Northwestern University
Master of Science in Data Science

By: Ali Gowani

Table of Contents

Task 1:

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

Figure 1: Scatter Plot of Cholesterol vs Fiber with Regression Line

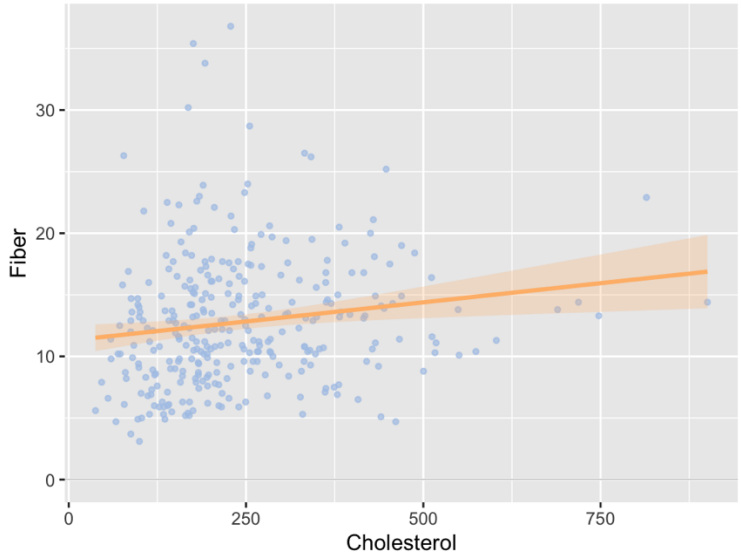
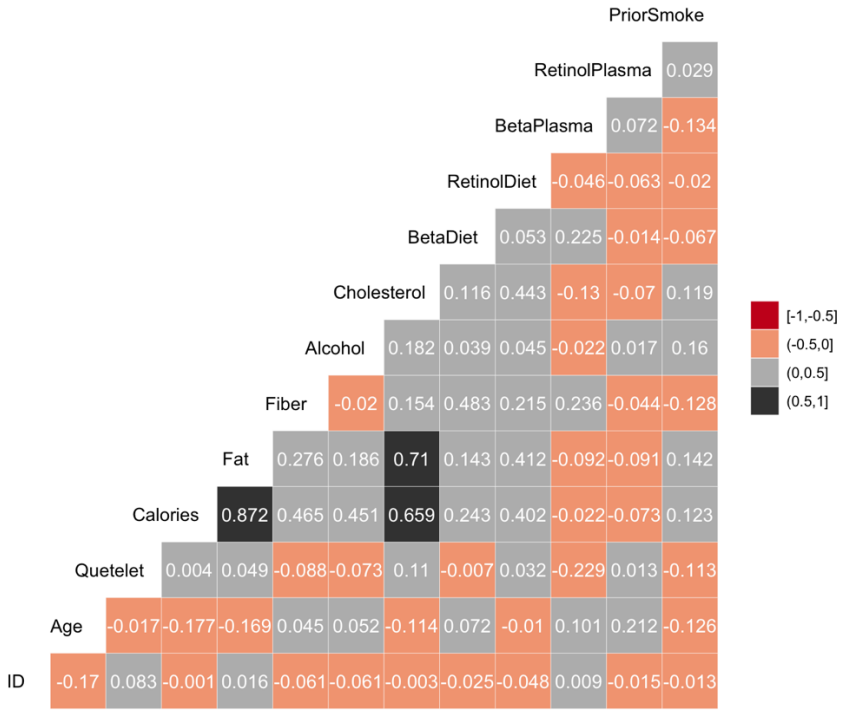


Figure 2: Correlation Plot of Data



There seems to be only a small, positive linear relationship between Cholesterol and Fiber. The correlation between these two variables is 0.154.

Task 2:

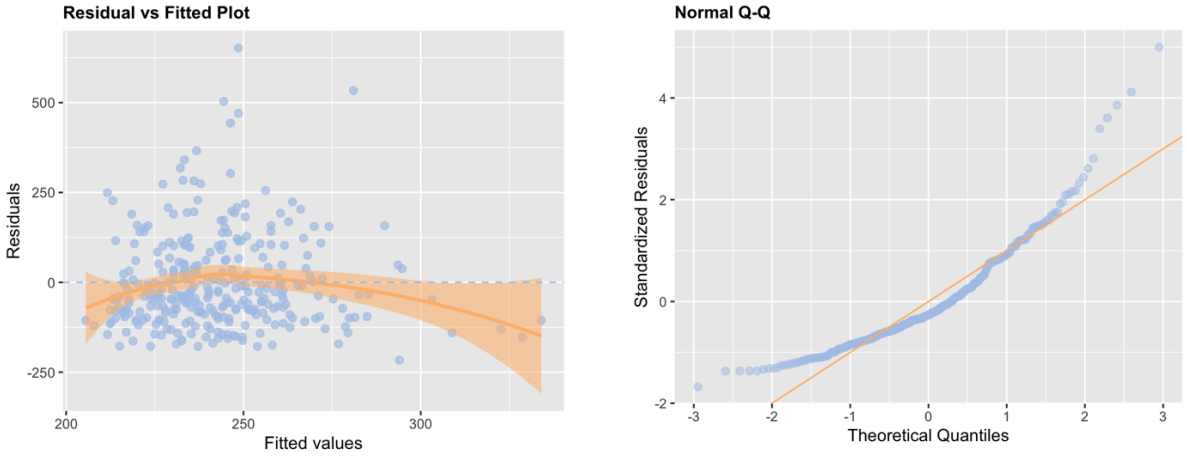
For the VITAMIN categorical variable, fit a simple linear model that uses the categorical variable to predict the response variable Y=CHOLESTEROL. Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Recode the VITAMIN categorical variable so that you have a different set of indicator values. For example, you could code it so that: 1=never, 2=occasional, 3=regular. Re-fit an OLS simple linear model using the new categorization. Report the model, interpret the coefficients, discuss test results, etc. What is going on here?

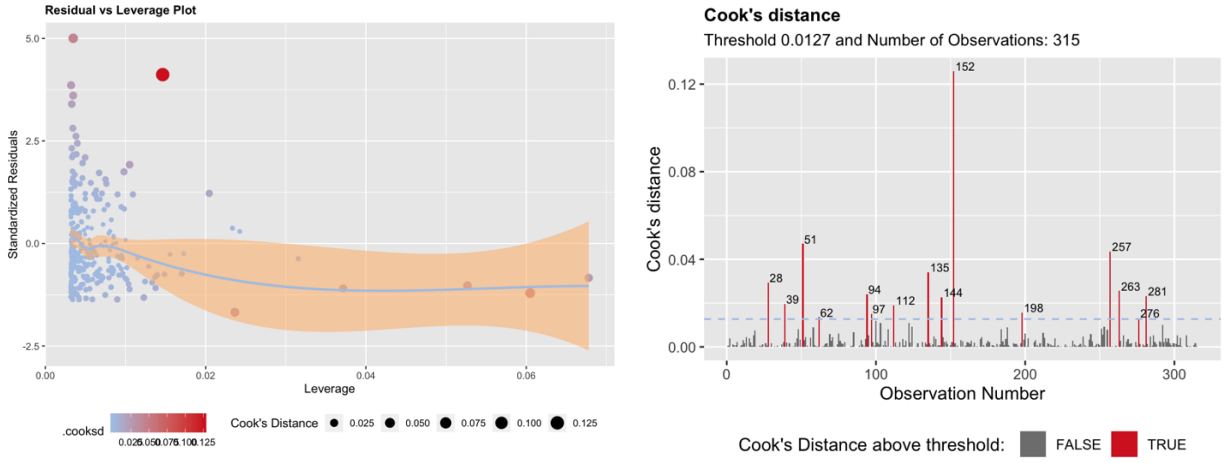
Model 1: lm(formula = Cholesterol ~ Fiber, data = mydata)

Note: full summary statistics in Appendix.

- $\hat{Y} = 193.701 + 3.813\beta_1$
- $R^2 = 0.02059$

Figure 3: Model 1 - Residual, QQ, Leverage and Cook's Distance Plots





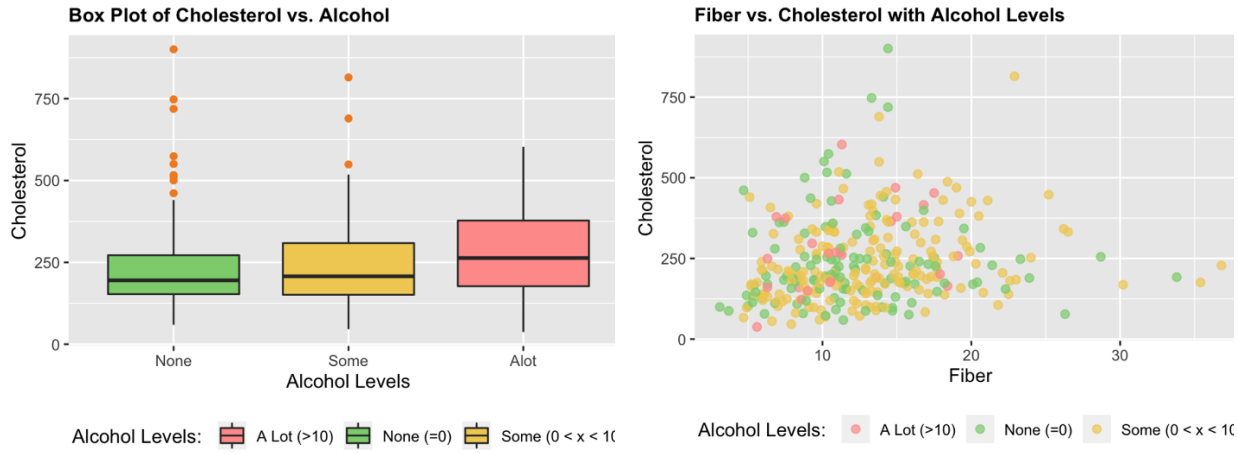
The y-intercept includes users who do not have any Fiber so if a person does not have any Fiber then their Cholesterol is 193.701. Fiber is positive so any 1 unit increase in Fiber means 3.813 increase in Cholesterol.

Task 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)?

The variable Alcohol was converted into a dummy coded variable (e.g.: AlcoholLevel_None, AlcoholLevel_Some, AlcoholLevel_Alot) in order to use it for predicting the response variable Cholesterol. It is interesting to note when looking at the box plot that the outliers for Cholesterol are for Alcohol Levels of “None” and “Some.” There are no outliers for Alcohol Level of “Alot.” This would mean that there are people who have high cholesterol that is outside of the range (i.e.: high outliers are above the $Q_3 + 1.5*IQR$).

Figure 4: Alcohol (Dummy Variable) with Cholesterol and Fiber



Model 2: lm(Cholesterol ~ Fiber + AlcoholLevel_Some + AlcoholLevel_Alot, data=mydata2)

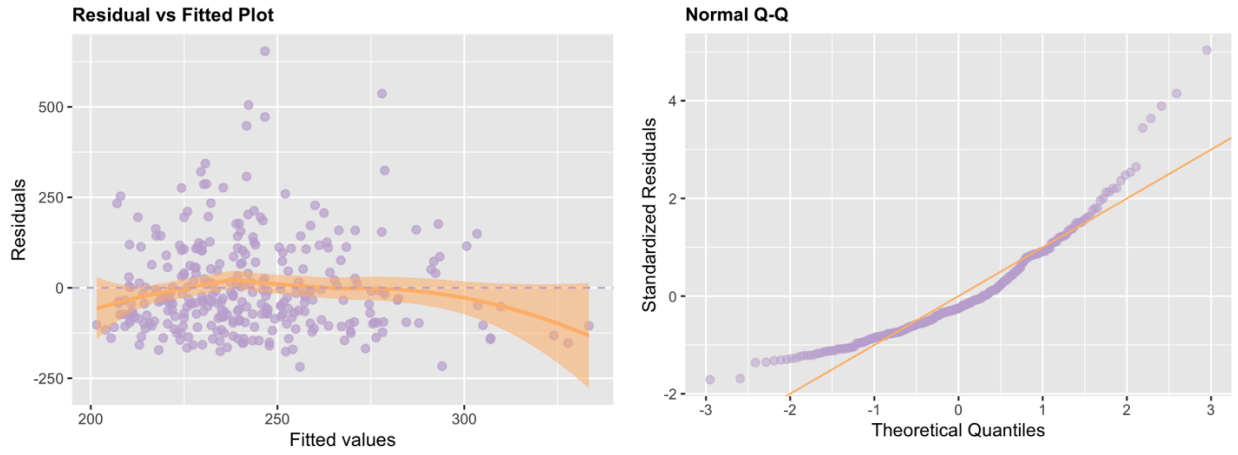
Note: full summary statistics

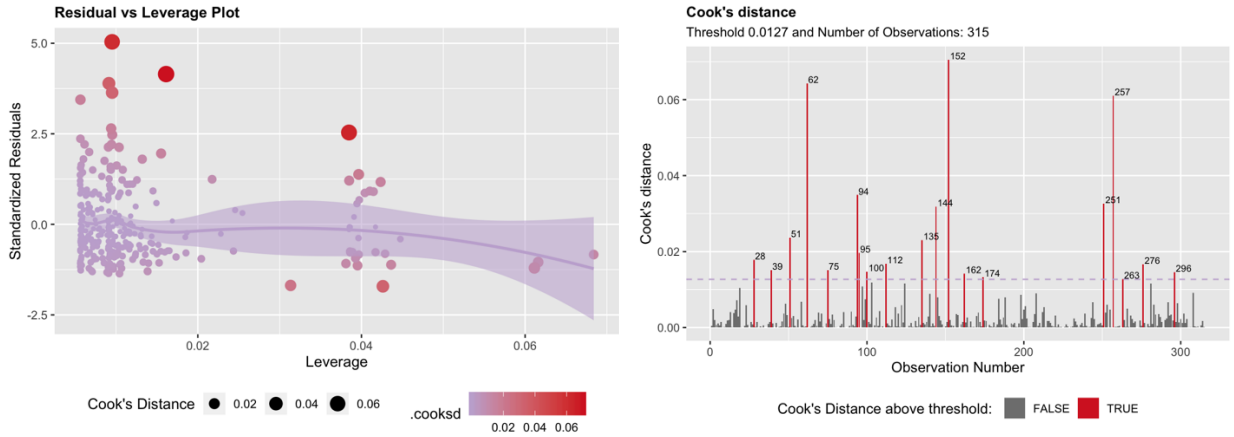
- $\hat{Y} = 189.266 + 3.984\beta_1 - 2.523\beta_2 + 44.429\beta_3$
- $R^2 = 0.03296$

The Omnibus Overall F-statistic for Model 2:

a. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2 or 3)

Figure 5: Model 2 - Residual, QQ, Leverage and Cook's Distance Plots





Since the F-statistic for Model 2 is 3.5331, which is greater than the critical F-statistic for Model 2 at 2.6336 and p-value is 0.01518 then we reject the Null Hypothesis (alpha = 0.05). This means that our model contains significant relationship between the explanatory variable and the response variable of Cholesterol.

Task 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 two figures below show a scatterplot of predicted values by Fiber and actual Cholesterol value by Fiber; both with color coding of Alcohol groups. The plot with the predicted values of Y seem to be linear in a positive slope, whereas the scatter plot of actual values of Cholesterol seems to be all over the place, without any linear relationship. It does not appear that the ANCOVA model fits well with the actual data, so it is not a good model to predict Y.

Figure 6: Fiber vs. Predicted Cholesterol with Alcohol Levels

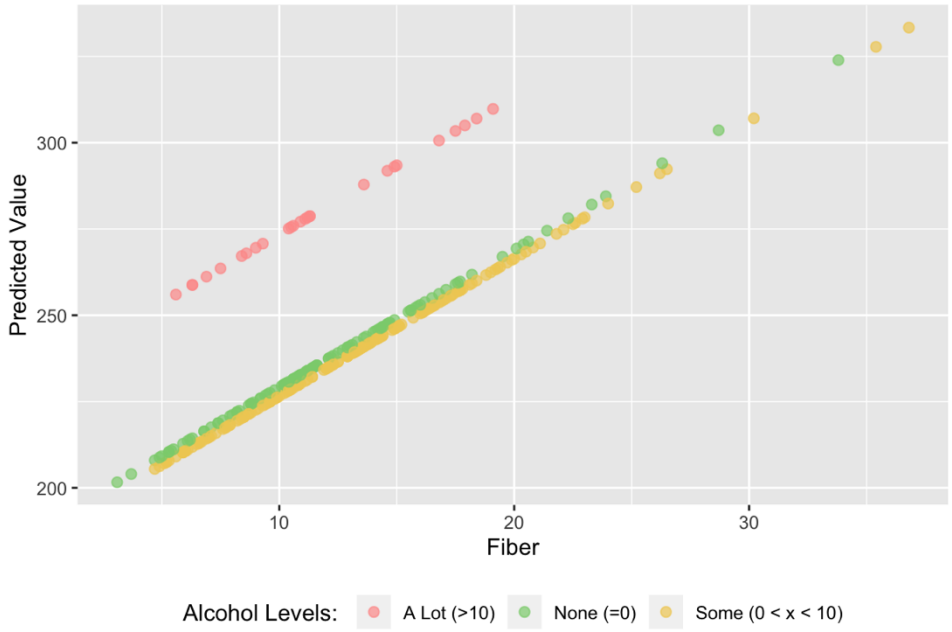
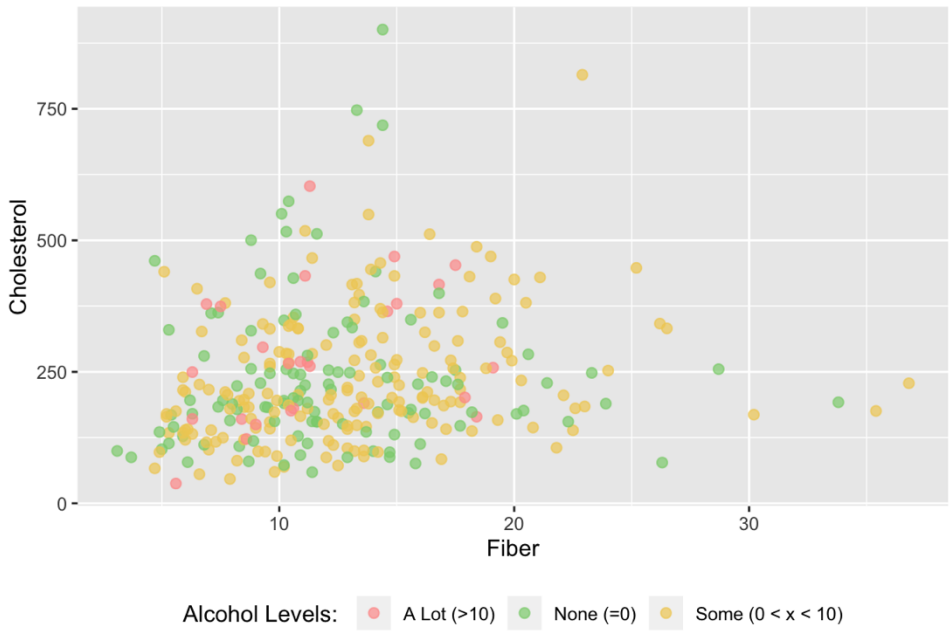


Figure 7: Fiber vs. Cholesterol with Alcohol Levels



Task 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.

Model 3: lm(Cholesterol ~ Fiber + AlcoholLevel_Some + AlcoholLevel_Alot + Fiber_AlcoSome + Fiber_AlcoAlot, data = mydata4)

Note: full summary statistics in Appendix.

- $\hat{Y} = 230.3434 + 0.6363\beta_1 - 62.8481\beta_2 - 63.3814\beta_3 + 4.7976\beta_4 + 9.0742\beta_5$
- $R^2 = 0.04366$

The Omnibus Overall F-statistic for Model 3:

c. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = ß_4 = ß_5 = 0$
d. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2, 3, 4 or 5)

Figure 8: Predicted Cholesterol vs. Fiber with Alcohol Levels

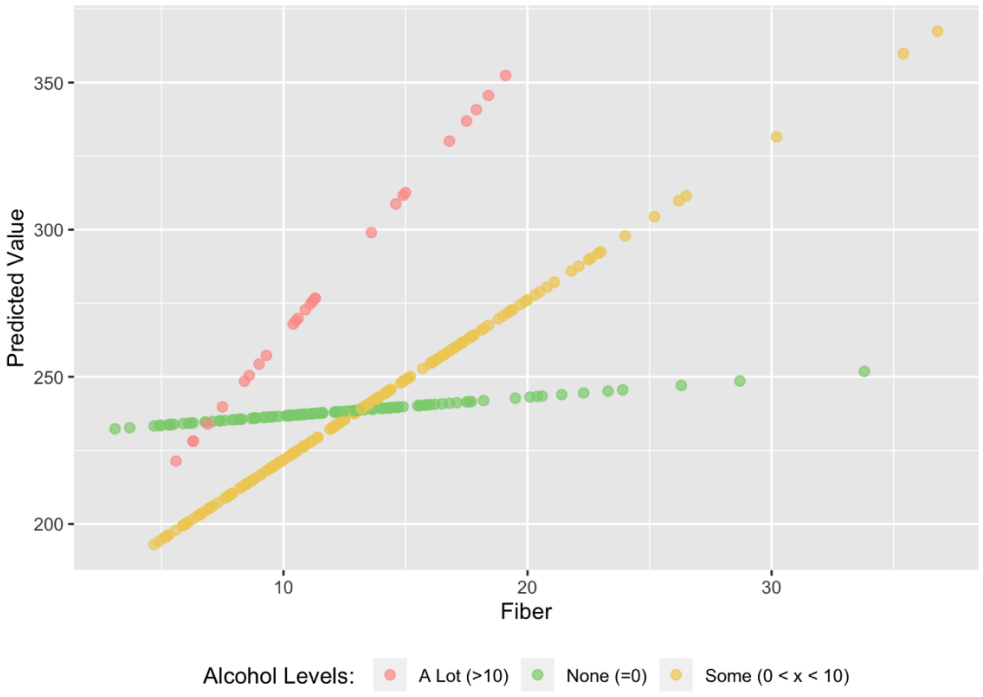
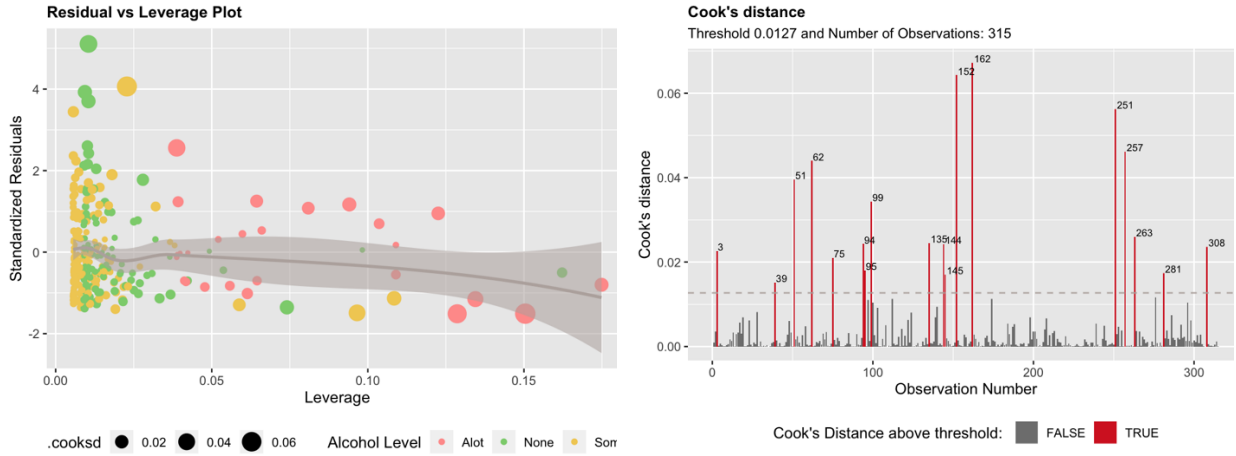


Figure 9: Model 3 – Residual vs Leverage and Cook's Distance Plots



Since the F-statistic for Model 3 is 2.821, which is greater than the critical F-statistic for Model 3 at 2.2432 and p-value is 0.01651 then we reject the Null Hypothesis (alpha = 0.05). This means that our model contains significant relationship between the explanatory variable and the response variable of Cholesterol. However, it is interesting to note the greater number of outliers in this model compared to the previous model and many of them having a high leverage as influencing points.

Task 6:

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.

Full Model: lm(Cholesterol ~ Fiber + AlcoholLevel_Some + AlcoholLevel_Alot + Fiber_AlcoSome + Fiber_AlcoAlot, data = mydata4)
Note: full summary statistics in Appendix.

- $\hat{Y} = 230.3434 + 0.6363\beta_1 - 62.8481\beta_2 - 63.3814\beta_3 + 4.7976\beta_4 + 9.0742\beta_5$
- $R^2 = 0.04366$

The Omnibus Overall F-statistic for Full Model:

e. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = ß_4 = ß_5 = 0$
f. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2, 3, 4 or 5)

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{5470440.852 - 5231591.648}{5}\right)}{\left(\frac{5231591.648}{315-5-1}\right)} = 2.8215$$

The critical F-statistic for Full Model is:

$$F_{i,n-k-p-1,1-a} = F_{5,315-5-1,0.95} = 2.2432$$

Since the F-statistic for Full Model is 2.8215, which is greater than the critical F-statistic at 2.2432 then we reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variables and the response variable of Cholesterol.

Reduced Model: lm(Cholesterol ~ Fiber + AlcoholLevel_Some + AlcoholLevel_Alot, data=mydata2)
Note: full summary statistics

- $\hat{Y} = 189.266 + 3.984\beta_1 - 2.523\beta_2 + 44.429\beta_3$
- $R^2 = 0.03296$

The Omnibus Overall F-statistic for Reduced Model:

g. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = 0$
h. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2 or 3)

We can calculate the F-test of the nested model by using the following formula:

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{5470440.852 - 5290147.432}{3}\right)}{\left(\frac{5290147.432}{315-3-1}\right)} = 3.5331$$

The critical F-statistic for Reduced Model is:

$$F_{i,n-k-p-1,1-a} = F_{4,315-4-1,0.95} = 2.6336$$

Since the F-statistic for Reduced Model is 3.5331, which is greater than the critical F-statistic at 2.6336 then we reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variables and the response variable of Cholesterol.

The Omnibus Overall F-statistic for Nested Model:

For a nested F-test, we use two models (Full Model and Reduced Model), these models are considered nested if they both have the same variables and one of the models (Full Model) has at least one additional variable. In our case, Reduced Model is nested within Full Model. Reduced Model is considered reduced and Full Model is considered complete. By conducting a nested F-test between these models, we will determine whether the additional explanatory variables in Full Model are more robust than the Reduced Model.

The values for i represent the additional variables added to our model.

a. Null Hypothesis (H_o): $ß_4 = ß_5 = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 4 or 5)

We can calculate the F-test of the nested model by using the following formula:

$$F = \frac{\frac{(SSE_R - SSE_C)}{s}}{\left(\frac{SSE_C}{n-k-p-1}\right)} = \frac{\frac{(238849.2 - 180293.4)}{2}}{\left(\frac{180293.4}{315-5-1}\right)} = 1.7293$$

The critical F-statistic value is:

$$F_{i,n-k-p-1,1-a} = F_{2,315-5-1,0.95} = 3.025$$

Since the F-statistic value of 1.7293 is less than the critical value of 3.025 at a confidence of 95%, then we fail to reject the null hypothesis that the Full Model is more robust than the Reduced Model. This means that the additional variables do not add significant information in predicting Cholesterol.

Task 7:

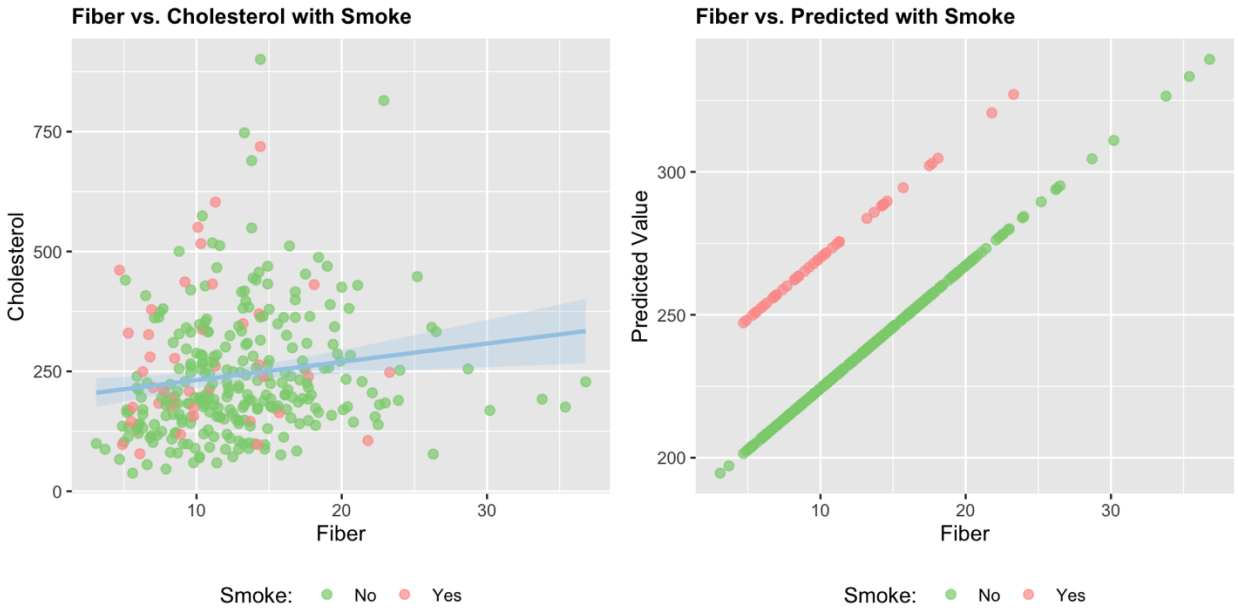
Now that you've been exposed to these modeling techniques, it is time for you to use them in practice. Let's examine more of the NutritionStudy data. Use the above practiced techniques to determine if SMOKE, VITAMINS, or GENDER interacts with the FIBER variable and influences the amount of CHOLESTEROL. 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.

Model 4 Smoke: lm(Cholesterol ~ Fiber + Smoke_Yes, data = mydata5)
Note: full summary statistics in Appendix.

Model 4 Smoke with Interaction: lm(Cholesterol ~ Fiber + Smoke_Yes + Fiber_Smoke_Yes, data = mydata5)

Note: full summary statistics in Appendix.

Figure 10: Model 4 Smoke – Scatter (Fiber and Cholesterol) and Fiber with Predicted Plots



Since the F-statistic value of 0.0587 is less than the critical value of 3.0248 at a confidence of 95%, then we fail to reject the null hypothesis that the model with interaction is more robust than the model without it. This means that the additional variables do not add significant information in predicting Cholesterol.

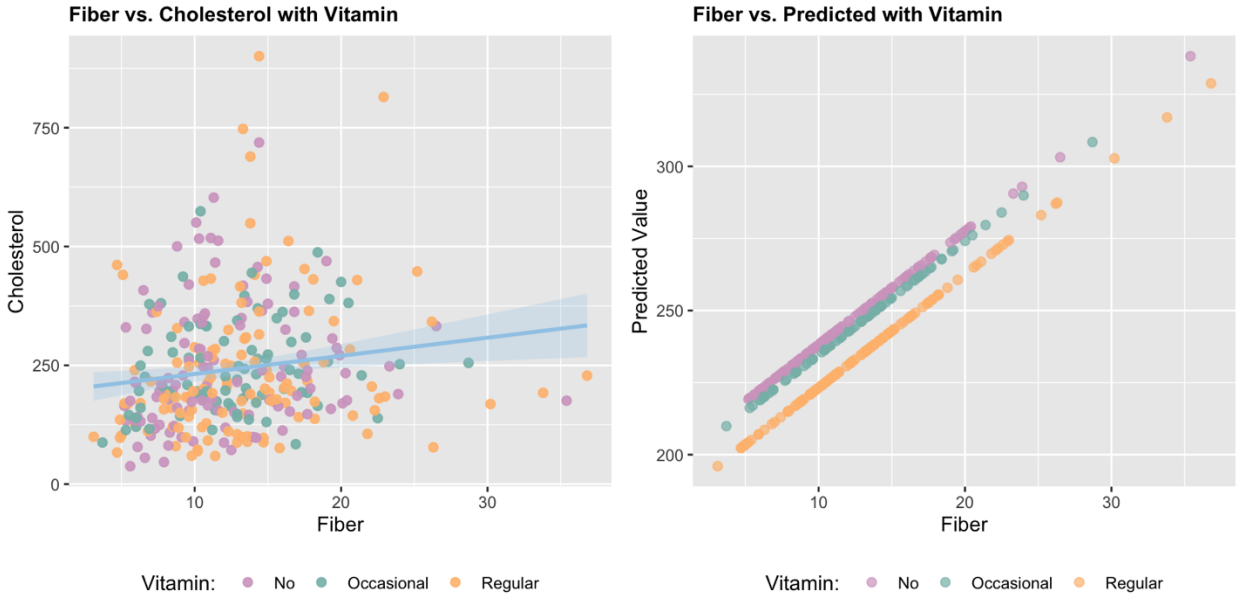
Model 4 Vitamin: lm(Cholesterol ~ Fiber + Vitamin_Reg + Vitamin_Occ, data = mydata5)

Note: full summary statistics in Appendix.

Model 4 Vitamin with Interaction: lm(Cholesterol ~ Fiber + Vitamin_Reg + Vitamin_Occ + Fiber_Vitamin_Reg + Fiber_Vitamin_Occ, data = mydata5)

Note: full summary statistics in Appendix.

Figure 11: Model 4 Vitamin – Scatter (Fiber and Cholesterol) and Fiber with Predicted Plots

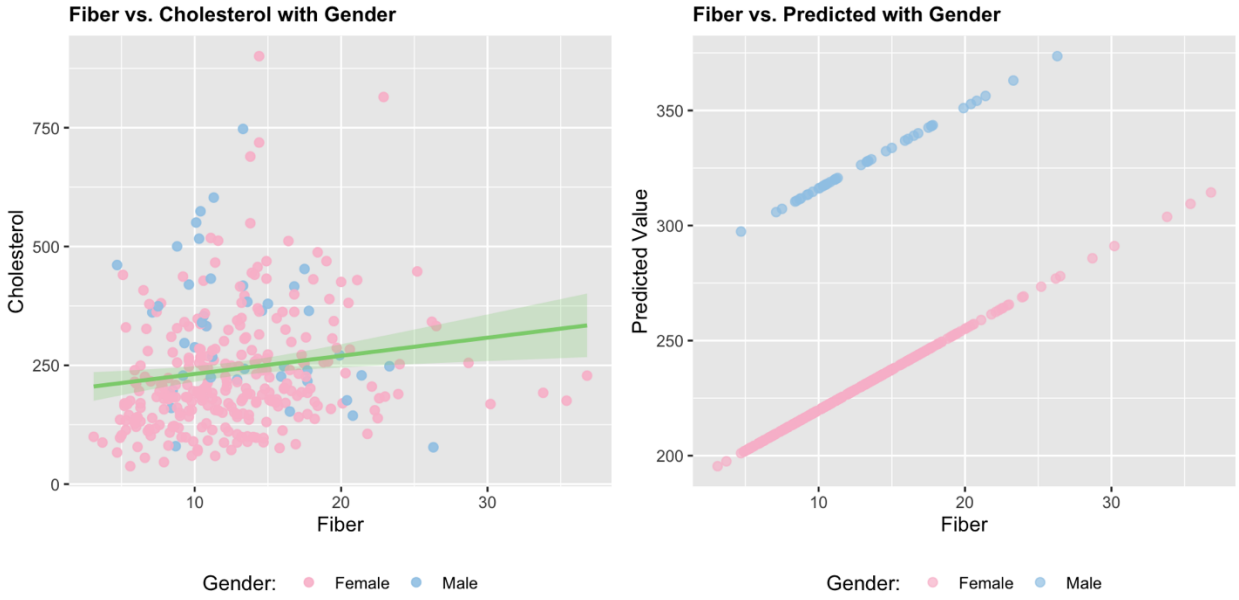


Since the F-statistic value of 0.0566 is less than the critical value of 2.6338 at a confidence of 95%, then we fail to reject the null hypothesis that the model with interaction is more robust than the model without it. This means that the additional variables do not add significant information in predicting Cholesterol.

Model 4 Gender: lm(Cholesterol ~ Fiber + Gender_Male, data = mydata5)
Note: full summary statistics in Appendix.

Model 4 Gender with Interaction: lm(Cholesterol ~ Fiber + Gender_Male + Fiber_Gender_Male, data = mydata5)
Note: full summary statistics in Appendix.

Figure 12: Model 4 Gender – Scatter (Fiber and Cholesterol) and Fiber with Predicted Plots



Since the F-statistic value of 7.2676 is greater than the critical value of 3.0248 at a confidence of 95%, then we reject the null hypothesis that the model with interaction is no more robust than the model without it. This means that the additional variables add significant information in predicting Cholesterol.

Based on the three variables (i.e.: Smoke, Vitamin and Gender) that we compared, it seems that Gender had an overwhelming effect on predicting the response variable Y (Cholesterol).

Task 8: Conclusion & Reflection

I enjoyed this assignment and seeing how categorical and continuous variables impact the response variable, as well as, seeing how variables can influence the various points in the model. I find the Cook's Distance plot quite intriguing. In addition, I utilized the dummy coded variables for categorical variable. We also implemented Analysis of Covariance Model (ANCOVA).

I found the task of seeing how Gender interacts with the Fiber variable and influences the amount of Cholesterol. I was surprised to see the high F-statistic value but a low r-squared value. I had to do further research to understand this. Based on our model, it seems that Gender plays a huge role in predicting Cholesterol compared to other variables. I am wondering that further analysis would need to be conducted in determining whether there is a causal relationship between Cholesterol and Gender.

The bonus task was fun for me to try few things and challenge my preconceived notion. I thought that Fat variable along with Gender would be a better predictor of Cholesterol, but I was wrong here. It seems that Fiber produces a more robust model, even though Fat may explain more of the variance in the model.

Learning these things and re-enforcing them on a weekly basis has helped me better understand how I can apply these things at my client. Overall, this was a nice assignment allowing me to reinforce previous concepts and learning new ones.

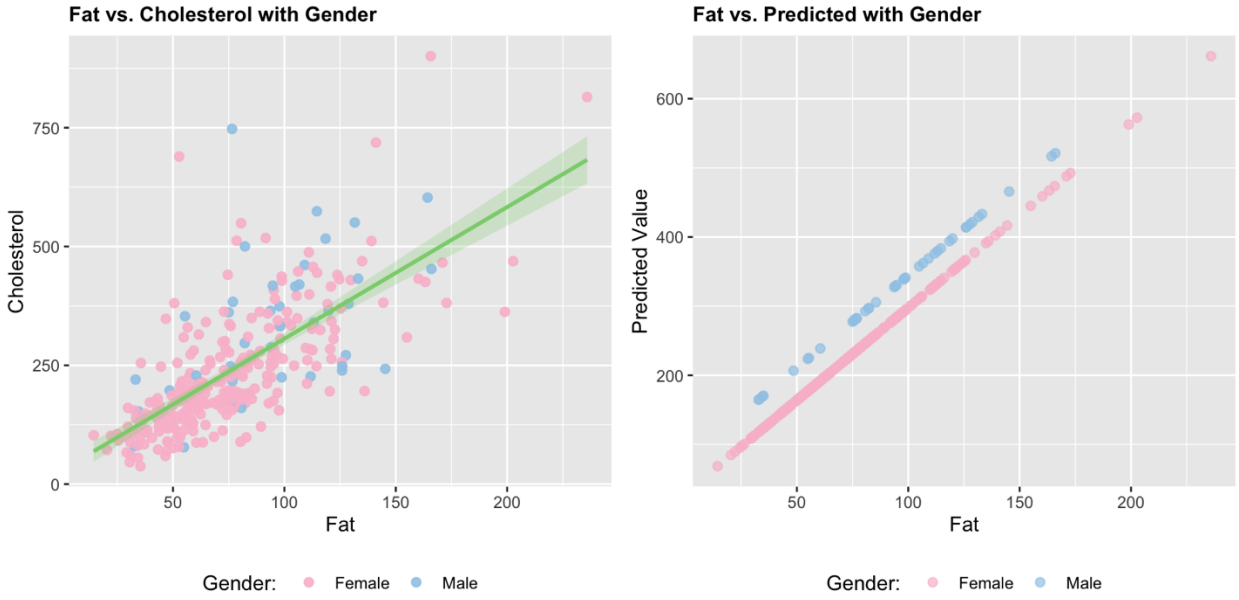
Task 9: Bonus

For the bonus section of the assignment, I want to see whether Fat along with Gender can provide a more robust model in predicting Cholesterol than just Gender alone. I have used two models for this: one with Gender_Male and Fat and the other with the interaction between Fiber and Gender_Male.

Model 4 Fat: lm(Cholesterol ~ Gender_Male + Fat, data = mydata5)
Note: full summary statistics in Appendix.

Model 4 Fat with Interaction: lm(Cholesterol ~ Gender_Male + Fat_Gender_Male + Fat, data = mydata5)
Note: full summary statistics in Appendix.

Figure 13: Model 4 Fat – Scatter (Fat and Cholesterol) and Fat with Predicted Plots



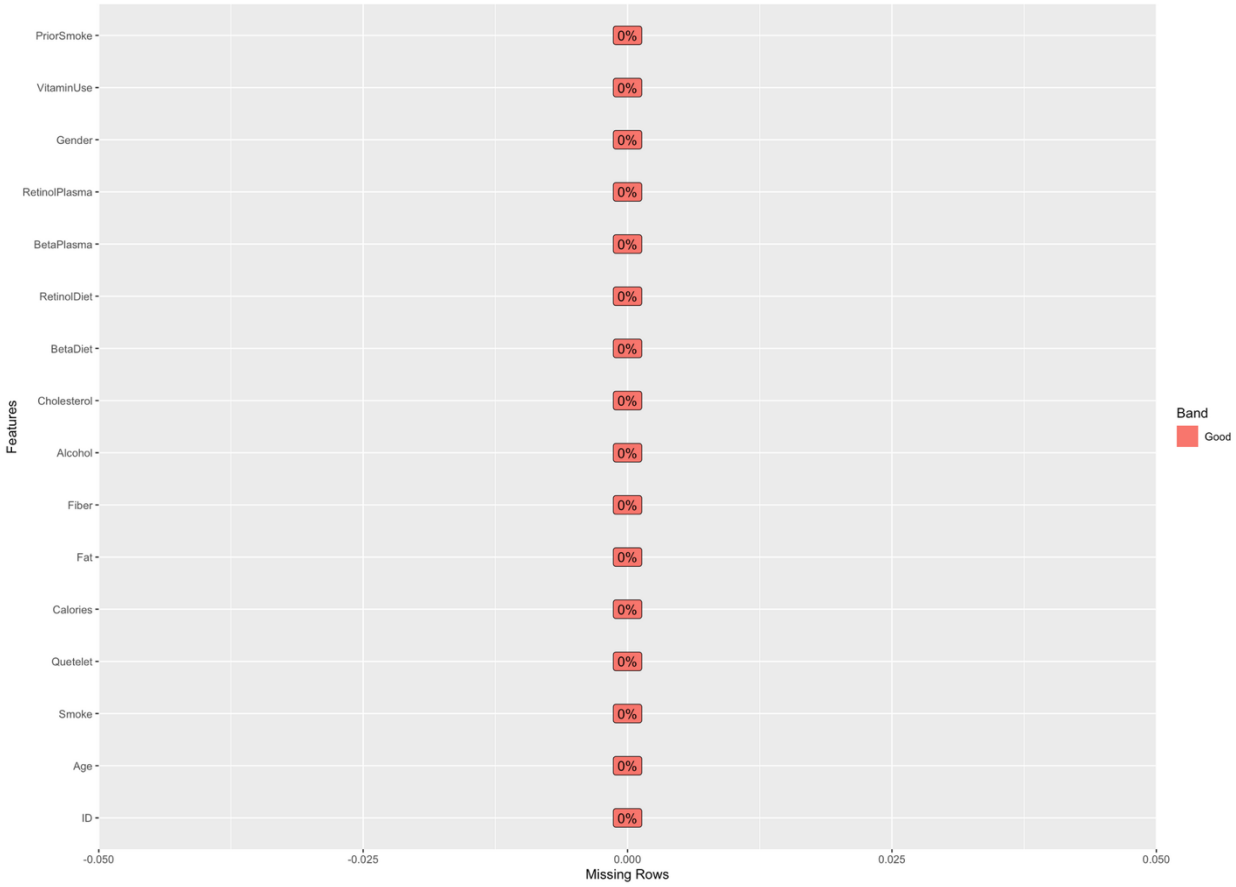
Since the F-statistic value of 6.7483 is greater than the critical value of 3.0248 at a confidence of 95%, then we reject the null hypothesis that the model with interaction is no more robust than the model without it. This means that the additional variables add significant information in predicting Cholesterol.

It seems that from our previous model, Fiber and Gender variables together provide the most robust model from what we have tried. However, the model of Fat and Gender is also robust in predicting Cholesterol. What caught my attention in this bonus attempt is that the Adjusted R-squared value of Fat and Gender model is 0.5335, compared to the Adjusted R-squared value of Fiber and Gender, which was 0.1177. While the F-statistic and t-values for Fiber and Gender model was higher, the Adjusted R-squared value was quite low. This suggests that this model is more accurate but does not do a good job in measuring the amount of variance in our response variable (Cholesterol) explained by the covariances.

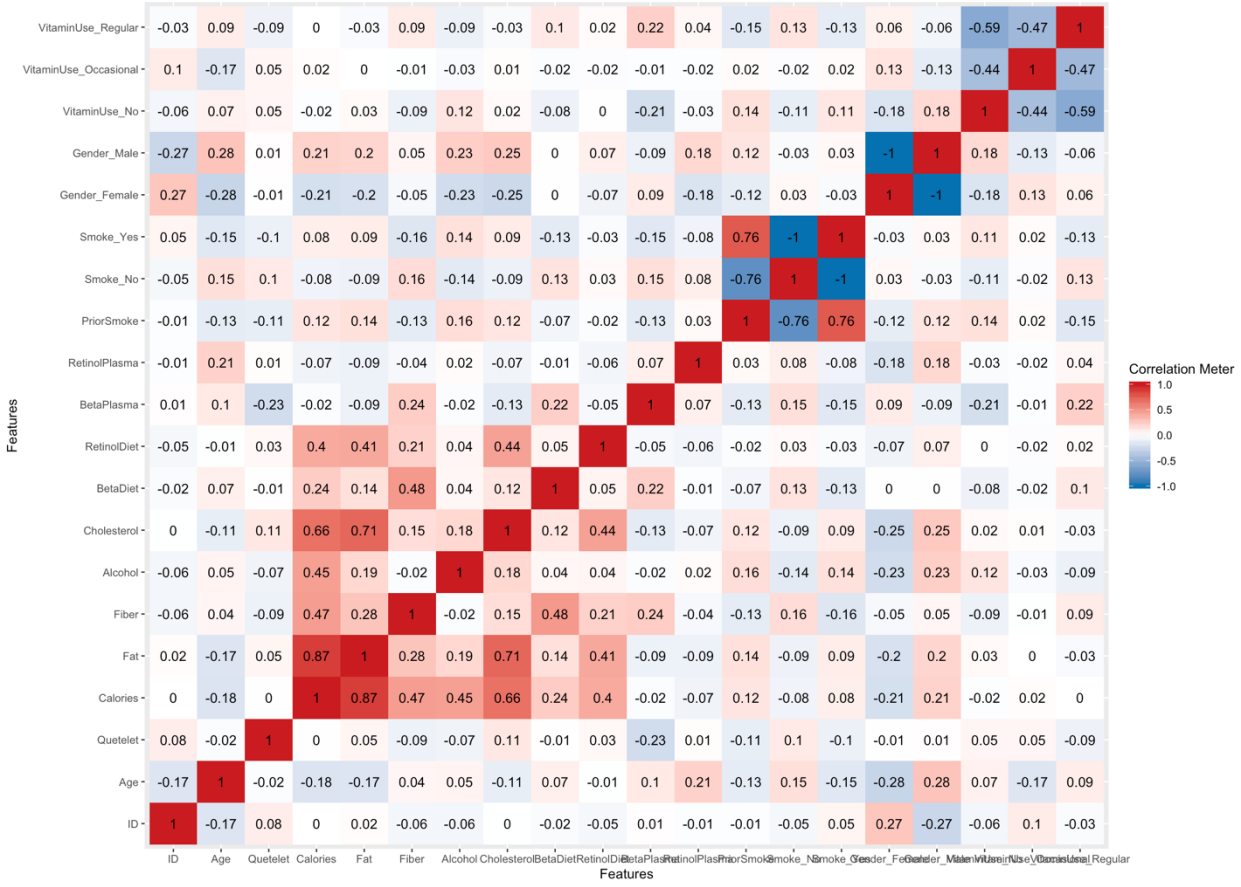
Appendix

A: Exploratory Data Analysis

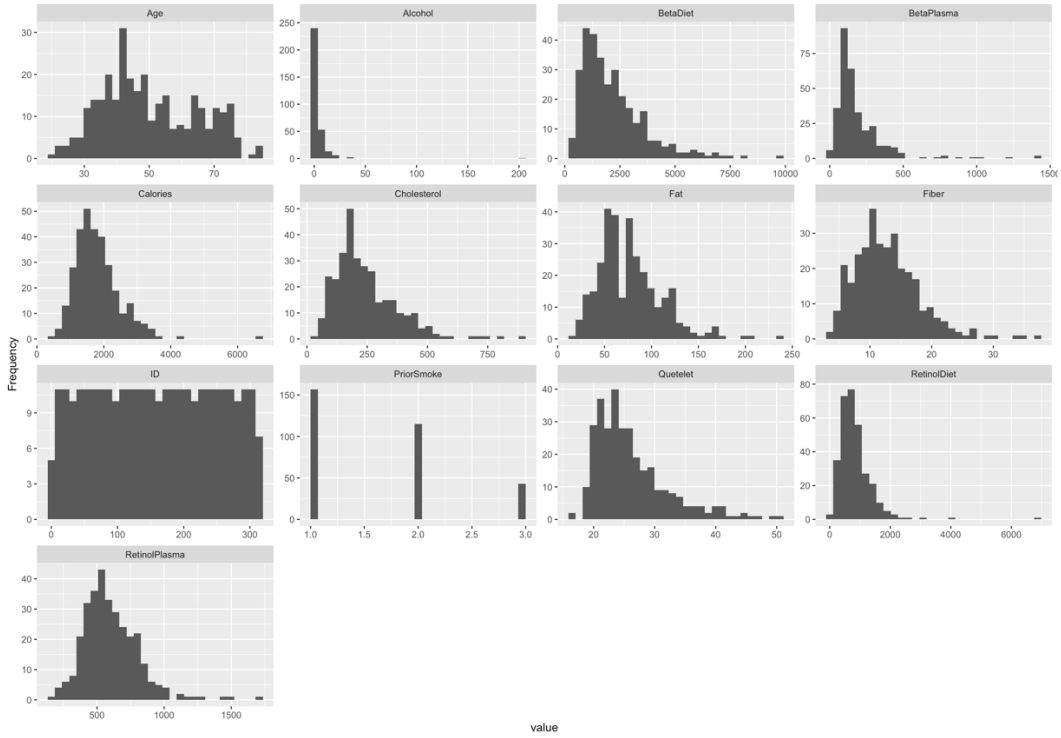
Missing Data



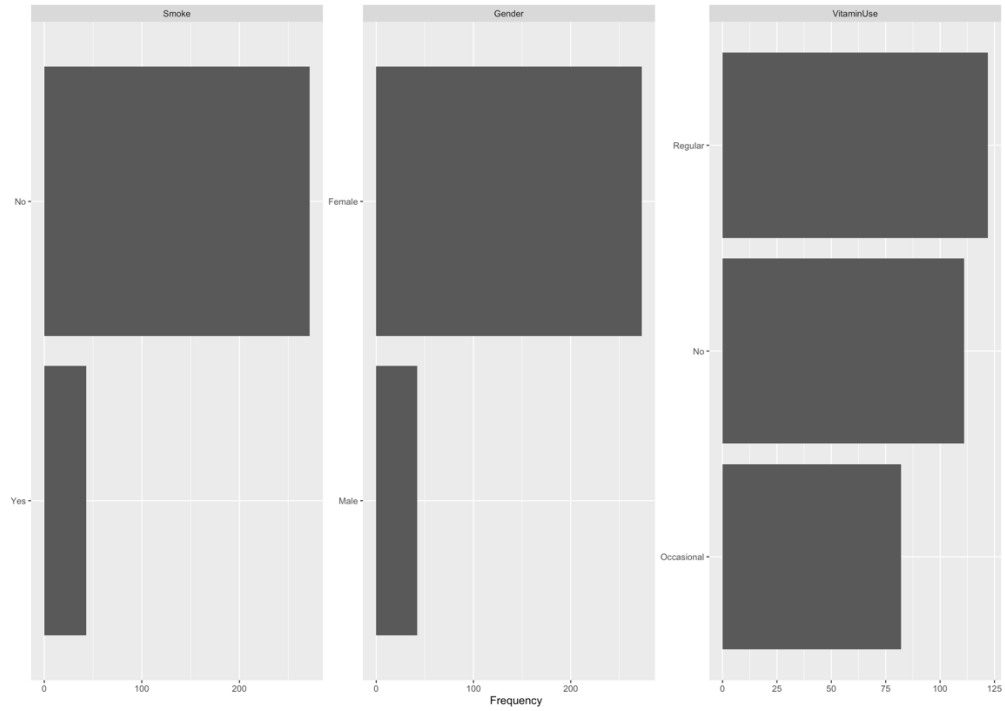
Correlation Analysis



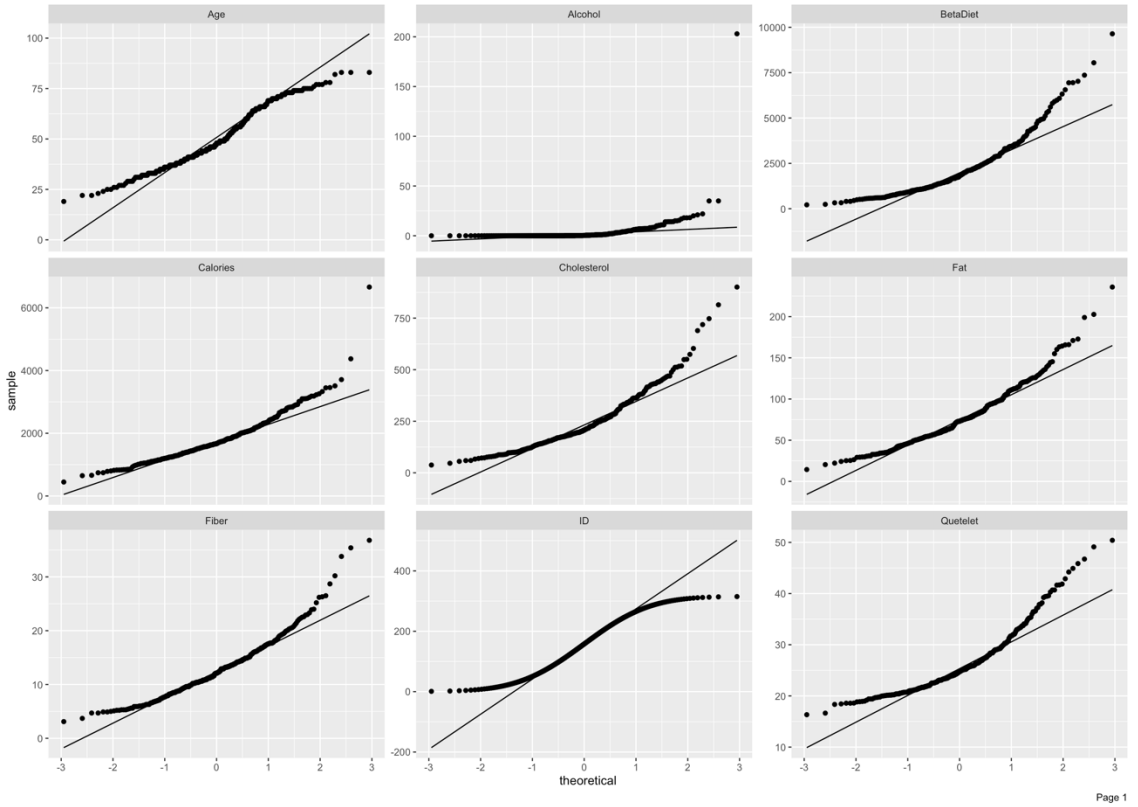
Univariate Distribution (Histograms)

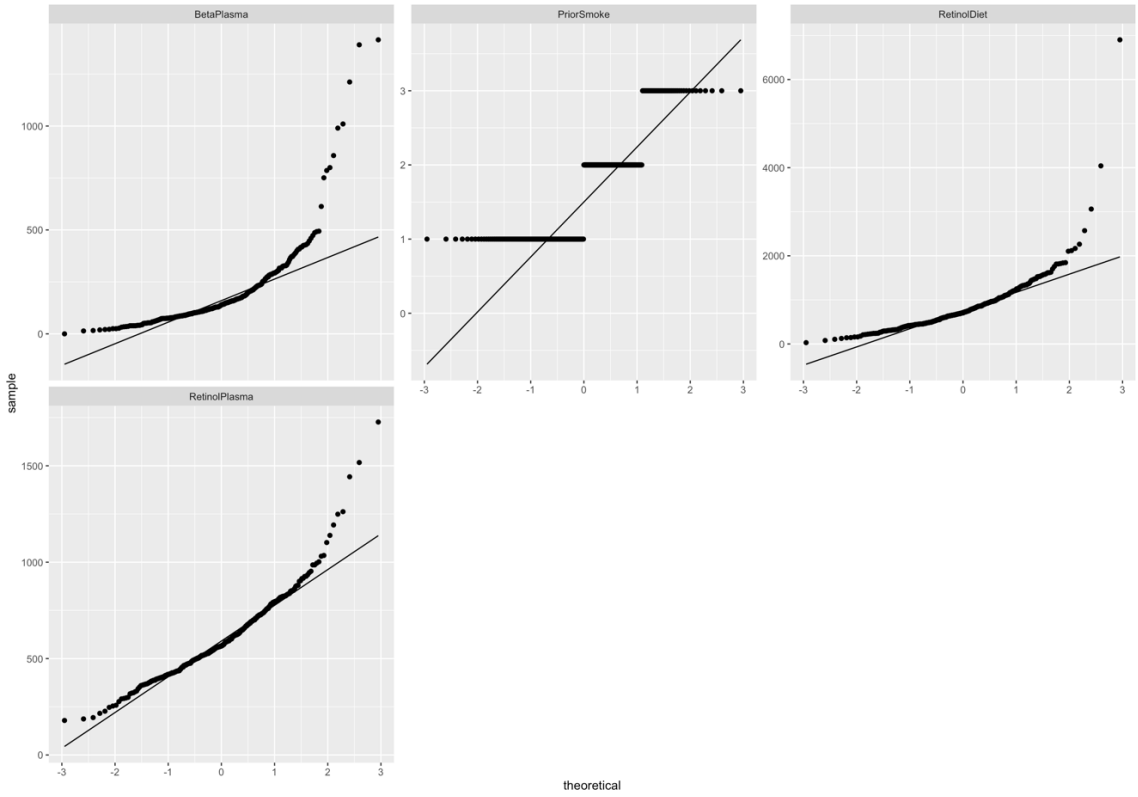


Frequency (Bar Chart)



QQ Plot


Age
Alcohol
BetaDiet
Calories
Cholesterol
Fat
Fiber
ID
Quetelet
sample
theoretical
Page 1


BetaPlasma
PriorSmoke
RetinolDiet
RetinolPlasma
sample
theoretical
Page 2

B: Model 1 Summary Statistics (Cholesterol and Fiber)

```
lm(formula = Cholesterol ~ Fiber, data = mydata)

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 < 0.0000000000000002 ***
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
```

C: Model 2 Summary Statistics (Cholesterol, Fiber with AlocholLevel)

```
lm(formula = Cholesterol ~ Fiber + AlcoholLevel_Some + AlcoholLevel_Alot,
    data = mydata2)

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 < 0.0000000000000002 ***
Fiber               3.984      1.389   2.868              0.00441 **
AlcoholLevel_Some  -2.523     15.836  -0.159              0.87352
AlcoholLevel_Alot  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
```

D: Model 3 Summary Statistics (Fiber, AlcoholLevels with Interactions)

```
lm(formula = Cholesterol ~ Fiber + AlcoholLevel_Some + AlcoholLevel_Alot +
    Fiber_AlcoSome + Fiber_AlcoAlot, data = mydata4)

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 0.00000000000241 ***
Fiber               0.6363     2.3655   0.269            0.788
AlcoholLevel_Some -62.8481    40.5528  -1.550            0.122
AlcoholLevel_Alot -63.3814    85.4549  -0.742            0.459
Fiber_AlcoSome      4.7976     2.9565   1.623            0.106
Fiber_AlcoAlot      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
```

E: Model 4 (Fiber and Smoke) Summary Statistics

```
lm(formula = Cholesterol ~ Fiber + Smoke_Yes, data = mydata5)

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 < 0.0000000000000002 ***
Fiber          4.296      1.394   3.081              0.00224 **
Smoke_Yes     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
```

F: Model 4 (Fiber & Smoke with Interaction) Summary

```
lm(formula = Cholesterol ~ Fiber + Smoke_Yes + Fiber_Smoke_Yes,
    data = mydata5)

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 0.00000000000000447 ***
Fiber             4.455      1.471   3.028              0.00267 **
Smoke_Yes        63.059     55.002   1.146              0.25248
Fiber_Smoke_Yes  -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
```

G: Model 4 (Fiber and Vitamin) Summary Statistics

```
lm(formula = Cholesterol ~ Fiber + Vitamin_Reg + Vitamin_Occ,
    data = mydata5)

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 < 0.0000000000000002 ***
Fiber          3.940      1.393   2.828              0.00498 **
Vitamin_Reg  -14.947     17.259  -0.866              0.38714
Vitamin_Occ   -3.401     19.074  -0.178              0.85861
---
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
```

H: Model 4 (Fiber & Vitamin with Interaction) Summary

```
lm(formula = Cholesterol ~ Fiber + Vitamin_Reg + Vitamin_Occ +
    Fiber_Vitamin_Reg + Fiber_Vitamin_Occ, data = mydata5)

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 0.000000000399 ***
Fiber               3.111      2.454   1.267          0.206
Vitamin_Reg       -29.942     43.947  -0.681          0.496
Vitamin_Occ       -19.453     52.883  -0.368          0.713
Fiber_Vitamin_Reg   1.196      3.188   0.375          0.708
Fiber_Vitamin_Occ   1.300      3.945   0.329          0.742
---
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
```

I: Model 4 (Fiber and Gender) Summary Statistics

```
lm(formula = Cholesterol ~ Fiber + Gender_Male, data = mydata5)

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)  184.490     18.681   9.876 < 0.0000000000000002 ***
Fiber          3.529      1.342   2.629              0.00898 **
Gender_Male   96.294     21.013   4.583           0.00000665 ***
---
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: 0.0000009149
```

J: Model 4 (Fiber & Gender with Interaction) Summary

```
lm(formula = Cholesterol ~ Fiber + Gender_Male + Fiber_Gender_Male,
    data = mydata5)

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)       162.359     19.188   8.462 0.00000000000000105 ***
Fiber               5.273      1.391   3.790            0.000181 ***
Gender_Male       311.514     60.083   5.185 0.00000039029294889 ***
Fiber_Gender_Male -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: 0.000000004028
```

J: Model 4 (Fat and Gender) Summary Statistics

```
lm(formula = Cholesterol ~ Gender_Male + Fat, data = mydata5)

Residuals:
    Min      1Q  Median      3Q     Max
-223.17  -49.65   -9.89   34.80  518.06

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  29.9715    12.9039   2.323              0.02084 *
Gender_Male  46.7640    15.5399   3.009              0.00283 **
Fat           2.6775     0.1564  17.119 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.94 on 312 degrees of freedom
Multiple R-squared:  0.5179,    Adjusted R-squared:  0.5148
F-statistic: 167.6 on 2 and 312 DF,  p-value: < 0.00000000000000022
```

K: Model 4 (Fat & Gender with Interaction) Summary

```
lm(formula = Cholesterol ~ Gender_Male + Fiber_Gender_Male +
    Fat, data = mydata5)

Residuals:
    Min      1Q  Median      3Q     Max
-223.22  -49.67   -9.98   32.81  518.01

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)       30.1251    12.6531   2.381             0.017874 *
Gender_Male      190.0619    41.8763   4.539            0.0000081 ***
Fiber_Gender_Male -10.6795     2.9070  -3.674             0.000281 ***
Fat                2.6754     0.1534  17.445 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90.15 on 311 degrees of freedom
Multiple R-squared:  0.5379,    Adjusted R-squared:  0.5335
F-statistic: 120.7 on 3 and 311 DF,  p-value: < 0.00000000000000022
```