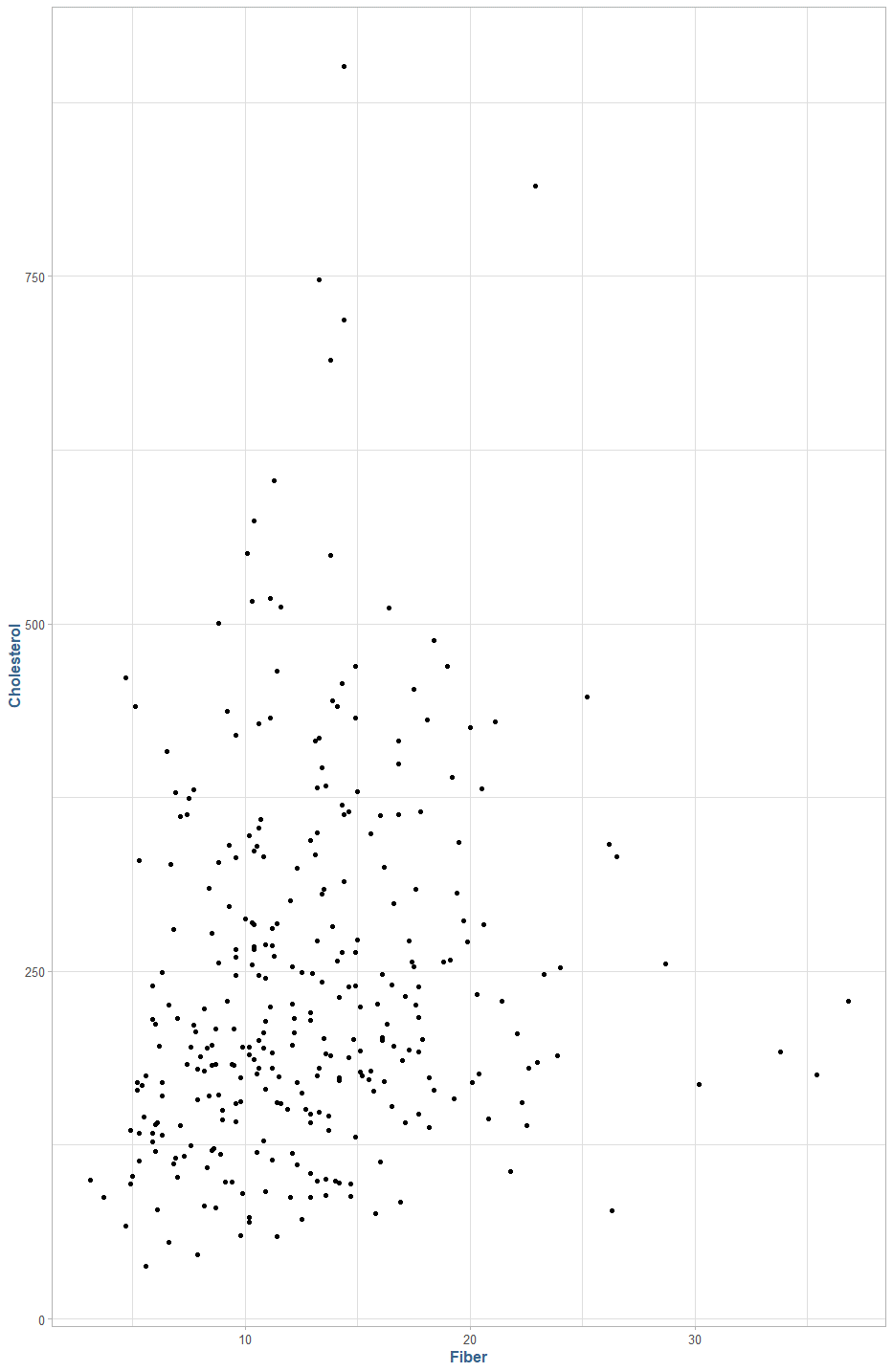
Computational Assigment #4

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.*

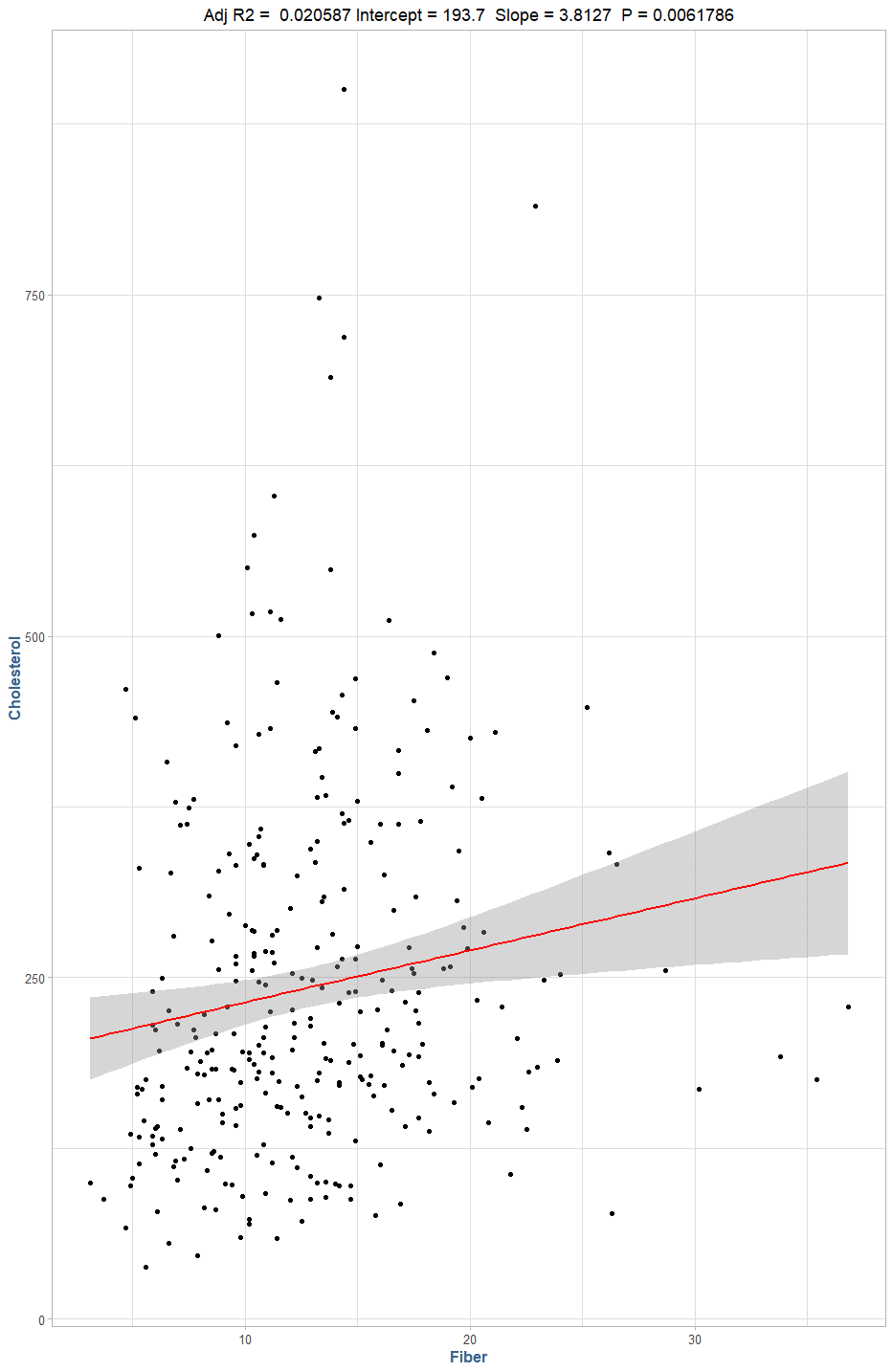
In the following diagram we see a lot of variance in the following scatterplot of Cholesterol ~ Fiber.



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

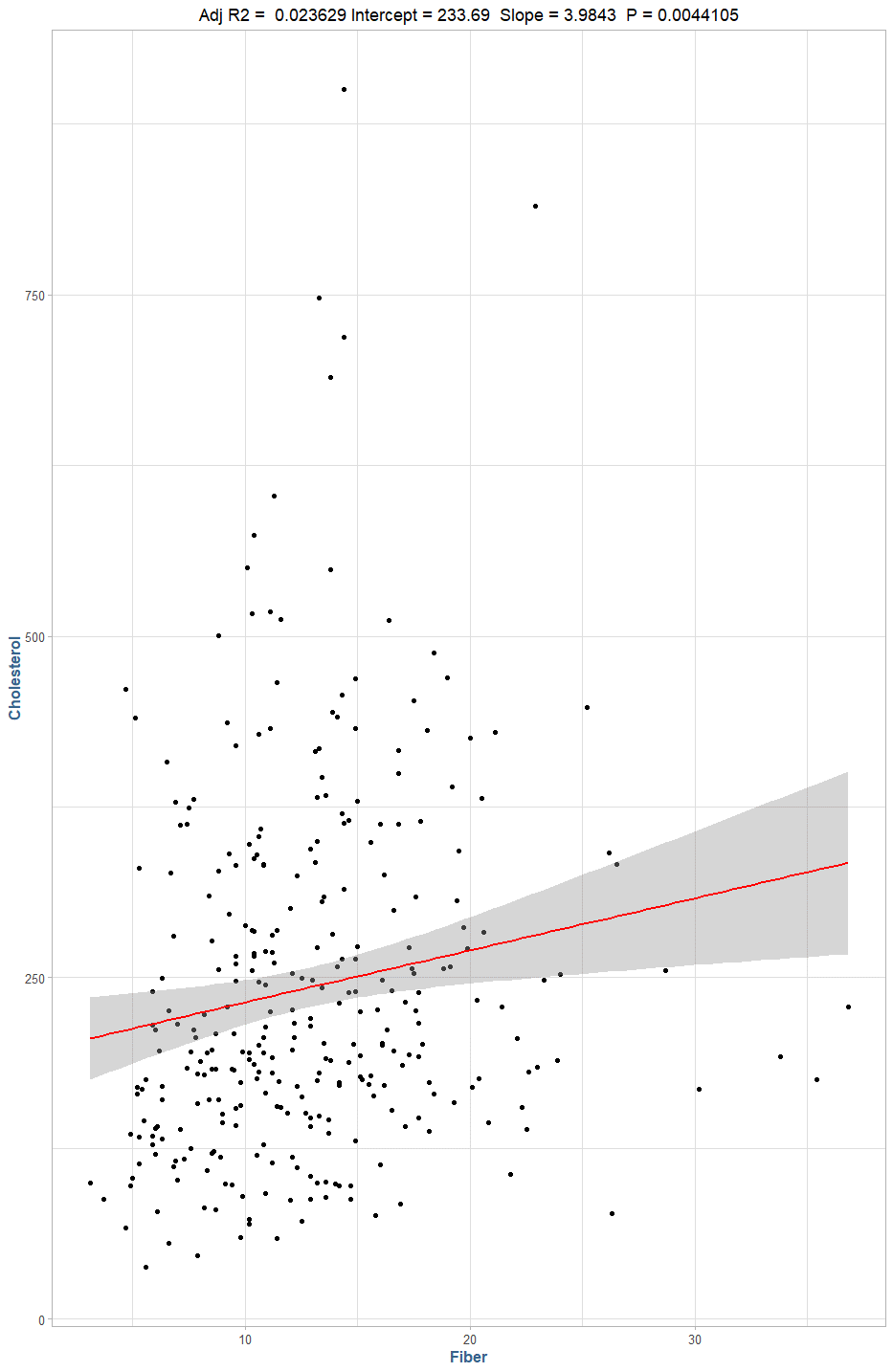
1. *Fit a simple linear regression model that uses FIBER to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss the goodness of fit.*

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. Additionally,



Model: 193.7014 + 3.8127 β1, where β1 is fiber.

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).*

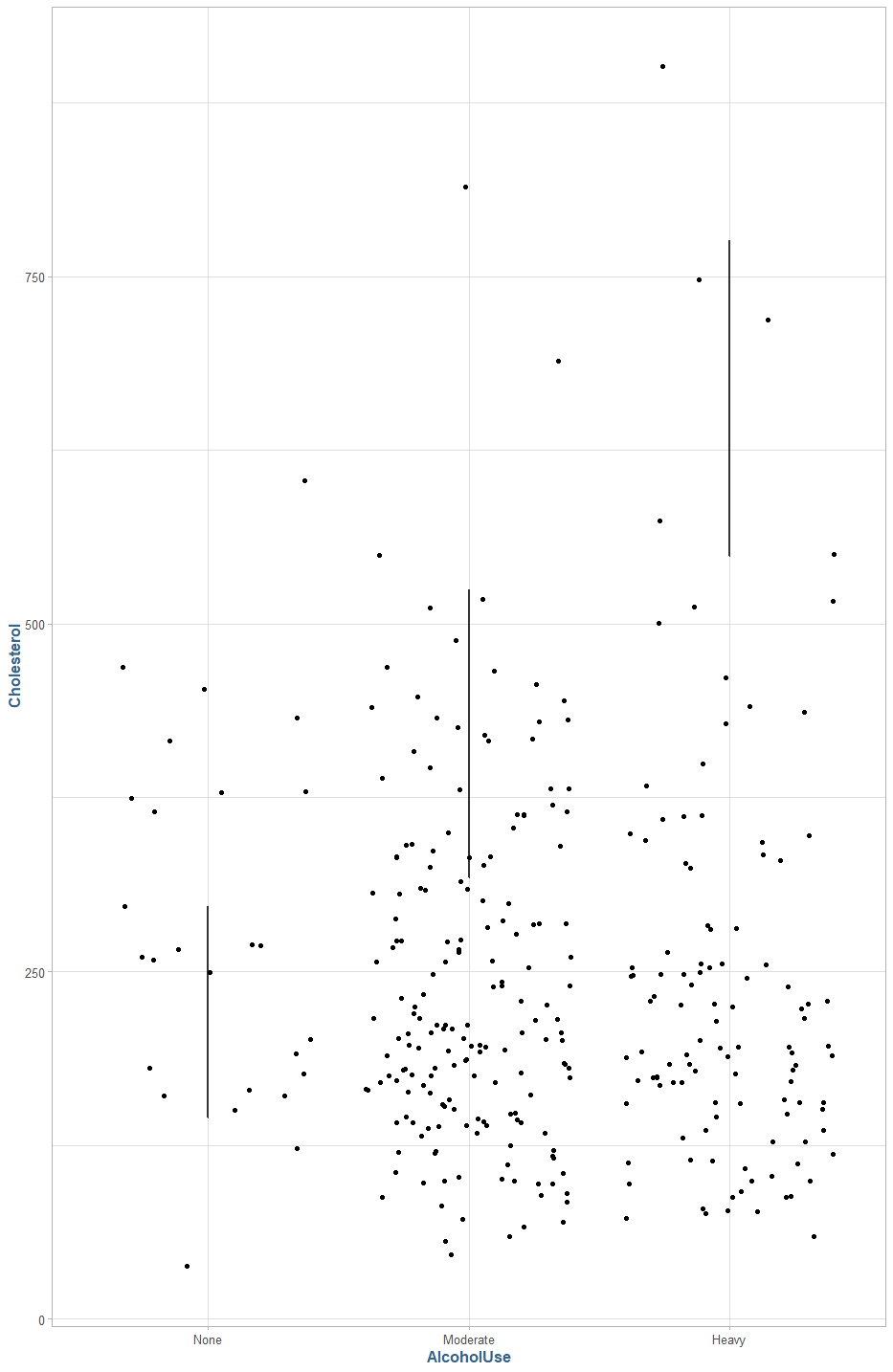


Model 2: 233.6946 + 3.9543β1 – 46.9518 β2 – 44.4289β3

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,

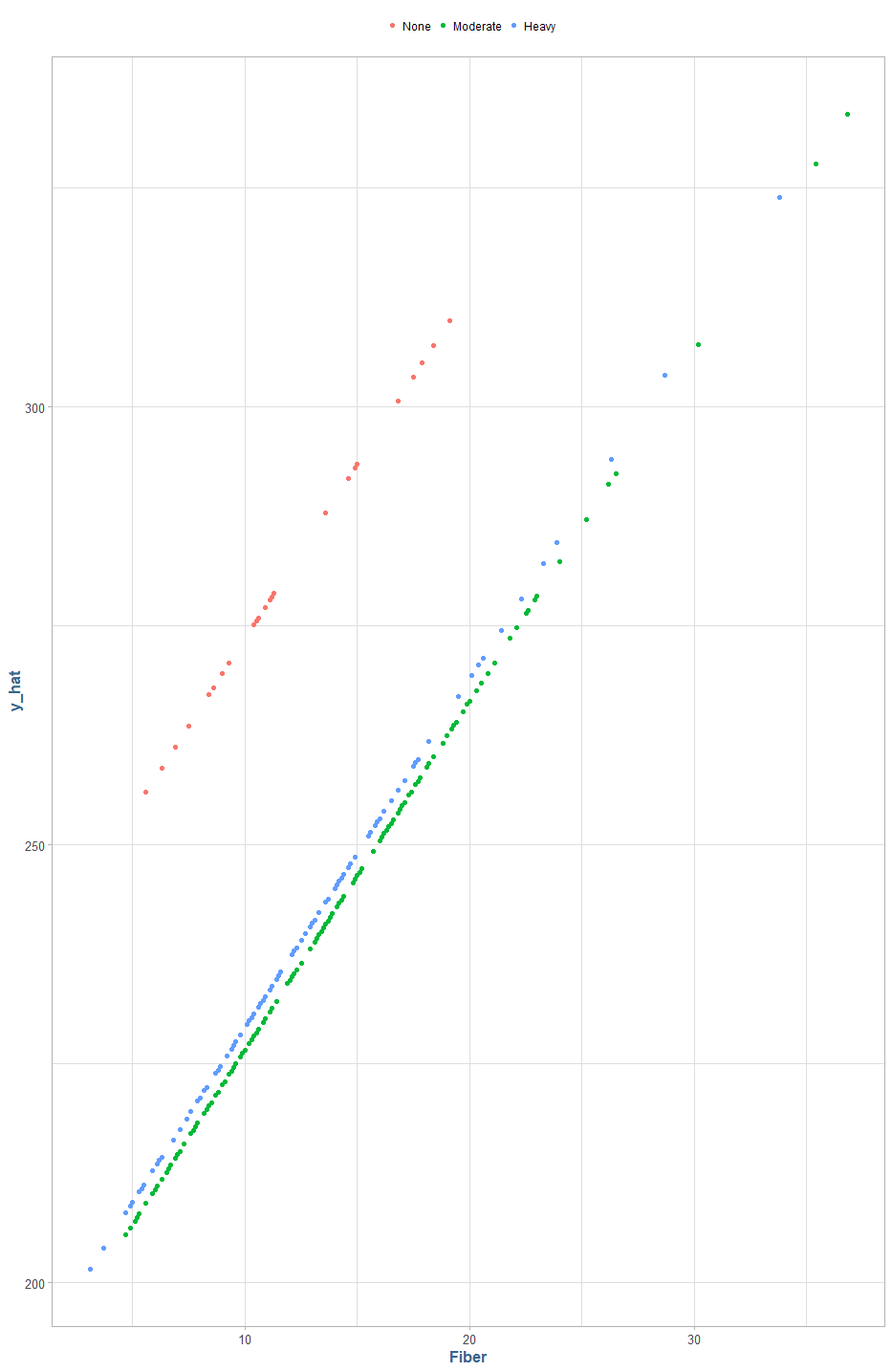
H0 : β1 = β2 = β3 = 0

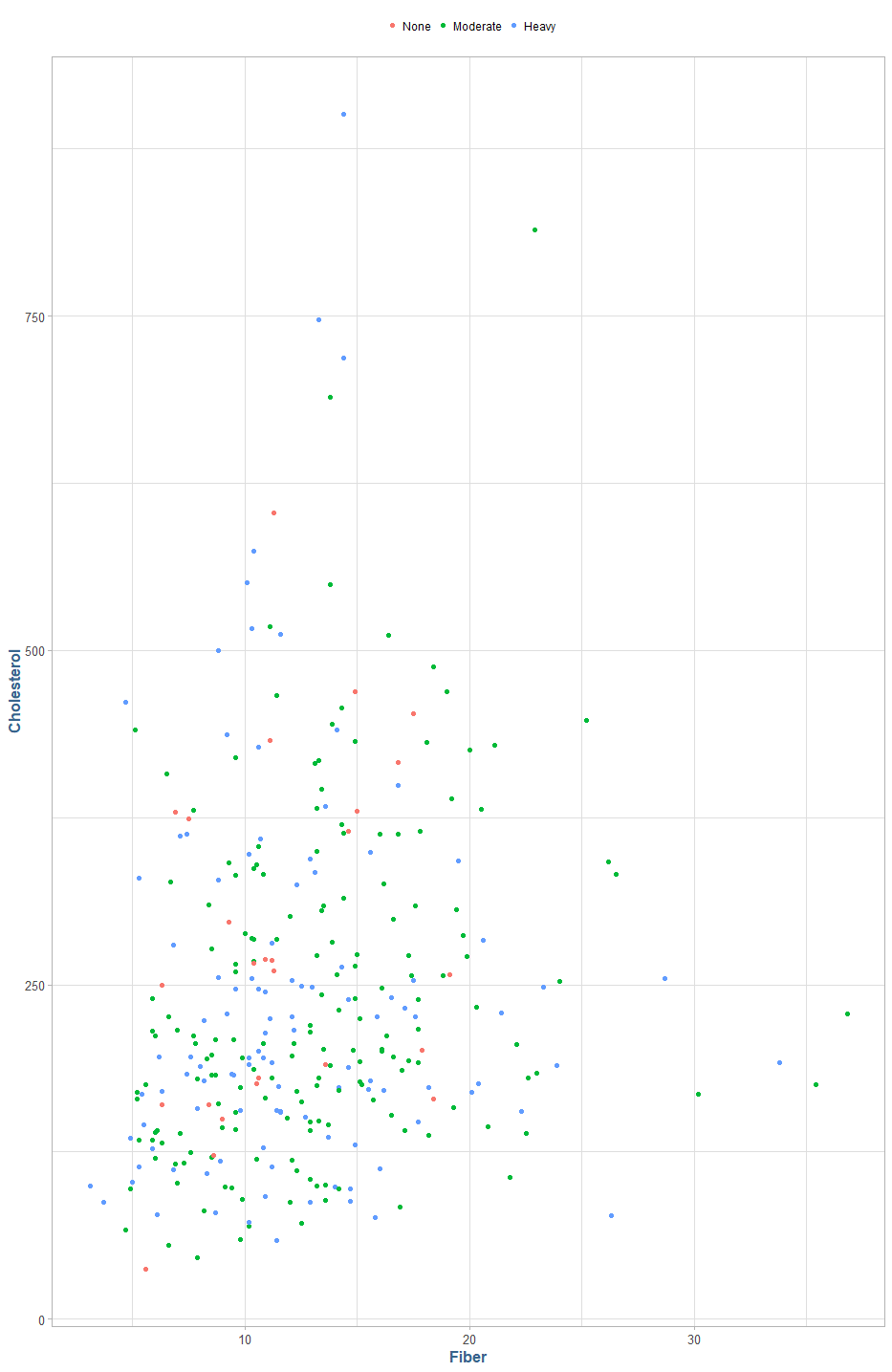
Or that there is no effect on the model using the beta coefficient derived from fiber consumption and the coded variables for alcohol consumption, against the alternative hypothesis that:

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.

### Research

### Conclusion