 on 310 degrees of freedom
Multiple R-squared:  0.008069,	Adjusted R-squared:  -0.00473
F-statistic: 0.6305 on 4 and 310 DF,  p-value: 0.6411
```

A nested model F-test was used on both the full and reduced models. The goal of doing so is to observe whether the addition of the interaction variables (as a set) significantly improved the prediction of Y given that the effect encoded Vitamin and Alcohol variables were already in the model. The equation for the partial F-test can be seen below in addition to the null and alternate hypotheses:

$$Null = H_0 : \beta_1^* = \beta_2^* = \beta_3^* = \beta_4^* = 0 \; in \; the \; full \; model$$

$$Alternate = H_a : \beta_i^* \neq 0 \; for \; at \; least \; 1 \; i \; in \; the \; model$$

$$F(X_1^*, X_2^*, X_3^*, X_4^* \mid X_1, X_2, X_3, X_4) = \frac{(\frac{SS(X_1^*, X_2^*, X_3^*, X_4^* \mid X_1, X_2, X_3, X_4)}{s})}{MS \; Residual \; (X_1^*, X_2^*, X_3^*, X_4^*, X_1, X_2, X_3, X_4)}$$

The equation to calculate the F-statistic can be rewritten in order to easily use values from the model ANOVA tables:

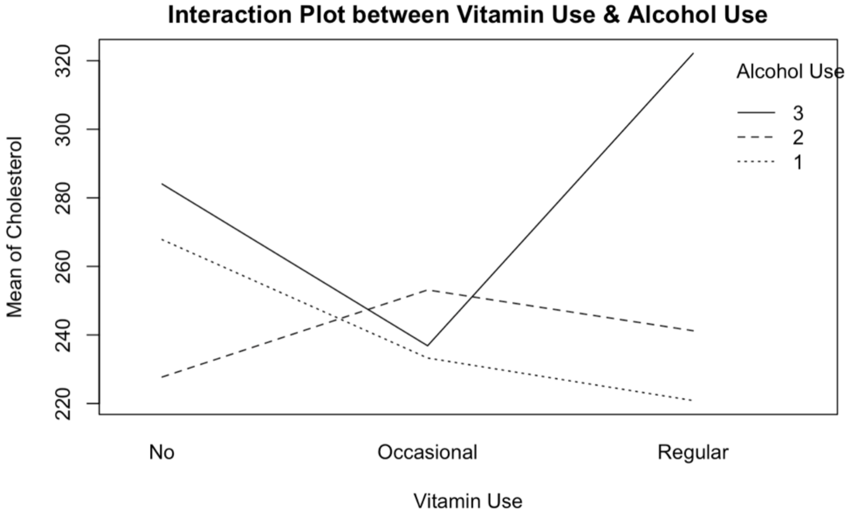
$$F(X_1^*, X_2^*, X_3^*, X_4^* \mid X_1, X_2, X_3, X_4) = \frac{(\frac{(Regression \; SS(full) - Regression \; SS(reduced))}{s})}{MS \; Residual(full)}$$

The values from the model ANOVA tables were inserted into the equation to generate a F-statistic of 1.204.

Meanwhile the critical F-value was computed from the following:

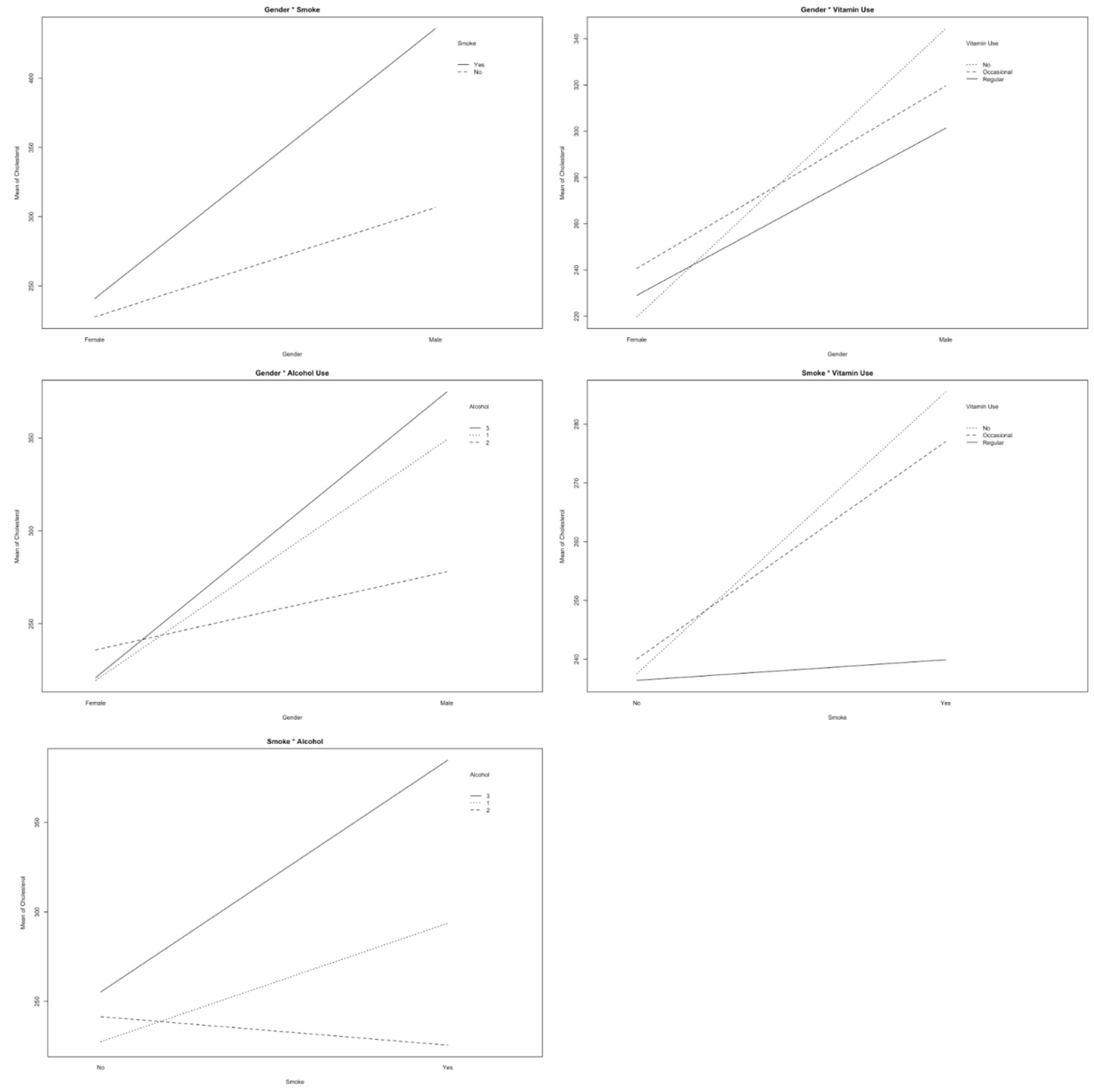
$$F_{s,\, n-q-s-1,\, 1-\alpha} = F_{4,\, 315-4-4-1,\, 1-0.05} = 2.4012$$

The F-statistic of 1.204 is well below our critical value of 2.4012, which indicates we fail to reject the null hypothesis. Hence, the interaction variables do not supply any significant information for predicting cholesterol.



We can see from the interaction plot that there are exchanges occurring between the variables. Despite this interaction effect, the impact on cholesterol does not achieve a level of statistical significance. Additionally, the F-test concluded that the interaction terms from the full model did not contribute any further information in predicting cholesterol than the reduced model.

7.



In observing Gender v. Smoke, the resulting F-statistic was calculated to be 2.0501, which is less than the computed critical F-value of 2.0248. These values indicate that we fail to reject the null hypothesis as it pertains to the relationship between these two features. Meanwhile, the observation of Gender v. VitaminUse, the resulting F-statistic was calculated to be 0.4545, which is less than the computed critical F-value of 2.6338. These values indicate that we fail to reject the null hypothesis as it pertains to the relationship between these two features. When discerning the interaction between Gender v. Alcohol, the generated F-statistic was calculated to be 1.6049, which is less than the computed critical F-value of 2.6338. These values indicate that we fail to reject the null hypothesis as it pertains to the relationship between these two features. When looking at the interaction between Smoke v. VitaminUse, the F-statistic was calculated to be 0.2274, which is less than the computed critical F-value of 2.6338. These values indicate that we fail to reject the null hypothesis as it pertains to the relationship between these two features. Lastly, in observing Smoke v. Alchohol, the F-statistic was calculated to be 1.8874, which is less than the computed critical F-value of 2.6338. These values indicate that we fail to reject the null hypothesis as it pertains to the relationship between these two features. If we were to observe the graphs above, there appears to be interactions among all combinations of variables. However, the partial F-tests show us that there is no significant value added in predicting cholesterol by any of these interaction terms.

8. Overall the concepts that were touched on in this assignment make me really feel like I am learning what supervised learning is all about. There are so many tools and methods available for building and rigorously testing generated models. I really like the methodical approach of checking assumptions, testing the model (coefficients and overall), investigating interactions, testing their impact, etc. I decided to write a custom function that would intake a full model object and a reduced model object and return the partial F-statistic as well as whether it was above or below the critical F-value. I wanted to write my own custom functions that wrapped around baseR built-in functions to calculate the different model diagnostics as a form of practice. This assignment was a great learning experience for me.