Michael Venit

MSDS 410

Computational Assignment #4

Objective

To use multiple regression to predict CHOLESTEROL using models with continuous and categorical variables. For these analyses, the response variable is Y=CHOLESTEROL, and the remaining variables will be considered explanatory (X's).

Formula/Hyptothesis Overview

Below are the formula definitions that will be used for various tests. The point of referencing these now is to avoid redundantly mentioning them throughout this analysis. The formula for calculating the F-statistic for overall model is:

$$F = \frac{Mean\ Squared\ Regression}{Mean\ Squared\ Residual} = \frac{(\frac{(SSY-SSE)}{k})}{(\frac{SSE}{(n-k-1)})}$$

The formula for calculating the critical F-statistic is:

$$F_{k,\, n-k-1,\, 1-\alpha}$$

The null and alternate hypotheses for the omnibus F-test is as follows:

$$Null = H_0 : \beta_i = 0\ for\ all\ i\ in\ the\ full\ model$$
$$Alternate = H_a : \beta_i \neq 0\ for\ at\ least\ 1\ i\ in\ the\ model$$

For reference, values will be investigated with Cook's Distance greater than 1 and leverage values greater than:

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

The formula for calculating a partial F-test can be seen below. Variables denoted as X^* represent the additional interaction variables added to the model. The value s represents the number of added independent variables:

$$F(X_1^*, X_2^*, \ldots, X_i^* \mid X_1, X_2, \ldots, X_j) = \frac{\left(\frac{SS(X_1^*, X_2^*, \ldots, X_i^* \mid X_1, X_2, \ldots, X_j)}{s}\right)}{MS\ Residual\ (X_1^*, X_2^*, \ldots, X_i^*,\ X_1, X_2, \ldots, X_j)}$$

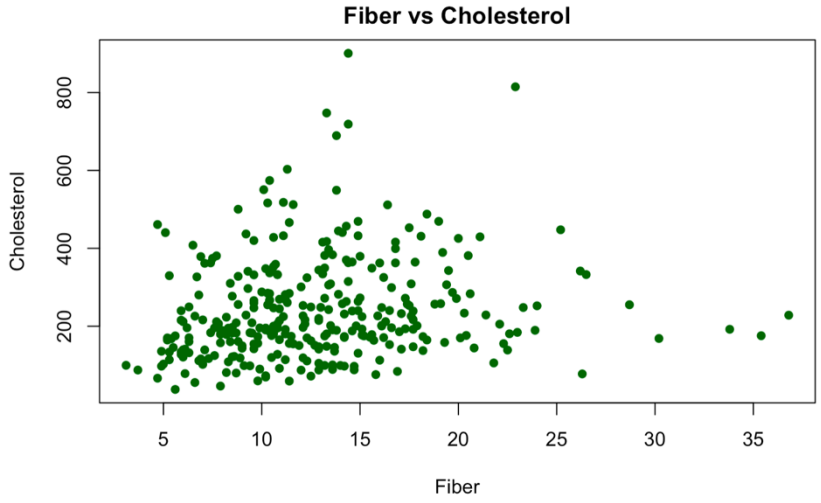
The null and alternate hypotheses are stated below:

$$Null = H_0 : \beta_i = 0\ for\ all\ i\ in\ the\ full\ model$$
$$Alternate = H_a : \beta_i \neq 0\ for\ at\ least\ 1\ i\ in\ the\ model$$

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

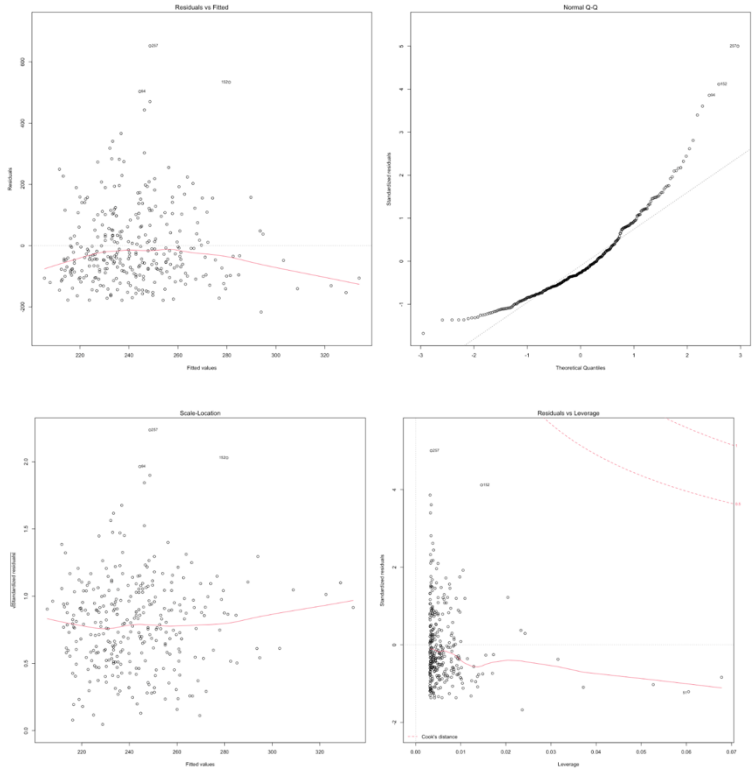
$$F(X_1^*, X_2^*, \ldots, X_i^* \mid X_1, X_2, \ldots, X_j) = \frac{\left(\frac{(Regression\ SS(full) - Regression\ SS(reduced))}{s}\right)}{MS\ Residual(full)}$$

1. We can see from the scatterplot below that there is possibly a very slight positive linear correlation between Fiber and Cholesterol. The Pearson correlation is 0.154, which confirms the slightly positive linear correlation observed below.



2.

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

Residuals:
    Min      1Q  Median      3Q     Max
-216.48  -88.58  -34.54   61.18  652.10

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  193.701     19.157  10.111  < 2e-16 ***
Fiber          3.813      1.383   2.757  0.00618 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.6 on 313 degrees of freedom
Multiple R-squared:  0.02371,   Adjusted R-squared:  0.02059
F-statistic:   7.6 on 1 and 313 DF,  p-value: 0.006179
```



From the regression coefficients, the summary table above indicates that the intercept term is 193.701, meaning that when Fiber is 0, Cholesterol is equal to 193.701. Meanwhile, the slope of the regression line is equal to 3.813, which indicates that a 1 unit increase in Fiber (when all else held constant) will result in an increase in Cholesterol of 3.813 units. Based on the t-statistics and the p-values above, we can reject the null hypothesis for both the intercept and the regression coefficient that they are equal to zero. The adjusted R-squared value for the model is 0.02059. which can be interpreted that Fiber only accounts for about 2% of the variance in Cholesterol and is indicative of a poor model. This is not surprising considering the scatterplot seen above in question 1 as there does not seem to be much of a relationship between these two variables. The F-statistic in the model summary above was calculated to be 7.6002 while the critical F-value resulted in a value of 3.8713. Since the F-statistic is greater than that of the critical F-value, we

can confidently reject the null hypothesis. There are also plots above relating to outliers, by observing leverage and Cook's Distance. From the plots and equations above there appear to be no outliers based on Cook's Distance, while there are 19 potential leverage outliers. Based on all the diagnostics of the model, Fiber is not a good predictor of Cholesterol on its own.

3. Below is the model summary, where the dummy variable, AlcNone ,was chosen to be left out.

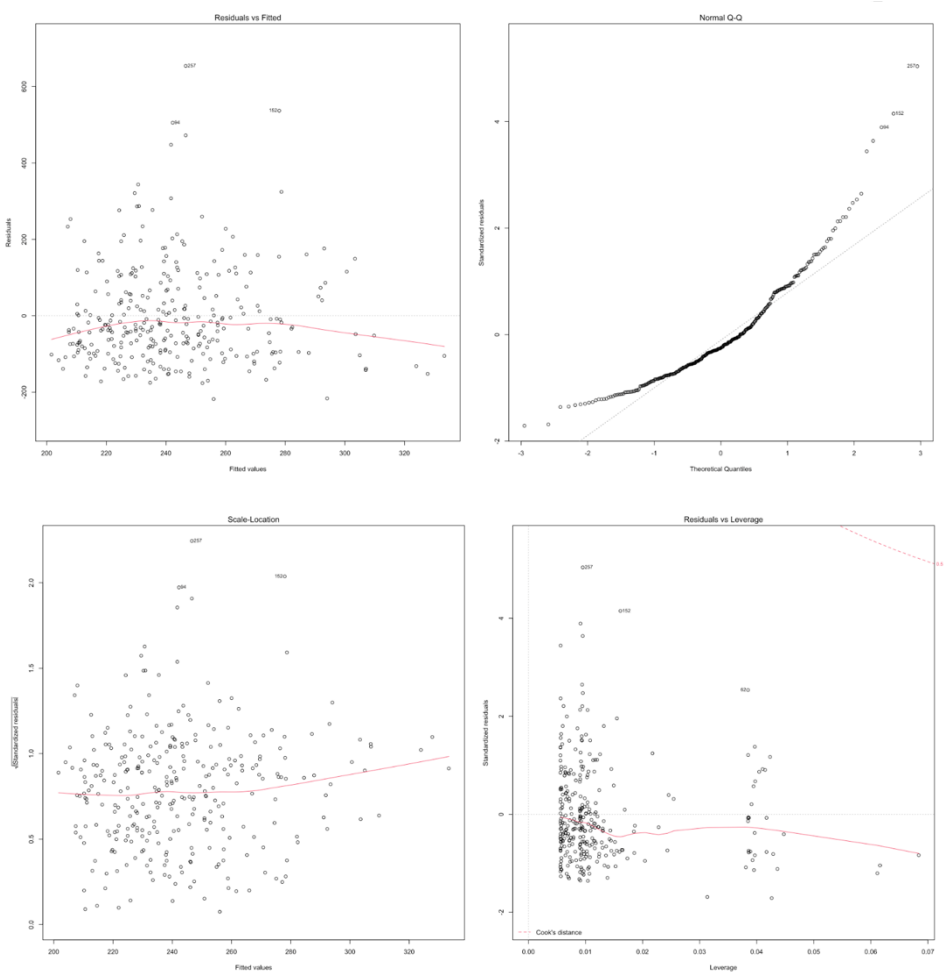
```
Call:
lm(formula = Cholesterol ~ Fiber + AlcLow + AlcHigh, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-218.31  -91.83  -32.24   64.65  654.06

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  189.266     21.065   8.985  < 2e-16 ***
Fiber          3.984      1.389   2.868  0.00441 **
AlcLow        -2.523     15.836  -0.159  0.87352
AlcHigh       44.429     28.429   1.563  0.11912
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 130.4 on 311 degrees of freedom
Multiple R-squared:  0.03296,   Adjusted R-squared:  0.02363
F-statistic: 3.533 on 3 and 311 DF,  p-value: 0.01518
```

The baseline model here consists of the intercept and the coefficient for Fiber. This is the regression equation when AlcNone is equal to 1 (since that is our baseline group):

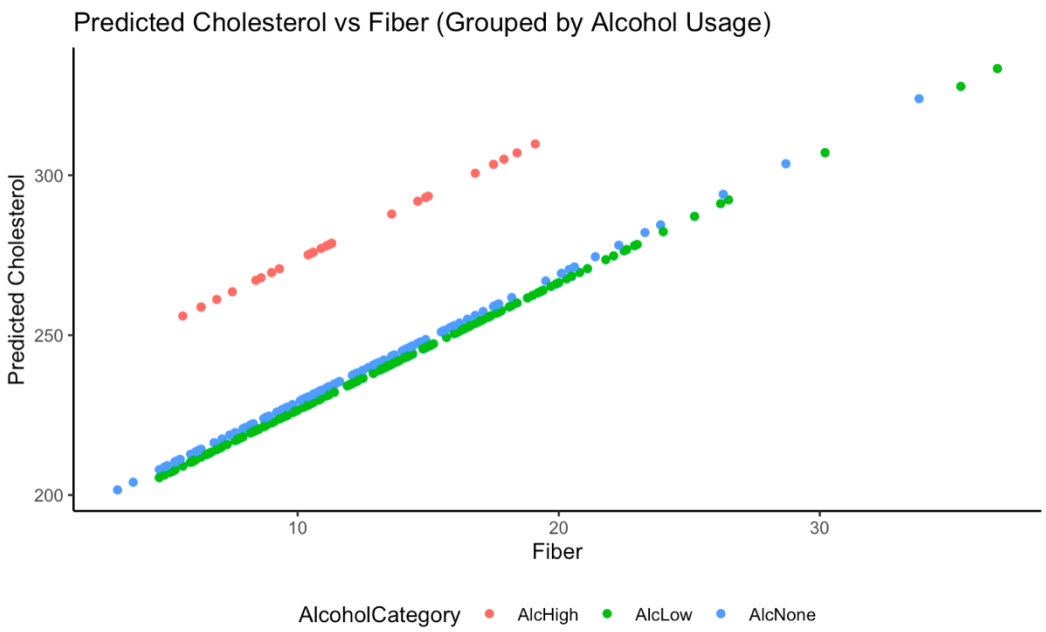
$$\hat{Y} = 189.266 + 3.9846X_1 - 2.523X_2 + 44.429X_3$$

The intercept of 189.266 is the Cholesterol value for the AlcNone group when all other variables are equal to zero, while the coefficient for Fiber (3.984) is the slope of our regression equation for the AlcNone group. It indicates that for a one unit increase in Fiber, we can expect a 3.984 unit increase in Cholesterol when all other variables are held constant. When observing AlcNone, then AlcLow, the Y-hat value will decrease by -2.523 units, indicating that the lines

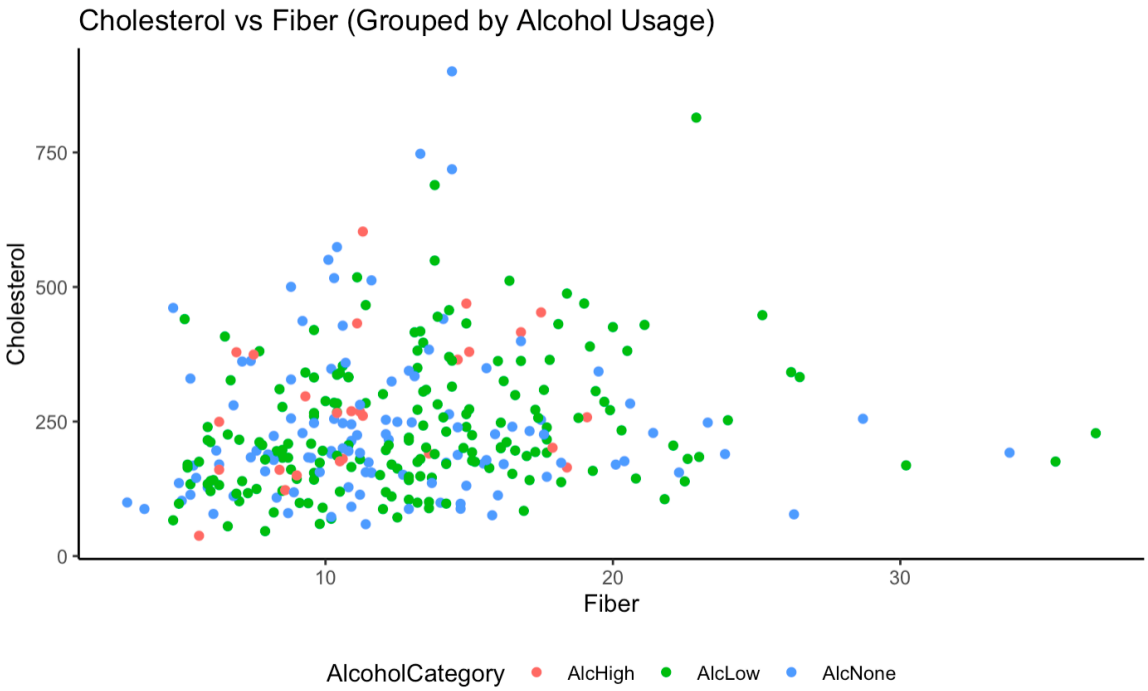
for AlcNone and AlcLow are parallel and 2.523 units apart. When moving from AlcNone to AlcHigh, we get an increase in Y-hat of 44.429. Overall, this indicates that all three lines are parallel with AlcLow being lower than AlcNone by 2.523 units and AlcHigh being higher than AlcNone by 44.429 units. From the table above, the F-statistic from the model summary was generated to be 3.533, while the critical F-value was calculated to be 2.6336. Since the F-statistic has a greater value than that of the critical F-value, we reject the null hypothesis. If we observe the diagnostic graphs, we can see several potential outliers on the Residuals vs Fitted graph (top left), while also being able to observe quite a few values on the Residuals vs Leverage graph (bottom right). Hence, the resultant 33 potential leverage data points were calculated below, though there were no outliers based on Cook's Distance.



4.



The patterns follow the parallel pattern indicated by their regression coefficients from task 3. We can see that AlcLow is slightly lower than AlcNone and AlcHigh is significantly higher than AlcNone.



In observing the scatterplot of Cholesterol vs Fiber, it is clear that ANCOVA model does not fit the data very well. There is a significant overlap in Cholesterol values for each of the Alcohol Groups.

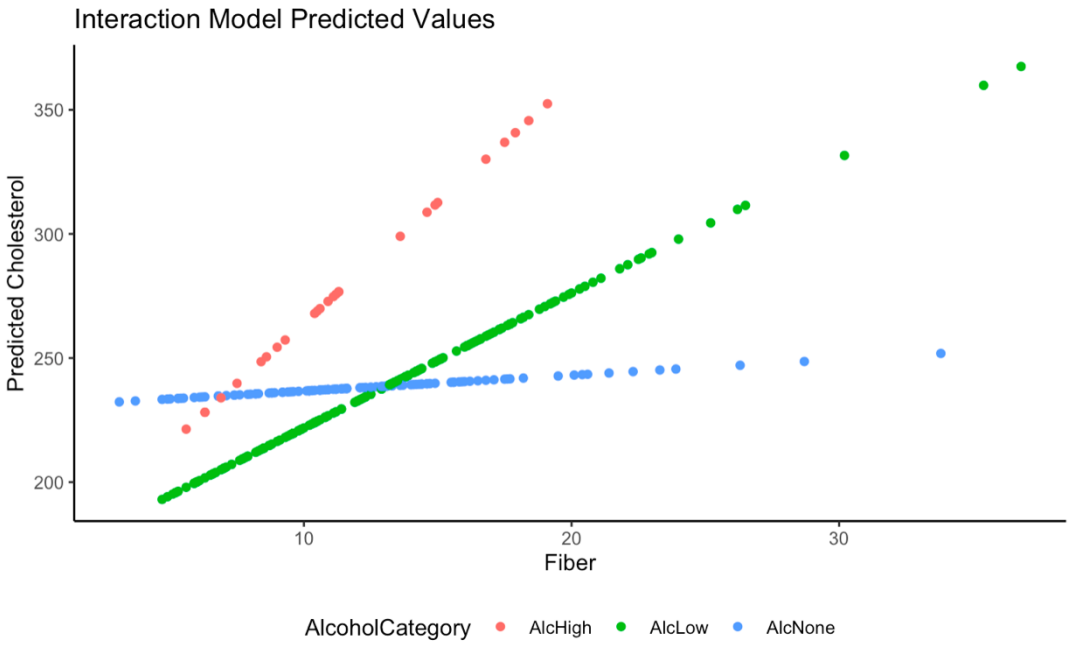
5. Below is the model summary in which AlcNone was left out as well as the interaction variable, AlcNone_Fiber, to establish a baseline:

```
Call:
lm(formula = Cholesterol ~ Fiber + AlcLow + AlcHigh + AlcLow_Fiber +
    AlcHigh_Fiber, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-184.25  -88.39  -25.85   64.40  661.19

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   230.3434    31.5413   7.303 2.41e-12 ***
Fiber           0.6363     2.3655   0.269    0.788
AlcLow        -62.8481    40.5528  -1.550    0.122
AlcHigh       -63.3814    85.4549  -0.742    0.459
AlcLow_Fiber    4.7976     2.9565   1.623    0.106
AlcHigh_Fiber   9.0742     6.8735   1.320    0.188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.1 on 309 degrees of freedom
Multiple R-squared:  0.04366,   Adjusted R-squared:  0.02819
F-statistic: 2.821 on 5 and 309 DF,  p-value: 0.01651
```

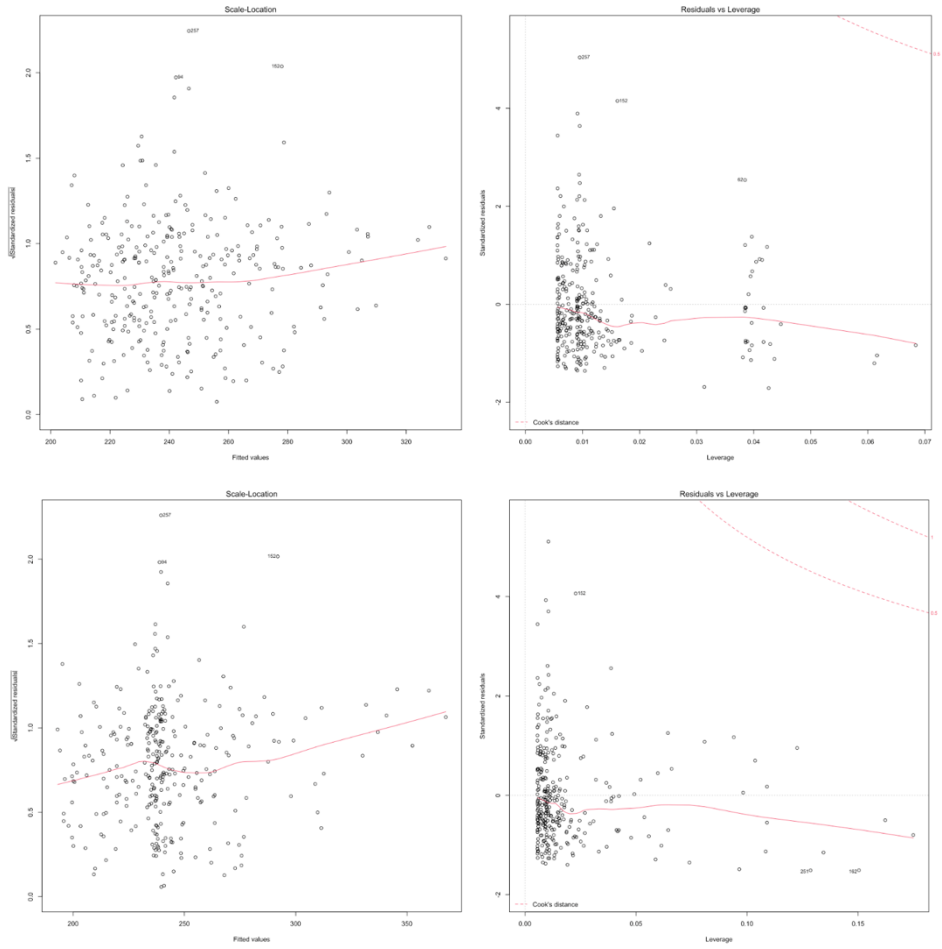
$$\hat{Y} = 230.3434 + 0.6363X_1 - 62.8481X_2 - 63.3814X_3 + 4.7976X_4 + 9.0742X_5$$

The intercept of 230.34 is representative of the Cholesterol level of the baseline group, AlcNone. The Fiber coefficient represents the slope of the regression equation for the AlcNone group and indicates that a one unit increase in Fiber will result in a 0.636 unit increase in Cholesterol when all else is held constant. The AlcLow coefficient (-62.85), represents the change in the intercept as we switch to the AlcLow group from AlcNone. The AlcLow_Fiber coefficient represents the change in slope as we shift from AlcNone to AlcLow, meaning that the intercept will decrease and the slope will get steeper as we transition to AlcLow from AlcNone. The transition from AlcNone to AlcHigh follows the same scenario with the intercept decreasing and the slope

increasing (getting steeper). This clearly indicates that the slopes are not the same for these Alcohol categories and that there is interaction occurring between the Fiber and Alcohol features and can more clearly be seen in the following graph:



It seems clear from this graph that the slopes are not equal and there is considerable interaction occurring between Fiber and Alcohol level, as all three of the Alcohol categories have different slopes. The graphic visualization of these differences really has helped with the interpretation of the coefficients of the interaction terms. As it pertains to the omnibus F-test, the F-statistic in the model summary table is 2.821, while the critical F-value was calculated 2.2432. The critical F-value is lower than the value of the F-statistic, meaning that we can reject the null hypothesis. If we observe the diagnostic graphs, we can see several potential outliers on the Residuals vs Fitted graph (top left), while also being able to observe quite a few values on the Residuals vs Leverage graph (bottom right). It appears that there are no outliers based on Cook's Distance, though there are 36 potential leverage outliers.



6. The null and alternate hypotheses have been stated above in the Formulas/Hypothesis overview portion of the report. The F-statistic is 1.7293, which is less than the critical F-value of 3.0254, would indicate that we fail to reject the null hypothesis. From the output, the addition of the interaction terms in the model can be seen to not add any significant information for predicting Cholesterol and indicates that we cannot say that our regression lines are not parallel. Since there was a failure to reject this hypothesis, ANCOVA is an appropriate model for this data, though there is not enough evidence to support saying that the slopes are unequal.

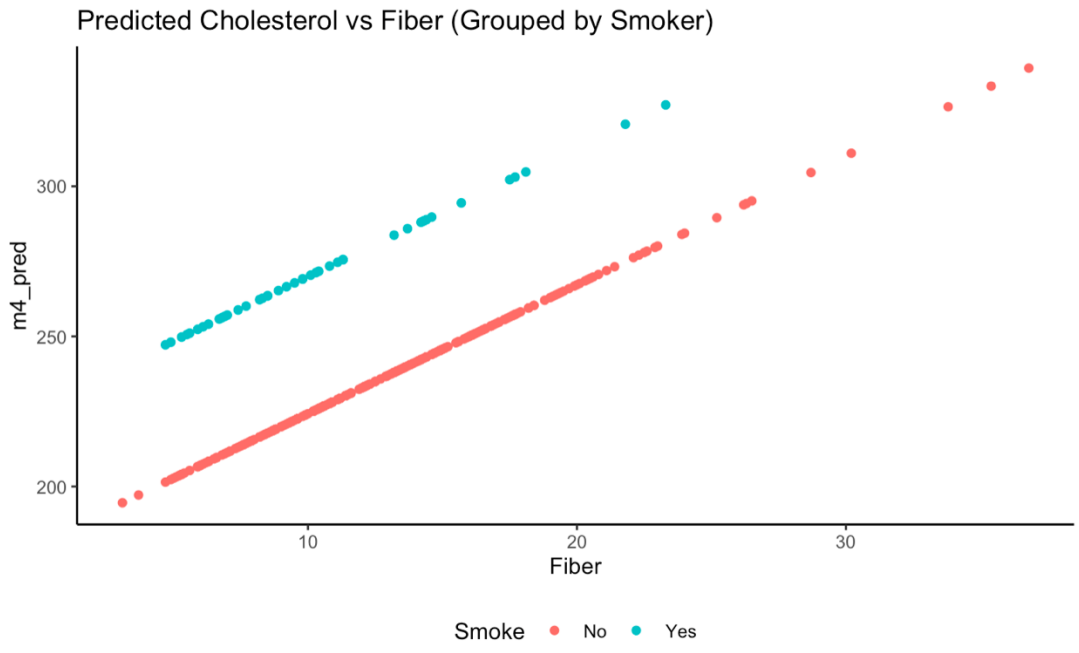
```
Model 1: Cholesterol ~ Fiber + AlcLow + AlcHigh + AlcLow_Fiber + AlcHigh_Fiber
Model 2: Cholesterol ~ Fiber + AlcLow + AlcHigh
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    309 5231592
2    311 5290147 -2    -58556 1.7293 0.1791
```

7. Below is the summary table for Model 4 (no interactions), which used the variable SmokeYes, with the Fiber feature to predict Cholesterol:

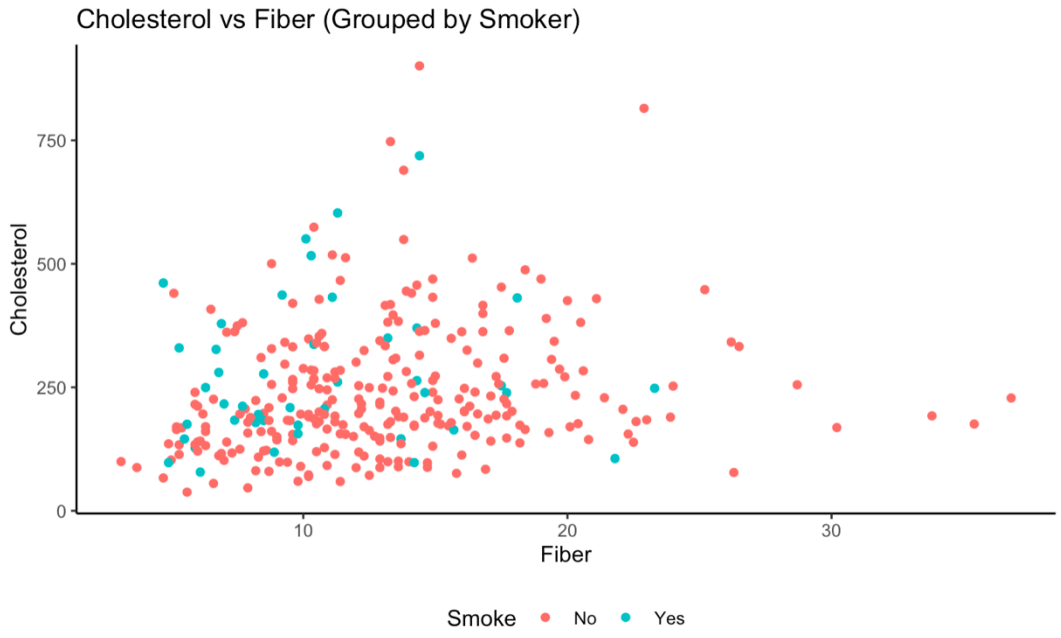
```
Call:
lm(formula = Cholesterol ~ Fiber + SmokeYes, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-216.77  -87.79  -35.81   65.54  657.56

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.274     19.936   9.093  < 2e-16 ***
Fiber          4.296      1.394   3.081  0.00224 **
SmokeYes      45.738     21.611   2.116  0.03510 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.9 on 312 degrees of freedom
Multiple R-squared:  0.03752,   Adjusted R-squared:  0.03135
F-statistic: 6.082 on 2 and 312 DF,  p-value: 0.002563
```

$$\hat{Y} = 181.274 + 4.296X_1 + 45.738X_2$$

The baseline represents the SmokeNo variable, while the intercept and Fiber coefficients are representative of the regression equation for the baseline group. The SmokeYes coefficient indicates an increase in Y-hat of 45.738 as we switch from SmokeNo to SmokeYes and can be observed in the following graph:



It appears, from the scatterplot below, that there is quite a bit of overlap in smokers vs non-smokers.



Below is the summary table for Model 5 (interaction terms), which used the variables SmokeYes and SmokeYes_Fiber with the Fiber feature to predict Cholesterol:

```
Call:
lm(formula = Cholesterol ~ Fiber + SmokeYes + SmokeYes_Fiber,
    data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-218.86  -87.71  -35.15   65.11  657.36

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    179.184     20.875   8.583 4.47e-16 ***
Fiber            4.455      1.471   3.028  0.00267 **
SmokeYes        63.059     55.002   1.146  0.25248
SmokeYes_Fiber  -1.597      4.661  -0.343  0.73218
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.1 on 311 degrees of freedom
Multiple R-squared:  0.03789,   Adjusted R-squared:  0.02861
F-statistic: 4.082 on 3 and 311 DF,  p-value: 0.007277
```

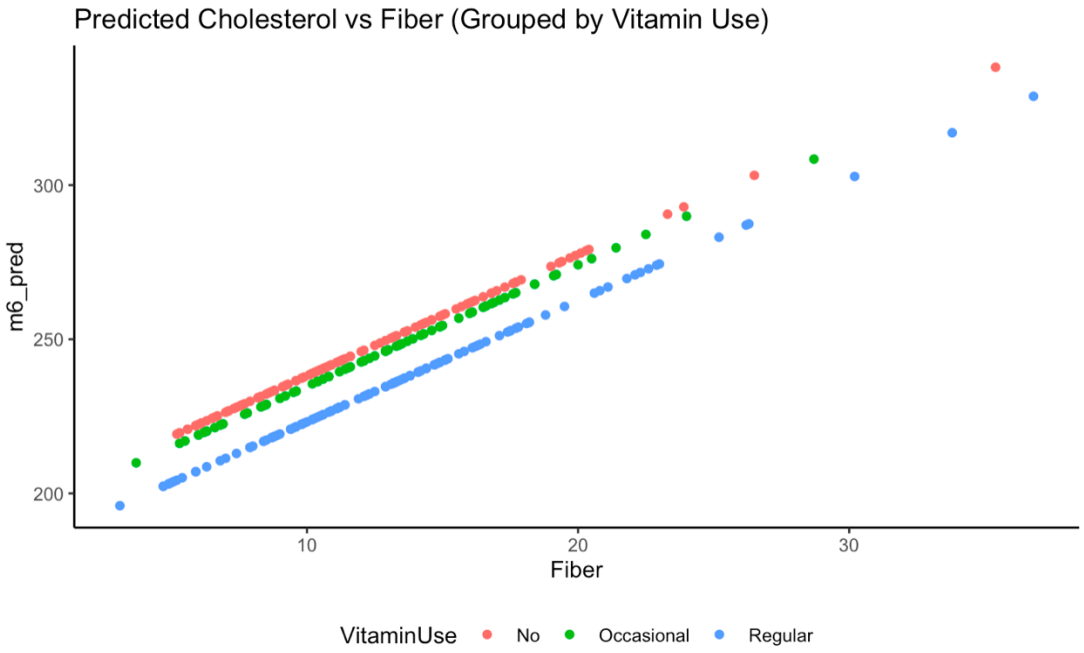
$$\hat{Y} = 179.184 + 4.455X_1 + 63.059X_2 - 1.597X_3$$

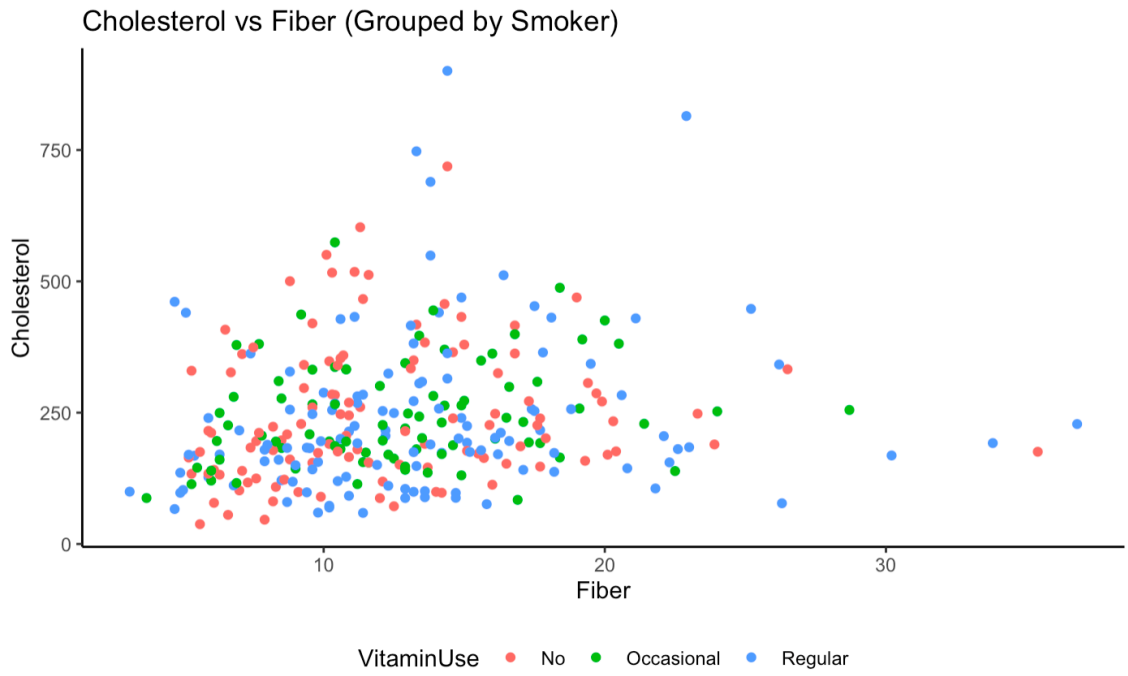
The SmokeYes_Fiber interaction coefficient indicates that as we move from SmokeNo to SmokeYes, the slope of the regression line will decrease slightly, resulting in less of an increase in Cholesterol for a one unit increase in Fiber. The F-statistic calculated using Models 4 and 5 resulted in a value of 0.0587, which is less than the critical F-value that was found to be 3.0248. Therefore, we fail to reject the null hypothesis and the partial F-test suggests that we cannot conclude that the lines are not parallel. Below is the summary table for Model 6 (no interactions), which used the variables VitaminOcc and VitaminReg with the Fiber feature to predict Cholesterol:

```
Call:
lm(formula = Cholesterol ~ Fiber + VitaminOcc + VitaminReg, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-209.93  -88.01  -35.04   62.02  660.16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  198.745     20.990   9.469  < 2e-16 ***
Fiber          3.940      1.393   2.828  0.00498 **
VitaminOcc    -3.401     19.074  -0.178  0.85861
VitaminReg   -14.947     17.259  -0.866  0.38714
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.9 on 311 degrees of freedom
Multiple R-squared:  0.02627,   Adjusted R-squared:  0.01688
F-statistic: 2.797 on 3 and 311 DF,  p-value: 0.04034
```

$$\hat{Y} = 198.745 + 3.940X_1 - 3.401X_2 - 14.947X_3$$

The baseline group for this model are those with no vitamin usage. The coefficients for VitaminOcc and VitaminReg indicate that Y-hat will decrease as we move from VitaminNone to the other groups, denoting that Cholesterol will be lower along the regression line for those that use Vitamins. We can see this in the graphic below:





It appears, from the scatterplot above, that there is quite a bit of overlap in occasional vs regular vs non-vitamin users. Below is the summary table for Model 7 (interaction terms), which used the variables VitaminOcc, VitaminReg, VitaminOcc_Fiber and VitaminReg_Fiber with the Fiber feature to predict Cholesterol:

```
Call:
lm(formula = Cholesterol ~ Fiber + VitaminOcc + VitaminReg +
    VitaminOcc_Fiber + VitaminReg_Fiber, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-214.64  -91.71  -33.55   63.36  659.80

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      208.821     32.308   6.463 3.99e-10 ***
Fiber              3.111      2.454   1.267    0.206
VitaminOcc       -19.453     52.883  -0.368    0.713
VitaminReg       -29.942     43.947  -0.681    0.496
VitaminOcc_Fiber   1.300      3.945   0.329    0.742
VitaminReg_Fiber   1.196      3.188   0.375    0.708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131.3 on 309 degrees of freedom
Multiple R-squared:  0.02681,   Adjusted R-squared:  0.01106
F-statistic: 1.702 on 5 and 309 DF,  p-value: 0.1338
```

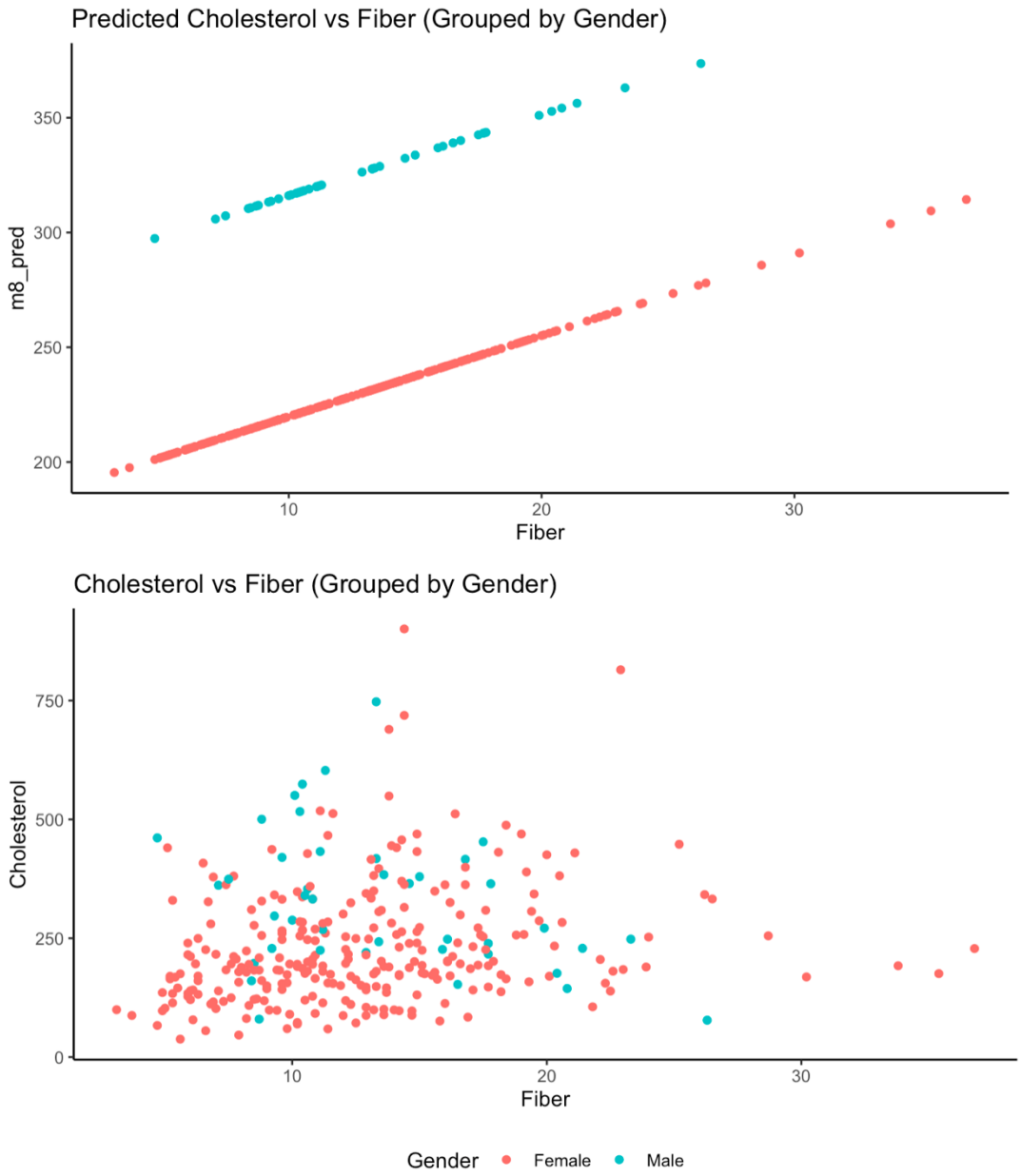
$$\hat{Y} = 208.821 + 3.111X_1 - 19.453X_2 - 29.942X_3 + 1.300X_4 + 1.196X_5$$

The interaction coefficients show us that there is a slight increase in the slope of the regression line as we move from group VitaminNone_Fiber to VitaminOcc_Fiber and VitaminReg_Fiber as both intercepts decrease while the slope increases slightly. The F-statistic calculated using Models 6 and 7 resulted in a value of 0.0566, which is less than the critical F-value that was found to be 2.6338. Therefore, we fail to reject the null hypothesis and the partial F-test suggests that we cannot conclude that the lines are not parallel. Below is the summary table for Model 8 (no interactions), which used the variable GenderFemale, with the Fiber feature to predict Cholesterol:

```
Call:
lm(formula = Cholesterol ~ Fiber + GenderFemale, data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-296.10  -85.19  -30.31   56.56  665.39

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   280.784     26.572  10.567  < 2e-16 ***
Fiber           3.529      1.342   2.629  0.00898 **
GenderFemale  -96.294     21.013  -4.583 6.65e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126.6 on 312 degrees of freedom
Multiple R-squared:  0.08527,   Adjusted R-squared:  0.07941
F-statistic: 14.54 on 2 and 312 DF,  p-value: 9.149e-07
```

$$\hat{Y} = 280.784 + 3.529X_1 - 96.294X_2$$

The baseline group in this model is GenderMale, while the GenderFemale coefficient (-96.294) indicates that as we switch from Male to Female, the predicted value of cholesterol will drop by almost 100 units from the designated baseline of 280.78. This difference can be seen graphically below:

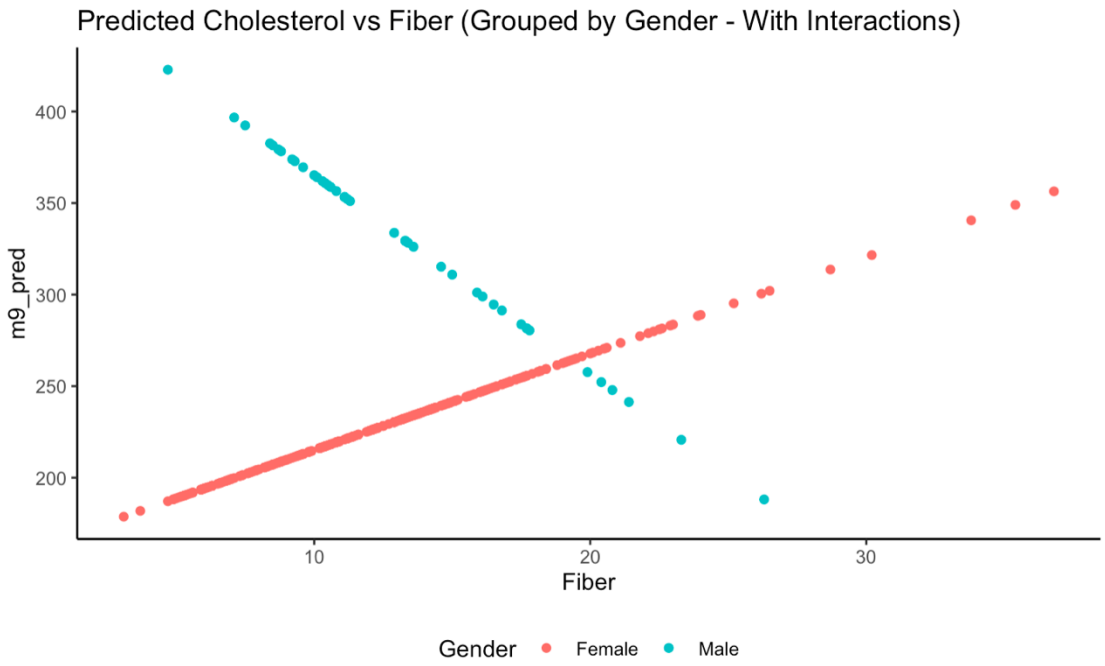


Below is the summary table for Model 9 (interaction terms), which used the variables GenderFemale and GenderFemale_Fiber with the Fiber feature to predict Cholesterol:

```
Call:
lm(formula = Cholesterol ~ Fiber + GenderFemale + GenderFemale_Fiber,
    data = n_df)

Residuals:
    Min      1Q  Median      3Q     Max
-299.55  -80.27  -25.28   53.23  662.41

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         473.873     56.936   8.323 2.75e-15 ***
Fiber               -10.865      3.998  -2.718 0.006939 **
GenderFemale       -311.514     60.083  -5.185 3.90e-07 ***
GenderFemale_Fiber   16.138      4.233   3.812 0.000166 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 124 on 311 degrees of freedom
Multiple R-squared:  0.1261,    Adjusted R-squared:  0.1177
F-statistic: 14.96 on 3 and 311 DF,  p-value: 4.028e-09
```

$$\hat{Y} = 473.873 - 10.865X_1 - 311.514X_2 + 16.138X_3$$

This model is quite interesting as the interaction between Gender and Fiber looks as if the slope changes as we switch from Male to Female. The intercept for Men is 473.87 and will reduce by 311.514 as we switch to female. The slope for men is negative and switches to positive for females, which indicates a strong interaction between Gender and Fiber. The F-statistic calculated using Models 8 and 9 resulted in a value of 7.2676, which is greater than the critical F-value that was found to be 3.0248. Therefore, we can reject the null hypothesis and the partial F-test suggests that there is a significant interaction between Gender and Fiber as it pertains to explaining additional variance in Cholesterol. The following graphic shows the interaction effect with the Genders having opposite slopes.



Overall, the only categorical variable that had a significant interaction with Fiber was Gender. We were able to show this through the partial F-test as Gender, in conjunction with Fiber, was the most predictive of Cholesterol. Despite this, the adjusted R-squared for the model was still extremely low indicating that there are other variables accounting for the variation in Cholesterol.

Conclusion

It took time for me to wrap my head around how ANCOVA works. The explanatory video in the module was very helpful in understanding how to interpret the regression coefficients, especially in the interaction terms. One thing that stood out was how it becomes increasingly difficult to interpret when more variables and interactions are added to a model. For these particular tasks, none of the models perform well in terms of accounting for variance in Cholesterol, but the purpose of this assignment was to get our feet wet with modeling a continuous, explanatory

variable along with a dummy encoded categorical variable. I look forward to experimenting further to build a good model that explains changes in cholesterol. Overall, the course is building up my repertoire of techniques and tools as it pertains to approaching modeling problems.