

Computational Assignment #4: OLS Regression Modeling with Continuous
and Categorical Variables
MSDS 410

This fourth computational assignment builds on your prior modeling and computing experiences with assignment #3. You may begin to work on this assignment anytime you wish.

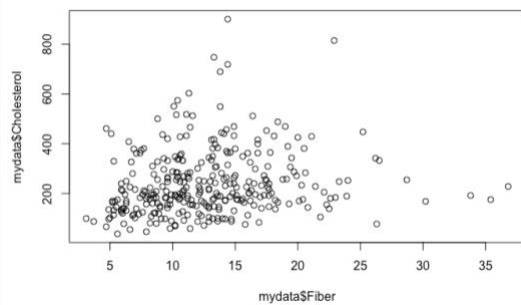
Data: The data for this assignment is the Nutrition Study data: NutritionStudy.CSV It is a 16 variable dataset with n=315 records. The data was obtained from medical record information and observational self-report of adults. The dataset consists of categorical, continuous, and composite scores of different types. A data dictionary is not available for this dataset, but the qualities measured can easily be inferred from the variable and categorical names for most of the variables. As such, higher scores for the composite variables translate into having more of that quality. The QUETELET variable is essentially a body mass index. It can be googled for more detailed information. It is the ratio of BodyWeight (in lbs) divided by (Height (in inch))^2. Then the ratio is adjusted with an adjustment factor so that the numbers become meaningful. Specifically, QUETELET above 25 is considered overweight, while a QUETELET above 30 is considered obese. There is no other information available about this data.

Objective: Use multiple regression to predict CHOLESTEROL using models with continuous and categorical variables. Please note: This assignment is not prescriptive of what you “should do” as an analysis. It is intended to give you experience conducting and reporting on different kinds of multiple regression models.

Tasks: To achieve the objective please complete the following tasks enumerated below. You are to use R to obtain any graphs or statistics requested.

For these analyses, let the response variable be: $Y = \text{CHOLESTEROL}$. The remaining variables will be considered explanatory variables, X 's.

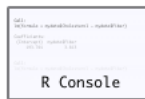
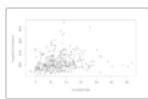
- 1) Consider the continuous variable, FIBER. Is this variable correlated with Cholesterol? Obtain a scatterplot and appropriate statistics to address this question.



```

{r }
plot(mydata$Fiber, mydata$Cholesterol)
model1=lm(mydata$Cholesterol~mydata$Fiber)
model1
summary(model1)
cor(mydata$Fiber,mydata$Cholesterol)

```



```
lm(formula = mydata$Cholesterol ~ mydata$Fiber)
```

```

Coefficients:
(Intercept) mydata$Fiber
193.701      3.813

```

```

Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber)

```

```

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 < 2e-16 ***
mydata$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

```

```
[1] 0.1539684
```

The R^2 value of this model is 0.02371, indicating that Fiber on its own is not a good predictor for cholesterol and that they do not correlate that much.

The correlation coefficient is 0.1539684, which is very low, indicating minimal correlation.

- 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 R^2 value of this model is 0.02371, indicating that Fiber on its own is not a good predictor for cholesterol and that they do not correlate that much.

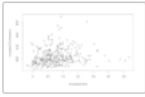

$$Y = 193.701 + 3.813 \cdot B1$$

The 193.701 is the intercept and 3.813 is the slope of the linear model, while B1 is the amount of fiber.

```

##{r}
plot(mydata$Fiber, mydata$Cholesterol)
model1=lm(mydata$Cholesterol~mydata$Fiber)
model1
summary(model1)
cor(mydata$Fiber,mydata$Cholesterol)

```

```

lm(formula = mydata$Cholesterol ~ mydata$Fiber)

Coefficients:
(Intercept)  mydata$Fiber
    193.701      3.813

Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber)

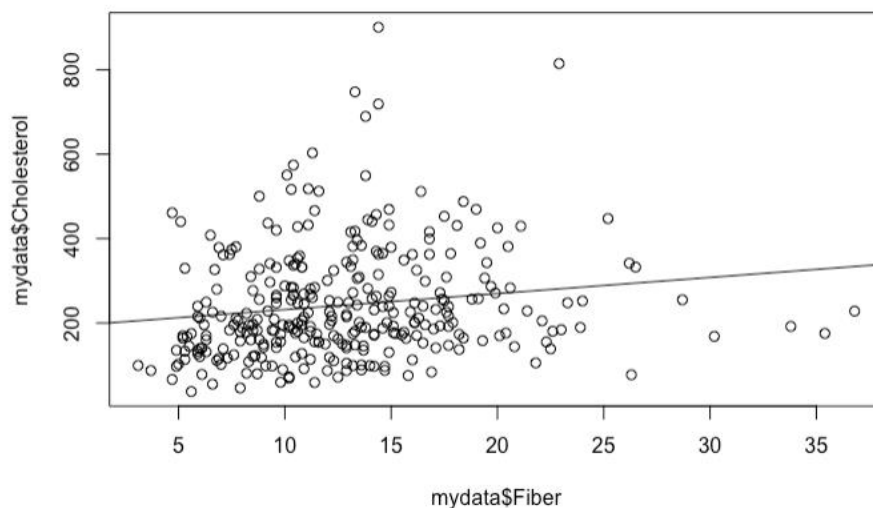
Residuals:
    Min       1Q   Median       3Q      Max
-216.48  -88.58  -34.54   61.18  652.10

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   193.701    19.157   10.111 < 2e-16 ***
mydata$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

[1] 0.1539684

```



- 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)

```

## {r}
model2=lm(mydata$Cholesterol~mydata$Fiber+mydata$AlcoholHi+mydata$AlcoholMid)
model2
summary(model2)

```

Call:
lm(formula = mydata\$Cholesterol ~ mydata\$Fiber + mydata\$AlcoholHi + mydata\$AlcoholMid)

Coefficients:
(Intercept) mydata\$Fiber mydata\$AlcoholHi mydata\$AlcoholMid
189.266 3.984 44.429 -2.523

Call:
lm(formula = mydata\$Cholesterol ~ mydata\$Fiber + mydata\$AlcoholHi + mydata\$AlcoholMid)

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 < 2e-16 ***
mydata\$Fiber 3.984 1.389 2.868 0.00441 **
mydata\$AlcoholHi 44.429 28.429 1.563 0.11912
mydata\$AlcoholMid -2.523 15.836 -0.159 0.87352

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

$$Y = 189.266 + 3.984 \cdot B1 + 44.429 \cdot B2 - 2.523 \cdot B3,$$

With 189.266 being the intercept, 3.984 being the impact of Fiber intake on cholesterol, 44.429 being the impact of High alcohol use (value above 10) and -2.523 being the impact of medium alcohol intake (value between 0 and 10).

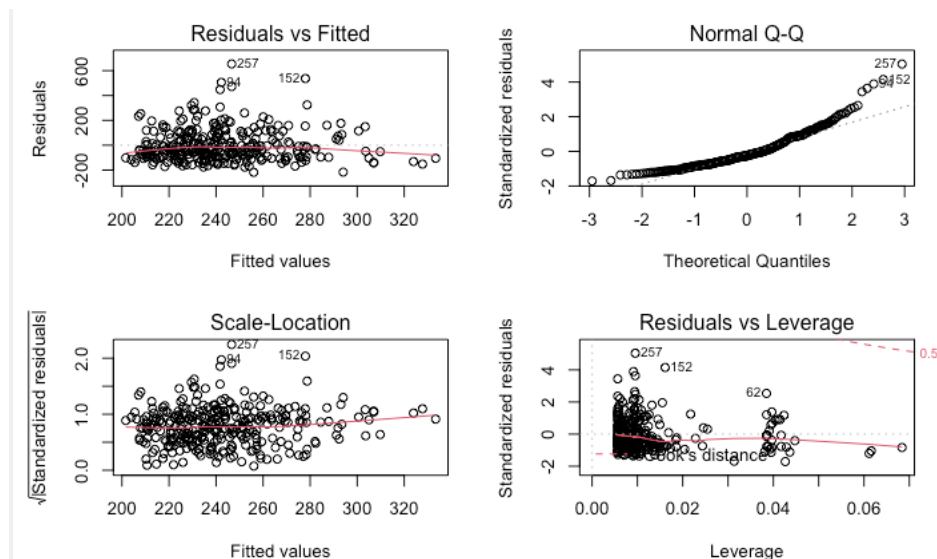
Hypothesis test:

Null: $B1=B2=B3=0$

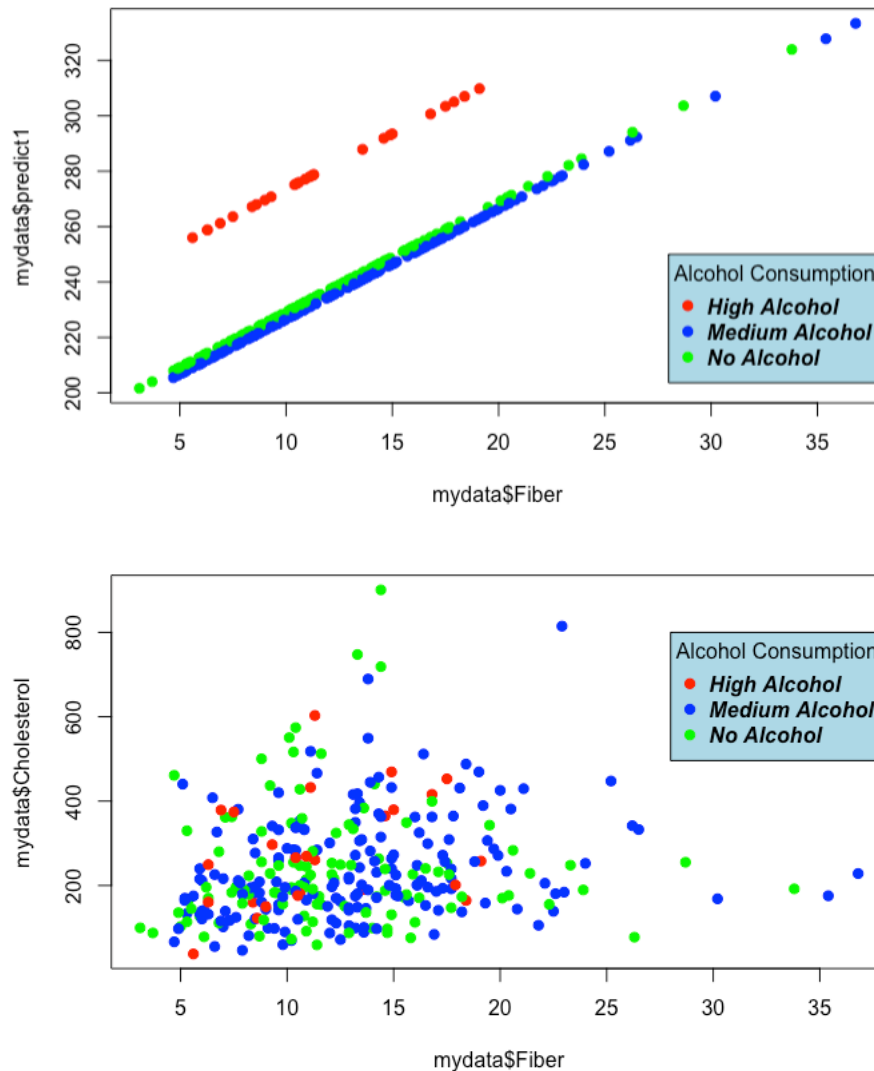
Alt: One of them does not equal 0

99% confidence interval test stat: ~ 1.97

Only the t-value for Fiber is greater than the test statistic, indicating it is significant in this case, while both the t values for the alcohol variables do not, indicating they do not have significant. The p-value for fiber is also the only one < 0.05 supporting the claim that it does not equal zero (rejecting null hypothesis).



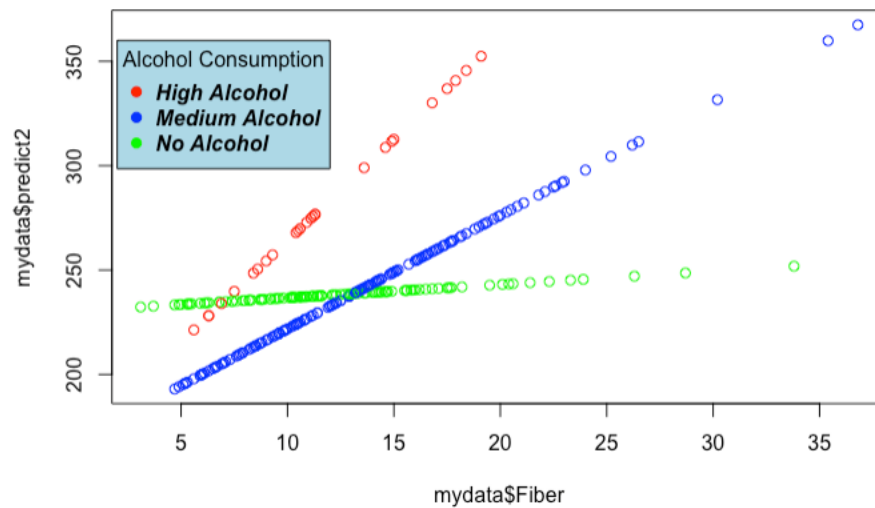
- 4) Use the ANCOVA model from task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FIBER (X), but color code the records for the different groups of ALCOHOL. What do you notice about the patterns in the predicted values of Y? Now, make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



I believe a more complex model is required to extract better meanings and correlations between the data due to linear trends being visible in the predicted model with separation of alcohol consumption.

- 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.



In the graph, I see that there are linear relationships between alcohol consumption and the prediction of cholesterol in individuals. The lower the alcohol consumption, the less cholesterol the individual seems to have in this case.

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber + mydata$AlcoholHi +
    mydata$AlcoholMid + mydata$FiberHiAlc + mydata$FiberMidAlc)

Coefficients:
    (Intercept)      mydata$Fiber  mydata$AlcoholHi  mydata$AlcoholMid  mydata$FiberHiAlc  mydata$FiberMidAlc
      230.3434         0.6363        -63.3814        -62.8481         9.0742         4.7976
```

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber + mydata$AlcoholHi +
    mydata$AlcoholMid + mydata$FiberHiAlc + mydata$FiberMidAlc)
```

```
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 2.41e-12 ***
mydata$Fiber      0.6363     2.3655   0.269  0.788
mydata$AlcoholHi -63.3814    85.4549  -0.742  0.459
mydata$AlcoholMid -62.8481    40.5528  -1.550  0.122
mydata$FiberHiAlc  9.0742     6.8735   1.320  0.188
mydata$FiberMidAlc 4.7976     2.9565   1.623  0.106
---
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
```

Analysis of Variance Table

```
Response: mydata$Cholesterol
              Df Sum Sq Mean Sq F value    Pr(>F)
mydata$Fiber    1 129684  129684  7.6597 0.005987 **
mydata$AlcoholHi 1  50178   50178  2.9637 0.086152 .
mydata$AlcoholMid 1    432     432  0.0255 0.873232
mydata$FiberHiAlc 1  13974   13974  0.8253 0.364332
mydata$FiberMidAlc 1  44582   44582  2.6332 0.105669
Residuals      309 5231592  16931
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Y = 230.3434 + 0.6363*B1 - 63.3814*B2 - 62.8481*B3 + 9.0742*B4 + 4.7976*B5$$

With 230.3434 being the intercept, 0.6363 being the impact of Fiber intake on cholesterol, - 63.3814 being the impact of High alcohol use (value above 10) and - 62.8481 being the impact of medium alcohol intake (value between 0 and 10), as well as 9.0742 being the impact of fiber along with high alcohol consumption and 4.7976 being the impact of fiber along with medium alcohol use.

Hypothesis test:

Null: $B1=B2=B3=B4=B5=0$

Alt: One of them does not equal 0

99% confidence interval test stat: ~ 1.97

None the t-values are greater than the test statistic, indicating it is not significant in this case, except the intercept's p-value. All of their p-values are less than 0.05, indicating no statistical significance. However, when combined, the p-value is 0.01651, which means all variables combined indicate statistical significance. You can see this through the f-statistic.

- 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.

Task 4 has the reduced model while task five has the full model.

Hypothesis test:

Null: $B4=B5=0$

Alt: One of them does not equal 0

Analysis of Variance Table

```
Model 1: mydata$Cholesterol ~ mydata$Fiber + mydata$AlcoholHi + mydata$AlcoholMid
Model 2: mydata$Cholesterol ~ mydata$Fiber + mydata$AlcoholHi + mydata$AlcoholMid +
  mydata$FiberHiAlc + mydata$FiberMidAlc
Res.Df    RSS Df Sum of Sq    F Pr(>F)
1      311 5290147
2      309 5231592    2      58556 1.7293 0.1791
```

Our f-statistic is 1.7293, with a p-value of 0.1791, which is greater than 0.05. This means we cannot reject our null hypothesis that $B4=B5=0$.

- 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.

```

Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber + mydata$Fiber *
  mydata$Male + mydata$Fiber * mydata$SmokeBin + mydata$Fiber *
  mydata$VitaminUseReg + mydata$Fiber * mydata$VitaminUseOcc)

Coefficients:
              (Intercept)              mydata$Fiber              mydata$Male              mydata$SmokeBin              mydata$VitaminUseReg              mydata$VitaminUseOcc
              154.3717              5.1400              313.8157              45.6770              -13.4119              3.6429
mydata$Fiber:mydata$Male      mydata$Fiber:mydata$SmokeBin      mydata$Fiber:mydata$VitaminUseReg      mydata$Fiber:mydata$VitaminUseOcc
              -16.2581              -0.1972              1.0078              0.8004

Residuals:
    Min       1Q   Median       3Q      Max
-287.02  -74.74  -28.00   58.67   671.21

Coefficients:
              (Intercept)              mydata$Fiber              mydata$Male              mydata$SmokeBin              mydata$VitaminUseReg              mydata$VitaminUseOcc
              154.3717              5.1400              313.8157              45.6770              -13.4119              3.6429
mydata$Fiber:mydata$Male      mydata$Fiber:mydata$SmokeBin      mydata$Fiber:mydata$VitaminUseReg      mydata$Fiber:mydata$VitaminUseOcc
              -16.2581              -0.1972              1.0078              0.8004

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 124.1 on 305 degrees of freedom
Multiple R-squared:  0.1413, Adjusted R-squared:  0.1159
F-statistic: 5.574 on 9 and 305 DF, p-value: 3.972e-07

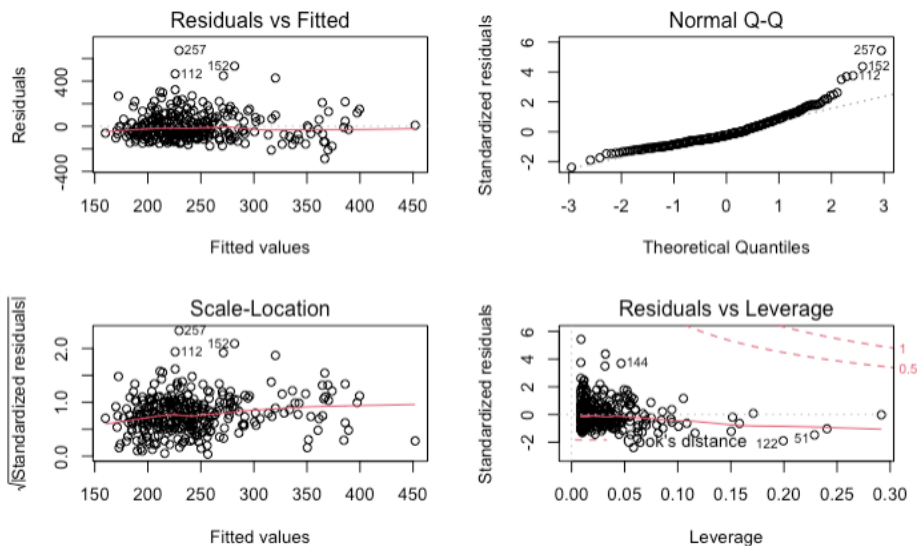
Analysis of Variance Table

Response: mydata$Cholesterol
              Df Sum Sq Mean Sq F value    Pr(>F)
mydata$Fiber      1 129684 129684  8.4198 0.0039819 ***
mydata$Male       1 336804 336804 21.8671 4.395e-06 ***
mydata$SmokeBin   1  62711  62711  4.0715 0.044868 *
mydata$VitaminUseReg  1  2488   2488  0.1615 0.6880280
mydata$VitaminUseOcc 1  8128   8128  0.5277 0.4681211
mydata$Fiber:mydata$Male 1 231103 231103 15.0044 0.0001313 ***
mydata$Fiber:mydata$SmokeBin 1  72    72  0.0047 0.9456514
mydata$Fiber:mydata$VitaminUseReg 1 1060 1060  0.0688 0.7931978
mydata$Fiber:mydata$VitaminUseOcc 1  689  689  0.0447 0.8326086
Residuals      305 4697702 15402
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The most significant variables in this model seem to be (from least significant to greatest) Smoking, Fiber, and equally being a Male and Fiber*Male.

This is because when conducting hypothesis testing, they have t-values greater than the threshold and very low p-values, meaning that we can reject the null hypothesis that their variable is equal to zero.




```

Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber * mydata$VitaminUseReg +
    mydata$Fiber * mydata$VitaminUseOcc)

Coefficients:
            (Intercept)              mydata$Fiber      mydata$VitaminUseReg      mydata$VitaminUseOcc      mydata$Fiber:mydata$VitaminUseReg      mydata$Fiber:mydata$VitaminUseOcc
                208.821                  3.111                 -29.942                 -19.453                  1.196                  1.300

Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber * mydata$VitaminUseReg +
    mydata$Fiber * mydata$VitaminUseOcc)

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 3.99e-10 ***
mydata$Fiber      3.111     2.454   1.267   0.206
mydata$VitaminUseReg -29.942    43.947  -0.681   0.496
mydata$VitaminUseOcc -19.453    52.883  -0.368   0.713
mydata$Fiber:mydata$VitaminUseReg  1.196     3.188   0.375   0.708
mydata$Fiber:mydata$VitaminUseOcc  1.300     3.945   0.329   0.742
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

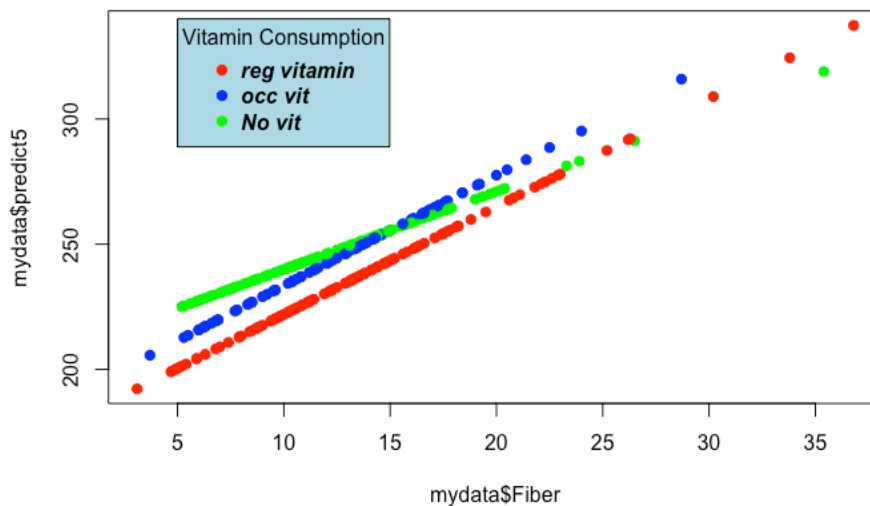
Residual standard error: 131.3 on 309 degrees of freedom
Multiple R-squared:  0.02681, Adjusted R-squared:  0.01106
F-statistic: 1.702 on 5 and 309 DF, p-value: 0.1338

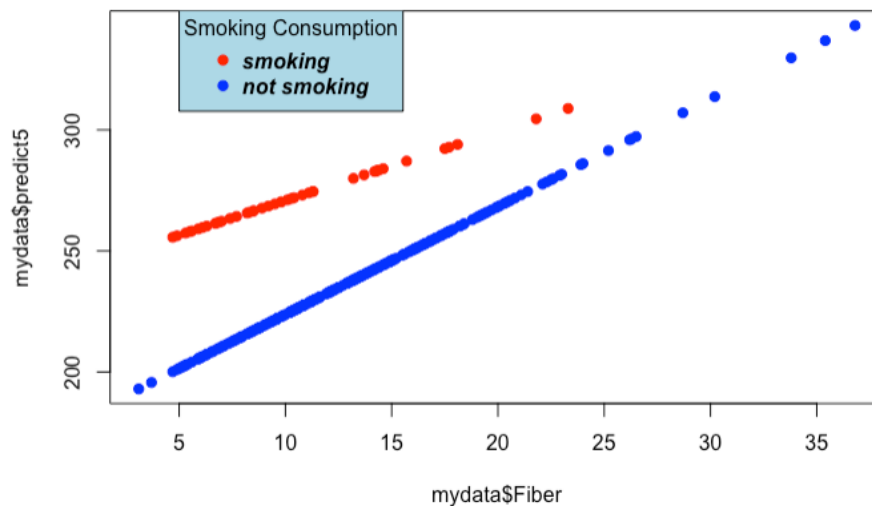
Analysis of Variance Table

Response: mydata$Cholesterol
              Df Sum Sq Mean Sq F value    Pr(>F)
mydata$Fiber    1 129684 129684  7.5270 0.006433 **
mydata$VitaminUseReg 1 13482 13482  0.7825 0.377055
mydata$VitaminUseOcc 1  544    544  0.0316 0.859021
mydata$Fiber:mydata$VitaminUseReg 1 1057 1057  0.0613 0.804578
mydata$Fiber:mydata$VitaminUseOcc 1 1870 1870  0.1085 0.742071
Residuals      309 5323804 17229
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

On their own, the vitamin use variables multiplied by fiber does not have statistically significant evidence to reject the null hypothesis, with both t and f tests.





Smoking shows to increase one's cholesterol as well. But people who do not smoke tend to have higher levels of fiber as well. However, the more fiber they have, the more likely they are to have higher cholesterol.

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber * mydata$SmokeBin)
```

```
Coefficients:
            (Intercept)              mydata$Fiber          mydata$SmokeBin mydata$Fiber:mydata$SmokeBin
               179.184                  4.455                 63.059                 -1.597
```

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber * mydata$SmokeBin)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-218.86  -87.71  -35.15   65.11  657.36
```

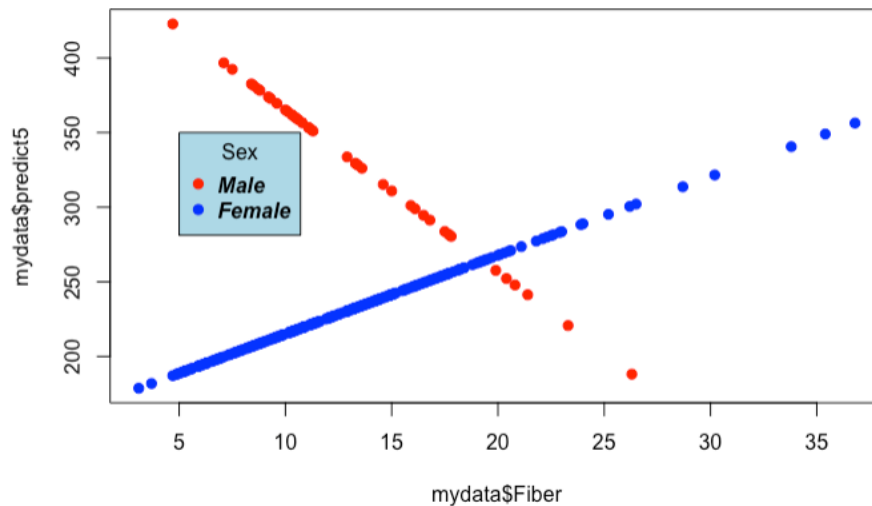
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    179.184    20.875   8.583 4.47e-16 ***
mydata$Fiber     4.455     1.471   3.028 0.00267 **
mydata$SmokeBin  63.059    55.002   1.146 0.25248
mydata$Fiber:mydata$SmokeBin -1.597     4.661  -0.343 0.73218
---
Signif. codes:  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
```

Analysis of Variance Table

```
Response: mydata$Cholesterol
              Df Sum Sq Mean Sq F value    Pr(>F)
mydata$Fiber    1  129684   129684   7.6630 0.005975 **
mydata$SmokeBin  1    75590    75590   4.4666 0.035360 *
mydata$Fiber:mydata$SmokeBin  1    1986    1986   0.1173 0.732179
Residuals      311 5263182   16923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once again, fiber is the only significantly significant variable out of the three. However, combined they have a low p-value indicating that a relation between smoking and cholesterol is present as well.



This was the most interesting relationship. As females increase their fiber intake, their cholesterol goes down, which is the opposite of men.

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber * mydata$Male)

Coefficients:
      (Intercept)          mydata$Fiber          mydata$Male mydata$Fiber:mydata$Male
            162.359              5.273             311.514             -16.138
```

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$Fiber * mydata$Male)
```

```
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 1.05e-15 ***
mydata$Fiber      5.273     1.391   3.790 0.000181 ***
mydata$Male     311.514    60.083   5.185 3.90e-07 ***
mydata$Fiber:mydata$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: 4.028e-09
```

Analysis of Variance Table

```
Response: mydata$Cholesterol
      Df Sum Sq Mean Sq F value    Pr(>F)
mydata$Fiber    1 129684 129684  8.4367 0.0039408 **
mydata$Male      1 336804 336804 21.9110 4.27e-06 ***
mydata$Fiber:mydata$Male 1 223427 223427 14.5352 0.0001659 ***
Residuals    311 4780527 15371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All variables appear to be statistically significant as well! They have t-values that are high, and low p-values, stating that the null hypothesis of their variables being equal to zero can be rejected.

8) Please write a reflection on your experiences.

Something surprising to me is that being a Male significantly increases your chances of having higher cholesterol with a higher fiber intake, and the complete reverse is the case for Females. This assignment has allowed me to become more familiar with creating linear models, and how sometimes

a simple linear model may not explain relationships well, and may in fact hide some in some cases. This is why more complex methods may be necessary to find correlations and causations, as they may not be so clear at first.

- 9) Extra Credit: Feel free to explore models that have other continuous variables, as well as interactions of categorical variables. The more you do, the more extra credit you can accumulate.