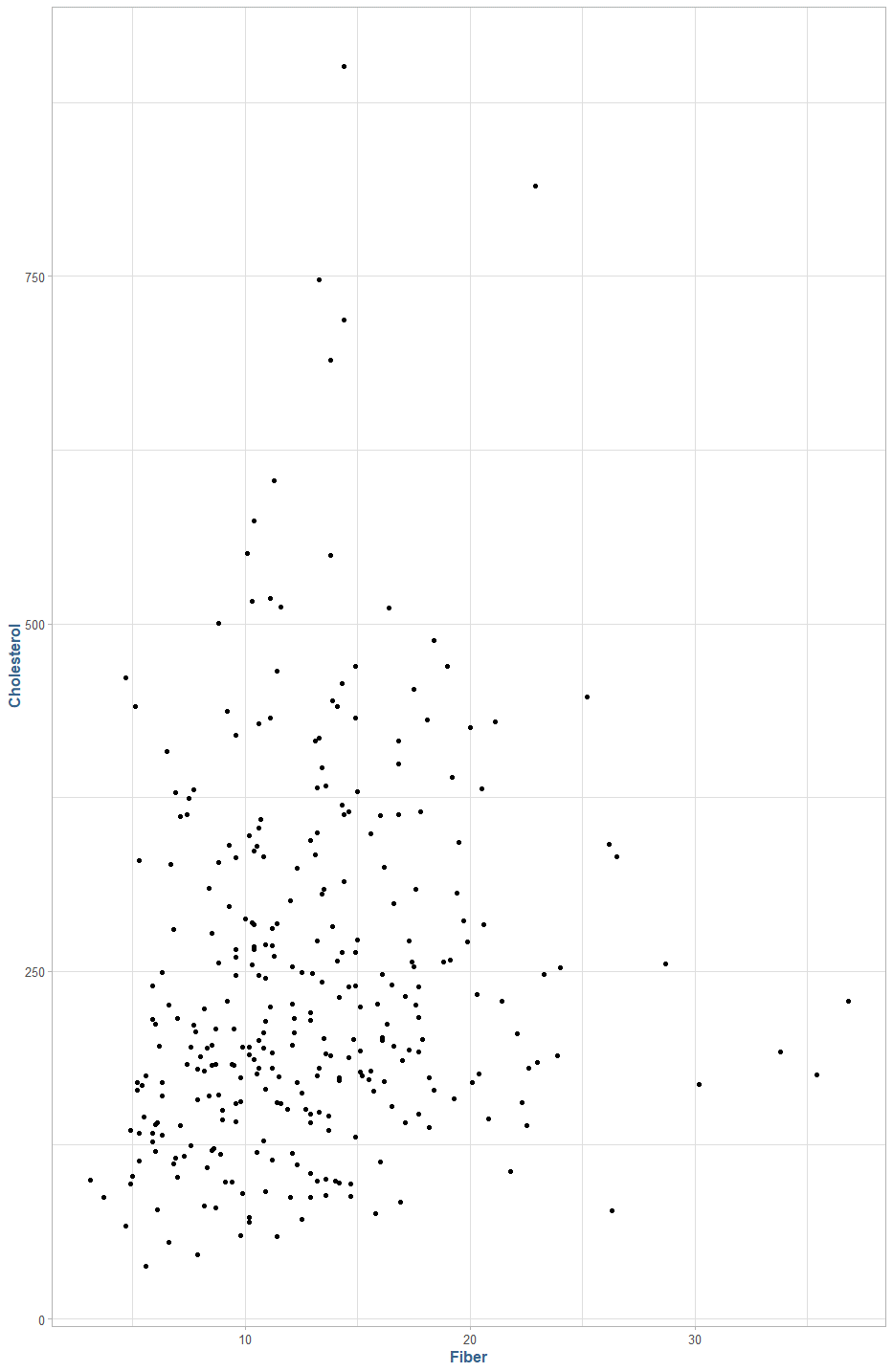
Computational Assigment #4

Brandon Moretz

### Introduction

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

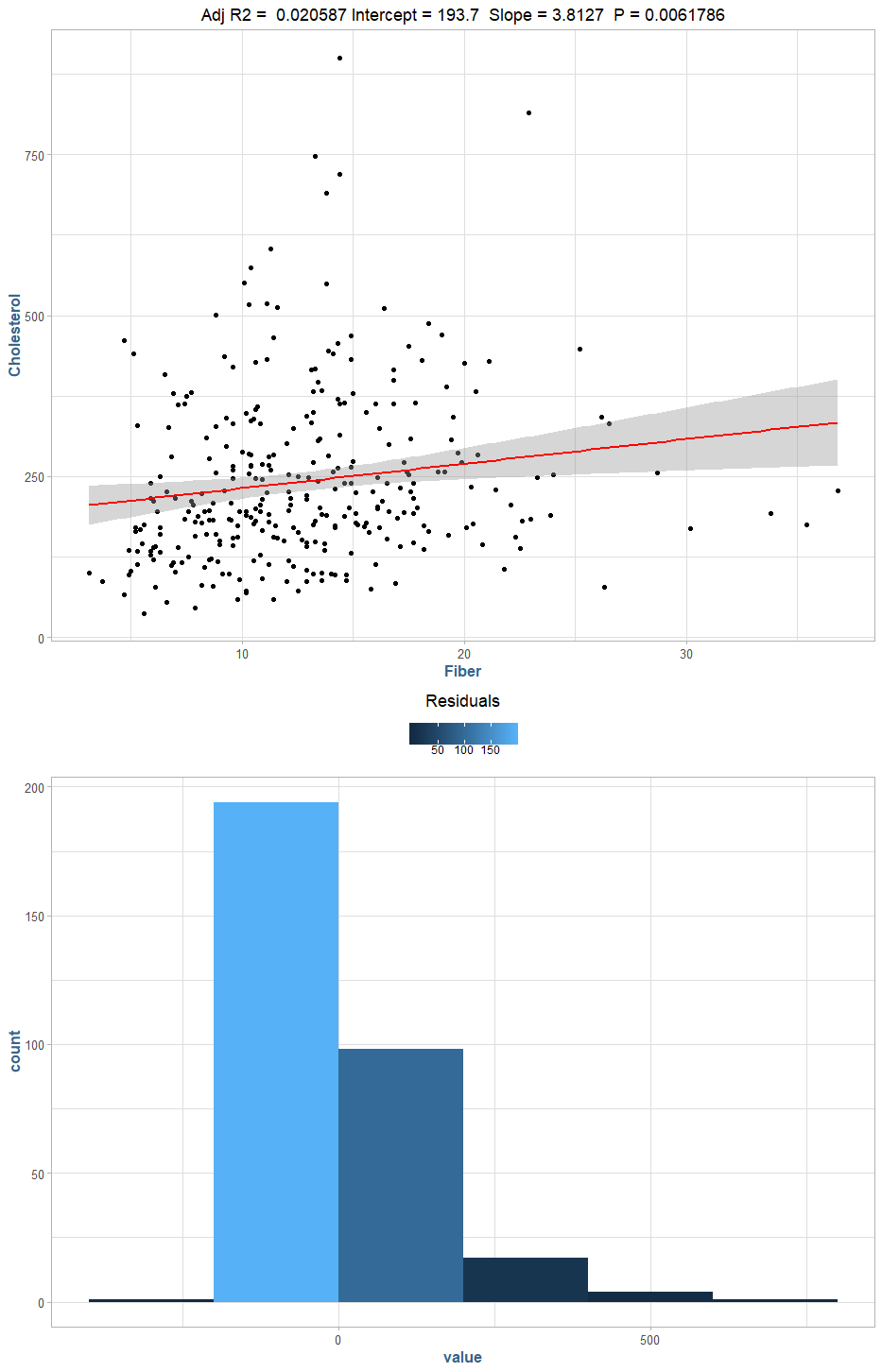
In the following diagram we see a lot of variance in the following scatterplot of Cholesterol ~ Fiber.



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

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

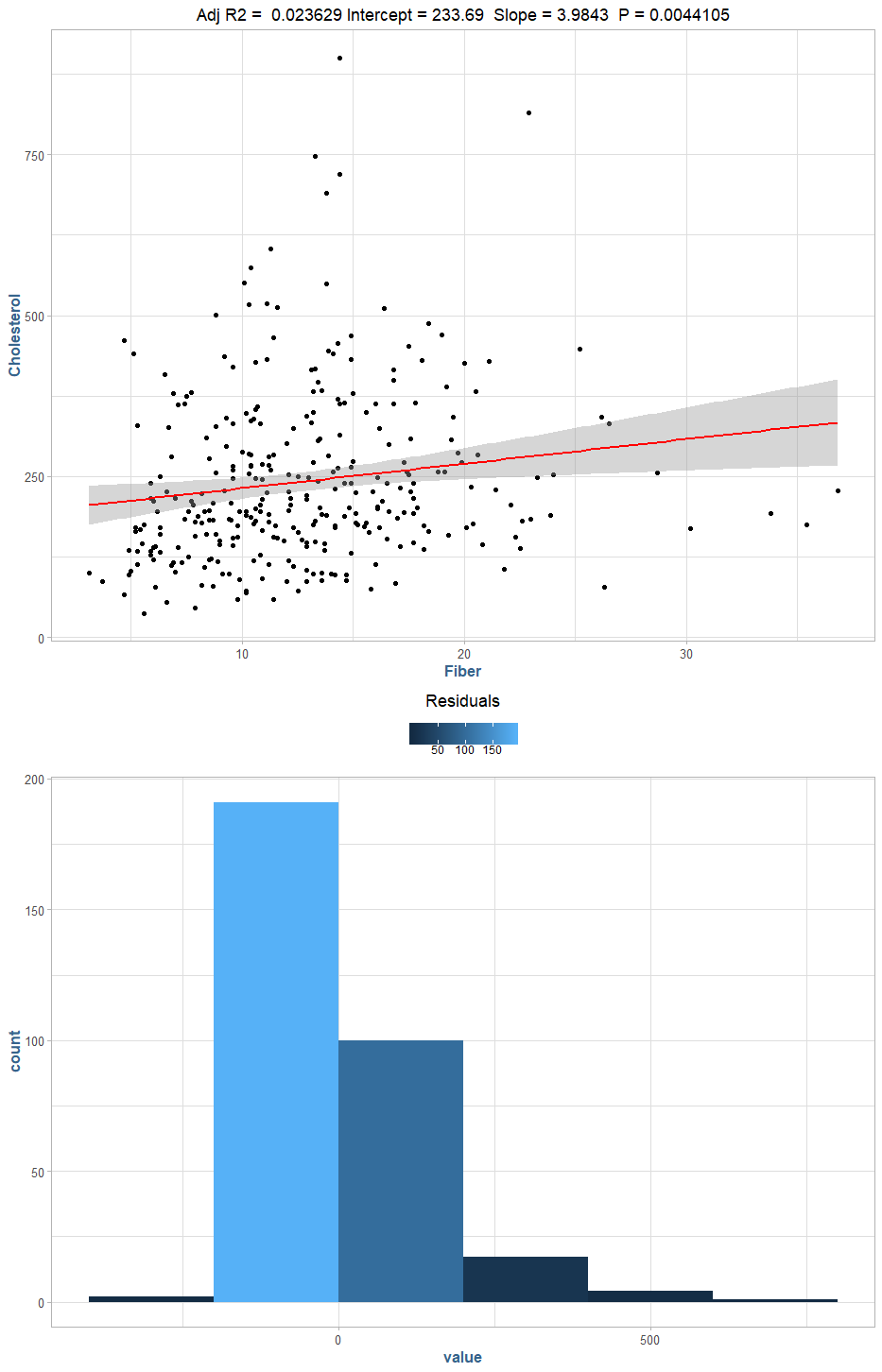
The Fiber variable appears to be a relatively poor indicator of cholesterol. In the graphic below, we can see an R2 value of 0.0237, indicating that approximately 2% of the variance in the data is explained by the Fiber variable.



Model: 193.7014 + 3.8127 β1, where β1 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).*

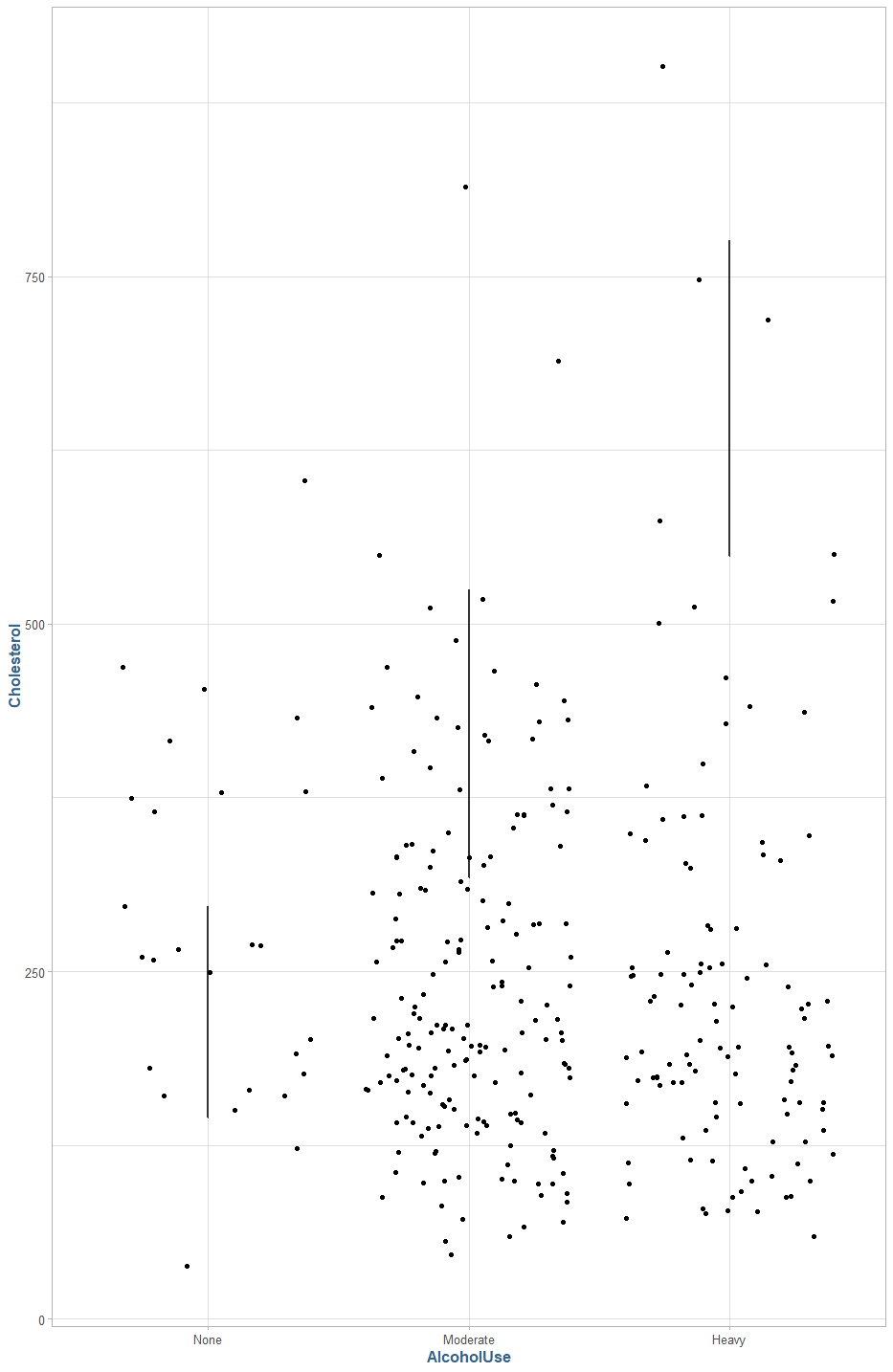


Model 2: 233.6946 + 3.9543β1 – 46.9518 β2 – 44.4289β3

Where β1 = Fiber, β2 = Alcohol (Moderate), β3 =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 R2 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,

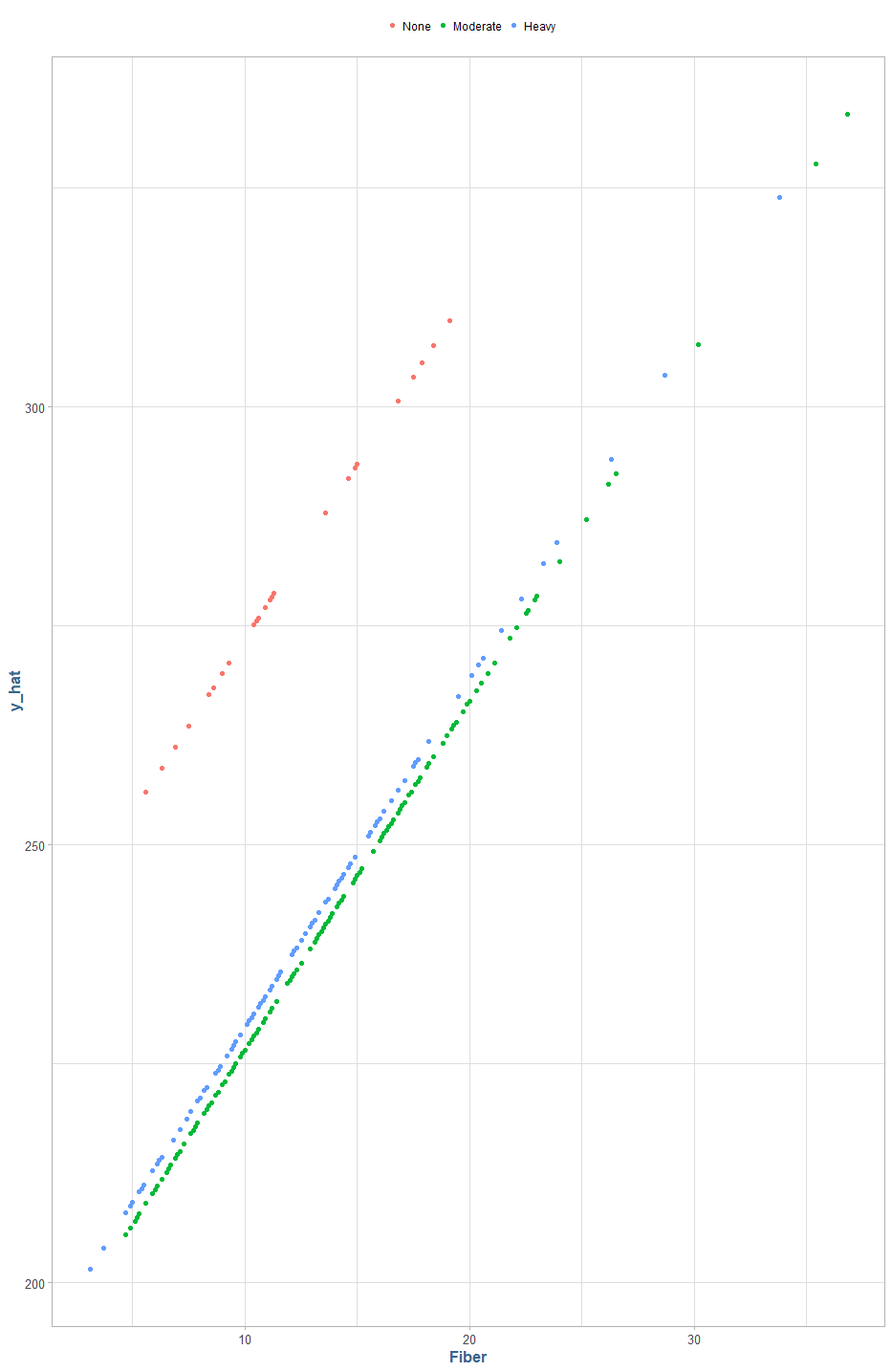
H0 : β1 = β2 = β3 = 0

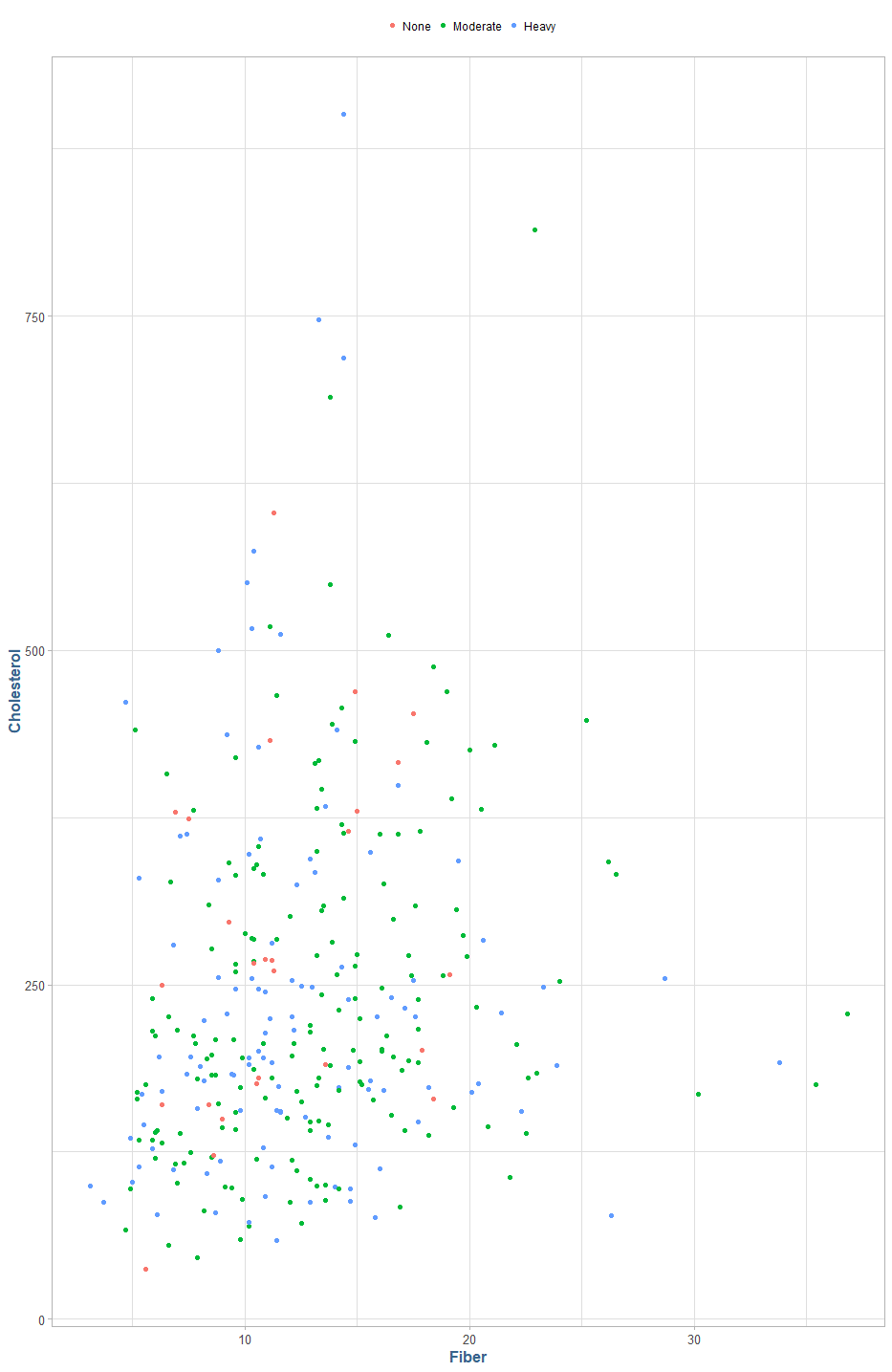
Or that there is no effect on the model using the beta coefficient derived from fiber consumption and the coded variables for alcohol consumption, against the alternative hypothesis that:

Ha : β1 = β2 = β3 ≠ 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.

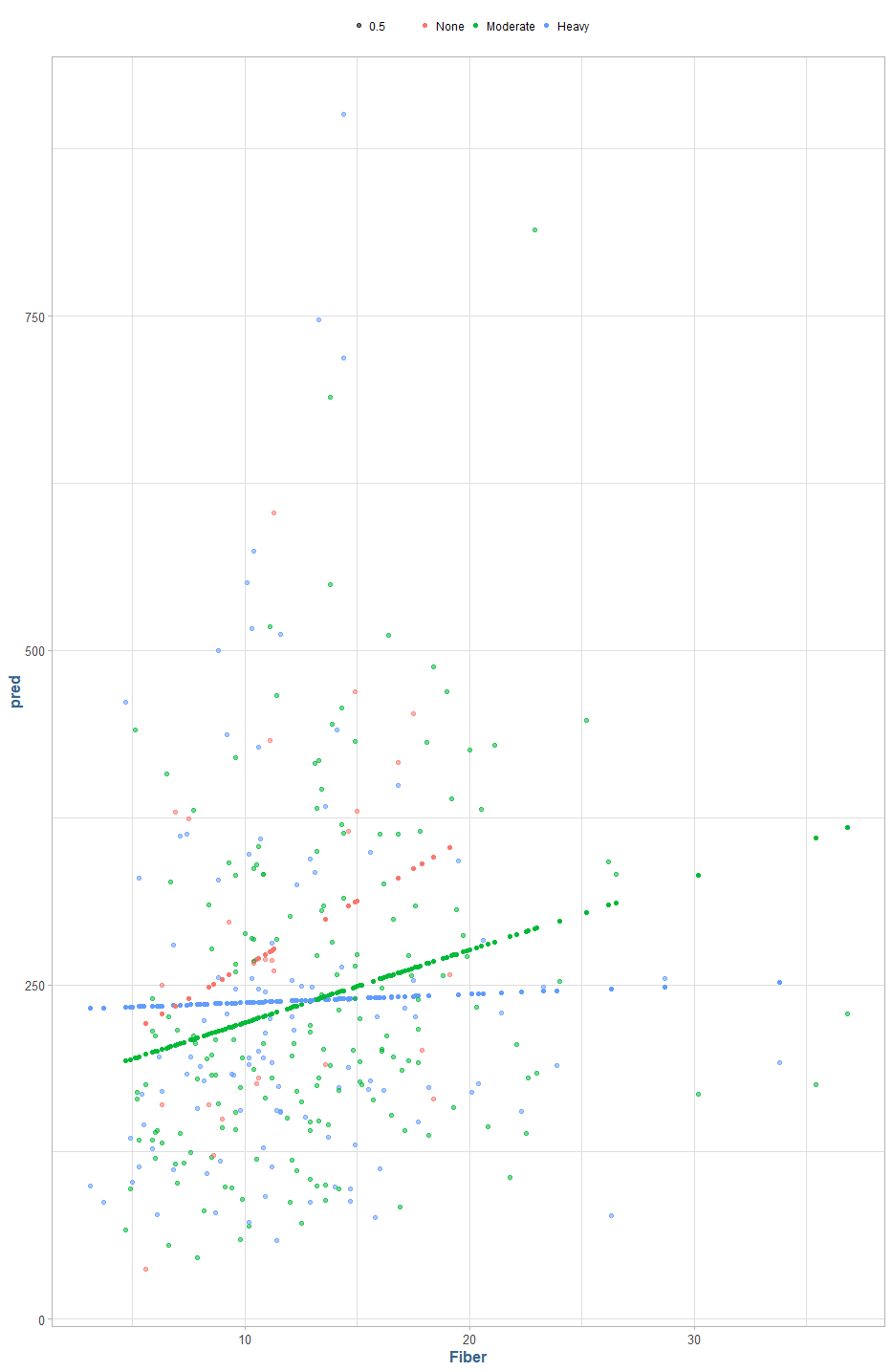
1. *Use the ANCOVA model from task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FIBER (X), but color code the records for the different groups of ALCOHOL. What do you notice about the patterns in the predicted values of Y? Now, make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

1. *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.*



Model 3: 166.9620 + 9.7105β1 + 0.5333β2 + 63.3814β3 -4.2737β4 – 9.0742β5

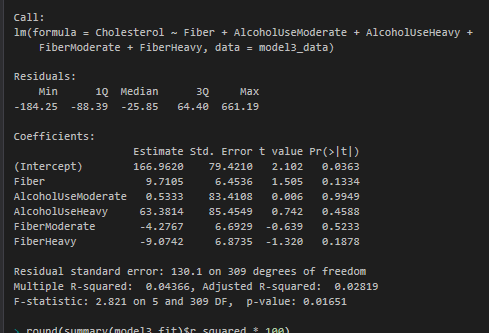
Where β1 = Fiber, β2 = Alcohol (Moderate) β3 = Alcohol (Heavy) β4 = Fiber + Moderate Alcohol interaction, β5 = Fiber + Heavy alcohol interaction. This model has an R2 of 0.0437 denoting that it accounts for approximately 4.3% of the overall variance in the data. We can see in the graph above the predicted model values (straight lines) plotted over the actual values, both of which are color coded to indicate the level of alcohol use per individual. This model appears to be a poor fit to the data.

The null hypothesis in this case would be,

H0 : β1 = β2 = β3 = β4= β5 = 0

Ha : βj ≠ 0, for at least one value of j (for j in 1,2,3,4,5)

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



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.

1. *You should be aware that the models of Task 4) and Task 5) are nested. Which model is the full and which one is the reduced model? Write out the null and alternative hypotheses for the nested F-test in this situation to determine if the slopes are unequal. Use the ANOVA tables from those two models you fit previously to compute the F-statistic for a nested F-test using Full and Reduced models. Conduct and interpret the nested hypothesis test. Are there unequal slopes? Discuss the findings.*

Model 3 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:

H0 : β5 = β6 = 0

Ha : βj ≠ 0, for at least one value of j (for j in 5, 6)

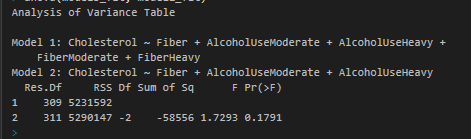
F = [ (SSER - SSEC) / (df2 – df1)] / ( SSEC / df1]

F = ((5,290,147 – 5,231,592) / 2) / [ 5,231,592/ 309]

= 29,277.89 / 16,930.72

= **1.7293**

Which is less than our critical value at 95% confidence, = F95, 309= **3.0254**, therefore we cannot reject the null hypothesis that fiber and alcohol have no interaction.



1. *Now that you’ve been exposed to these modeling techniques, it is time for you to use them in practice. Let’s examine more of the NutritionStudy data. Use the above practiced techniques to determine if SMOKE, VITAMINS, or GENDER interacts with the FIBER variable and influences the amount of CHOLESTEROL. Formulate hypotheses, construct essential variables (as necessary), conduct the analysis and report on the results. Which categorical variables are most predictive of CHOLESTEROL, in conjunction with FIBER.*

### Research

### Conclusion