Assignment #5

May 2021

1)

Vitamin: 1 = never, 2 = occasional, 3 + Regular

Gender: 1 = Male, 2 = Female

Smoke: 1 = Yes, 2 = No

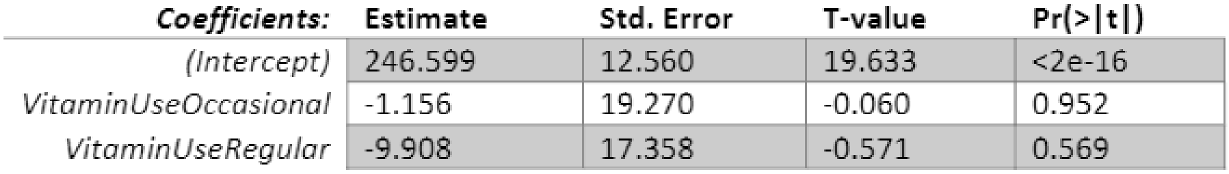
2)

Y_hat = 246.599 – 1.156X1 – 9.908X2

The intercept at 0 is y = -247. The slopes are all negative meaning occasional and regular vitamins usage would lower cholesterol levels. More regular vitamin usage shows a great change in slope meaning that more regular vitamin usage would be best for lowering cholesterol level.

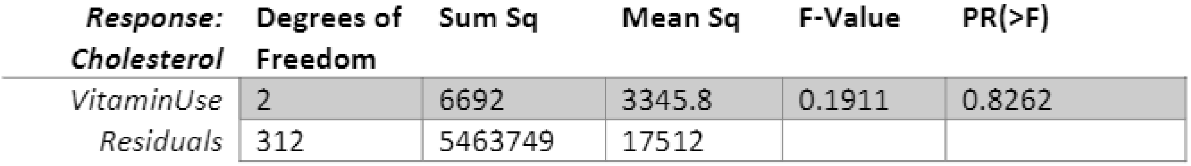
R-squared = .001223, meaning only .1223% of variability is explained in the model.

Coefficients Table:



Coefficients:	**Estimate**	**Std. Error**	**T-value**	**Pr(>|t|)**
(Intercept)	246.599	12.560	19.633	<2e-16
VitaminUseOccasional	-1.156	19.270	-0.060	0.952
VitaminUseRegular	-9.908	17.358	-0.571	0.569

Anova Table:



Response: Cholesterol	**Degrees of Freedom**	**Sum Sq**	**Mean Sq**	**F-Value**	**PR(>F)**
VitaminUse	2	6692	3345.8	0.1911	0.8262
Residuals	312	5463749	17512		

T-test for B1/ Occasional Vitamin Usage

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

T-statistic = -.060, meaning we would not reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Regular Vitamin Usage

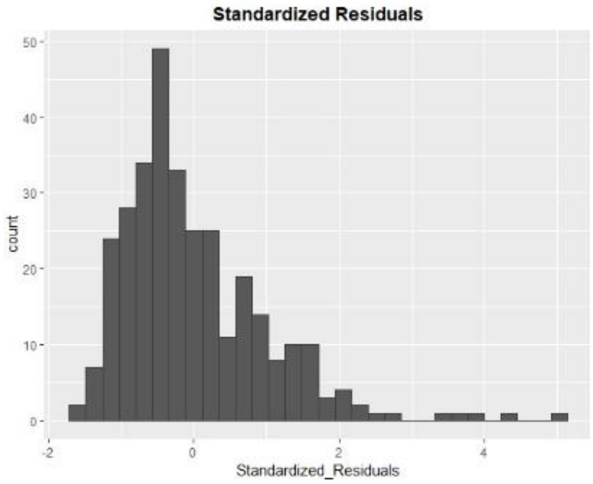
H0: B2 = 0 vs Ha:B2 !=0

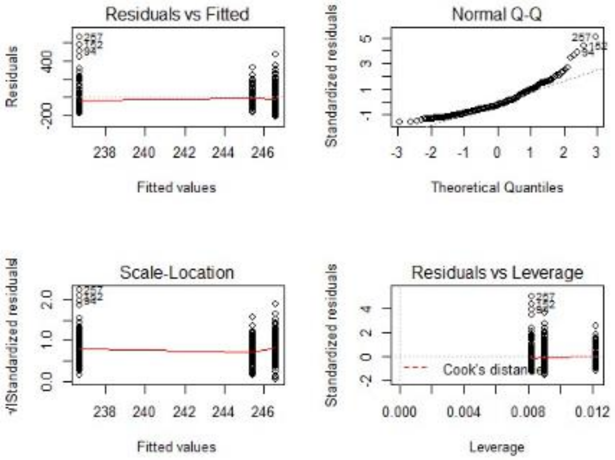
T-statistic = -.571, meaning we would not reject the null hypothesis and the variable is significant to the overall model

F-Test

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

F-statistic = .1911, 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.

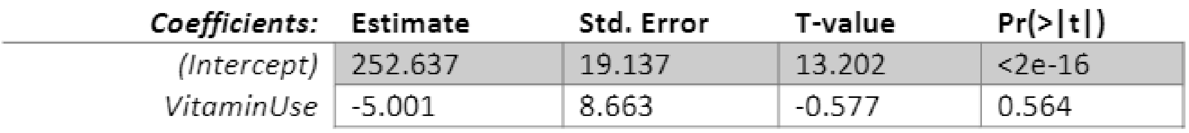
Vitamin Numerical Value

Y_hat = 232.634 – 5.001X1

The intercept at 0 is the starting predicted cholesterol value. The slopes changes show that every change in x, the predicted cholesterol value drops by 5 points partly due to the categorical variables being applied as numerical variables for changes in x for a simple linear regression model.

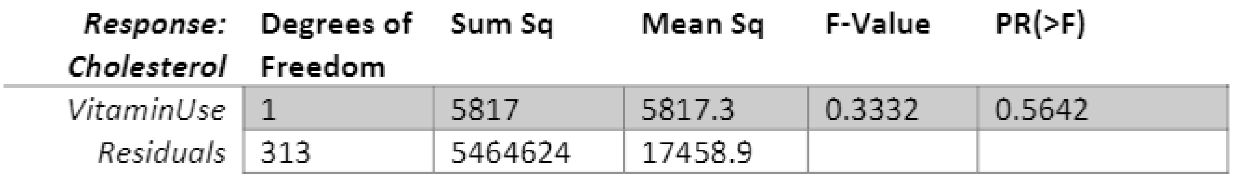
R-squared = .001063, meaning only .1063% of variability is explained in the model.

Coefficients Table:



Coefficients:	Estimate	Std. Error	T-value	Pr(>|t|)
(Intercept)	252.637	19.137	13.202	<2e-16
VitaminUse	-5.001	8.663	-0.577	0.564

Anova table:



Response: Cholesterol	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
VitaminUse	1	5817	5817.3	0.3332	0.5642
Residuals	313	5464624	17458.9		

T-test for B1/ Occasional Vitamin Usage

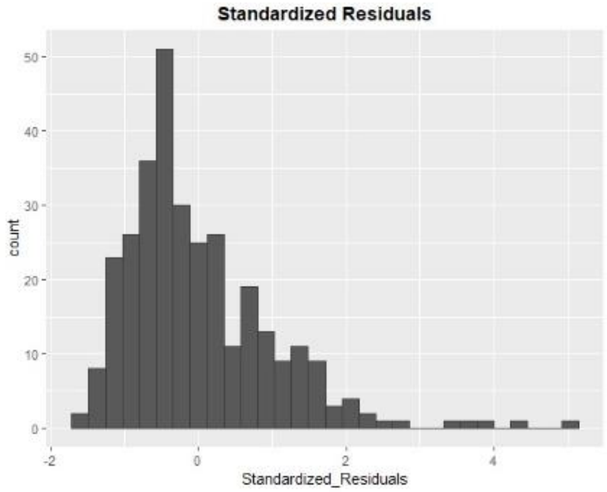
H0: B1 = 0 vs Ha:B1 !=0

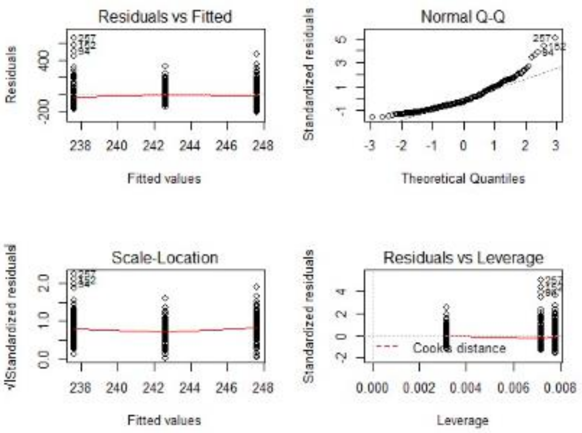
T-statistic = -.577, meaning we would not reject the null hypothesis and the variable is significant to the overall model

F-Test

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

F-statistic = .3332, 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.

Unfortunately, neither model is a good fit, and thus using the vitamin variable is not the best explanatory variable to use for cholesterol. The first model had a higher r-squared score but both models had very low r-squared values. The fits did not change for either of the three categories we had set, never, occasional, or regular.

3)

Y_hat = 246.599 – 1.156X1 – 9.908X2

The intercept at 0 is y = -247. The slopes are all negative meaning occasional and regular vitamins usage would lower cholesterol levels. More regular vitamin usage shows a great change in slope meaning that more regular vitamin usage would be best for lowering cholesterol level.

R-squared = .001223, meaning only .1223% of variability is explained in the model.

Coefficients Table:

Coefficients:	Estimate	Std. Error	T-value	Pr(>|t|)
(Intercept)	246.599	12.560	19.633	<2e-16
V_O	-1.156	19.270	-0.060	0.952
V_R	-9.908	17.358	-0.571	0.569

Anova Table:

Response: Cholesterol	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
V_O	1	986	985.8	0.0563	0.8126
V_R	1	5706	5705.7	0.3258	0.5685
Residuals	312	5463749	17512		

T-test for B1/ Occasional Vitamin Usage

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

T-statistic = -.060, meaning we would not reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Regular Vitamin Usage

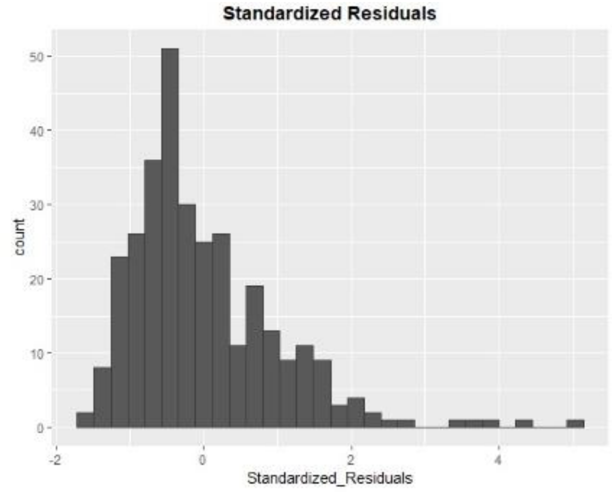
H0: B2 = 0 vs Ha:B2 !=0

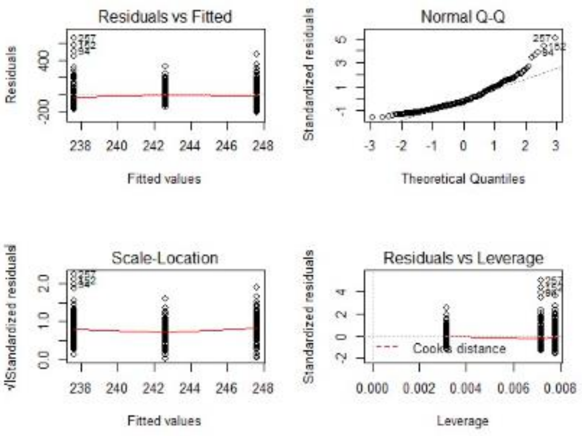
T-statistic = -.571, meaning we would not reject the null hypothesis and the variable is significant to the overall model

F-Test

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

F-statistic = .1911, 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.

In comparison to the first model created, the dummy variables had a very negligible effect. There seems to be some crossover with the dummy variables and the variables for model one and not much change can be noted for occasional and regular vitamin use. There are differences in the anova table but, this model is still not a good fit for predicting cholesterol.

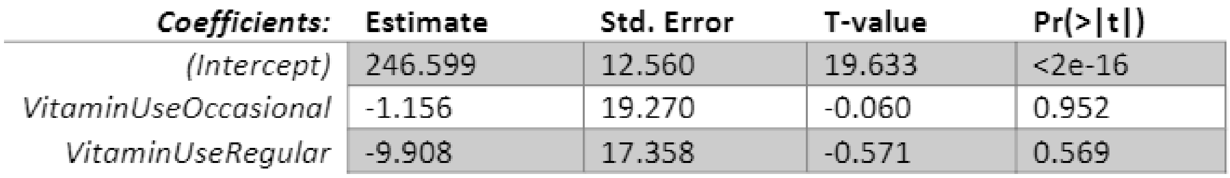
4)

Y_hat = 246.599 – 1.156X1 – 9.908X2

The intercept at 0 is y = -247. The slopes are all negative meaning occasional and regular vitamins usage would lower cholesterol levels. More regular vitamin usage shows a great change in slope meaning that more regular vitamin usage would be best for lowering cholesterol level.

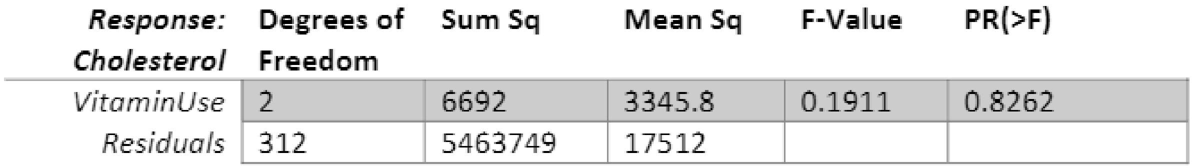
R-squared = .001223, meaning only .1223% of variability is explained in the model.

Coefficients Table:



Coefficients:	**Estimate**	**Std. Error**	**T-value**	**Pr(>|t|)**
(Intercept)	246.599	12.560	19.633	<2e-16
VitaminUseOccasional	-1.156	19.270	-0.060	0.952
VitaminUseRegular	-9.908	17.358	-0.571	0.569

Anova Table:



Response: Cholesterol	**Degrees of Freedom**	**Sum Sq**	**Mean Sq**	**F-Value**	**PR(>F)**
VitaminUse	2	6692	3345.8	0.1911	0.8262
Residuals	312	5463749	17512		

T-test for B1/ Occasional Vitamin Usage

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

T-statistic = -.060, meaning we would not reject the null hypothesis and the variable is significant to the overall model

T-test for B2/ Regular Vitamin Usage

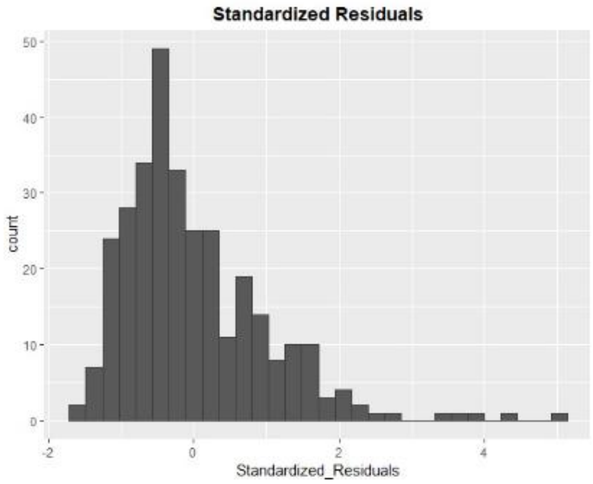
H0: B2 = 0 vs Ha:B2 !=0

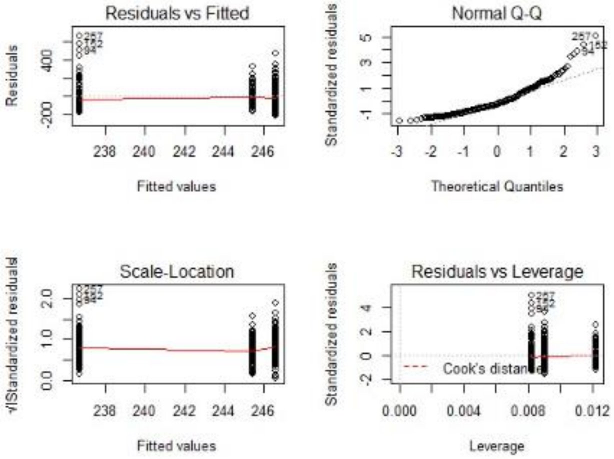
T-statistic = -.571, meaning we would not reject the null hypothesis and the variable is significant to the overall model

F-Test

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

F-statistic = .1911, 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.

Unfortunately, there are no changes from this model to the dummy variable model from the previous question. The dummy coding is easier to break out ant code as actual data values are being used instead of the new ones in the occasional vs regular category. The explanatory variables wind up being good matches for the actual models.

5)

I applied our 3 categories to the usage of alcohol as an indicator for alcohol effect and it had a similar effect as vitamin usage to cholesterol. With alcohol = 0, to test fo high and low usage, we can use it to do anova testing.

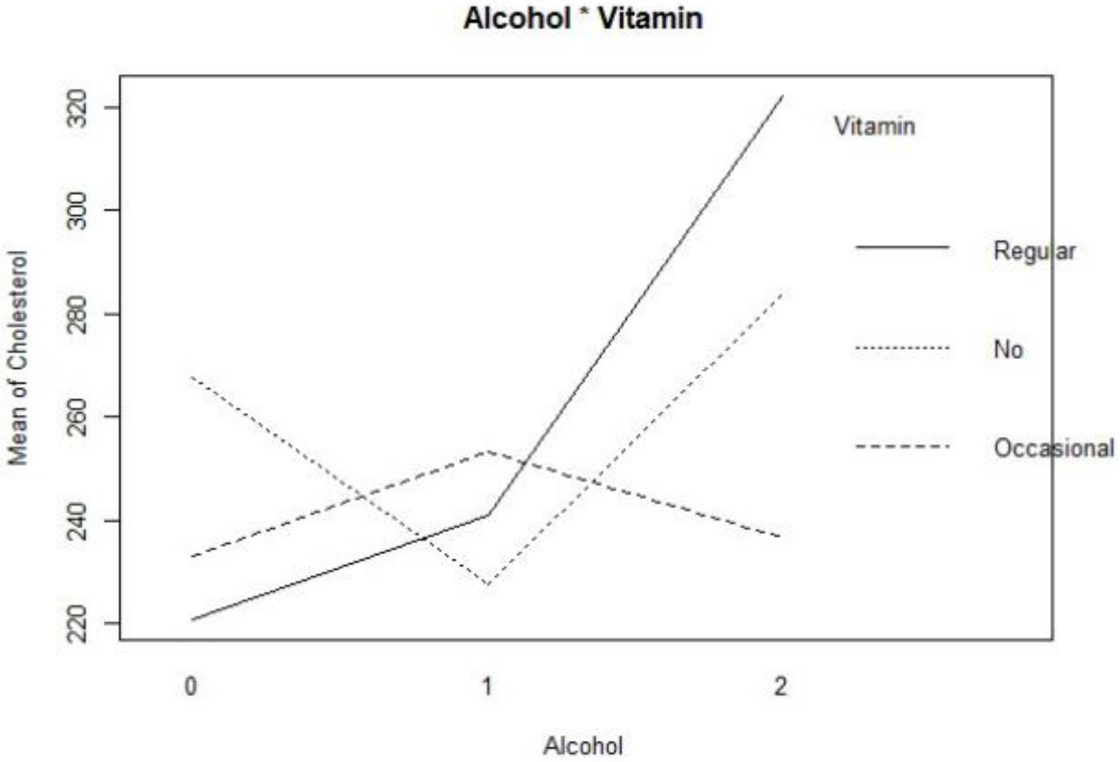
6)

F-test

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

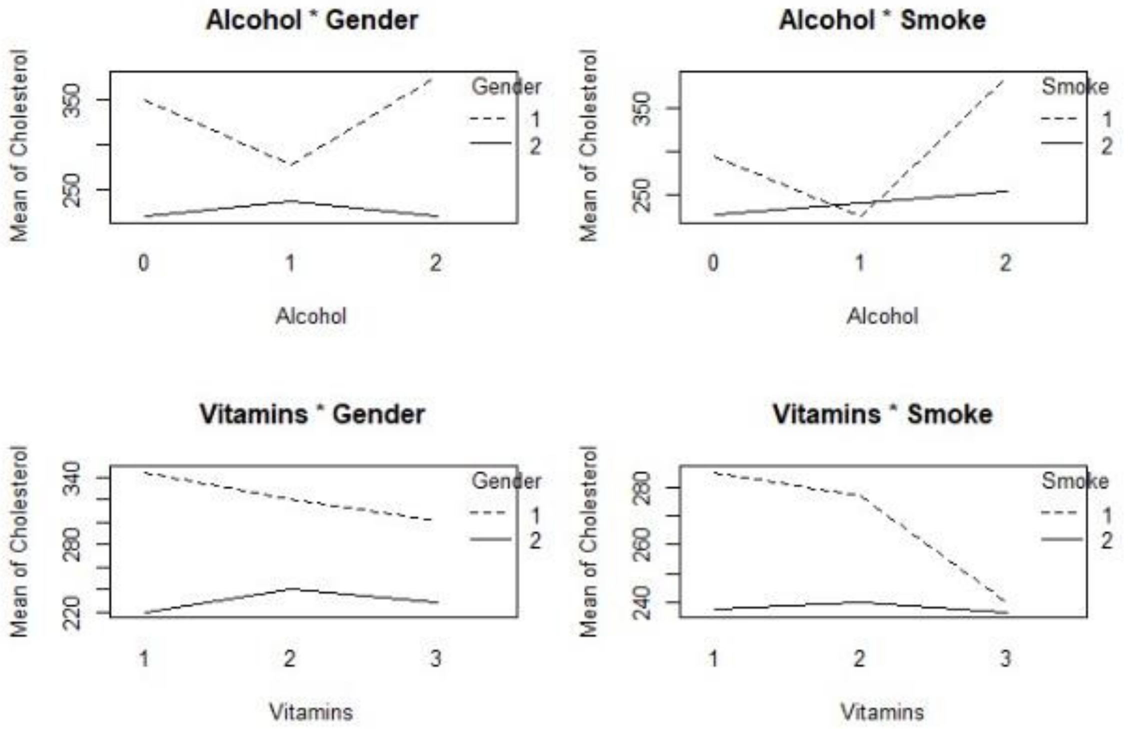
Nested F Test = ((SSE1 – SSE2)/(p2 – p1)/((SSE2)/(n-p2)): ((5426297 – 5342216)/(9-4))/((5342216/(315-9)) = 1.204

We reject the null hypothesis here because the f-test value is greater than 0, which means the additional interactional variables model contribute more info about the association between cholesterol and the 8 prediction variables.



After reviewing the results of the plot, there are two interactions that are not useful with only one being relevant. Regular vitamin usage has an interaction with low alcohol usage with both cholesterol levels lower than the mean. No vitamin usage has decreasing cholesterol value between alcohol usage and low alcohol usage. Occasional vitamin usage has decreasing level of cholesterol between low alcohol usage and high alcohol usage. There are low r-squared values for this model which helps explain the logical discrepancies that were reported.

7)



After viewing the results we can see that women have lower cholesterol levels on their own in both alcohol & gender, and gender 7 vitamin plots; Non-smokers have lower cholesterol levels, regardless if they drink or take vitamins; Smokers who take vitamins showed a strong interaction as the mean of cholesterol levels decrease with increased vitamin usage; Smokers who drink a high amount have a sharp increase in mean cholesterol levels.

Nested F-tests

Gender = Gender + Smoking = 2.2587

Gender + Smoking = Gender + Smoking + Alcohol = 4.6091

Gender + Smoking + Alcohol = Gender + Smoking + Alcohol + Vitamins = .1626

After viewing the results, we should retain gender, smoking and alcohol as variables. They each contribute additional information about the association between cholesterol and each variable. We should rejected the null when including vitamins because all the models we tested including vitamins as a variable proved not to be a good fit.

8)

My greatest learning point from this assignment was how important the dummy variables are in coding. There was a stark difference in the models that accounted for the variables directly as to those which had dummy coding account for the variables. I was surprised that vitamins did not become a good fit as it would logically make sense that vitamins would affect health. However, this allowed us to narrow in on the variables that had the most effect, in our case gender smoking and alcohol. Dummy coding and effect coding proved useful in this assignment and was a skill I picked up here. I plan to use them both in future assignments!