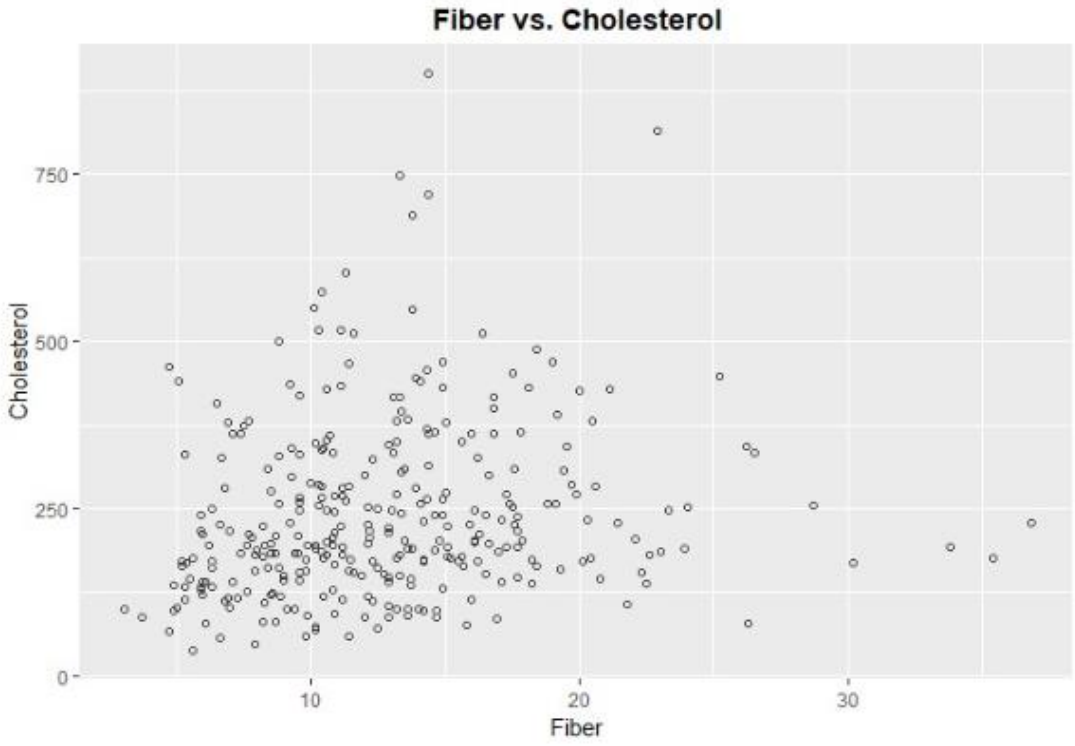
Assignment # 6

May 2021

1)



Correlation = .1540

After viewing the results, there does not seem to be a strong correlation between these two variables, but this might change once we add more variables. The slope value is positive meaning every 1 unit intake of fiber, the predicted cholesterol levels are expected to increase.

2)

Y_hat = 193.701 + 3.813X1

The intercept at 0 is y = -194.

R-squared = .02371, meaning only 2.371% of variability is explained in the model.

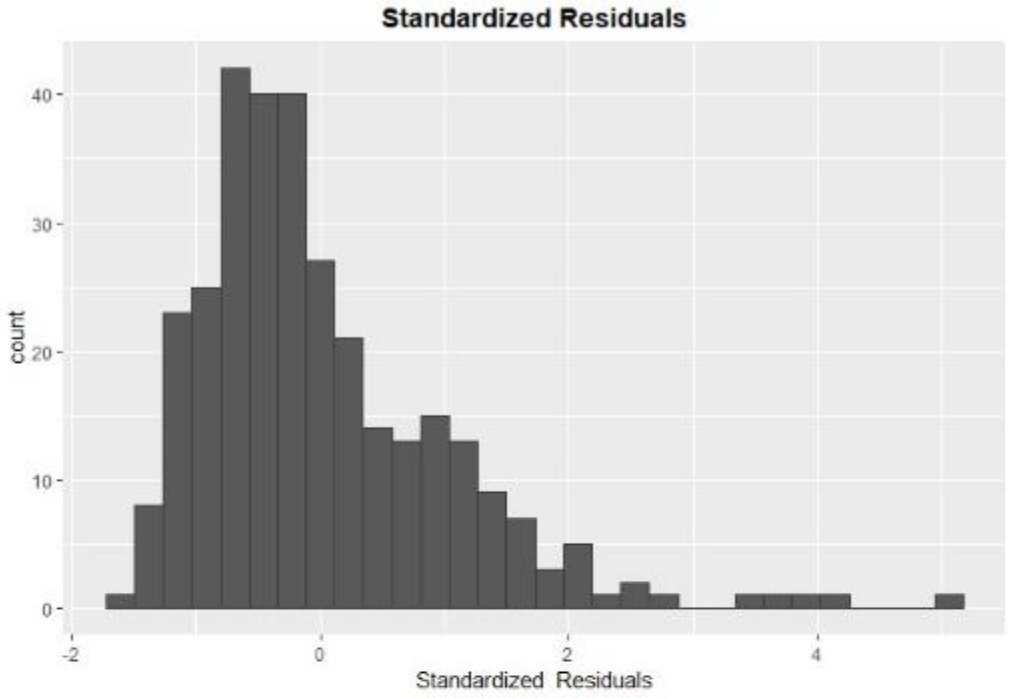
Coefficients Model:

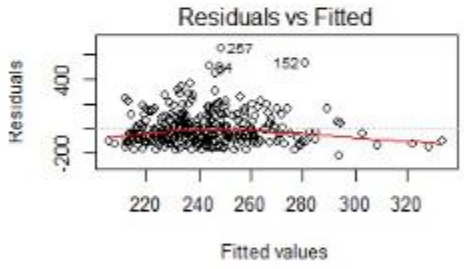
Coefficients:	**Estimate**	**Std. Error**	**T-value**	**Pr(>|t|)**
(Intercept)	193.701	19.157	10.111	< 2e-16
Fiber	3.813	1.383	2.757	0.00618

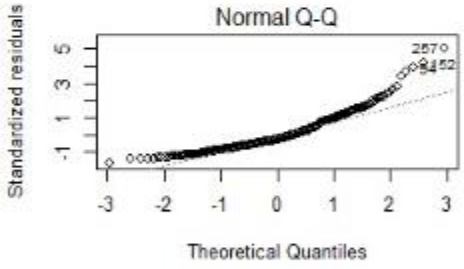
Anova Table:

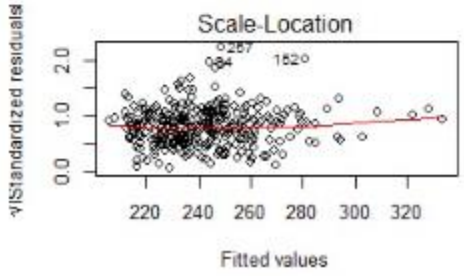
Response: Cholesterol	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
Fiber	1	129684	129684	7.6002	0.006179
Residuals	313	5340757	17063		

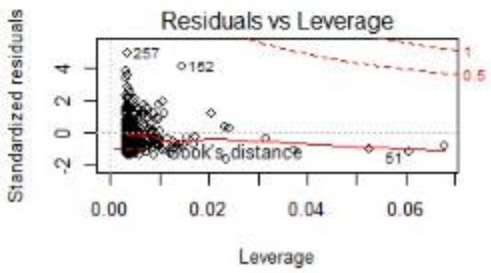
Underlying Fit:











The models has a right skew in standardized residuals and in the Q-Q plot meaning the model is not a good fit.

3)

Y_hat = 190.284 + 4.021X1 -10.371X2 + 43.238X3

The intercept at 0 is y = -191. From the results, it seems both fiber and regular alcohol usage have a positive relationship with cholesterol, while occasional alcohol usage has a negative relationship, with a 1 unit increase in fiber leading to increase in cholesterol in 4.021, 1 unit increase in occasional alcohol usage decreases cholesterol by 10.371, 1 unit increase in regular alcohol usage increases cholesterol level by 43.328.

R-squared = .0342, meaning only 3.42% of variability is explained in the model.

Coefficients Model:

Coefficients:	**Estimate**	**Std. Error**	**T-value**	**Pr(>|t|)**
(Intercept)	190.824	19.895	9.592	< 2e-16
Fiber	4.021	1.386	2.902	0.00397
A_2	-10.371	15.926	-0.651	0.51540
A_3	43.238	27.066	1.597	0.11117

Anova table:

Response: Cholesterol	**Degrees of Freedom**	**Sum Sq**	**Mean Sq**	**F-Value**	**PR(>F)**
Fiber	1	129684	129684	7.6337	0.00607
A_2	1	14028	14028	0.8257	0.36421
A_3	1	43354	43354	2.5520	0.11117
Residuals	311	5283375	16988		

T-test for B1/ Fiber

H0: B1 = 0 vs Ha:B1 !=0

T-statistic = 2.902, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Occasional Alcohol Usage

H0: B2 = 0 vs Ha:B2 !=0

T-statistic = -.651, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

T-test for B3/ Regular Alcohol Usage

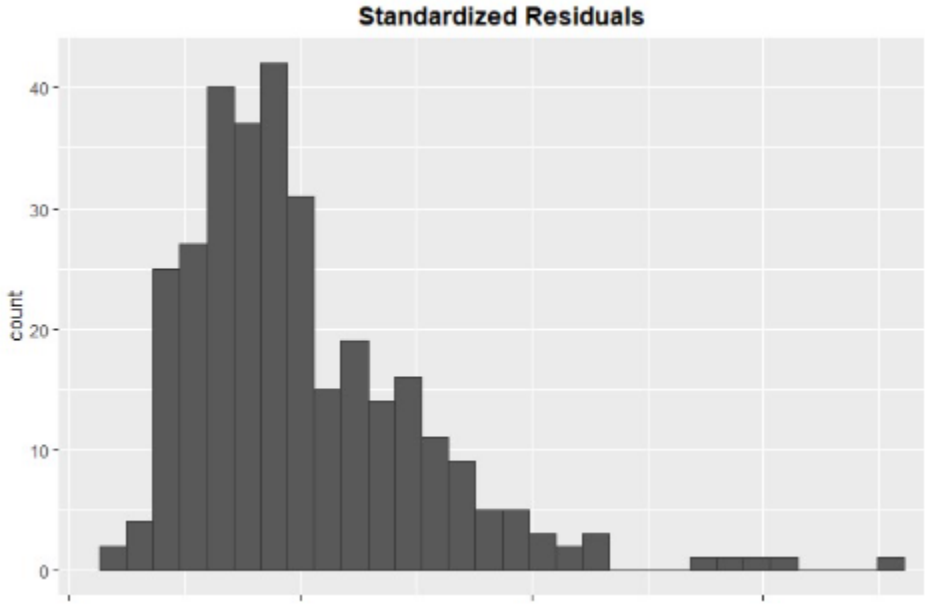
H0: B3 = 0 vs Ha:B3 !=0

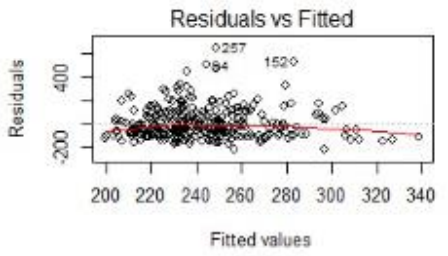
T-statistic = -.571, meaning we would reject the null hypothesis and the variable is not significant to the overall model

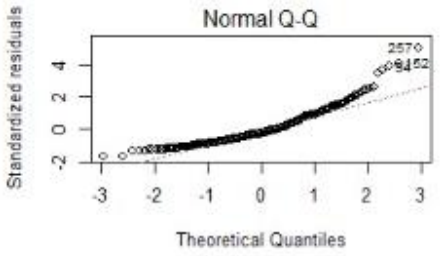
F-Test

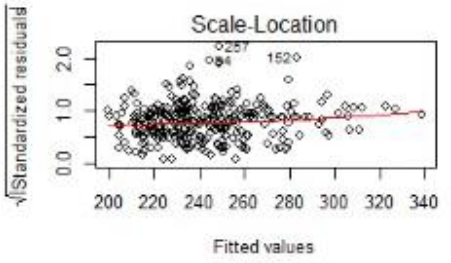
H0:B1 = B2 = B3 = 0 vs Ha: B1 = 0 for E(1, 2, 3)

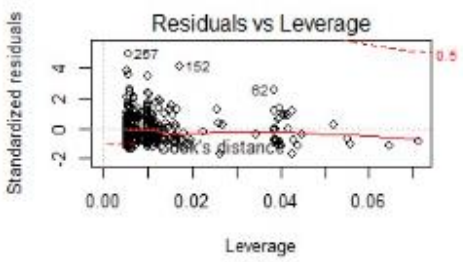
F-statistic = 3.67, meaning we can reject the null hypothesis and state that the model is not a better fit than an intercept only model.





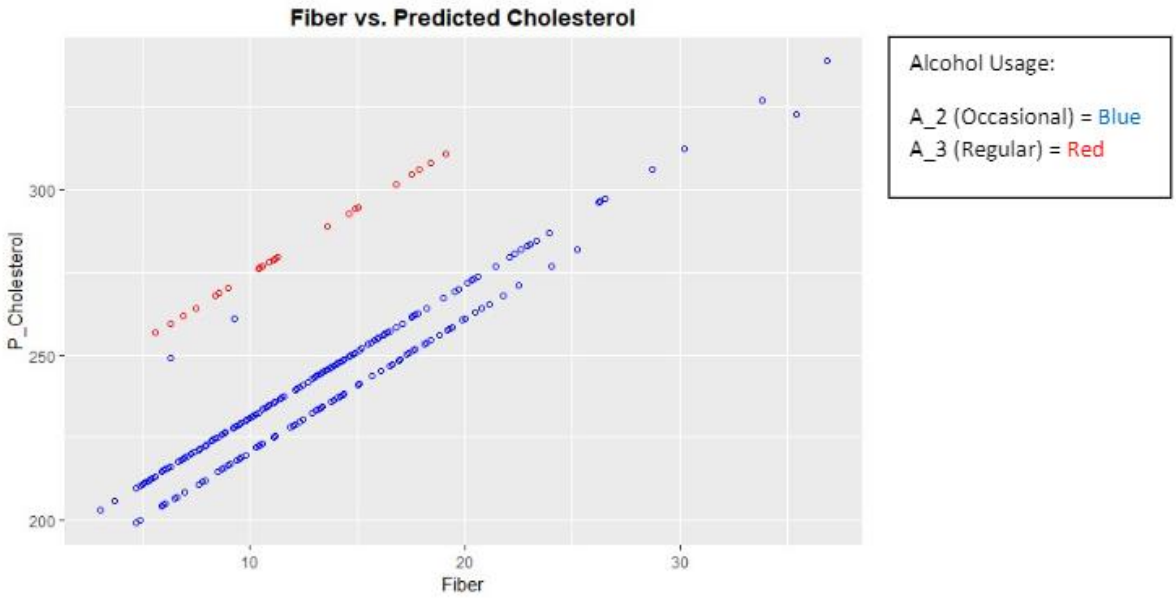






The models has a right skew in standardized residuals and in the Q-Q plot meaning the model is not a good fit.

4)



The results show the base level for regular alcohol usage is greater than that of occasional alcohol usage which comes as no surprise, with regular usage at 250 whole occasional alcohol is at 200. This difference in cholesterol only increases as fiber usage increases but there are two notable outliers.

R-squared = .03972, meaning only 3.972% of variability is explained in the model.

Coefficients Model:

Coefficients:	**Estimate**	**Std. Error**	**T-value**	**Pr(>|t|)**
(Intercept)	205.753	23.342	8.815	< 2e-16
Fiber	2.807	1.701	1.651	0.0998
A_2	-49.511	42.698	-1.160	0.2471
A_3	-30.767	82.270	-0.374	0.7087
Int2	3.064	3.034	1.010	0.3133
Int3	6.384	6.660	0.959	0.3385

Anova Table:

Response: Cholesterol	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
Fiber	1	129684	129684	7.6283	0.00609
A_2	1	14028	14028	0.8252	0.36438
A_3	1	43354	43354	2.5502	0.11130
Int2	1	14621	14621	0.8600	0.35446
Int3	1	15621	15621	0.9188	0.33853
Residuals	309	5253134	17000		

T-test for B1/ Fiber

H0: B1 = 0 vs Ha:B1 !=0

T-statistic = 1.651, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Occasional Alcohol Usage

H0: B2 = 0 vs Ha:B2 !=0

T-statistic = -1.651, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

T-test for B3/ Regular Alcohol Usage

H0: B3 = 0 vs Ha:B3 !=0

T-statistic = -.374, meaning we would reject the null hypothesis and the variable is not significant to the overall model

T-test for B4/ Fiber * Occasional Alcohol Usage

H0: B4 = 0 vs Ha:B4 !=0

T-statistic = 1.010, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B5/ Fiber * Regular Alcohol Usage

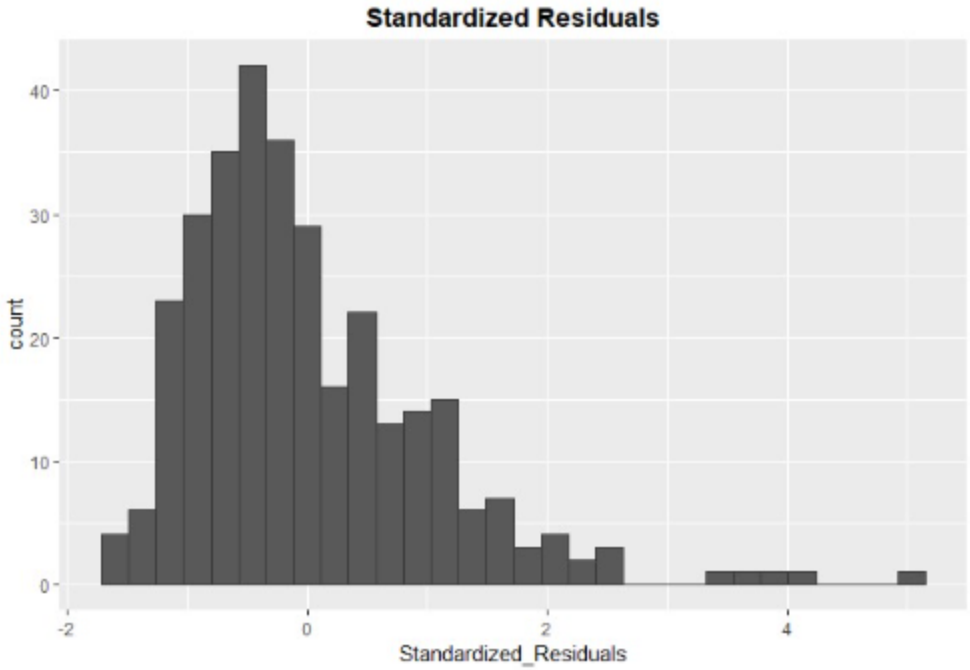
H0: B5 = 0 vs Ha:B5 !=0

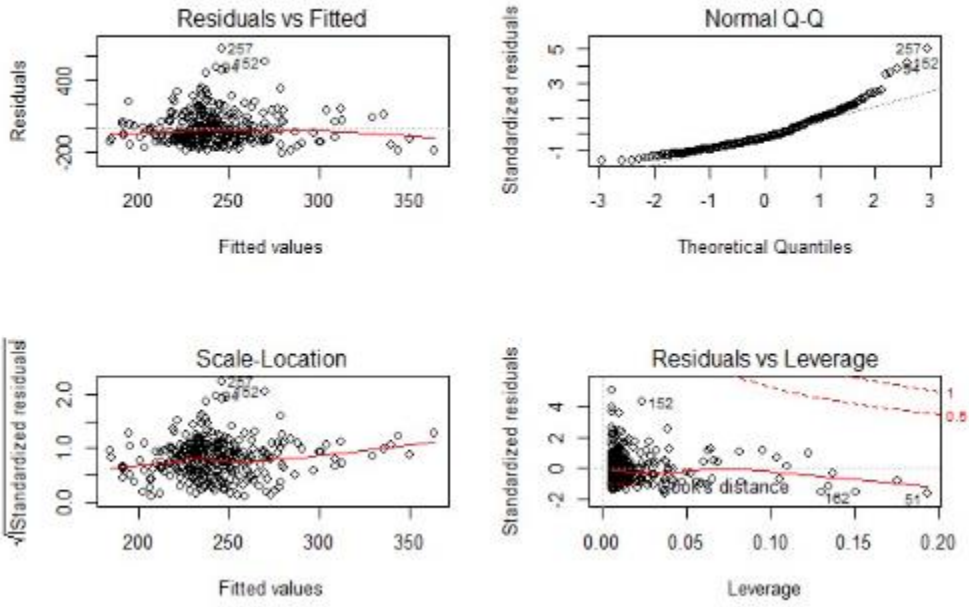
T-statistic = .959, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

F-Test

H0:B1 = B2 = B3 = B4 = B5 = 0 vs Ha: B1 = 0 for E(1, 2, 3, 4, 5)

F-statistic = 2.556, meaning we can reject the null hypothesis and state that the model is not a better fit than an intercept only model.





The models has a right skew in standardized residuals and in the Q-Q plot meaning the model is not a good fit.

6)

From the results, it shows the model of fiber and alcohol is nested within the model of fiber, alcohol, and each interaction because fiber and alcohol are included within bullet point 5 with the additional variables being the interactions.

F-Test

H0:B4 = B5 0 vs Ha: B1 ! = 0 for E(4, 5)

Nested F Test: ((SSE1 – SSE2)/(p2-p1)/((SSE2)/(n-p2)) = ((5283375 – 523134)/(6-4)/((523134)/(315-6)) = 1.204, meaning we cannot reject the null hypothesis which means the additional interaction variables do not contribute additional info about the association between cholesterol and the set of 5 predictors. Based on the results, there are unequal slopes after adding the two new interaction variables to the model.

7)

Smoke:

Y_hat = 179.184 + 4.55X1 + 63.059X@ - 1.1597X3

The interaction between fiber and smoking is 1.1597 meaning the interaction increases the slope, and thus increases the predicted cholesterol values at a higher rate than fiber without smoking.

R-squared = .03789, meaning only 3.789% of variability is explained in the model.

Coefficients Model:

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
S_2	63.059	55.002	1.146	0.25248
I_2	-1.597	4.661	-0.343	0.73218

Anova Table:

Response: ***Cholesterol***	**Degrees of Freedom**	**Sum Sq**	**Mean Sq**	**F-Value**	**PR(>F)**
Fiber	1	129684	129684	7.6630	0.005975
S_2	1	75590	75590	4.4666	0.035360
I_2	1	1986	1986	0.1173	0.732179
Residuals	311	5263182	16923		

T-test for B1/ Fiber

H0: B1 = 0 vs Ha:B1 !=0

T-statistic = 3.208, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Occasional Alcohol Usage

H0: B2 = 0 vs Ha:B2 !=0

T-statistic = 1.146, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

T-test for B3/ Regular Alcohol Usage

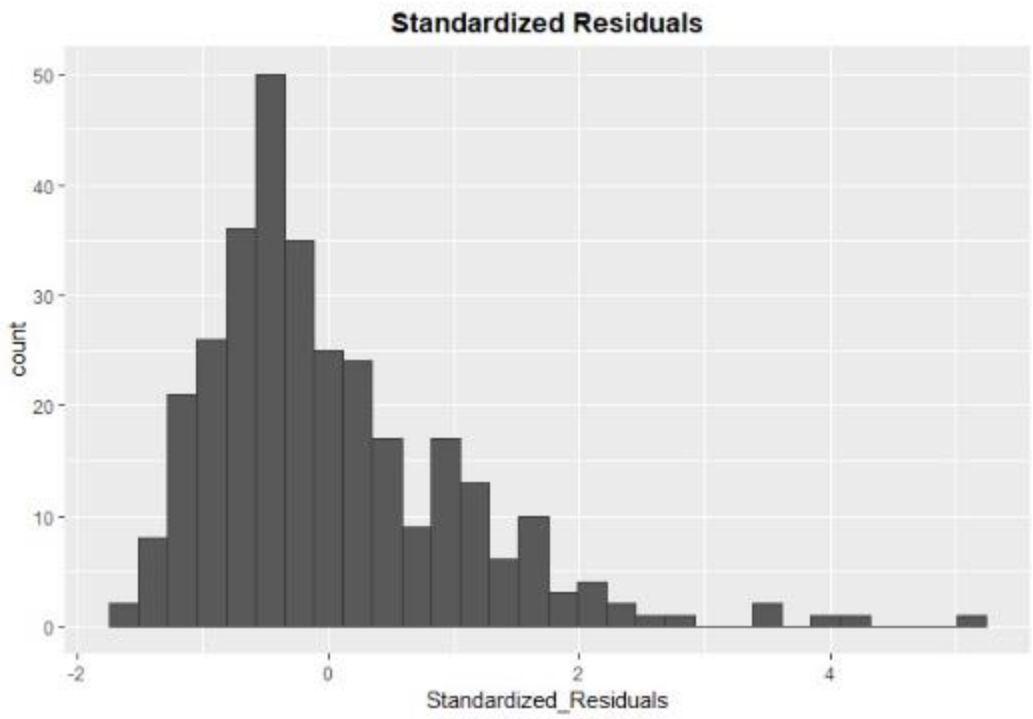
H0: B3 = 0 vs Ha:B3 !=0

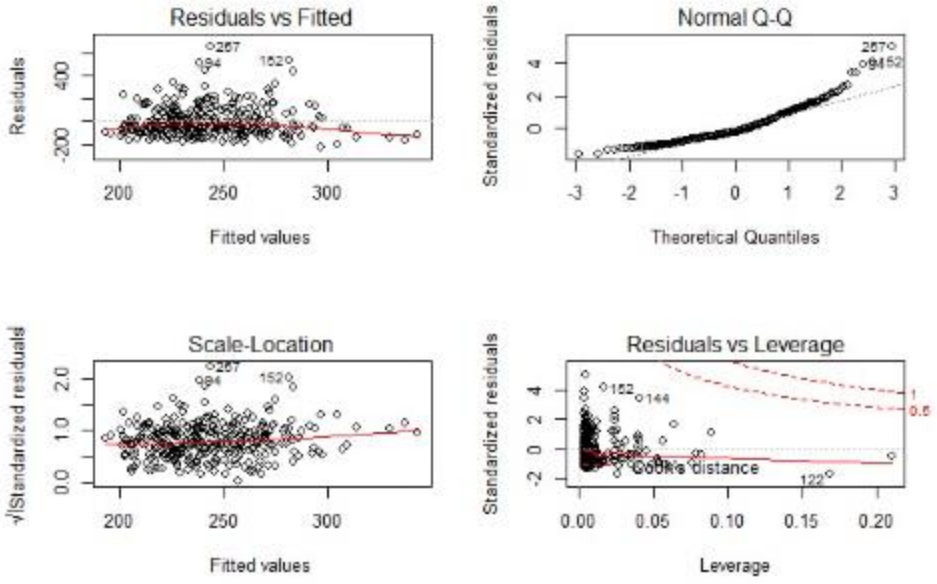
T-statistic = .1173, meaning we would reject the null hypothesis and the variable is not significant to the overall model

F-Test

H0:B1 = B2 = B3 = 0 vs Ha: B1 = 0 for E(1, 2, 3)

F-statistic = 4.208, meaning we can reject the null hypothesis and state that the model is not a better fit than an intercept only model.





The models has a right skew in standardized residuals and in the Q-Q plot meaning the model is not a good fit.

Vitamins:

Y_hat = 208.821 + 3.111X1 – 19.453X2 – 29.942X3 + 1.3X4 + 1.196X5

The interaction between fiber and occasional vitamins is 1.3 meaning the interaction increases the slope, and thus increases the predicted cholesterol values at a higher rate than fiber without occasional

vitamins. The interaction between fiber and regular vitamins is 1.196 meaning the interaction increases the slope, and thus increases the predicted cholesterol values at a higher rate than fiber without regular vitamins.

R-squared = .02681, meaning only 2.681% of variability is explained in the model.

Coefficients Model:

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
V_O	-19.453	52.883	-0.368	0.713
V_R	-29.942	43.947	-0.681	0.496
I_2	1.300	3.945	0.329	0.742
I_3	1.196	3.188	0.375	0.708

Anova Table:

Response: Cholesterol	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
Fiber	1	129684	129684	7.5270	0.006433
V_O	1	1181	1181	0.0686	0.793631
V_R	1	12846	12846	0.7456	0.388547
I_2	1	501	501	0.0291	0.864663
I_3	1	2425	2425	0.1407	0.707808
Residuals	309	5323804	17229		

T-test for B1/ Fiber

H0: B1 = 0 vs Ha:B1 !=0

T-statistic = 1.267, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Occasional Alcohol Usage

H0: B2 = 0 vs Ha:B2 !=0

T-statistic = -.368, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

T-test for B3/ Regular Alcohol Usage

H0: B3 = 0 vs Ha:B3 !=0

T-statistic = -.681, meaning we would reject the null hypothesis and the variable is not significant to the overall model

T-test for B4/ Fiber * Occasional Alcohol Usage

H0: B4 = 0 vs Ha:B4 !=0

T-statistic = .329, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B5/ Fiber * Regular Alcohol Usage

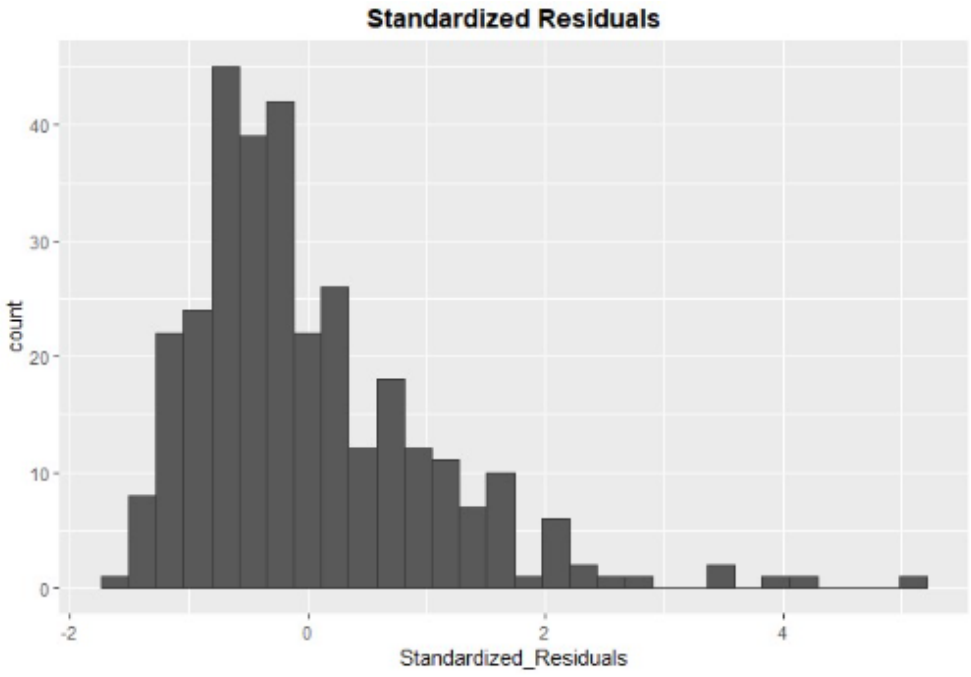
H0: B5 = 0 vs Ha:B5 !=0

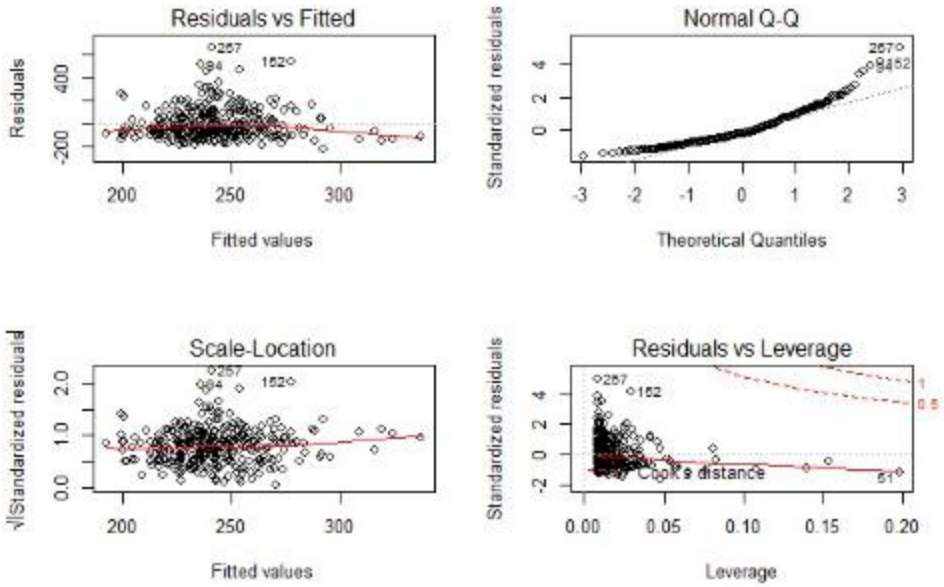
T-statistic = .375, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

F-Test

H0:B1 = B2 = B3 = B4 = B5 = 0 vs Ha: B1 = 0 for E(1, 2, 3, 4, 5)

F-statistic = 1.702, meaning we can reject the null hypothesis and state that the model is not a better fit than an intercept only model.





The models has a right skew in standardized residuals and in the Q-Q plot meaning the model is not a good fit.

Gender:

Y_hat = 473.873 – 10.865X1 -311.514X2 + 16.138X3

The interaction between fiber and the gender category of female with a 16.138 beta value which increases the slope meaning there is a large interaction between cholesterol and gender.

R-squared = .1261, meaning only 12.61% of variability is explained in the model.

Coefficients Model:

Coefficients:	**Estimate**	**Std. Error**	**T-value**	**Pr(>|t|)**
(Intercept)	473.873	56.936	9.323	2.75e-15
Fiber	-10.865	3.998	-2.718	0.006939
G_2	-311.514	60.083	-5.185	3.90e-07
I_2	16.138	4.233	3.812	0.000166

Anova table:

Response: Cholesterol	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
Fiber	1	129684	129684	8.4367	0.0039408
G_2	1	336804	336804	21.9110	4.27e-06
I_2	1	223427	223427	14.5352	0.0001659
Residuals	311	4780527	15371		

T-test for B1/ Fiber

H0: B1 = 0 vs Ha:B1 !=0

T-statistic = -2.718, meaning we would reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Female

H0: B2 = 0 vs Ha:B2 !=0

T-statistic = -5.185, meaning we would not reject the null hypothesis and the variable is not significant to the overall model

T-test for B3/ Fiber * Females

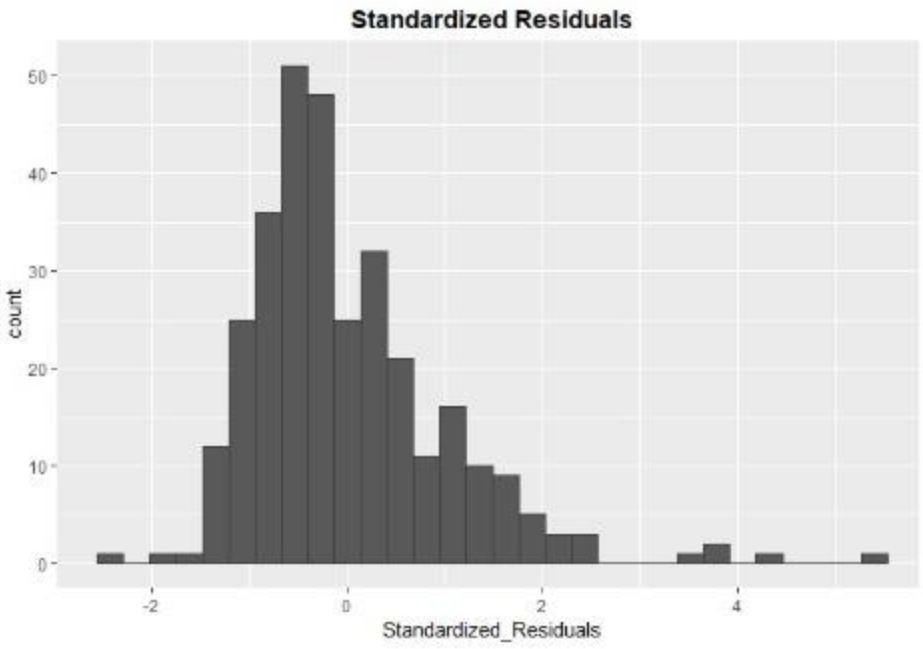
H0: B3 = 0 vs Ha:B3 !=0

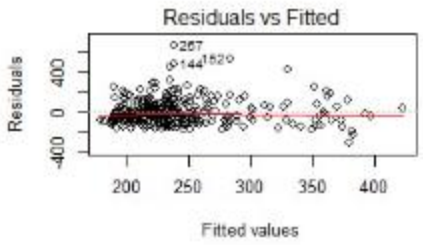
T-statistic = 3.812, meaning we would reject the null hypothesis and the variable is not significant to the overall model

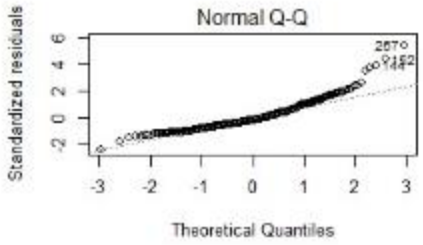
F-Test

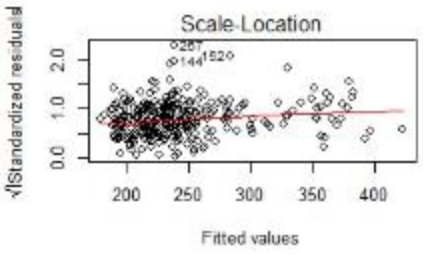
H0:B1 = B2 = B3 = 0 vs Ha: B1 = 0 for E(1, 2, 3)

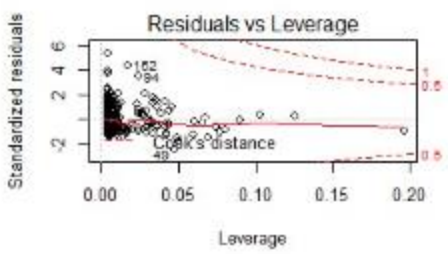
F-statistic = 14.96, meaning we can reject the null hypothesis and state that the model is not a better fit than an intercept only model.











The models has a right skew in standardized residuals and in the Q-Q plot meaning the model is not a good fit.

8)

This was another interesting and engaging assignment and it was good to test more variables against cholesterol and perform more t and f tests to further my skill set and knowledge of these statistical tests. The multiple scatterplots were also a good tool to use to visualize the variables and view the varying slopes and their changes. I got to see how the different variables fit against the cholesterol variable and how point 4 and point 5 were nested into one another. It was also good to see the mix of

continuous and categorical variables and how they affect one another when tested against the variable of cholesterol.