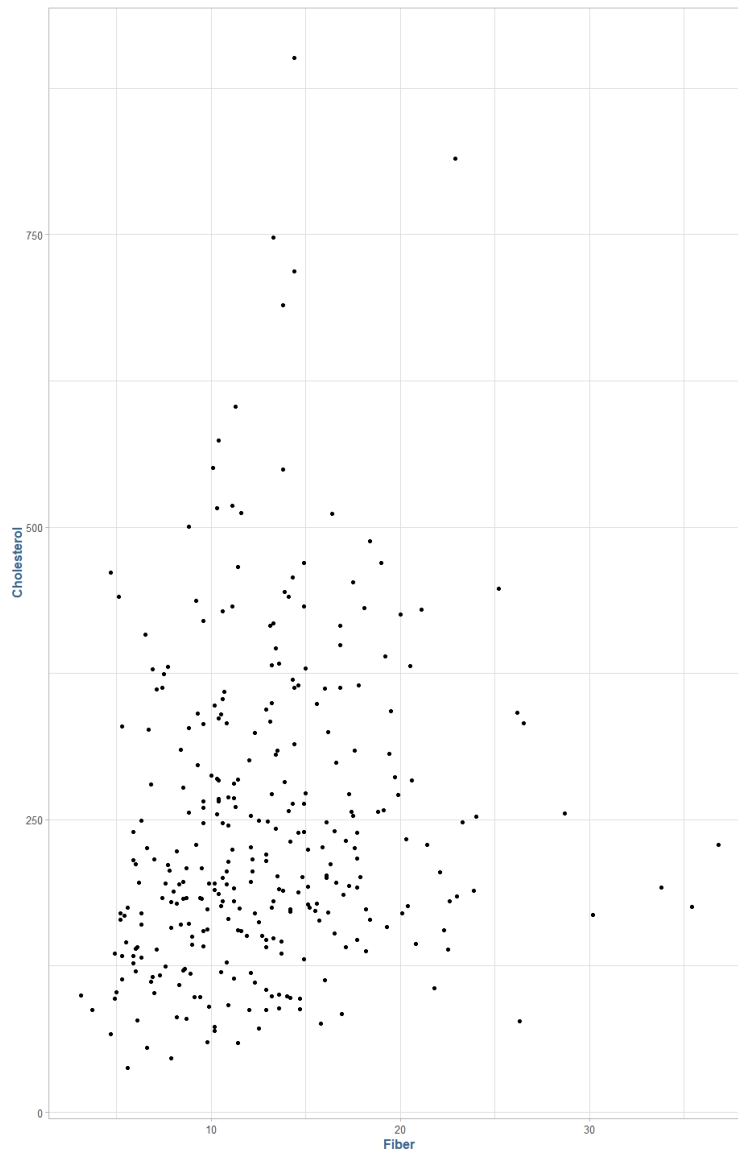


## INTRODUCTION

- 1) Consider the continuous variable, FIBER. Is this variable correlated with Cholesterol? Obtain a scatterplot and appropriate statistics to address this question.

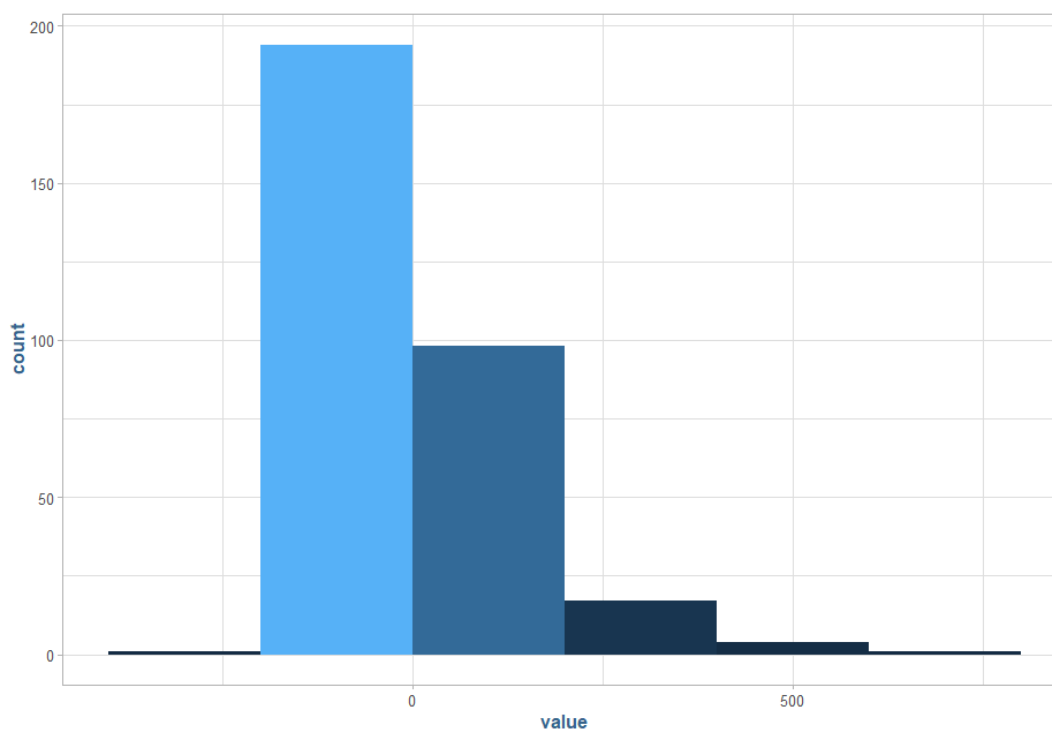
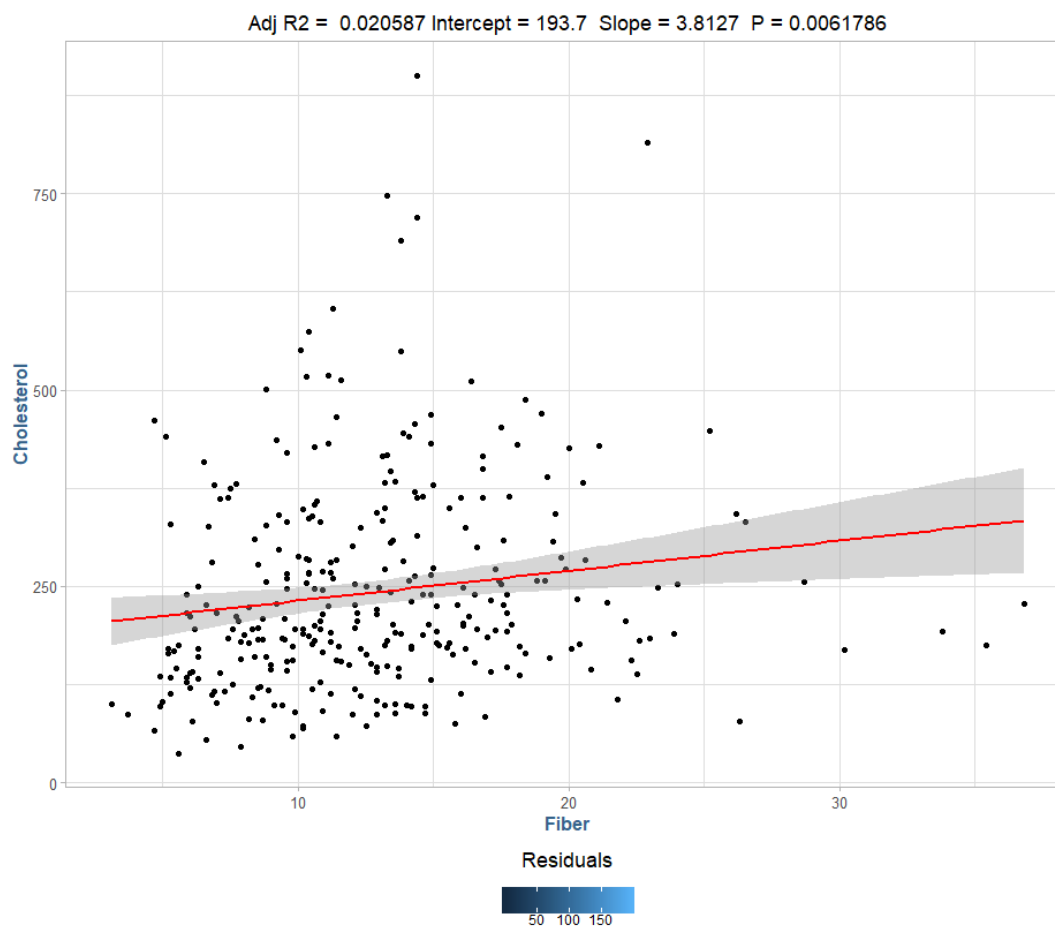
In the following diagram we see a lot of variance in the following scatterplot of Cholesterol ~ Fiber.



There is a liner correlation of .1539 between Cholesterol and Fiber.

- 2) Fit a simple linear regression model that uses FIBER to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss the goodness of fit.

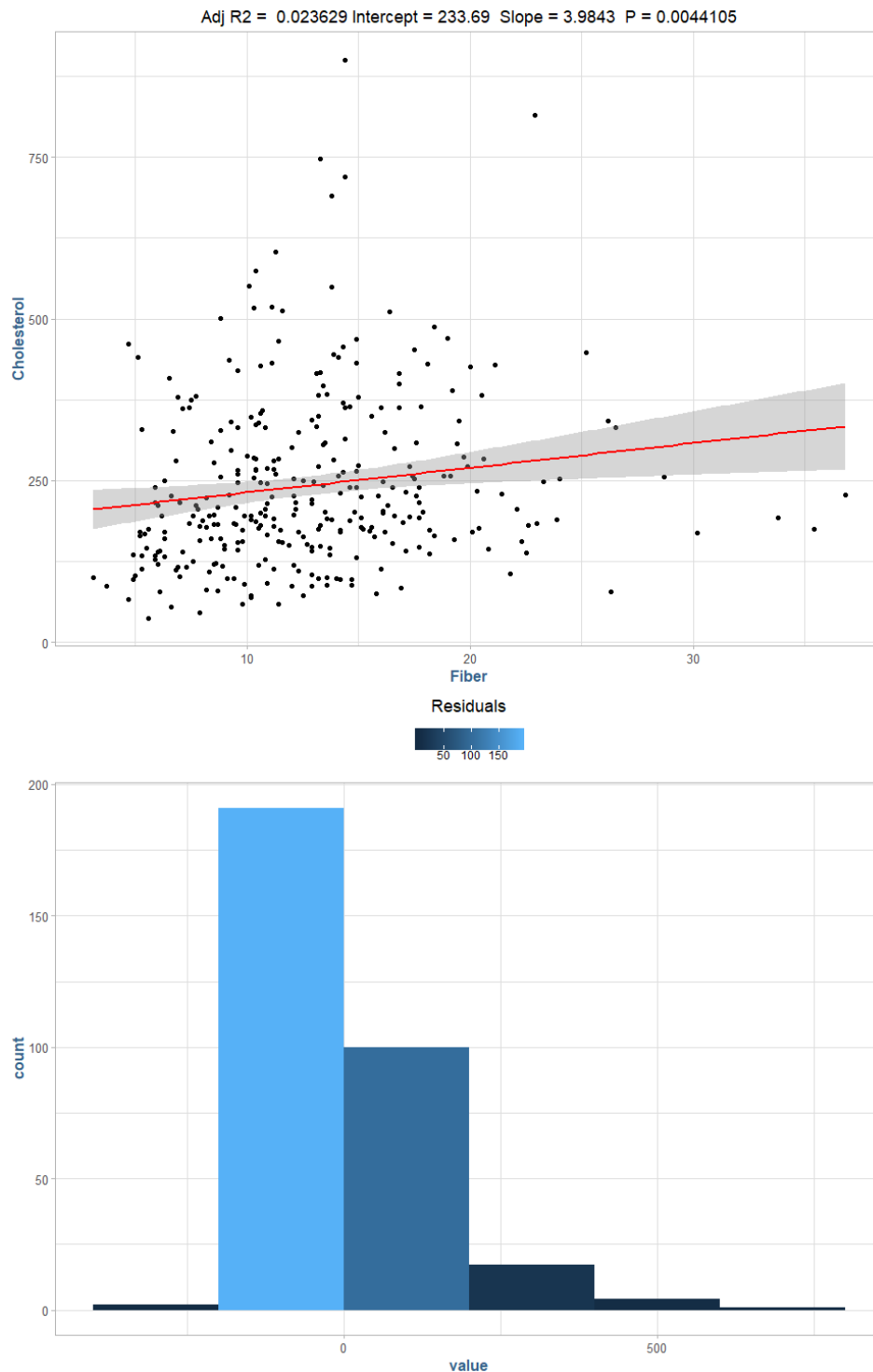
The Fiber variable appears to be a relatively poor indicator of cholesterol. In the graphic below, we can see an  $R^2$  value of 0.0237, indicating that approximately 2% of the variance in the data is explained by the Fiber variable.



Model:  $193.7014 + 3.8127 \beta_1$ , where  $\beta_1$  is fiber.

The intercept term here, 193.701, is the baseline cholesterol level for an individual, and for every 1 unit increase in fiber consumed it should account for an approximate 3.8 unit increase in an individual's cholesterol level.

3.) For the ALCOHOL categorical variable, create a set of dummy coded (0/1) indicator variables. Fit a multiple linear model that uses the FIBER continuous variable and the ALCOHOL dummy coded variables to predict the response variable  $Y=CHOLESTEROL$ . Remember to leave one of the dummy coded variables out of the model so that you have a basis of interpretation for the constant term. Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. This is called an Analysis of Covariance Model (ANCOVA).

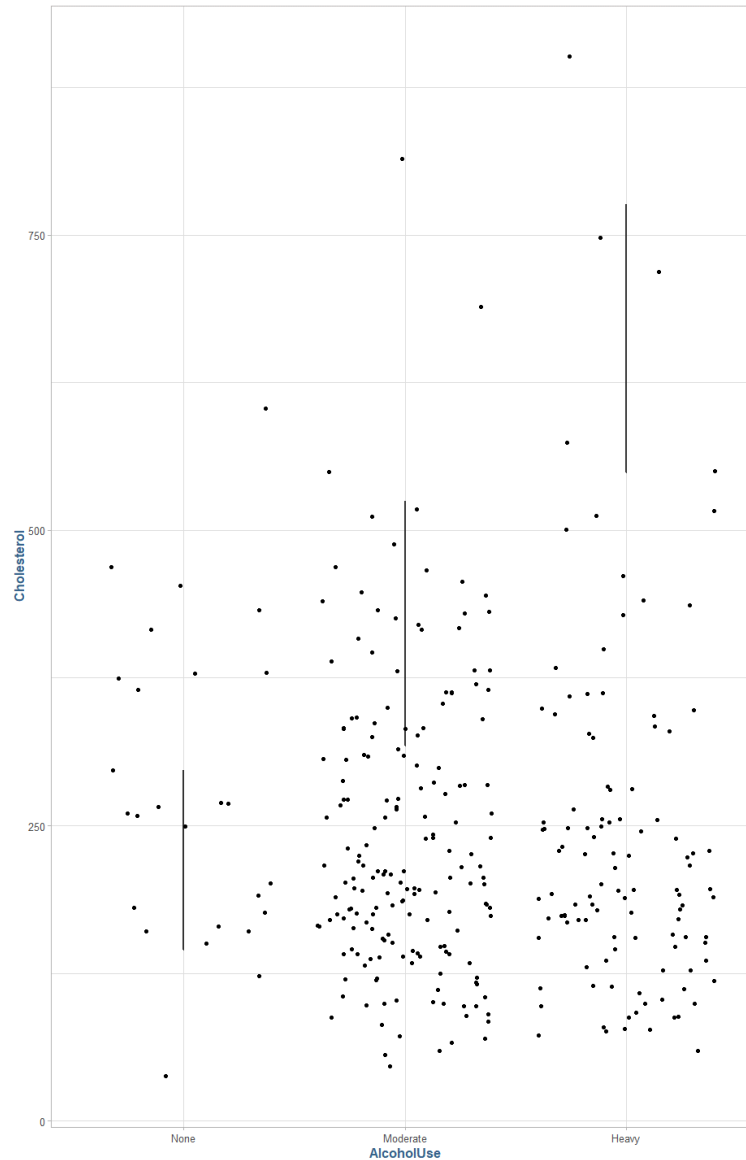


$$\text{Model 2: } 233.6946 + 3.9543\beta_1 - 46.9518\beta_2 - 44.4289\beta_3$$

Where  $\beta_1$  = Fiber,  $\beta_2$  = Alcohol (Moderate),  $\beta_3$  = Alcohol (Heavy)

This model will output a predicted value of cholesterol for an individual given their self-reported fiber intake and alcohol consumption, where each unit of fiber consumption increases their cholesterol by 3.6 points per unit, and alcohol will reduce the predicted cholesterol by either 46.96 points or 44.43 points depending on if they consume moderate or heavy amounts of alcohol, respectively. The  $R^2$  denotes that about 3.3% of the total variance in the data is explained by the model.

We should note that there is an uneven distribution of subjects reported with no alcohol consumption (26), relative to those who report either moderate (178) or heavy (111) consumption. Additionally, there is a great deal of variance in the cholesterol levels by alcohol use:



The null hypothesis in this case would be,

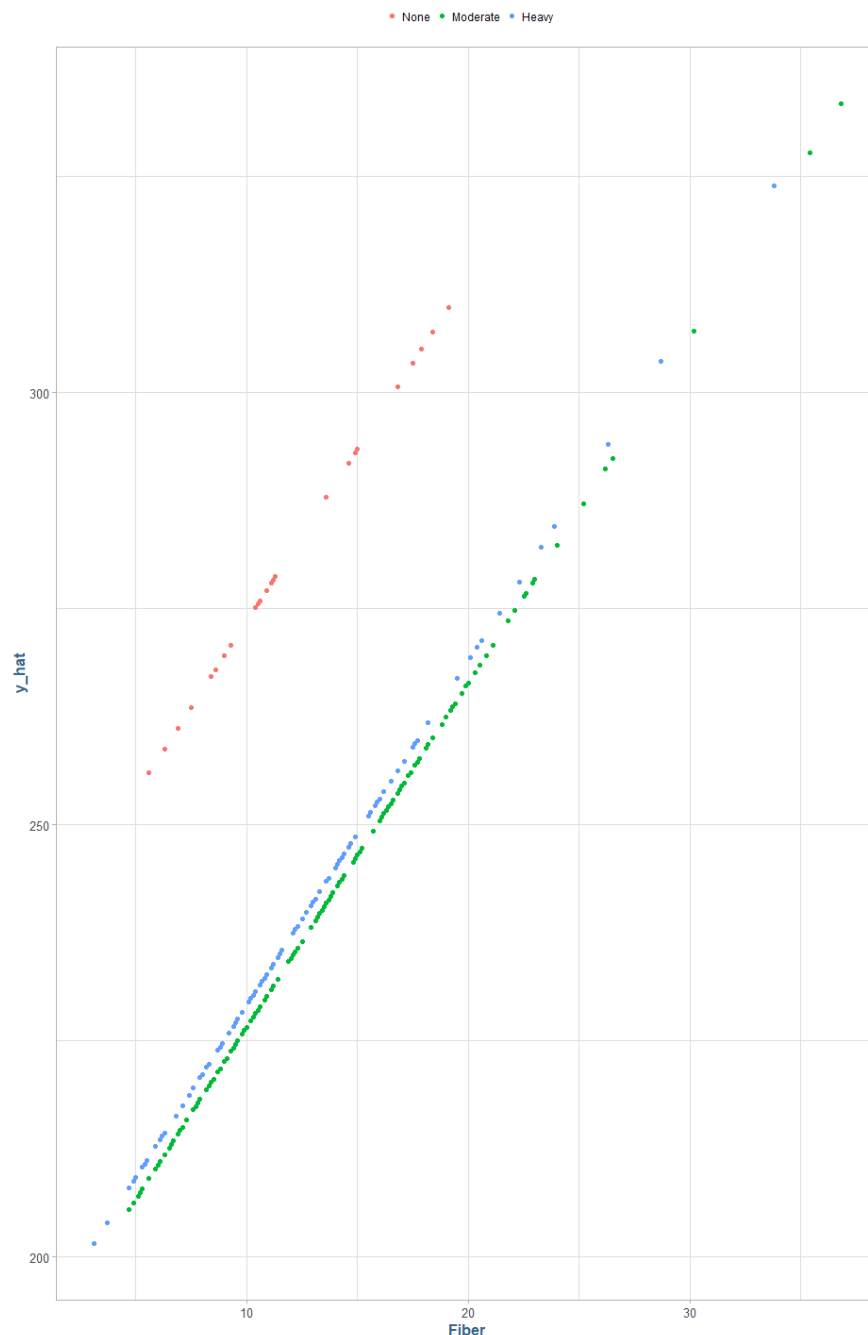
$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

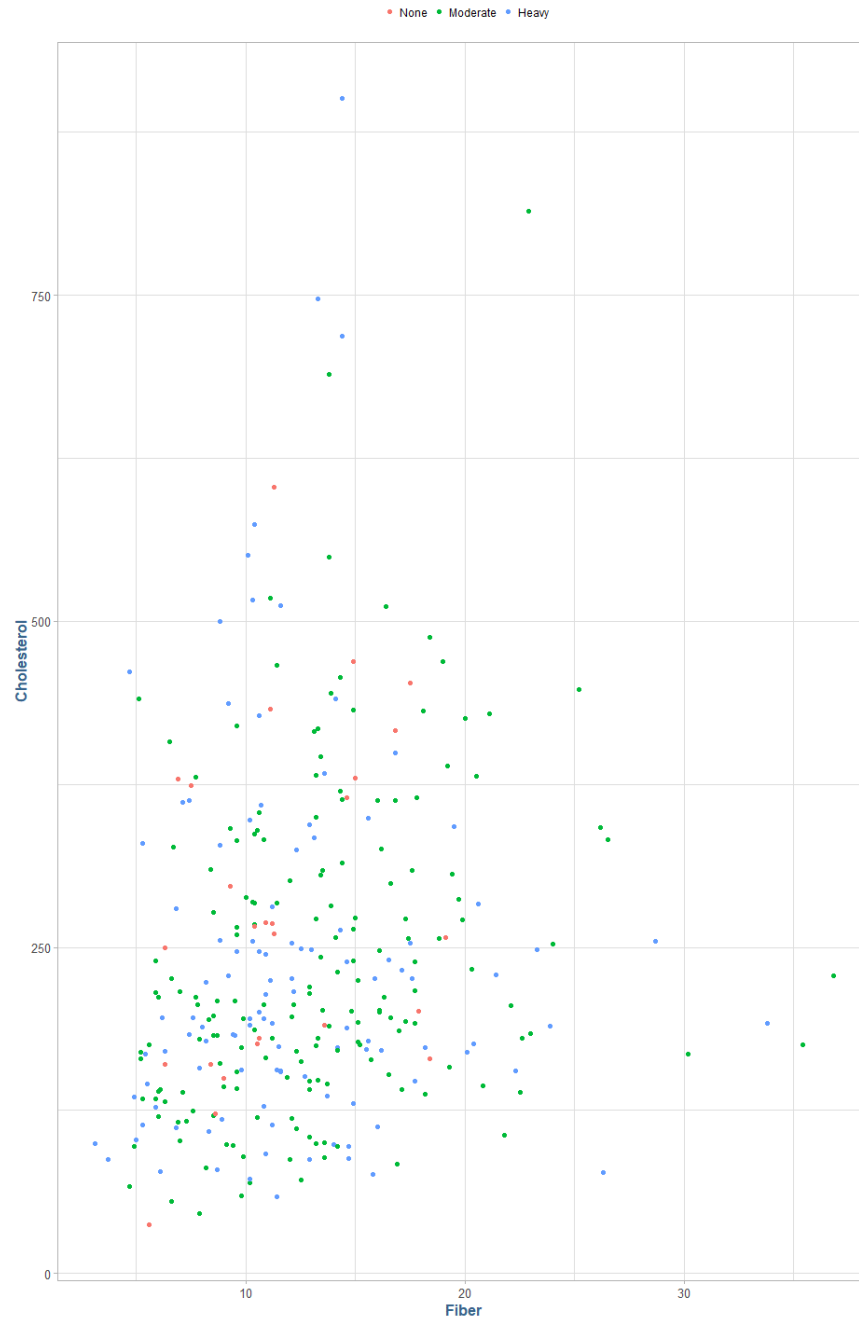
Or that there is no effect on the model using the beta coefficient derived from fiber consumption and the coded variables for alcohol consumption, against the alternative hypothesis that:

$$H_a: \beta_1 = \beta_2 = \beta_3 \neq 0$$

Or that there is additional variance explained in the data by including the beta coefficients. In our model summary, the p-value of  $< 0.05$  for our fiber variable suggests that there is statistically significant difference when using the beta1 coefficient. However, the alcohol variables have p-values  $> 0.5$ , which suggests that they are not statistically significant.

- 4.) Use the ANCOVA model from task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FIBER (X), but color code the records for the different groups of ALCOHOL. What do you notice about the patterns in the predicted values of Y? Now, make a scatterplot of the actual values of CHOLESTEROL(Y) by FIBER (X), but color code by the different groups of the ALCOHOL variable. If you compare the two scatterplots, does the ANCOVA model appear to fit the observed data very well? Or, is a more complex model needed?





The predicted values of cholesterol all fill in parallel straight lines separated by alcohol consumption, whereas the actual values are clustered together in pockets in the lower left quadrant of the graph. It does not appear that a linear model is reflective enough of the actual data to provide meaningful predictions.

- 5.) Create new interaction variables by multiplying the dummy coded variables for ALCOHOL by the continuous FIBER(X) variable. Save these product variables to your dataset. Now, to build the model, start with variables in your ANCOVA model from task 4) and add the interaction variables you just created into the multiple regression model. Don't forget, there is one category that is the basis of interpretation. DO NOT include any interaction term that is associated with that category. This is called an Unequal Slopes Model. Fit this model, and save the predicted values. Plot the predicted values for CHOLESTEROL (Y) by FIBER(X). Discuss what you see in this graph. In addition, report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics.



$$\text{Model 3: } 166.9620 + 9.7105\beta_1 + 0.5333\beta_2 + 63.3814\beta_3 - 4.2737\beta_4 - 9.0742\beta_5$$

Where  $\beta_1$  = Fiber,  $\beta_2$  = Alcohol (Moderate)  $\beta_3$  = Alcohol (Heavy)  $\beta_4$  = Fiber + Moderate Alcohol interaction,  $\beta_5$  = Fiber + Heavy alcohol interaction. This model has an  $R^2$  of 0.0437 denoting that it accounts for approximately 4.3% of the overall variance in the data. We can see in the graph above the predicted model values (straight lines) plotted over the actual values, both of which are color coded to indicate the level of alcohol use per individual. Heavy alcohol use appears to have an unequal slope to that of the non/moderate users.

The null hypothesis in this case would be,

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_a: \beta_j \neq 0, \text{ for at least one value of } j \text{ (for } j \text{ in } 1, 2, 3, 4, 5)$$

Or, simply that at least one of the interaction coefficients in the model is not zero, and that they would help explain the variance in the data greater than the intercept alone.

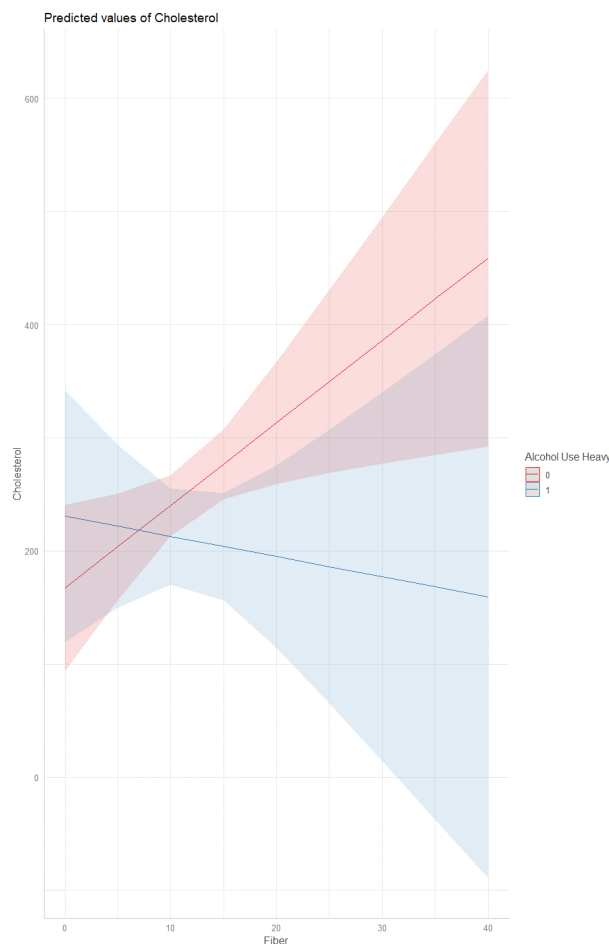
```
Call:
lm(formula = Cholesterol ~ Fiber + AlcoholUseModerate + AlcoholUseHeavy +
    Fiber * AlcoholUseModerate + Fiber * AlcoholUseHeavy, data = model3_data)

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)    166.9620    79.4210   2.102  0.0363
Fiber           9.7105     6.4536   1.505  0.1334
AlcoholUseModerate  0.5333     83.4108   0.006  0.9949
AlcoholUseHeavy  63.3814     85.4549   0.742  0.4588
Fiber:AlcoholUseModerate -4.2767     6.6929  -0.639  0.5233
Fiber:AlcoholUseHeavy  -9.0742     6.8735  -1.320  0.1878

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
```

Looking at the summary of the model, only the intercept p-value value falls below the standard 0.05 threshold and it does appear that the null hypothesis in this case cannot be rejected. Both of the interaction terms have relatively high p-values; however, it does appear visually that there is some difference with the group of heavy alcohol users and the overall model is significant at  $p < 0.5$ :





Model 3 (from task 4) with the additional interaction variables is the “full” model, and model 2 (from task 4) is the reduced model. The null and alternative hypothesis in this case would be:

$$H_a: \beta_j \neq 0, \text{ for at least one value of } j \text{ (for } j \text{ in } 5, 6)$$

**= 1.7293**

```

Analysis of Variance Table

Model 1: Cholesterol ~ Fiber + AlcoholUseModerate + AlcoholUseHeavy +
  FiberModerate + FiberHeavy
Model 2: Cholesterol ~ Fiber + AlcoholUseModerate + AlcoholUseHeavy
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     309 5231592
2     311 5290147 -2     -58556 1.7293 0.1791
>

```

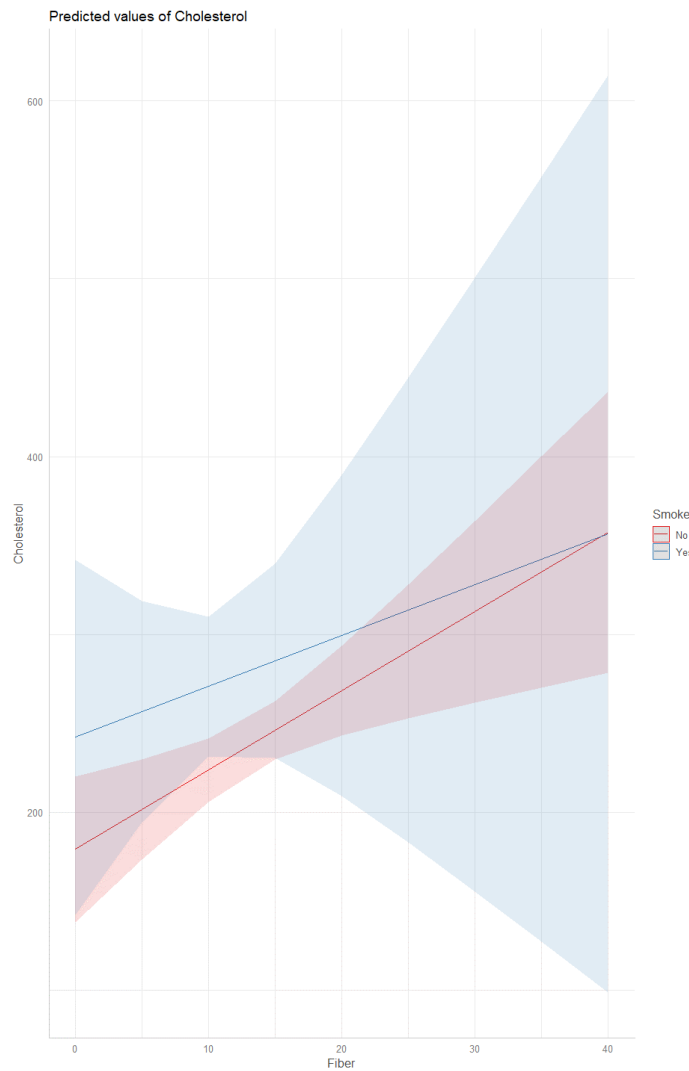
```
Call:
lm(formula = smoke, data = data.nutrition)

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 0.000000000000000447
Fiber         4.455     1.471   3.028    0.00267
SmokeYes      63.059    55.002   1.146    0.25248
Fiber:SmokeYes -1.597     4.661  -0.343    0.73218

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
```

In the preceding model, we note that the significant terms are the intercept and fiber, it appears that smoking and the interaction between fiber and smoking are not statistically significant. We can visualize this interaction in the following graph, noting the roughly equal lines and the large variances.



For vitamin use and fiber, we can the following terms generated from the model.

```
Call:
lm(formula = vitamin, data = data.nutrition)

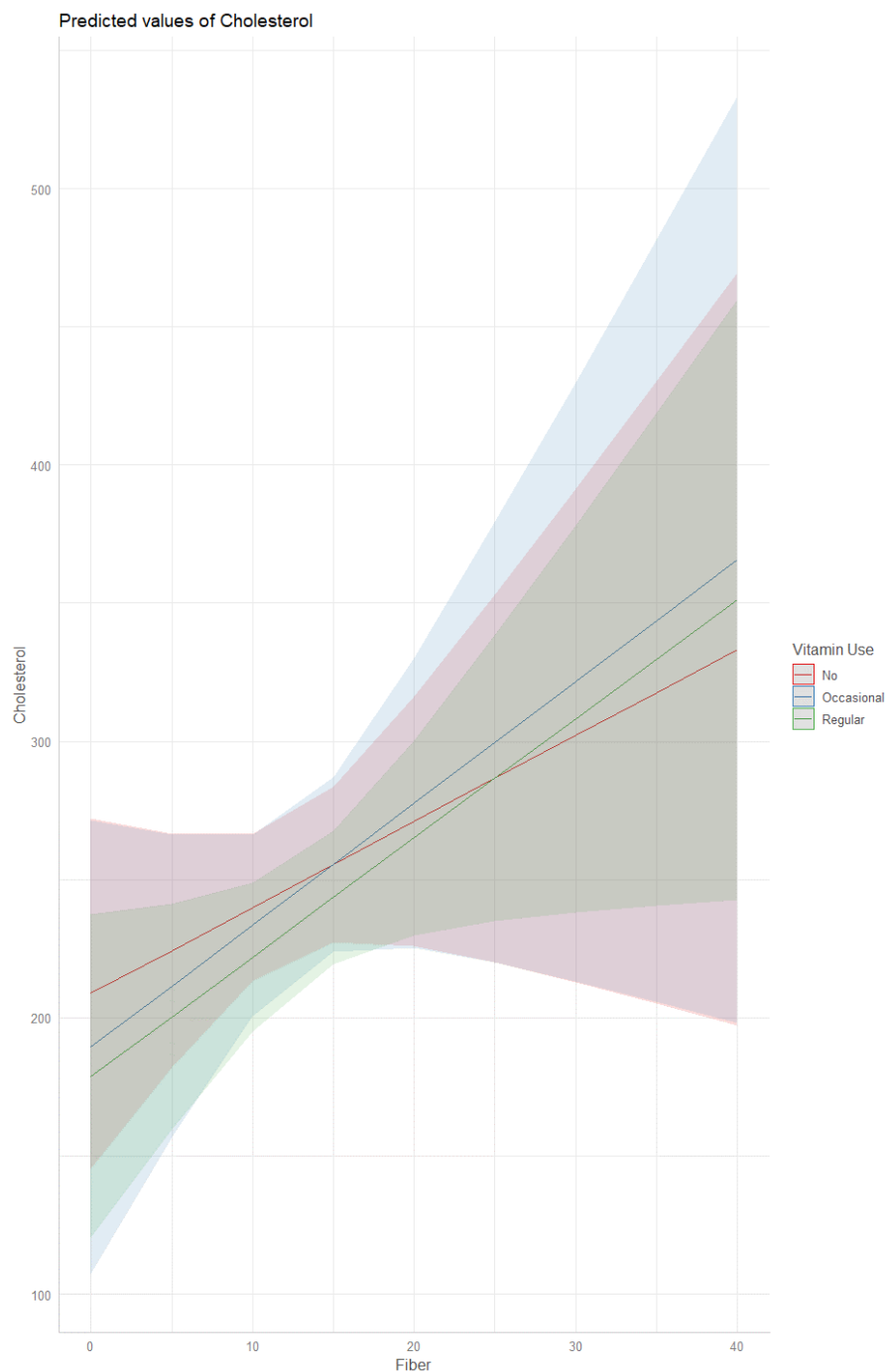
Residuals:
    Min       1Q   Median       3Q      Max
-214.64  -91.71  -33.55   63.36  659.80

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    208.821    32.308   6.463 0.00000000399
Fiber           3.111     2.454   1.267    0.206
VitaminUseOccasional -19.453    52.883  -0.368    0.713
VitaminUseRegular   -29.942    43.947  -0.681    0.496
Fiber:VitaminUseOccasional  1.300     3.945   0.329    0.742
Fiber:VitaminUseRegular    1.196     3.188   0.375    0.708

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

>
```

When fiber is used in conjunction with the various vitamin usage levels, it does not appear to have any statistically significant effect on cholesterol levels.



We can also confirm with the results of a type II anova test:

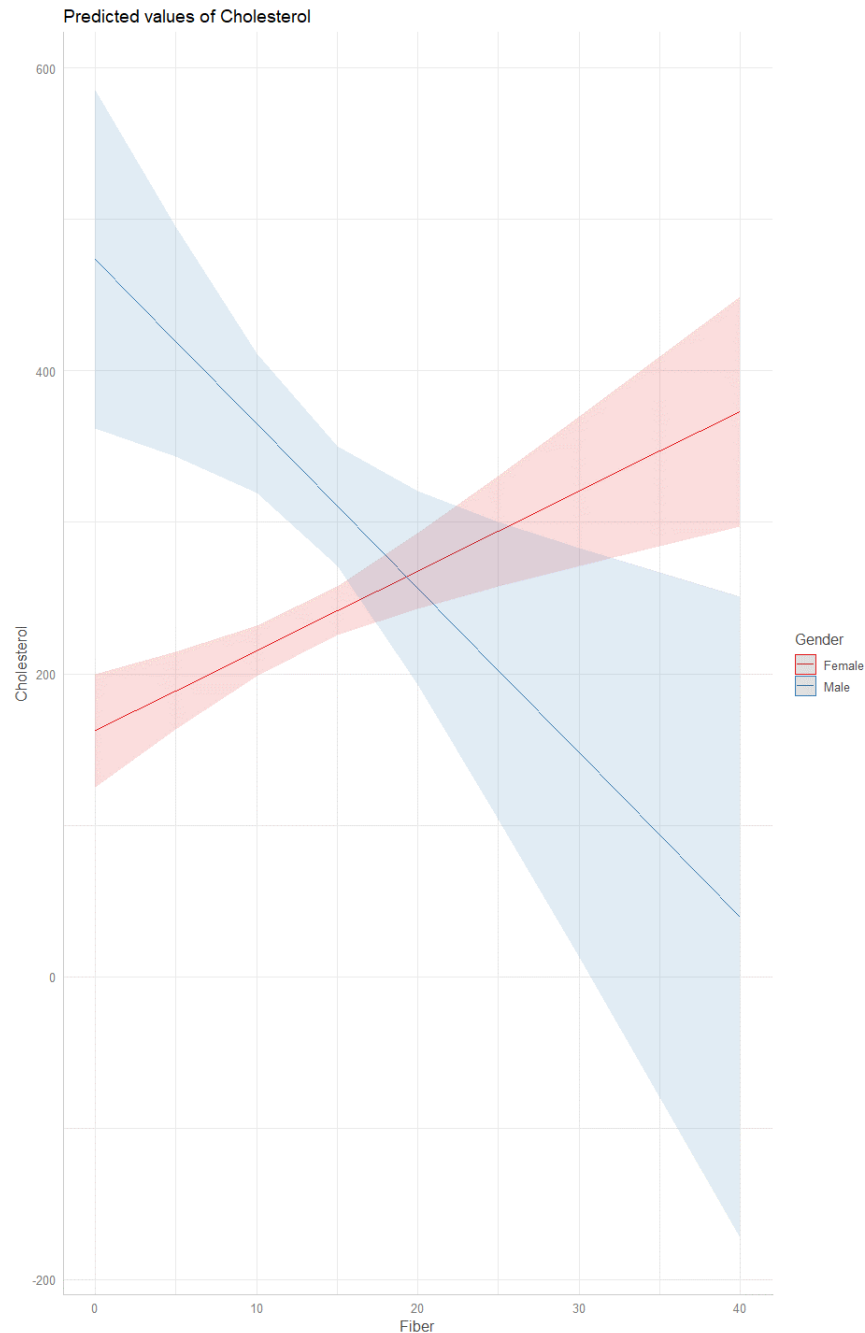
```

> anova(model2_fit, type = "II")
Anova Table (Type II tests)

Response: Cholesterol
      Sum Sq Df F value    Pr(>F)
Fiber    137019  1  7.9528 0.005112
VitaminUse  14027  2  0.4071 0.665955
Fiber:VitaminUse  2926  2  0.0849 0.918610
Residuals  5323804 309
>

```

The next variable under consideration is the gender variable in conjunction with the fiber variable. The male and female lines in the following figure clearly exhibit different behavior than the preceding variables.



We can clearly see an effect on the cholesterol levels with this interaction considered. For males, their cholesterol levels decrease sharply with increasing levels of fiber consumption, and a female's cholesterol appear to increase with increasing amounts of fiber. We can also execute a type II ANOVA test to confirm our visual inspection:

```
> Anova(model6_fit)
Anova Table (Type II tests)

Response: Cholesterol
      Sum Sq Df F value    Pr(>F)
Fiber   110868  1  7.2126 0.0076280
Gender   336804  1 21.9110 0.00000427
Fiber:Gender 223427  1 14.5352 0.0001659
Residuals 4780527 311
>
```

---

## CONCLUSION

In this lab we explore the statistical technique known as ANCOVA, where we assess the impact of categorical values in conjunction with continuous variables. We used the nutritional data collected from an observational study to assess the impacts of both types of variables of predicting an individual's cholesterol levels.

We started by looking at the correlation between fiber and cholesterol, given that one of the underlying assumptions for ANCOVA models is that there is a linear relationship between the dependent variable (cholesterol) and the covariate (fiber), we found a modest colinear association. We then fitted a simple linear model to the fiber variable and confirmed the normality of the residuals from the model.

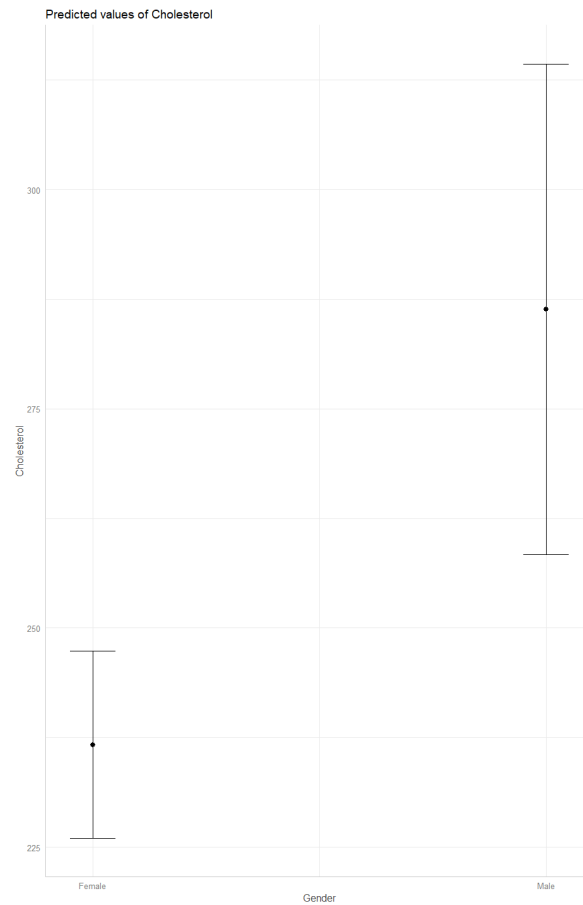
Our first ANCOVA model was generated using cholesterol as the dependent variable, and fiber in conjunction with different predetermined cut-offs in alcohol use which were used to define a categorical variable. This model generated similar slopes for the predicted values and did not exhibit characteristics of having an impact on our dependent variable. Our second model introduced interaction terms between the covariates to help explain individual's cholesterol levels. We noted that the interactions between alcohol and fiber did not appear to be statistically significant, although the heavy alcohol use group did display distinct behavior when examining cholesterol. The nested f-test of these two models did not suggest that the additional interaction terms in the more complex model added significantly to the unexplained variance in the data.

Lastly, we looked at three additional categorical variables: smoking, gender and vitamin usage and their interactions with the fiber variable. Of the three additional variables we considered, only the gender variable appeared to have a statistically significant impact on the individual's cholesterol levels.

The ANVOCA methods deployed in this lab were both informative and useful, and these techniques will undoubtedly come in useful in further research and analysis.

## BONUS

The first model we will explore is based upon the premise of using an individual's diet and sex to predict their cholesterol levels. For this model, we will use the following variables: calories, fat, fiber and gender. From the preceding work, we have a strong indication that sex will have a large influence on an individual's cholesterol.



The model summary can be seen below:

```
lm(formula = bonus1, data = data.nutrition)

Residuals:
    Min       1Q   Median       3Q      Max
-211.58  -49.72  -10.53   29.40   500.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.74904    16.16823   1.407   0.16043
Calories      0.04305     0.01741   2.473   0.01393
Fat           1.98241     0.31978   6.199 0.00000000181
Fiber        -1.26100     1.19748  -1.053   0.29314
GenderMale    179.02513    44.30498   4.041 0.00006730124
Fiber:GenderMale -10.11389     3.08801  -3.275   0.00118

Residual standard error: 89.56 on 309 degrees of freedom
Multiple R-squared:  0.5469, Adjusted R-squared:  0.5396
F-statistic: 74.6 on 5 and 309 DF, p-value: < 0.0000000000000022
```

Bonus Model 1:  $22.7490 + 0.043\beta_1 + 1.9824\beta_2 - 1.261\beta_3 + 179.0251\beta_4 - 10.1139\beta_5$

Where,  $\beta_1$  = calories,  $\beta_2$  = fat,  $\beta_3$  = fiber,  $\beta_4$  = male,  $\beta_5$  = Gender/Fiber interaction

This model can be interpreted as the intercept  $\beta_0$ , which is 22.7494, is the predicted value of cholesterol when all the coefficients are zero is such a low value compared to the actuals, with a mean of 242.4606, this value is a simple placeholder for the rest of the model.

For the first coefficient  $\beta_1$ , for every 1 unit increase in calories consumed, we see a .0431 increase in predicted cholesterol, indicating roughly more calories consumed is going to yield higher cholesterol values.

For the second coefficient  $\beta_2$ , we see that for every 1 unit in fat consumed, we see a 1.1924 increase in and individuals predicted cholesterol, indicating roughly that the heavier the diet is in fat the higher we would expect an individual's cholesterol to be.

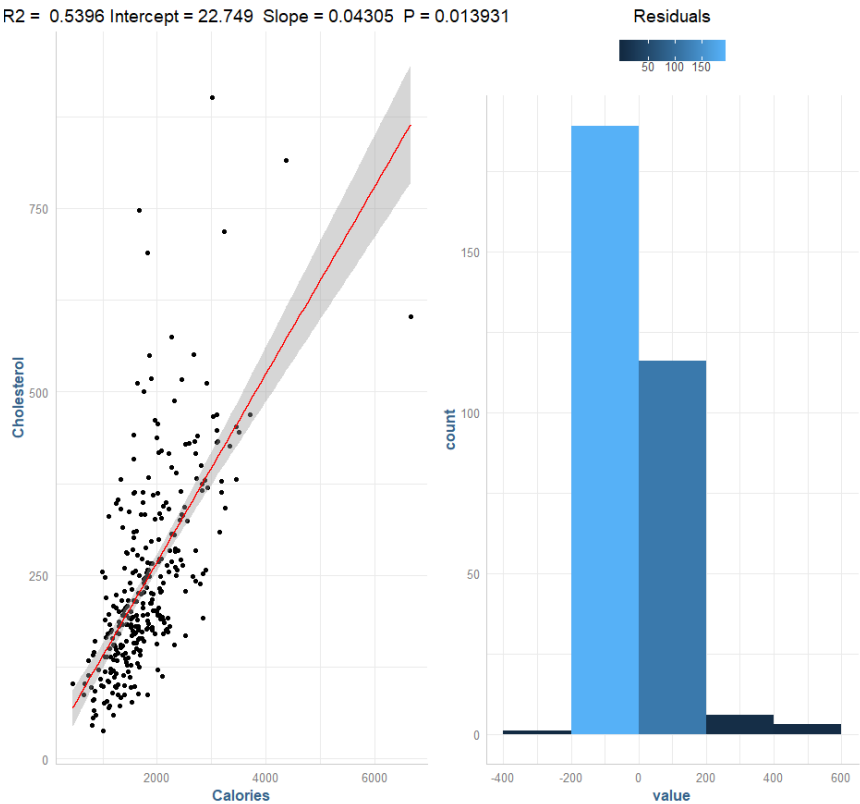
With the third coefficient,  $\beta_3$ , fiber we see that for every 1 unit of fiber increased, we would expect to see a 1.261 decrease in an individual's cholesterol level.

We also note the cholesterol levels vary greatly by gender, so the coefficient  $\beta_4$  represents an additional predicted increase in cholesterol levels for males, with an additional 179.025 cholesterol added into account for the variance in males.

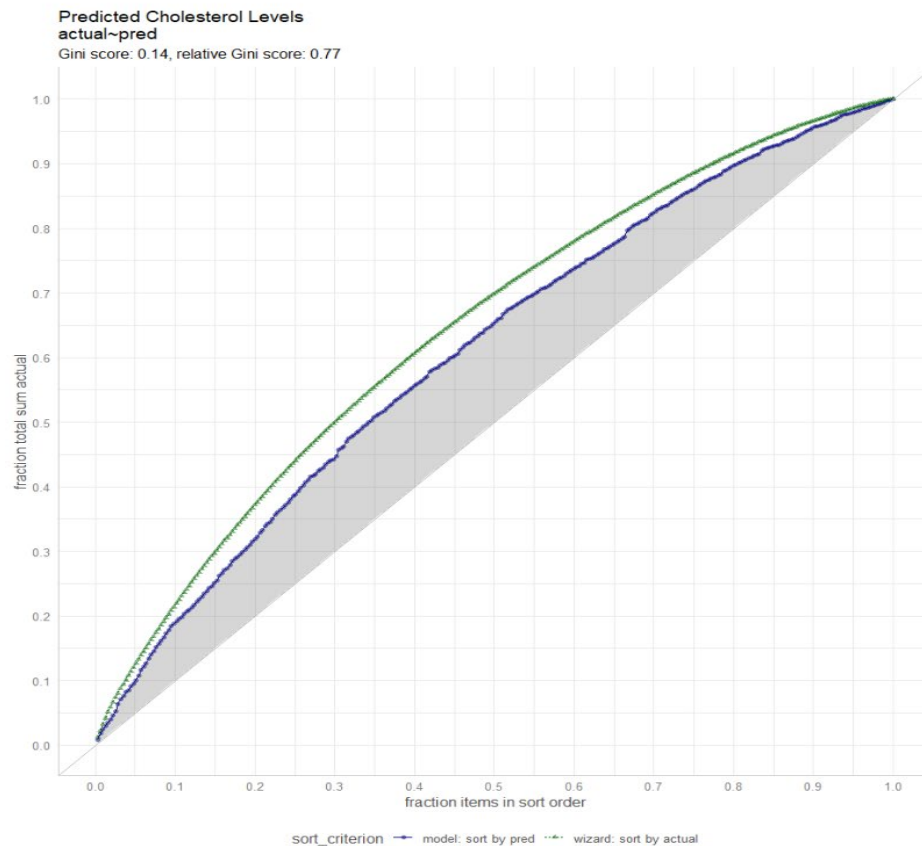
Finally, we have determined previously that there is a significant interaction between fiber usage in males and females, where it appears fiber has a higher impact on male cholesterol levels, represented by a 10.113 decrease per unit of fiber in males.

We note the model has a  $R^2$  value of 0.5469, denoting that approximately 55% of the variance in the data is explained by the model and a Gini score of .77, denoting a relatively high prediction accuracy. We can also look at a gain curve plot, which shows a green curve of a model with 100% prediction accuracy plotted against the area of that curve that our model accounts for with the blue line, of predicted values vs the actual values:

Standard model fit, residuals and gain curve can be seen below:



The gain curve plot of predicted vs actuals for the above model can be seen below.



The null hypothesis in this case would be,

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_a : \beta_j \neq 0, \text{ for at least one value of } j \text{ (for } j \text{ in } 1, 2, 3, 4, 5)$$

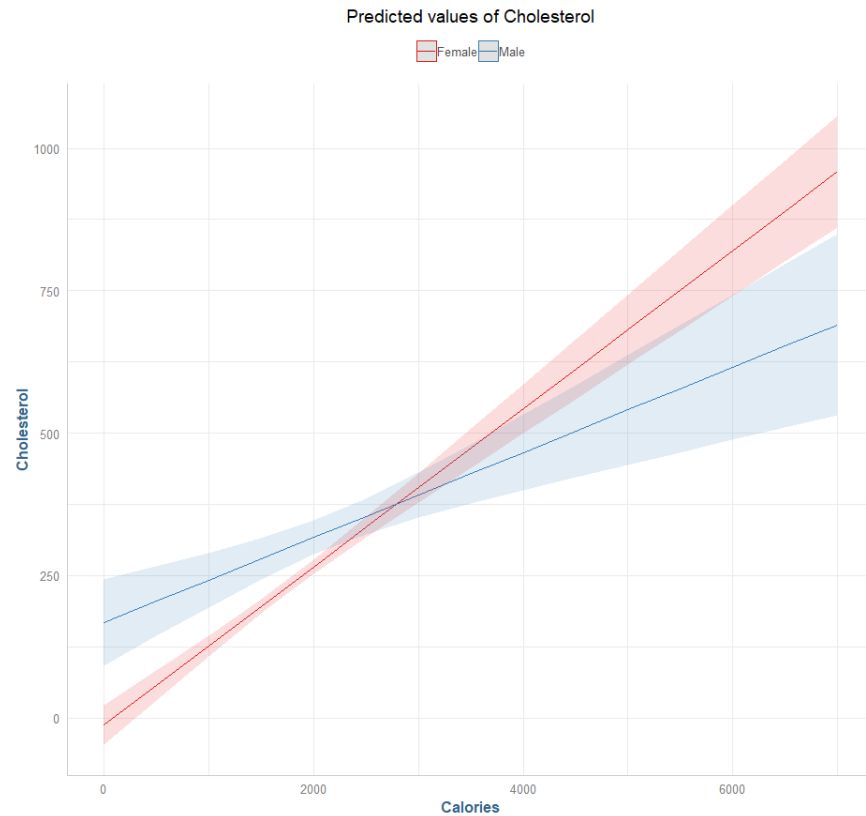
Or, simply that at least one of the interaction coefficients in the model is not zero, and that they would help explain the variance in the data greater than the intercept alone. The F-statistic of 91.4 denotes full model significance.

The overall p-value for the model denotes significance, as do the four out of the five coefficient terms, indicating that this model helps explain more variance than the intercept alone, and that we can therefore reject the null hypothesis.

For the next model, we will do the preliminary analysis of the interactions between gender and two of our previous model terms, fat and calories. The intuition here is that males typically have higher caloric intake than females, suggested by several independent studies.



The interaction between an individual's calorie consumption by gender, and the interaction between the two terms can be seen in the following diagram:



```
Call:
lm(formula = calories_gender, data = data.nutrition)

Residuals:
    Min       1Q   Median       3Q      Max
-284.22  -58.80  -13.78   32.87  494.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -12.381236   17.467758  -0.709  0.478976
Calories        0.138775    0.009451  14.683 < 2e-16 ***
GenderMale     179.429766   42.296214   4.242 2.92e-05 ***
Calories:GenderMale -0.064057    0.018993  -3.373 0.000838 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.68 on 311 degrees of freedom
Multiple R-squared:  0.4686, Adjusted R-squared:  0.4634
F-statistic: 91.4 on 3 and 311 DF, p-value: < 2.2e-16
```

We note the model 7, here as:  $-12.3814 + 0.1388\beta_1 + 179.43\beta_2 - 0.06\beta_3$

Where the intercept being negative is an impossible value to achieve with zero coefficient terms, it is obviously just used as a placeholder in this model. The  $\beta_1$  coefficient denotes a predicted .14 increase in cholesterol per additional calorie consumed,  $\beta_2$  denotes a 179.43 increase in cholesterol levels for male subjects, and the interaction term  $\beta_3$  denotes a reduction in 0.06 units of cholesterol for every increased unit in calories consumed for males, suggesting that there is less of an impact on male calorie consumption compared to females.

The null hypothesis for this model would be,

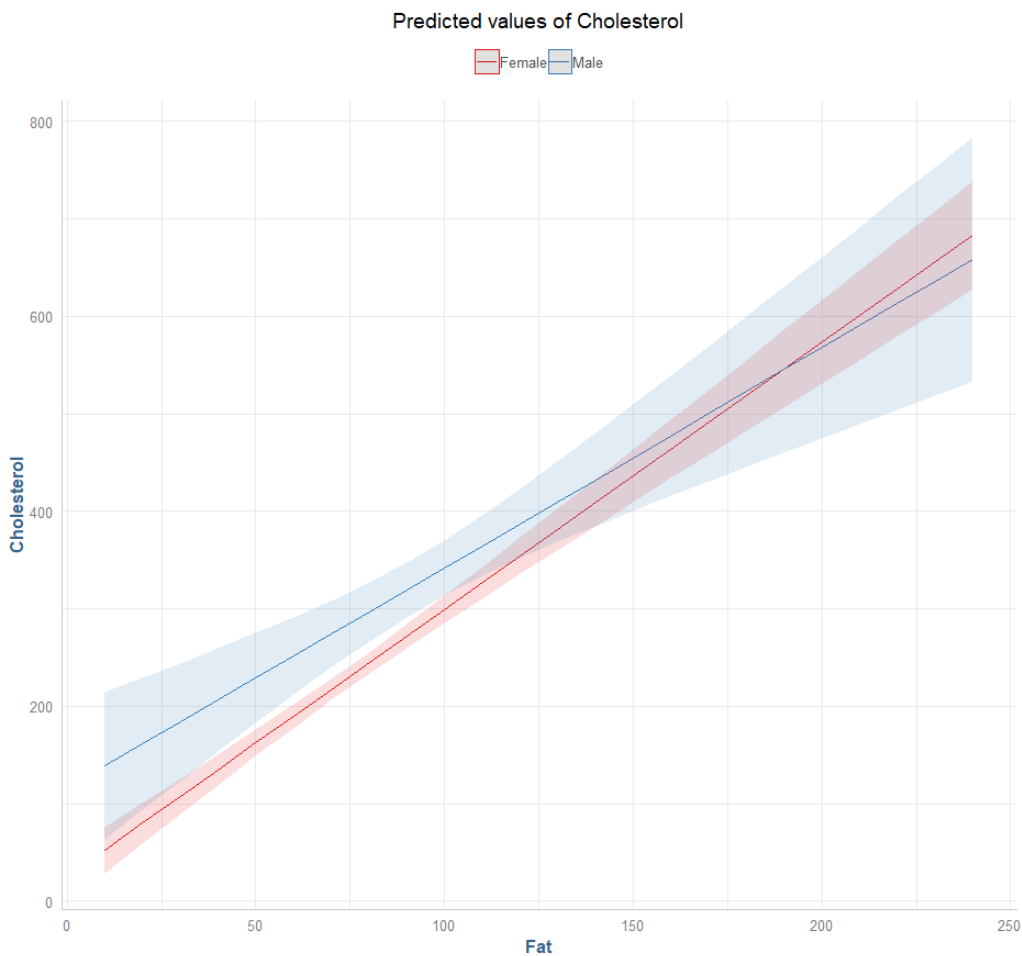
$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_a: \beta_j \neq 0, \text{ for at least one value of } j \text{ (for } j \text{ in } 0, 1, 2, 3)$$

Or the explained variance in the data represented by the model is better than the simple mean of the data, given that the intercept term in this model is not a useful metric in explaining cholesterol levels for individuals without the additional coefficient terms.

Given the p-value of the overall model, the explained variance of about 46.86%, and the significance of all three of the coefficient terms, we can reject the null hypothesis that this model does not help explain variance in the data.

The next interaction we will examine is between fat consumed in males vs females. We can see the visual interaction in the following diagram:



We note the slopes of the two lines seem to be running not parallel but do have a convergence pattern in the overall diagram.

The model diagnostics confirm our visual observations in that there is statistical significance in the two previously examined terms, fat and gender, however the interaction between these two terms does not appear to be significant with a p-value of 30%.

```
Call:
lm(formula = fat_gender, data = data.nutrition)

Residuals:
    Min       1Q   Median       3Q      Max
-208.26  -49.71   -9.76    34.62   519.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.1472    13.6927   1.837   0.0672 .
Fat          2.7423     0.1681  16.316 <2e-16 ***
GenderMale   90.7850    44.6411   2.034   0.0428 *
Fat:GenderMale -0.4823     0.4585  -1.052   0.2937
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.93 on 311 degrees of freedom
Multiple R-squared:  0.5196, Adjusted R-squared:  0.515
F-statistic: 112.1 on 3 and 311 DF, p-value: < 2.2e-16
```

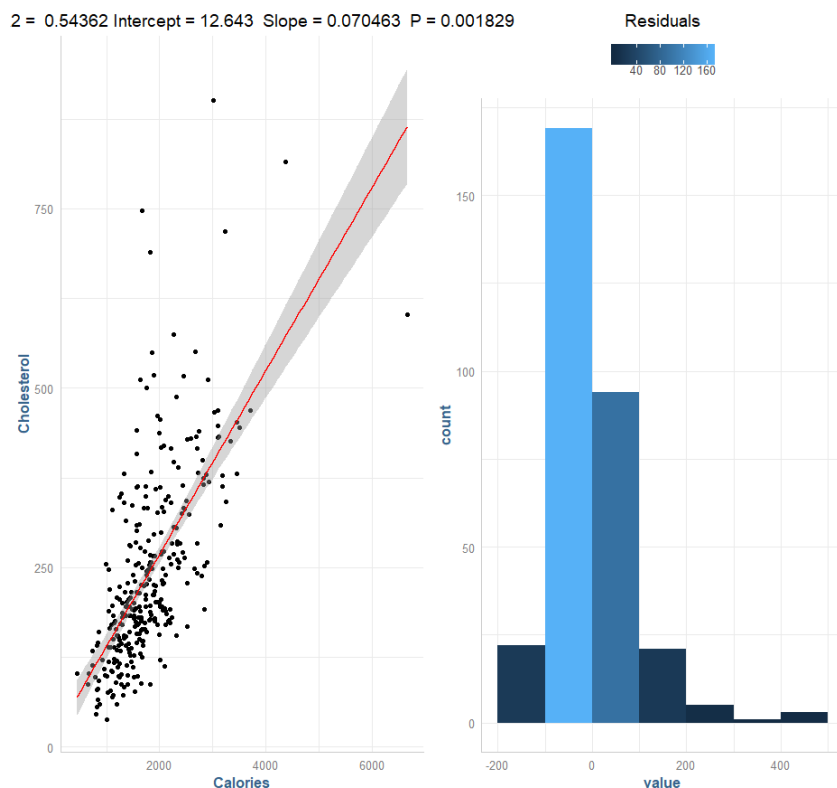
The null hypothesis in this case would be

$$H_0: \beta_3 = 0$$

$$H_a: \beta_3 \neq 0$$

Where  $\beta_3$  represents the interaction between fat and gender. Given the high p-value of this term, it does not appear to have any significance on the predicted cholesterol levels and therefore we cannot reject the null hypothesis. Additionally, we will exclude this interaction term from future models.

The next model is a superset of the previous model, adding in the additional interaction term between gender and calories.



```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.64264    16.92855   0.747  0.45574
Calories         0.07046     0.02241   3.144  0.00183 **
Fat             1.67287     0.35653   4.692 4.07e-06 ***
Fiber          -2.41047     1.33285  -1.809  0.07150 .
GenderMale     242.96490    55.17584   4.403 1.47e-05 ***
Fiber:GenderMale -8.79753     3.14932  -2.793  0.00554 **
Calories:GenderMale -0.03994     0.02071  -1.929  0.05464 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.17 on 308 degrees of freedom
Multiple R-squared:  0.5523, Adjusted R-squared:  0.5436
F-statistic: 63.34 on 6 and 308 DF, p-value: < 2.2e-16

```

$$\text{Model: } 12.6426 + 0.0704\beta_1 + 1.6728\beta_2 - 2.41\beta_3 + 242.96\beta_4 - 8.798\beta_5 - 0.04\beta_6$$

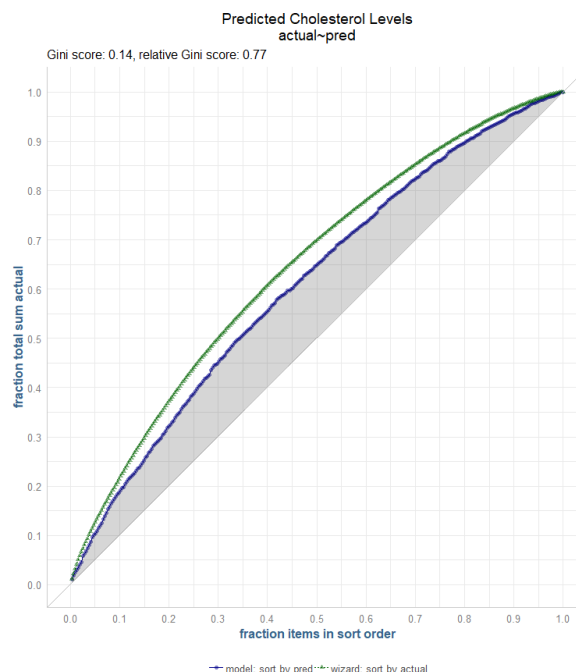
We can summarize the model by noting that the intercept term we see here is so far below any reasonable value of cholesterol level that it must be a simple placeholder value in the model. The beta1 coefficient term here denotes that for every unit of caloric intake, we see an expected increase in 0.07 cholesterol level. The beta2 coefficient denotes that for every one unit of fat we can expect to see a 1.672 increase in cholesterol, and the beta3 coefficient denotes an expected decrease of 2.4 units of cholesterol per unit increase in fiber in an individuals diet. The interaction terms denote that males will have approximately 8.79 and 0.04 units decrease in cholesterol levels per unit of fiber and calories respectively.

The null hypothesis for the model would be,

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

$$H_a: \beta_j \neq 0, \text{ for at least one value of } j \text{ (for } j \text{ in } 0, 1, 2, 3, 4, 5, 6)$$

Given the overall p-value denotes statistical significance in the model, as well as four of the 6 terms denote statistical significance in explaining variance in the observed data, as well as a  $R^2$  value of 55.23%, we can reject the null hypothesis that the model does not explain further variance in the data. The F-statistic of 63.24 denotes the significance of the full model with all parameters.



The gain curve plot shows a Gini score of .14 and a relative score of .77, the same as our previous model without the additional interaction term.

Next, we will conduct a nested f-test given that bonus model 1 is nested inside of bonus model 2:

The null hypothesis for the model would be,

$$H_0 : \beta_6 = 0$$

$$H_a : \beta_6 \neq 0$$

$$F = [(SSE_R - SSE_C) / (df_2 - df_1)] / (SSE_C / df_1)$$

$$F = ((2,478,476 - 2,44,887) / 1) / [2,44,887 / 308]$$

$$= 29,588.9 / 7,950.93$$

$$= \mathbf{3.7214}$$

Which is less than our critical value at 95% confidence,  $= F_{95, 308} = \mathbf{3.8718}$ , therefore we cannot reject the null hypothesis that the caloric interaction between gender does not explain further variance in the data.

