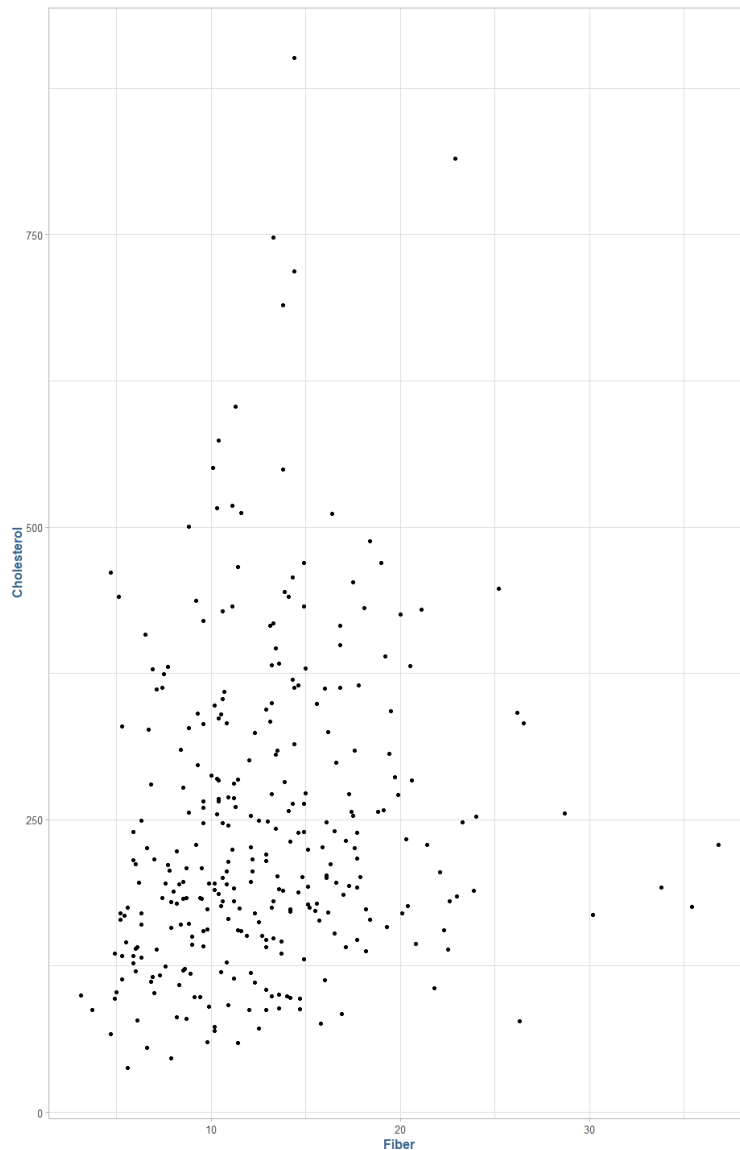


## INTRODUCTION

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

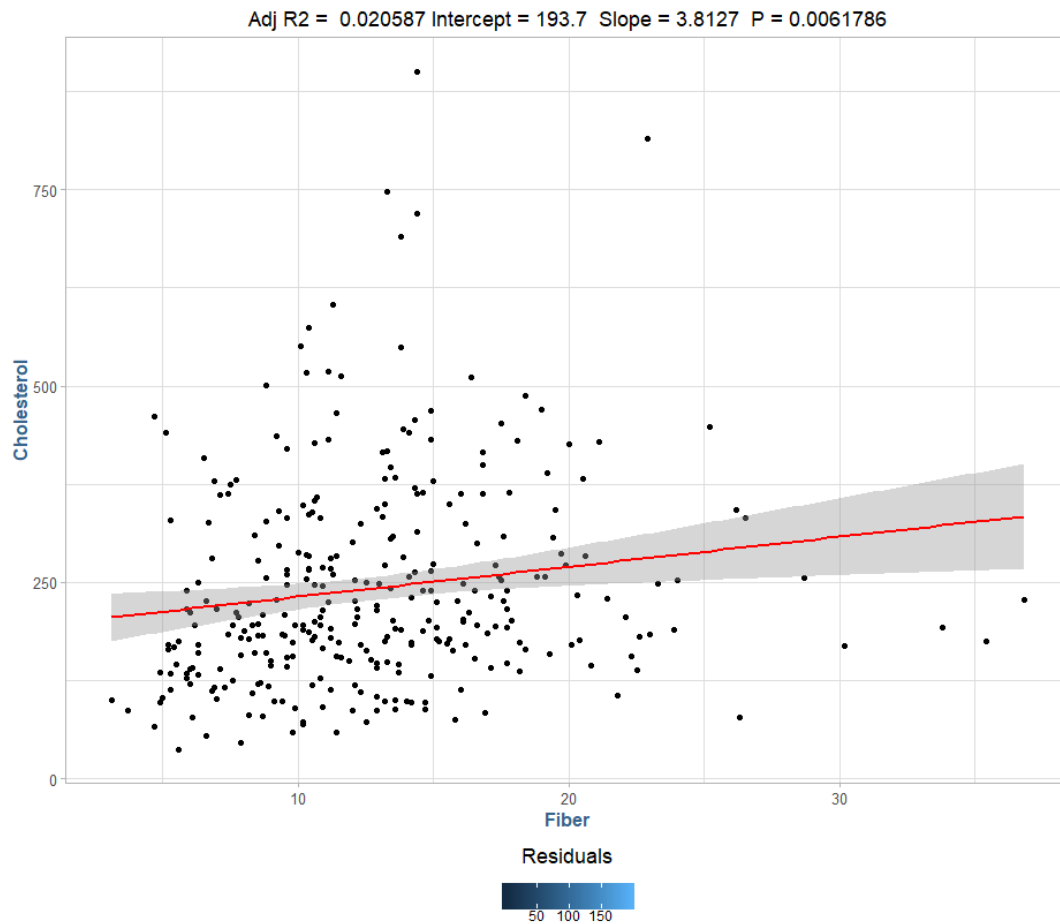
In the following diagram we see a lot of variance in the following scatterplot of Cholesterol ~ Fiber.



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.

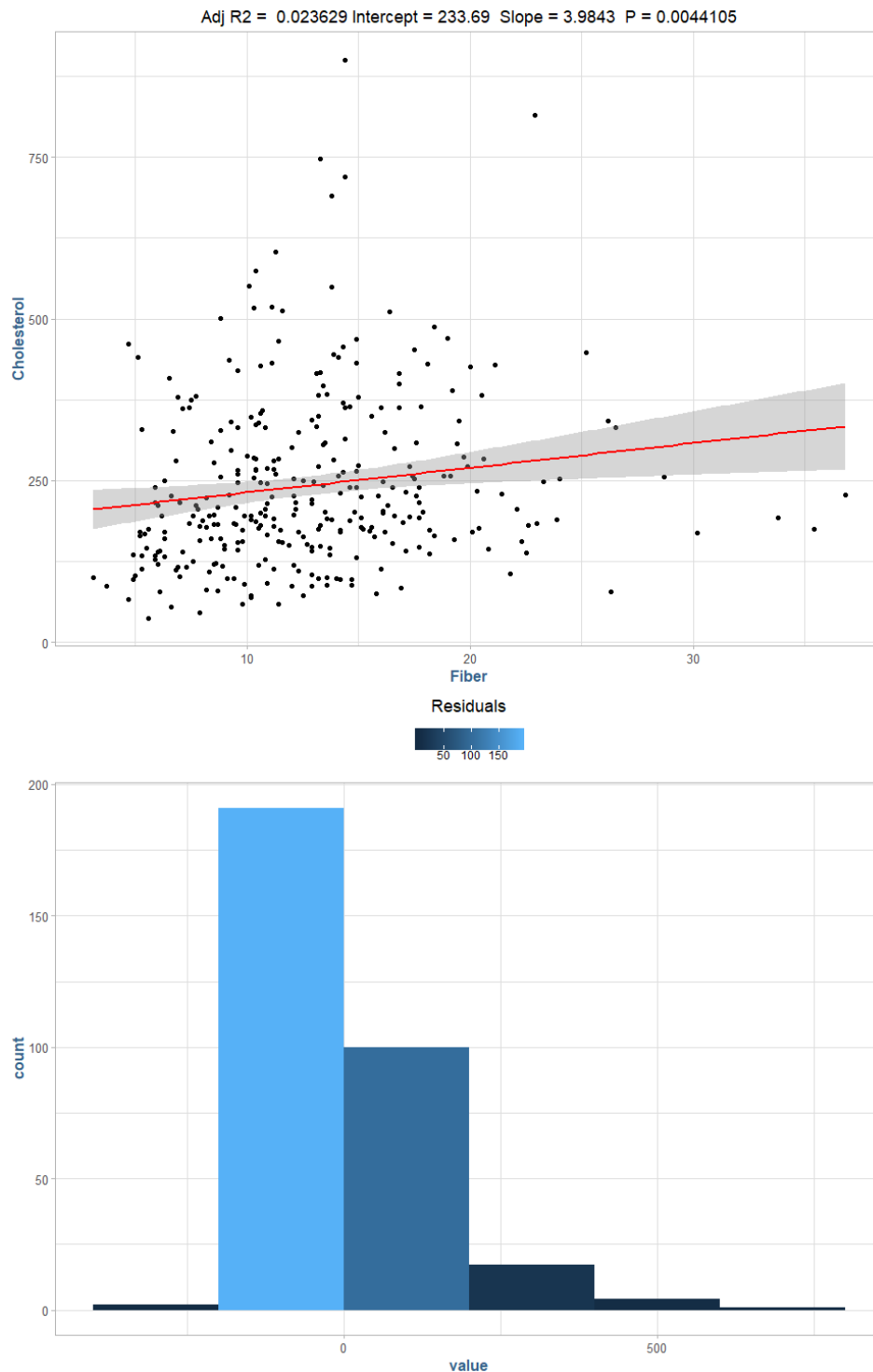
The Fiber variable appears to be a relatively poor indicator of cholesterol. In the graphic below, we can see an  $R^2$  value of 0.0237, indicating that approximately 2% of the variance in the data is explained by the Fiber variable.



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).

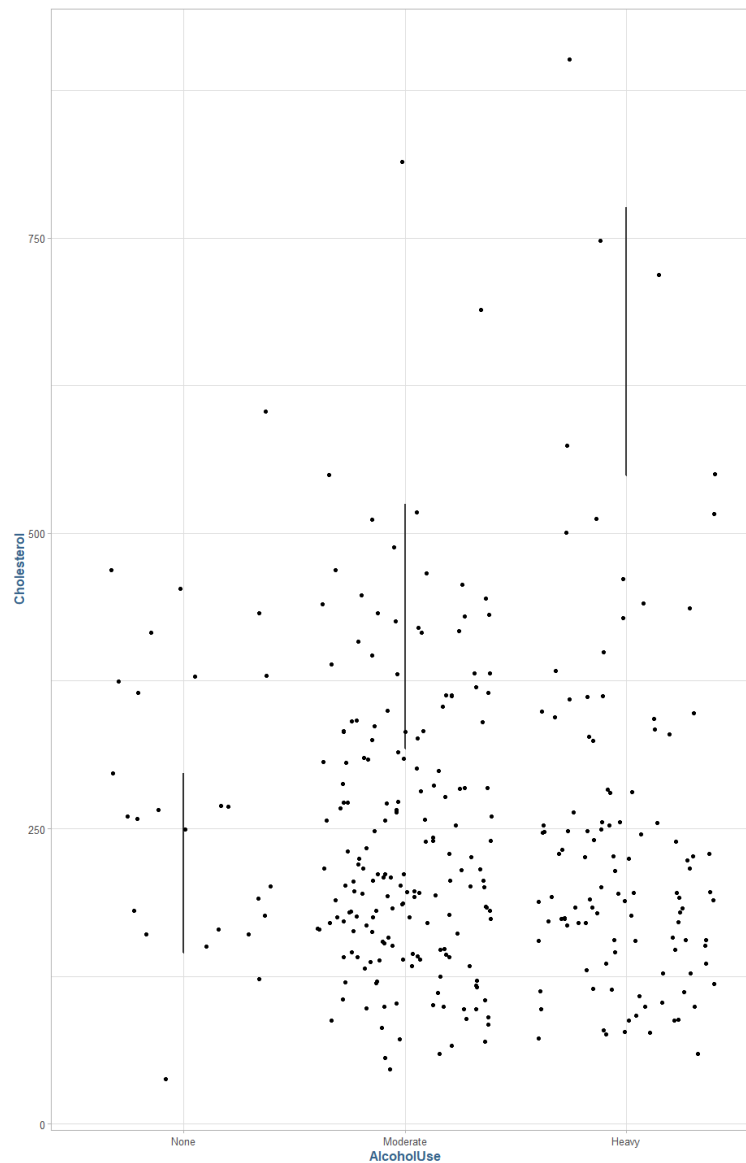


$$\text{Model 2: } 233.6946 + 3.9543\beta_1 - 46.9518\beta_2 - 44.4289\beta_3$$

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,

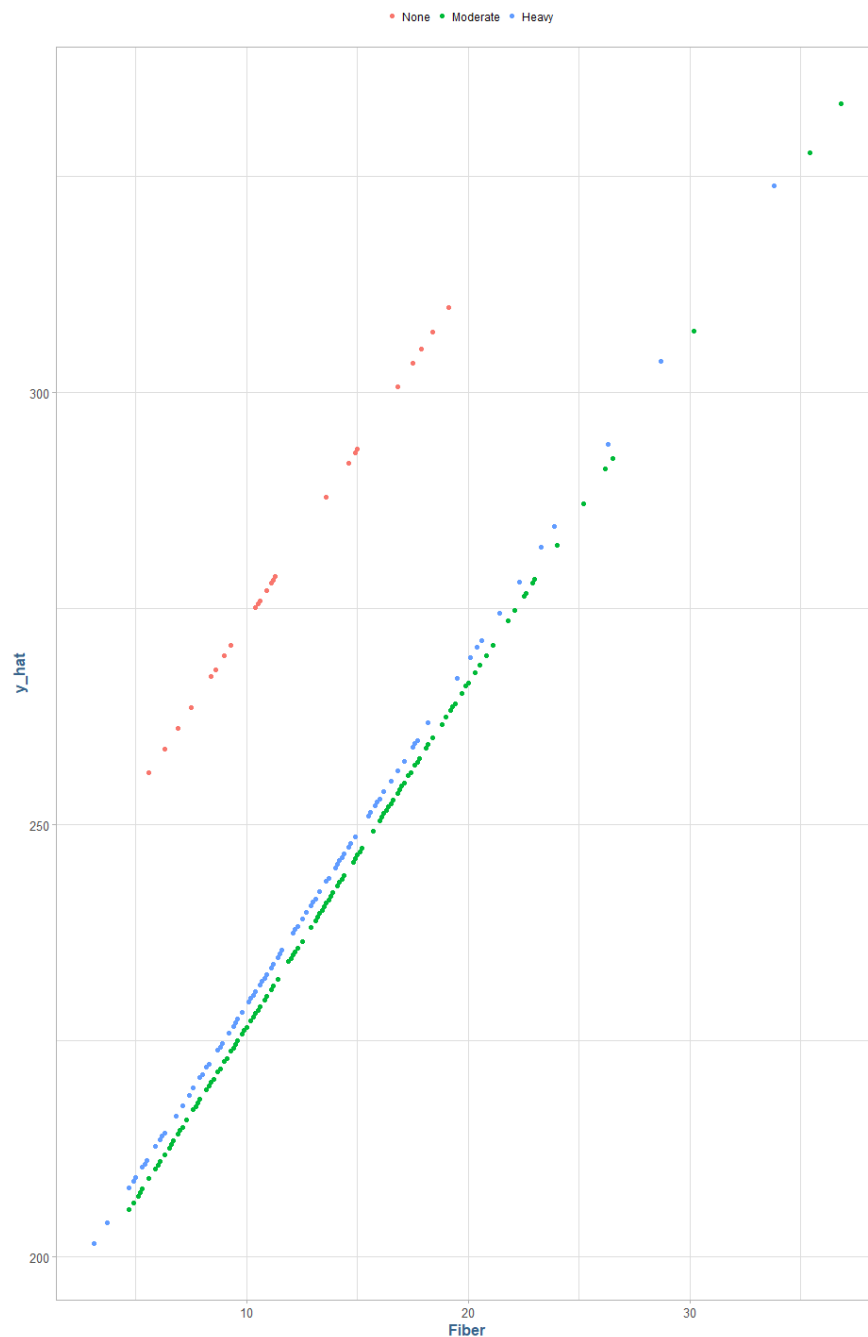
$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

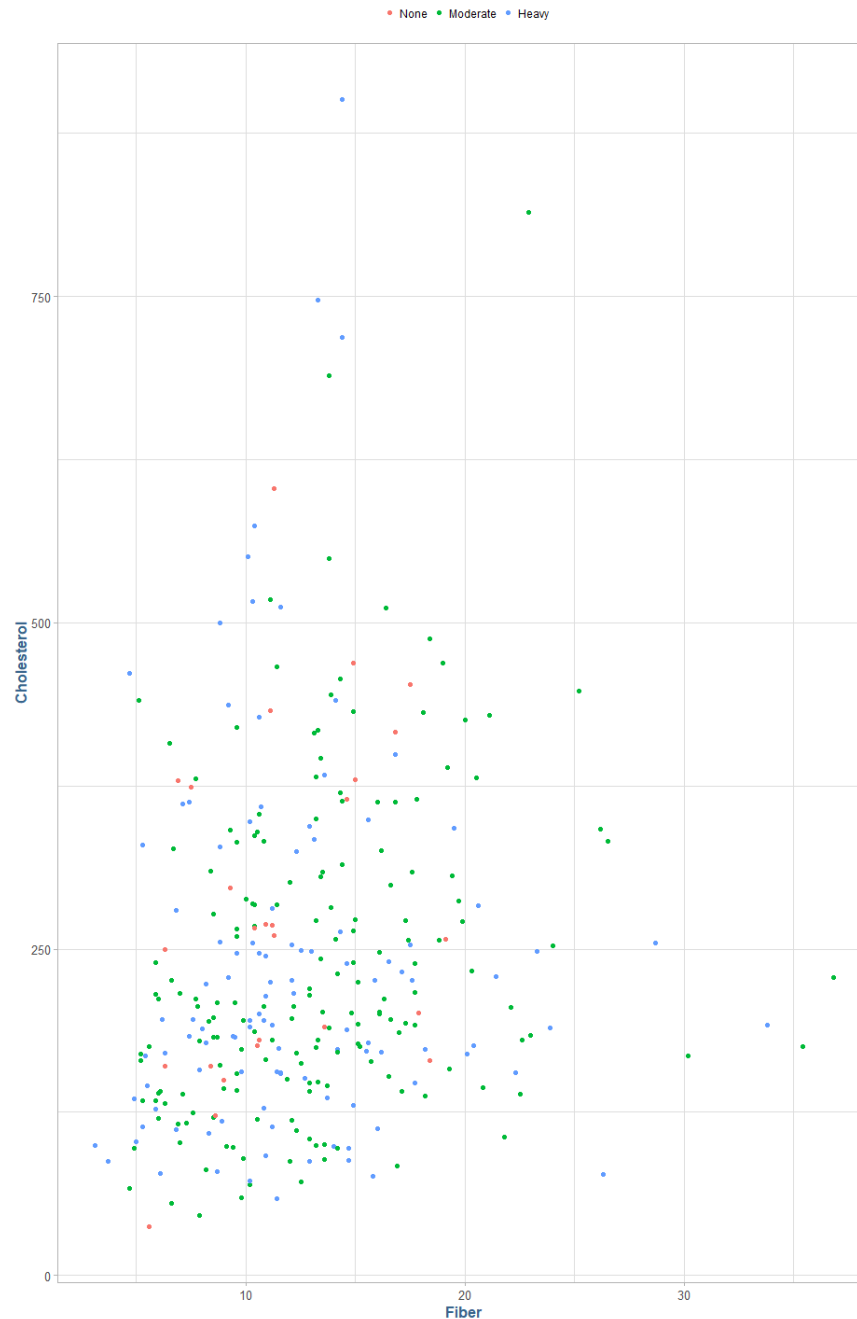
Or that there is no effect on the model using the beta coefficient derived from fiber consumption and the coded variables for alcohol consumption, against the alternative hypothesis that:

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

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.

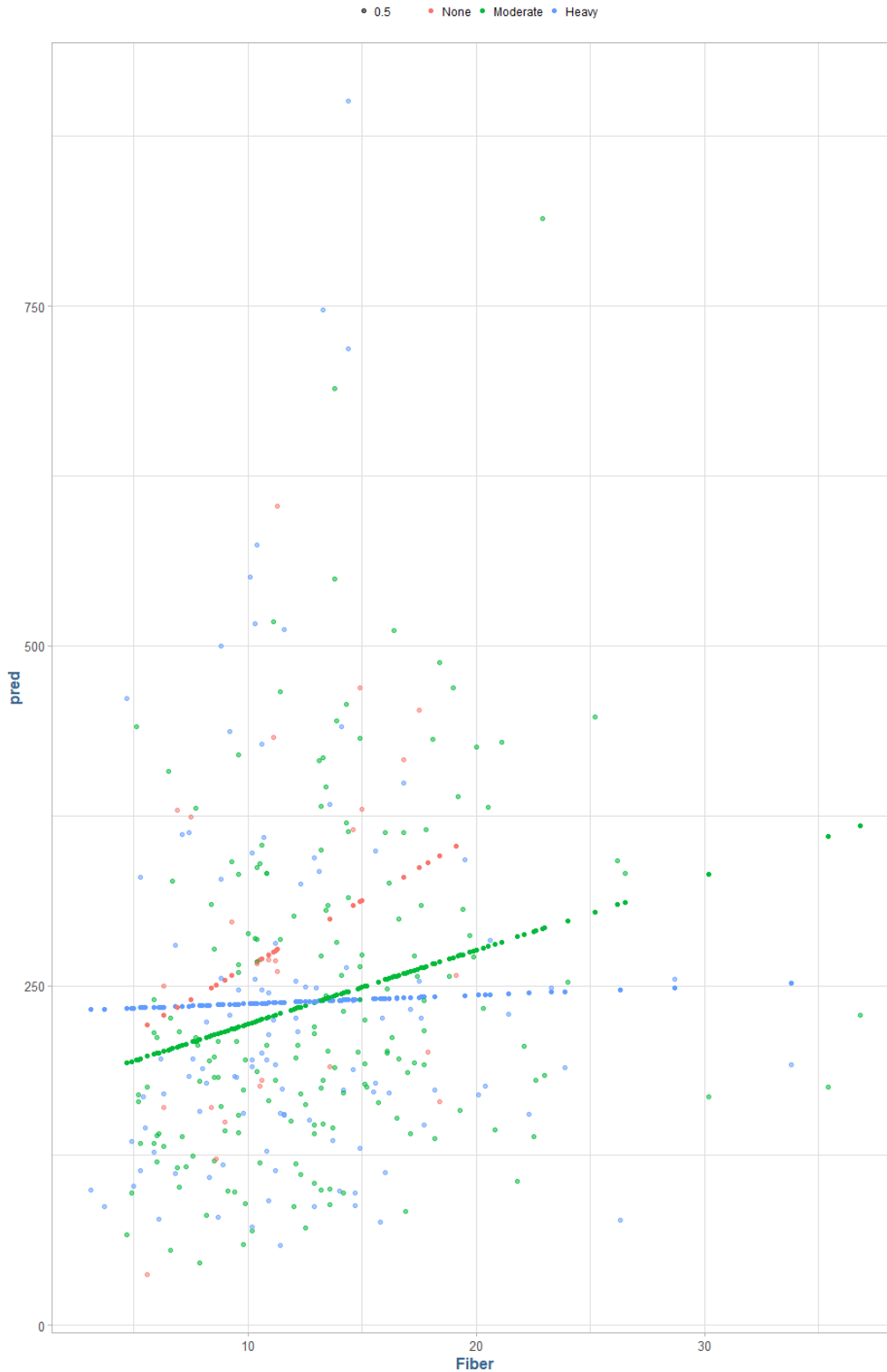
- 4.) Use the ANCOVA model from task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FIBER (X), but color code the records for the different groups of ALCOHOL. What do you notice about the patterns in the predicted values of Y? Now, 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.

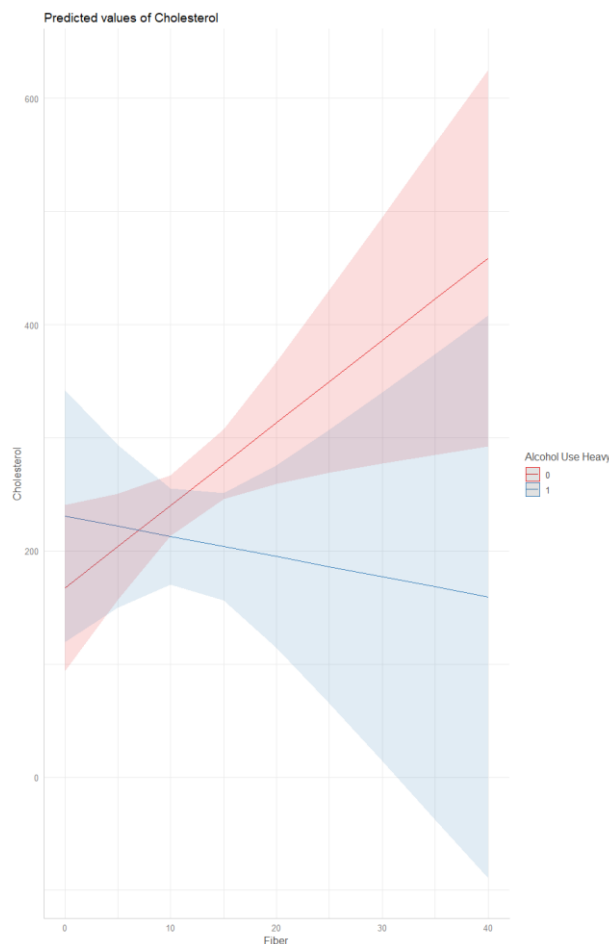
```
Call:
lm(formula = Cholesterol ~ Fiber + AlcoholUseModerate + AlcoholUseHeavy +
    Fiber * AlcoholUseModerate + Fiber * AlcoholUseHeavy, data = model3_data)

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**

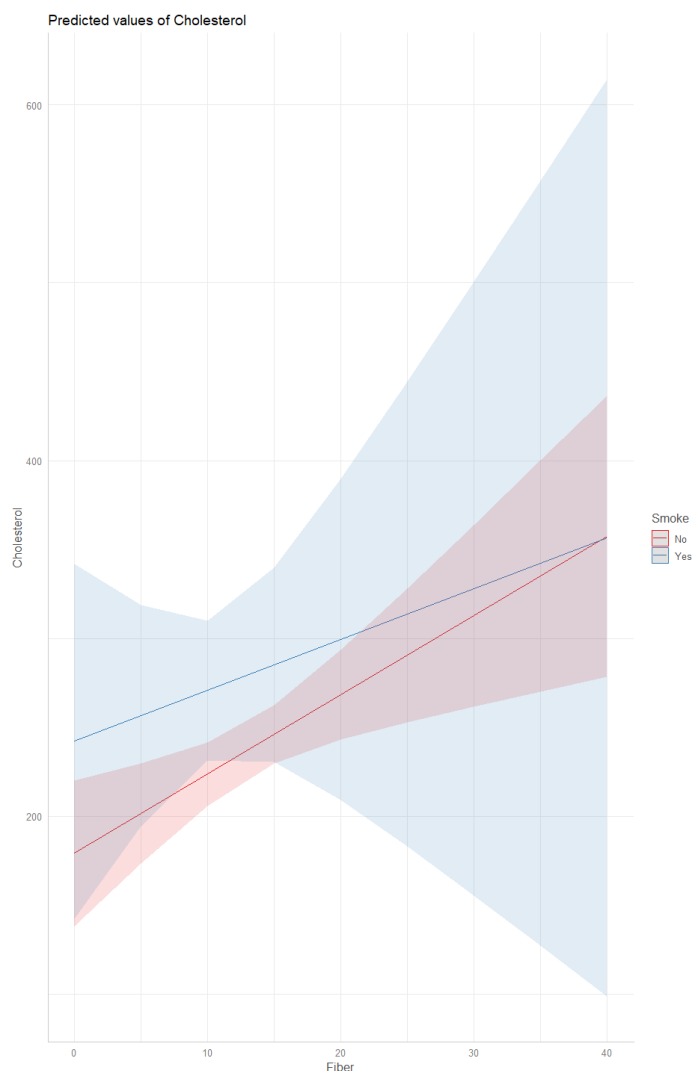
```
Analysis of Variance Table

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
```

First, let's examine the relationship between cholesterol levels and fiber/smoking.

9

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)

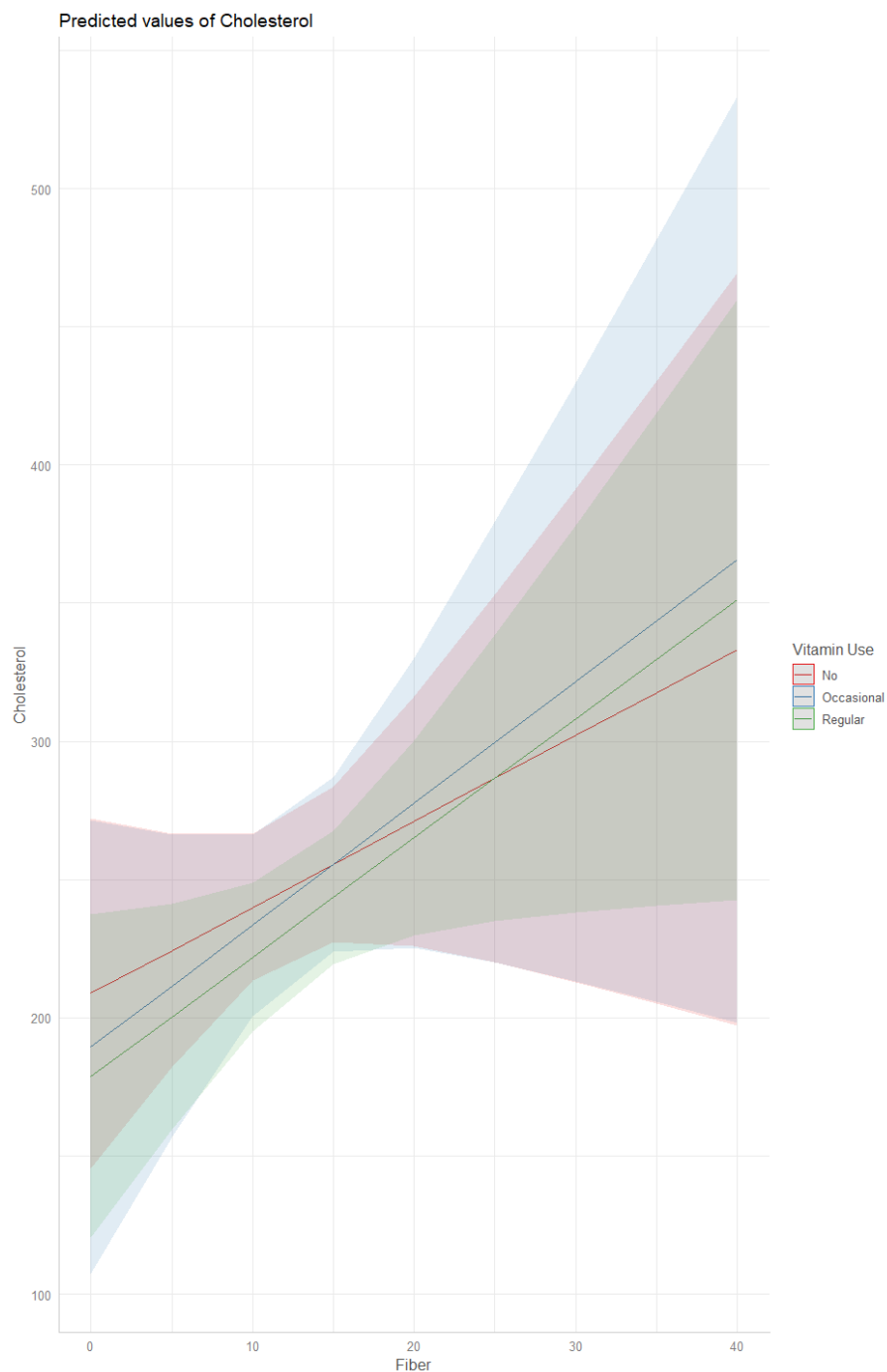
Residuals:
    Min       1Q   Median       3Q      Max
-214.64  -91.71  -33.55   63.36  659.80

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.

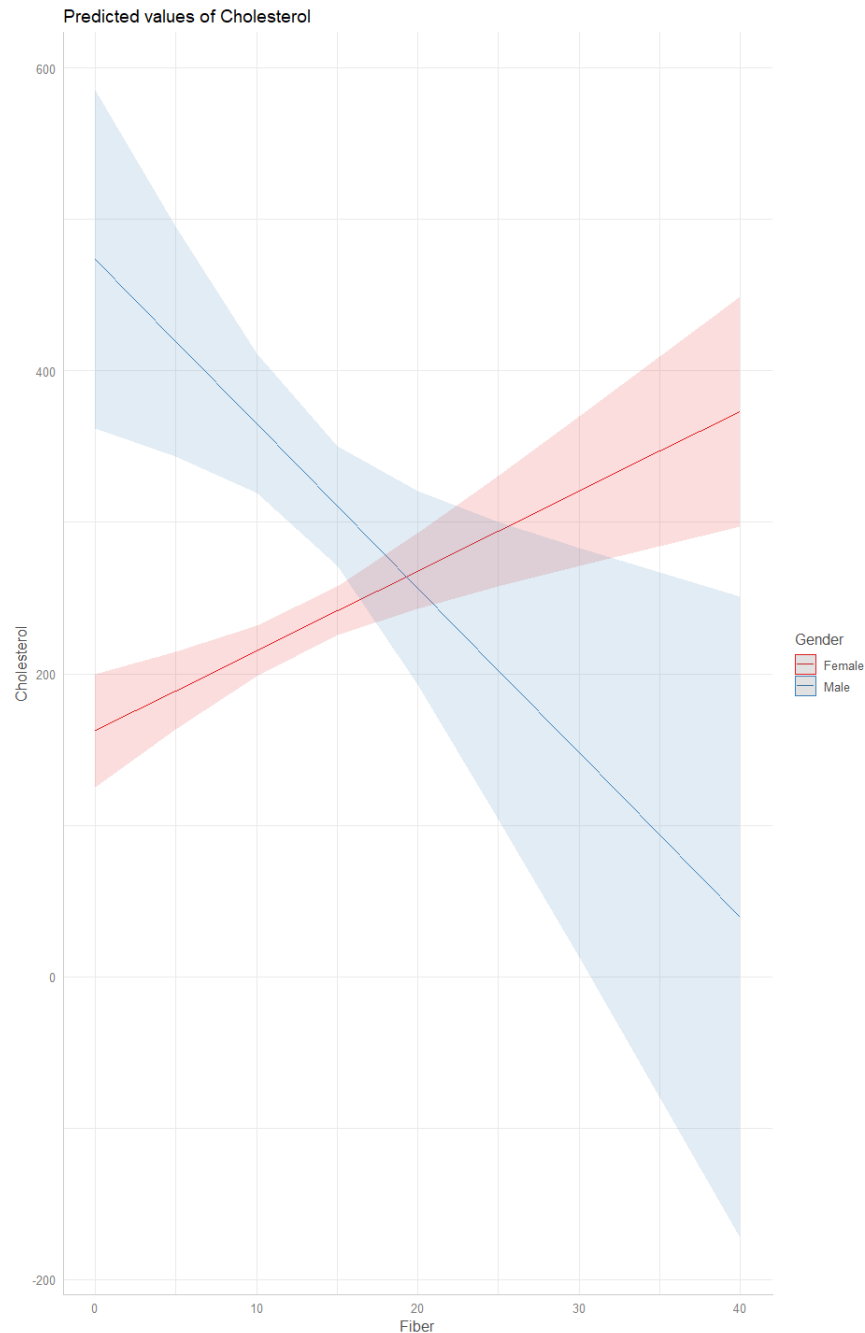


We can also confirm with the results of a type II anova test:

```
> 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.

---

## RESEARCH