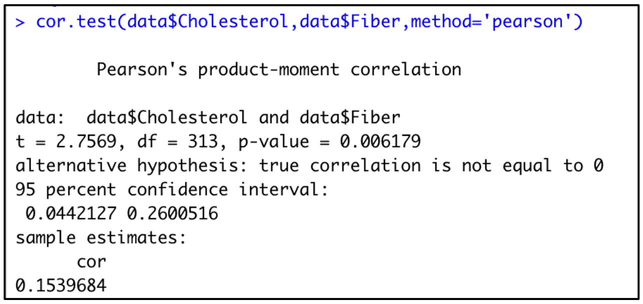
Pooja Deshpande

Computational Assignment 4
MSDS – 410 Data Modeling for Supervised Learning,
Summer 2020
Northwestern University

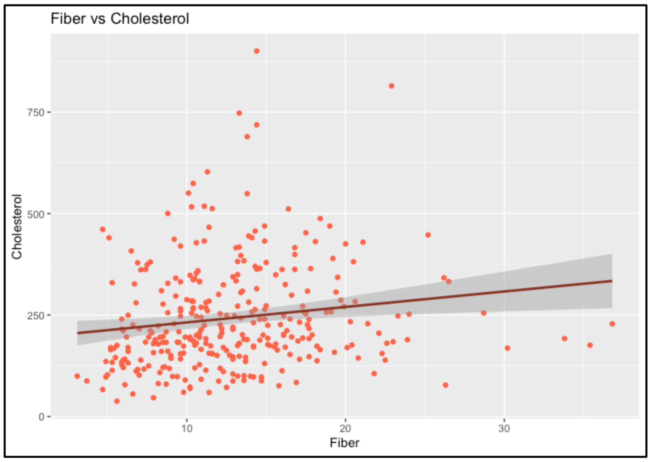
Contents:

Q.1

There is a slight positive correlation between *Fiber* and *Cholesterol*. The correlation coefficient, r = 0.1539

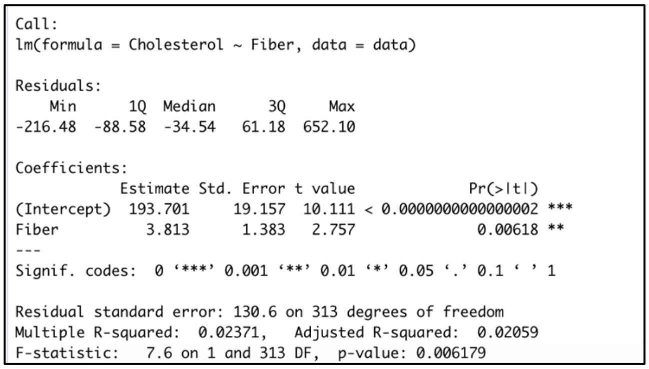


```
> cor.test(data$Cholesterol,data$Fiber,method='pearson')

	Pearson's product-moment correlation

data:  data$Cholesterol and data$Fiber
t = 2.7569, df = 313, p-value = 0.006179
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0442127 0.2600516
sample estimates:
      cor 
0.1539684 
```

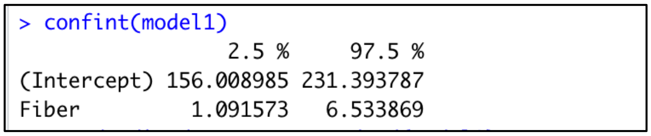


Q.2

Model 1 for *Cholesterol* with *Fiber* as the Regressor:



```
Call:
lm(formula = Cholesterol ~ Fiber, data = data)

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
```



```
> confint(model1)
                 2.5 %     97.5 %
(Intercept) 156.008985 231.393787
Fiber         1.091573   6.533869
```

Model equation and Interpretation:

*Cholesterol = 193.701 + 3.813*Fiber*

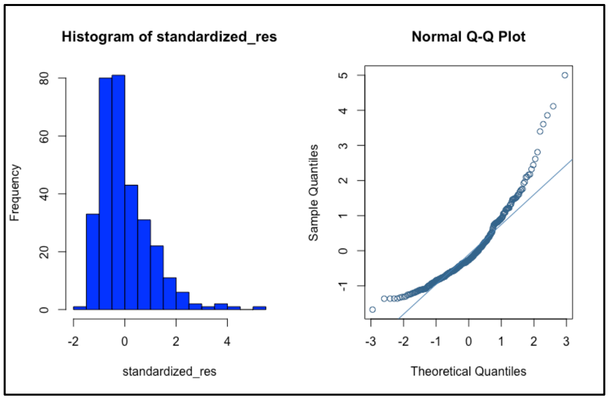
For every unit increase in fiber, the cholesterol of the subject increases by 3.813.

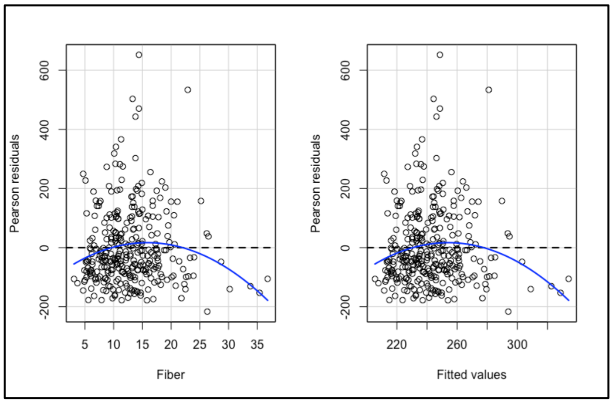
The p-value associated with the coefficient estimate is 0.00618 which suggests that there is a significant relationship between fiber and cholesterol.

R-square: 0.02371

2.371% of the variance in cholesterol can be explained by fiber alone.

Residual plots:

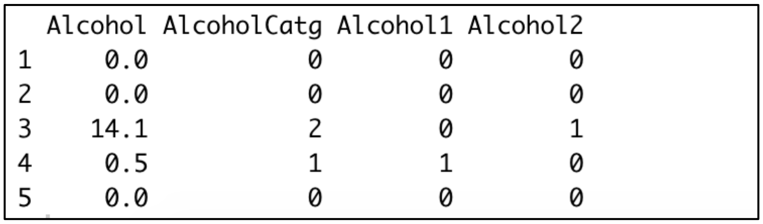




Residuals are non-normal, positively skewed. There is presence of heteroscedasticity.

Q.3

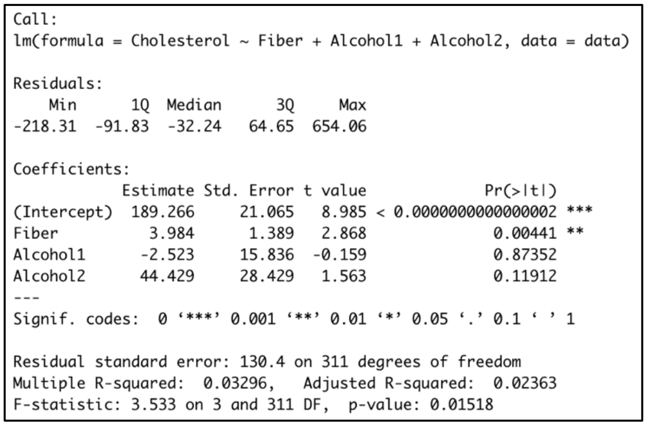
Dummy coded *Alcohol* categorical variables:



```
  Alcohol AlcoholCatg Alcohol1 Alcohol2
1     0.0           0        0        0
2     0.0           0        0        0
3    14.1           2        0        1
4     0.5           1        1        0
5     0.0           0        0        0
```

	Category	Dummy coded Alcohol1	Dummy coded Alcohol2
Alcohol == 0	0	0	0
Alcohol>0 & Alcohol < 10	1	1	0
Alcohol >= 10	2	0	1

Model 2 for *Cholesterol* with *Fiber* and dummy coded *Alcohol* categorical variables:



```
Call:
lm(formula = Cholesterol ~ Fiber + Alcohol1 + Alcohol2, data = data)

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 ** 
Alcohol1      -2.523     15.836  -0.159             0.87352    
Alcohol2      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
```

Model equations and Interpretation:

Cholesterol = 189.266 + 3.984*Fiber – 2.523*Alcohol1 + 44.429*Alcohol2

Estimated average cholesterol when Alcohol category is 0:

Cholesterol = 189.266 + 3.984*Fiber

Estimated average cholesterol when Alcohol is between 0 and 10:

Cholesterol = 186.743 + 3.984*Fiber

Estimated average cholesterol when Alcohol is equal or above 10:

Cholesterol = 233.695 + 3.984*Fiber

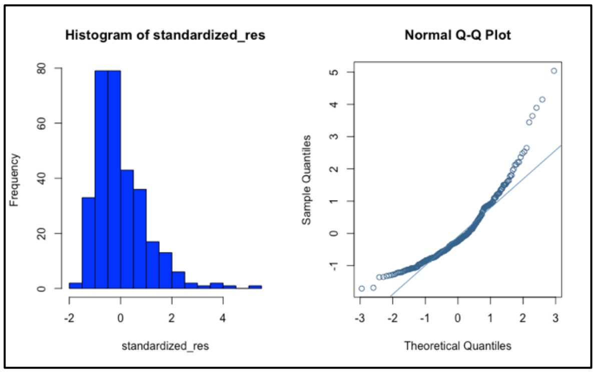
The above model is a same slopes, different intercepts model.

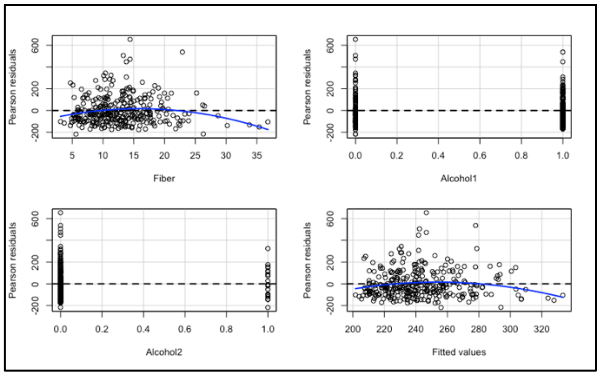
The p-values for Alcohol1 and Alcohol2 are higher than 0.05 which suggests that there isn’t enough evidence to conclude that there is a significant relationship between these variables and cholesterol.

R-square: 0.03296.

3.296% of the variance in cholesterol is explained by fiber and alcohol categories.

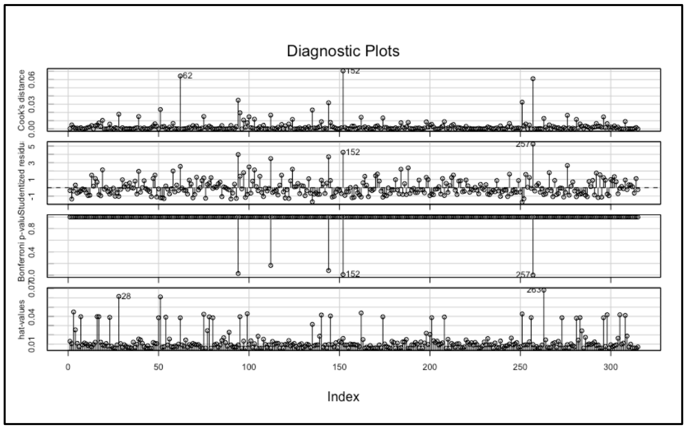
Residual plots:





Residuals are not normal. Evidence of heteroscedasticity.

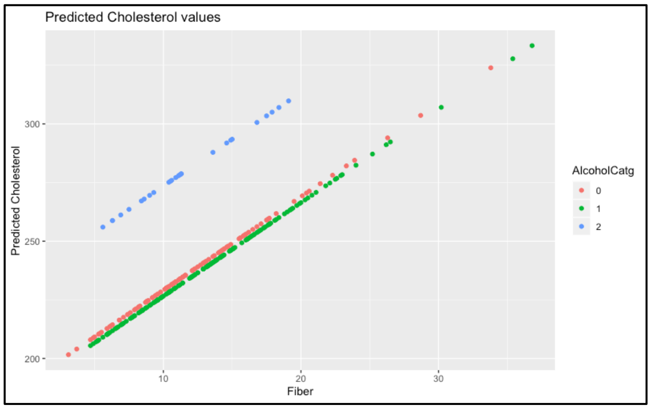
Influential points:



Observations 152, 62 highly influential. Observations 263, 28 have high leverage values.

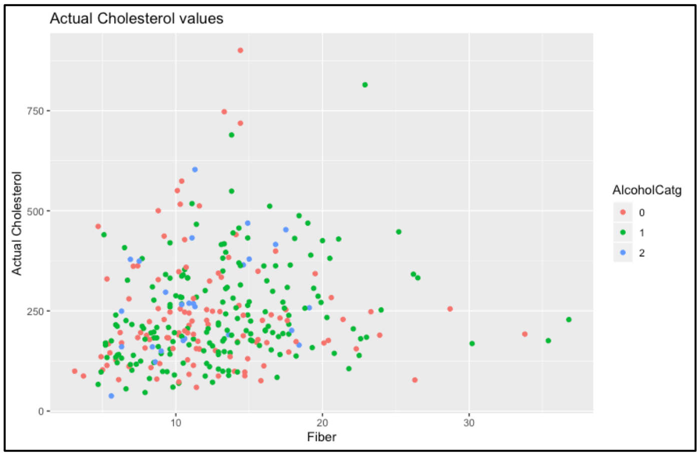
Q.4

Scatter plot of predicted values of Cholesterol, regressor Fiber, color coded by the Alcohol category:



Since we built a *"same slopes, different intercepts"* model in Q.3, we can clearly observe from the plot that for each alcohol category, the predicted values lie along the same line. There is a clear distinction between the predicted values of Cholesterol for all three alcohol categories.

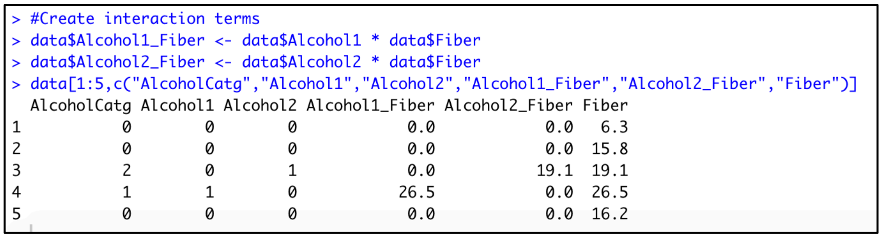
Scatter plot of Actual Cholesterol values:



The plot of the actual cholesterol values with regressor fiber and color coded by alcohol category seems to be more scattered and complex. There is no clear separation between the cholesterol values for the different alcohol categories. In order to accurately fit the data, the model needs to be more complicated.

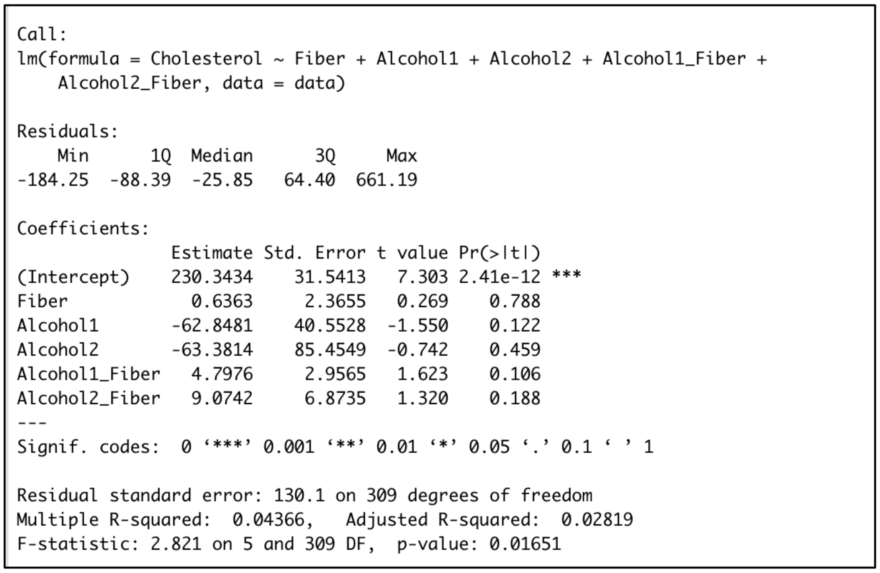
Q.5

Creating interaction variables:



```
> #Create interaction terms
> data$Alcohol1_Fiber <- data$Alcohol1 * data$Fiber
> data$Alcohol2_Fiber <- data$Alcohol2 * data$Fiber
> data[1:5,c("AlcoholCatg","Alcohol1","Alcohol2","Alcohol1_Fiber","Alcohol2_Fiber","Fiber")]
  AlcoholCatg Alcohol1 Alcohol2 Alcohol1_Fiber Alcohol2_Fiber Fiber
1           0        0        0            0.0            0.0   6.3
2           0        0        0            0.0            0.0  15.8
3           2        0        1            0.0           19.1  19.1
4           1        1        0           26.5            0.0  26.5
5           0        0        0            0.0            0.0  16.2
```

Model with interaction terms:



```
Call:
lm(formula = Cholesterol ~ Fiber + Alcohol1 + Alcohol2 + Alcohol1_Fiber +
    Alcohol2_Fiber, data = 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)    230.3434    31.5413   7.303 2.41e-12 ***
Fiber            0.6363     2.3655   0.269    0.788
Alcohol1       -62.8481    40.5528  -1.550    0.122
Alcohol2       -63.3814    85.4549  -0.742    0.459
Alcohol1_Fiber   4.7976     2.9565   1.623    0.106
Alcohol2_Fiber   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
```

Model equation:

*Cholesterol = 230.3434 + 0.6363*Fiber - 62.8481*Alcohol1 – 63.3814*Alcohol2 + 4.7976*Alcohol1_Fiber + 9.0742*Alcohol2_Fiber*

Estimated average cholesterol where alcohol is zero:

*Cholesterol = 230.3434 + 0.6363*Fiber*

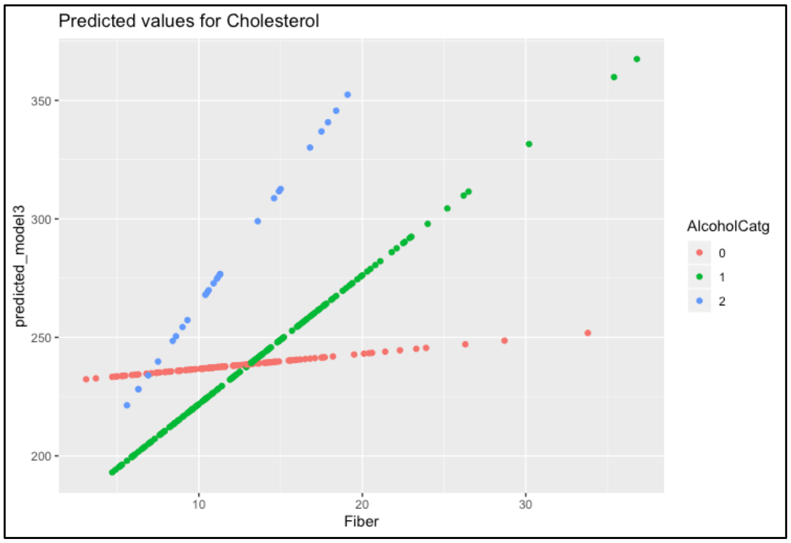
Estimated average cholesterol where alcohol is between 0 and 10:

*Cholesterol = 167.4953 + 0.6363*Fiber + 4.7976*Alcohol1_Fiber*

Estimated average cholesterol where alcohol is greater than 10:

*Cholesterol = 166.962 + 0.6363*Fiber + 9.0742*Alcohol2_Fiber*

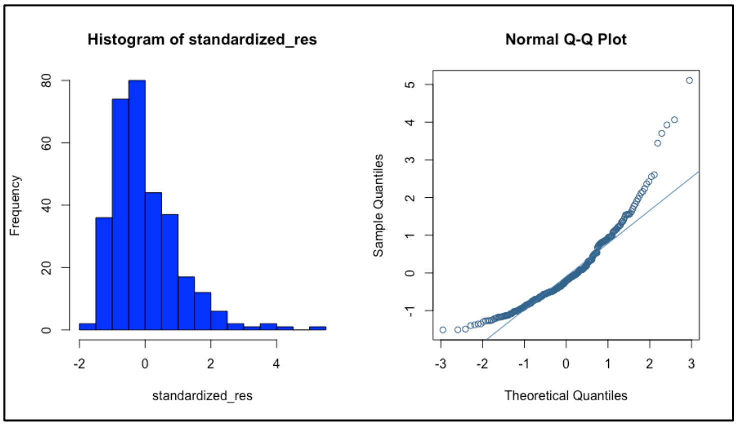
This is a model with different intercepts and different slopes for each alcohol category. The model's t-test reveals that none of the predictors have a significant relationship with the response variable. However, F-test statistic reveals that the combination of these predictors has a significant relationship with the response variable.

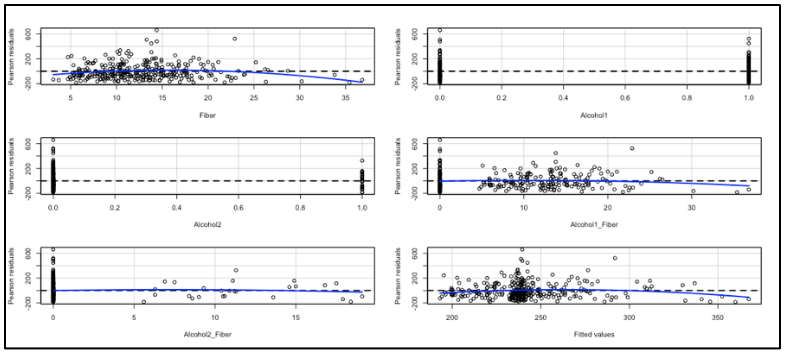


R-square: 0.04366.

4.366% of the variance in Cholesterol can be explained by Fiber, Alcohol Catg and the interactions between them.

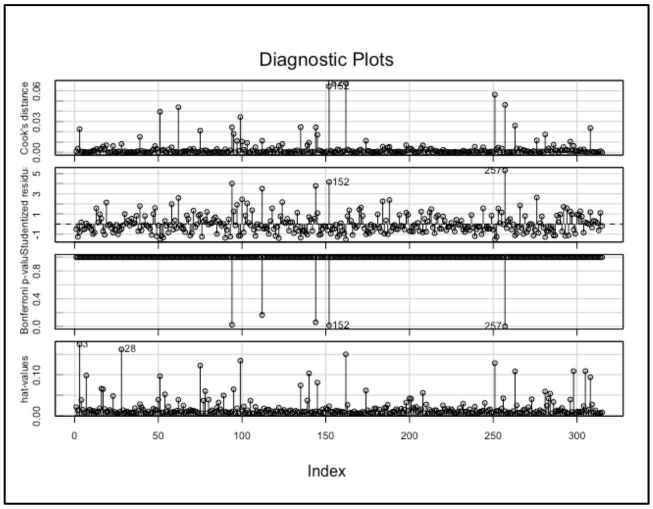
Residual plots:





Residuals are positively skewed and there is evidence of violation of homoscedasticity.

Influential points:



Observations 62, 152 are influential/outliers, 3 and 28 have high leverage values.

Q. 6

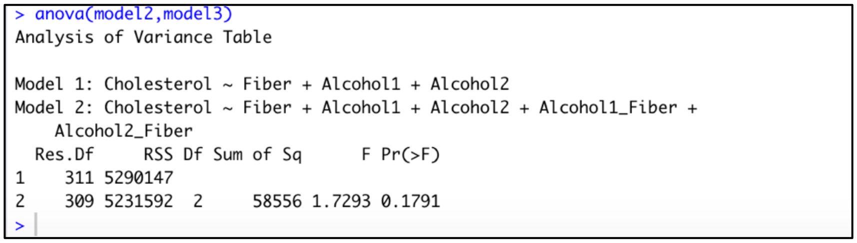
Models 2 and 3 are nested. The model without the interaction terms is the reduced/restricted model and the one with the interaction terms is the full model.

Ho: Reduced model is adequate

or

Coefficients of the interaction terms, *Alcohol1_Fiber* and *Alcohol2_Fiber* are zero or the slopes of model 1 and model 2 are equal.

Ha: Full model is adequate. Slopes of model 1 and model 2 are unequal.

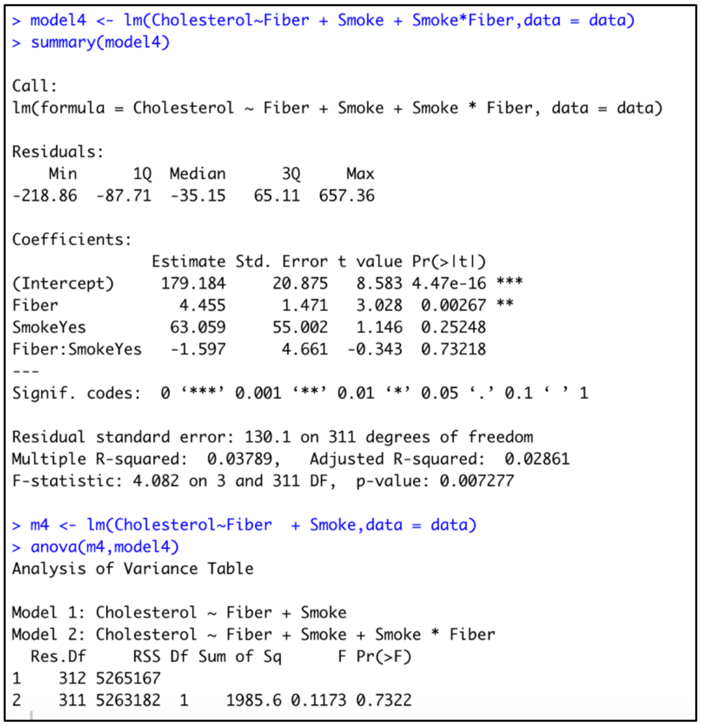


```
> anova(model2,model3)
Analysis of Variance Table

Model 1: Cholesterol ~ Fiber + Alcohol1 + Alcohol2
Model 2: Cholesterol ~ Fiber + Alcohol1 + Alcohol2 + Alcohol1_Fiber +
    Alcohol2_Fiber
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    311 5290147
2    309 5231592  2     58556 1.7293 0.1791
>
```

F-test results suggest that the reduced model is adequate and that the interaction terms are insignificant. A model without unequal slopes is adequate.

Q.7

Building models for Cholesterol using Fiber and each of the following categorical variables: Smoke, VitaminUse, Gender



```
> model4 <- lm(Cholesterol~Fiber + Smoke + Smoke*Fiber,data = data)
> summary(model4)

Call:
lm(formula = Cholesterol ~ Fiber + Smoke + Smoke * Fiber, data = data)

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 ***
Fiber             4.455      1.471   3.028  0.00267 ** 
SmokeYes         63.059     55.002   1.146  0.25248    
Fiber:SmokeYes   -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

> m4 <- lm(Cholesterol~Fiber  + Smoke,data = data)
> anova(m4,model4)
Analysis of Variance Table

Model 1: Cholesterol ~ Fiber + Smoke
Model 2: Cholesterol ~ Fiber + Smoke + Smoke * Fiber
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    312 5265167                           
2    311 5263182  1    1985.6 0.1173 0.7322
```

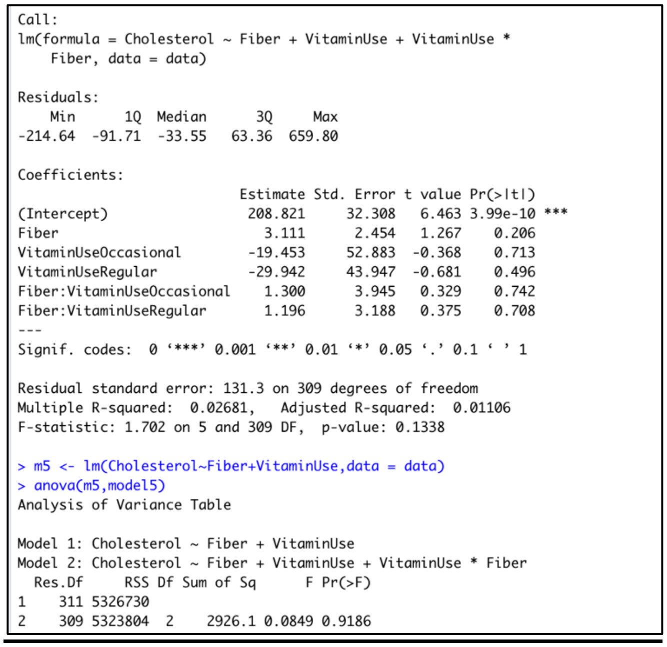
Ho: Reduced model is adequate

or

Coefficients of the interaction term *SmokeYes*Fiber* is zero, slopes of both the models are equal

Ha: Full model is adequate, slopes of both the models are unequal.

p-value: 0.7322 suggests that the reduced model is adequate. No significant interaction between Smoke and Fiber.



```
Call:
lm(formula = Cholesterol ~ Fiber + VitaminUse + VitaminUse *
    Fiber, data = data)

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 ***
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
---
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

> m5 <- lm(Cholesterol~Fiber+VitaminUse,data = data)
> anova(m5,model5)
Analysis of Variance Table

Model 1: Cholesterol ~ Fiber + VitaminUse
Model 2: Cholesterol ~ Fiber + VitaminUse + VitaminUse * Fiber
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    311 5326730
2    309 5323804  2    2926.1 0.0849 0.9186
```

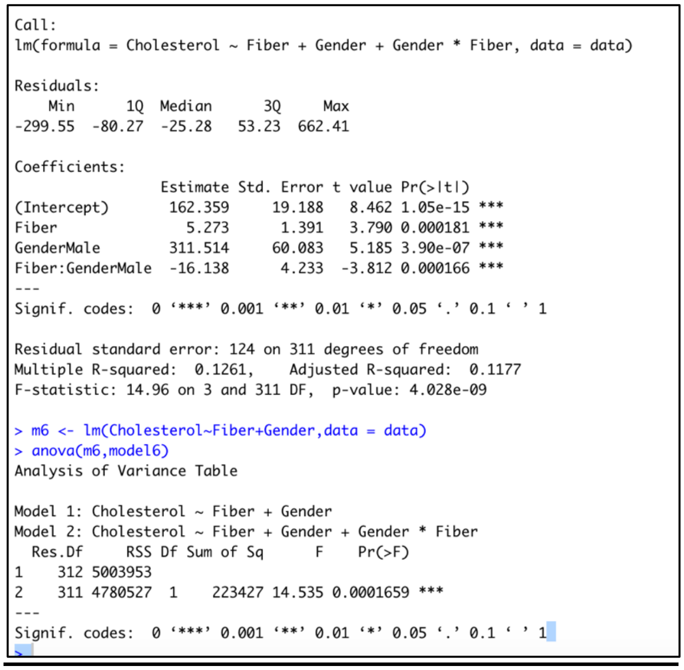
Ho: Reduced model is adequate

or

Coefficients of the interaction term *VitaminUse*Fiber* is zero, slopes of both the models are equal.

Ha: Full model is adequate, slopes of both the models are unequal.

p-value: 0.9186 suggests that the reduced model is adequate. No significant interaction between VitaminUse and Fiber.



```
Call:
lm(formula = Cholesterol ~ Fiber + Gender + Gender * Fiber, data = data)

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 ***
Fiber               5.273      1.391   3.790 0.000181 ***
GenderMale        311.514     60.083   5.185 3.90e-07 ***
Fiber:GenderMale  -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

> m6 <- lm(Cholesterol~Fiber+Gender,data = data)
> anova(m6,model6)
Analysis of Variance Table

Model 1: Cholesterol ~ Fiber + Gender
Model 2: Cholesterol ~ Fiber + Gender + Gender * Fiber
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
1    312 5003953                                  
2    311 4780527  1    223427 14.535 0.0001659 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Ho: Reduced model is adequate

or

Coefficients of the interaction term *Gender*Fiber* is zero, slopes of both the models are equal.

Ha: Full model is adequate, slopes of both the models are unequal.

p-value: 0.000165 suggests that the full model is adequate. There is significant interaction between Gender and Fiber. Different intercepts and different slopes model are required when considering both Gender and Fiber.

Q.8

This assignment helped me understand that sometimes we need to engineer a new feature when there is evidence that the interaction between two features might produce various outcomes. In the above problem, it was found that there is significant interaction between *Fiber* and *Gender* while determining *Cholesterol.* One can always start with a simple model, examine the results and then develop complex models with additional features to finally find a model that best fits the data. It may be a good idea to begin with a different intercept model and then move to different intercept and different slopes model. On a side note, I recently worked at Amaral Lab, Northwestern where we built similar hierarchical models to determine if an individual's state membership has an effect on his final voting decision.