Northwestern University
Master of Science in Data Science

By: Ali Gowani

Table of Contents

Task 1:

For all of the categorical variables in the dataset, recode the text based categories into numerical values that indicate group. For example, for the VITAMIN variable, you could code it so that: 1=regular, 2=occasional, 3=never. Save the categorical variables to the dataset.

Figure 1: Display of categorical variables and their values

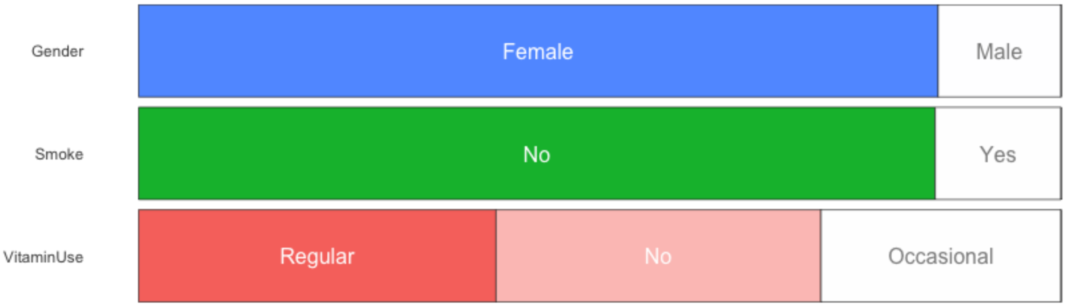
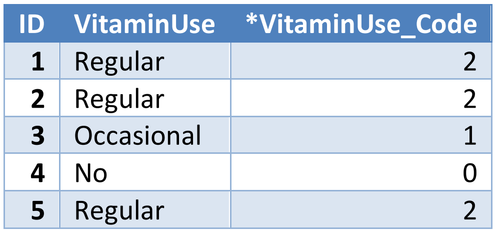


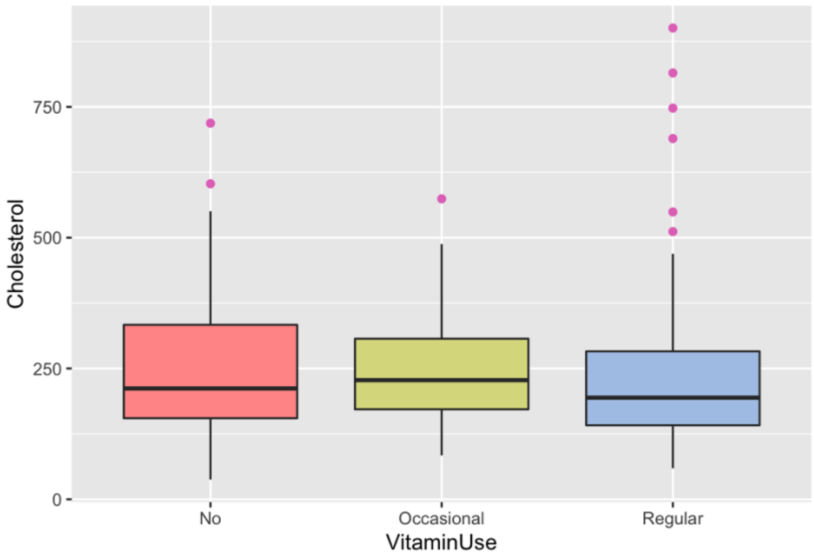
Figure 2: Table showing the first 5 observations of VitaminUse and VitaminUse_Code



ID	VitaminUse	*VitaminUse_Code
1	Regular	2
2	Regular	2
3	Occasional	1
4	No	0
5	Regular	2

*VitaminUse_Code: 0=No, 1=Occasional, 2=Regular)

Figure 3: Boxplot showing VitaminUse and outliers



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 ~ VitaminUse, data = n_df)

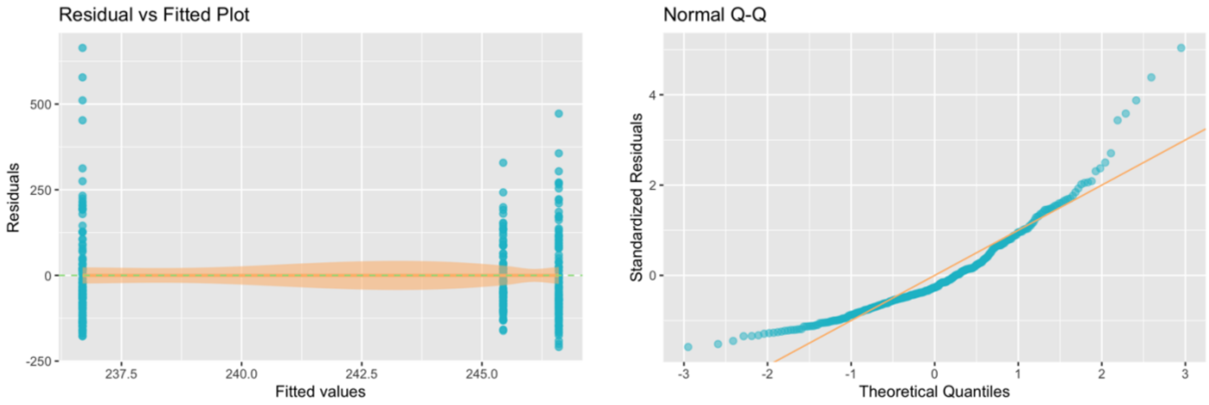
Note: full summary statistics in Appendix.

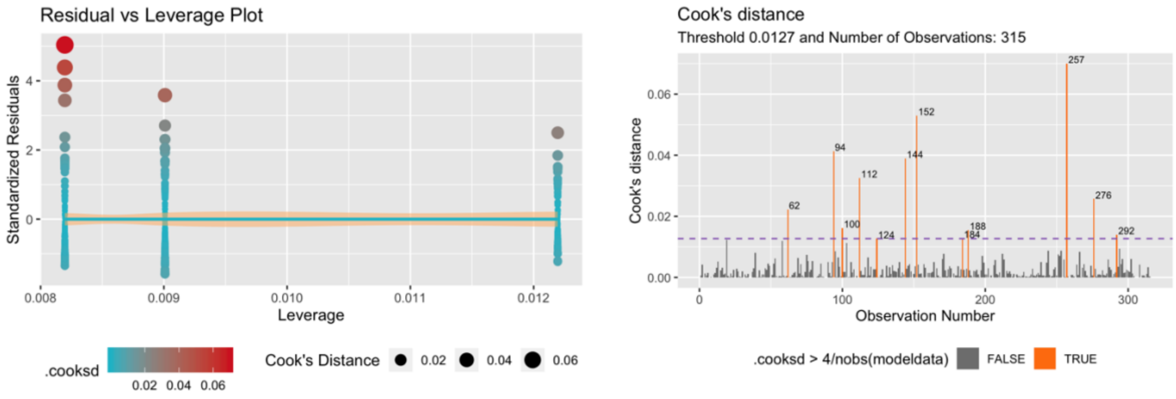
- $\hat{Y} = 246.599 - 1.156\beta_1 \; - 9.908\beta_2$
- $R^2 = -0.005179$

The Omnibus Overall F-statistic for Model 1:

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

Figure 4: Model 1 - Residual, QQ, Leverage and Cook's Distance Plots





The y-intercept includes users who do not take vitamins so if a person does not take any vitamins then their Cholesterol is 246.599. VitaminUseRegular and VitaminUseOccasional are negative so for every 1 unit increase then the persons Cholesterol will decrease. Since the F-statistic for Model 1 is 0.1911, which is less than the critical F-statistic for Model 1 at 3.0247 and p-value is 0.8262 then we fail to reject the Null Hypothesis (alpha = 0.05). This means that our model does not contain significant relationship between the explanatory variable and the response variable of Cholesterol. In addition, the R-Squared value is low, which means that our dependent variable is unable to explain the variance in Cholesterol.

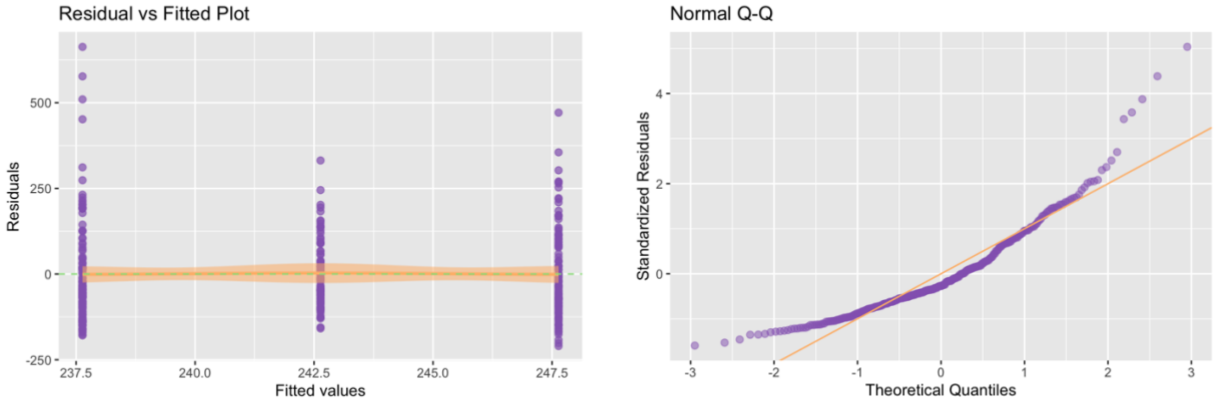
Model 2: lm(formula = Cholesterol ~ VitaminUse_Code, data = n_df)
Note: full summary statistics in Appendix.

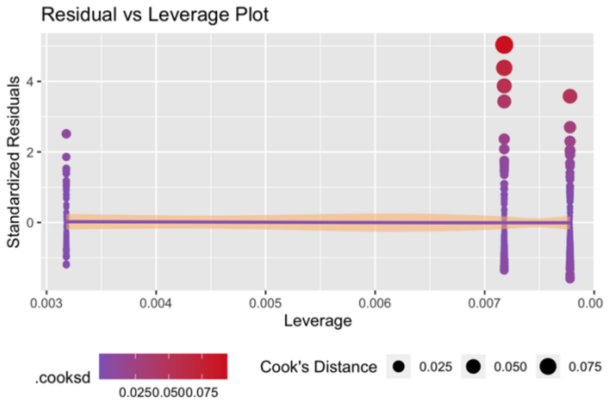
- $\hat{Y} = 247.636 - 5.001\beta_1$
- $R^2 = -0.002128$

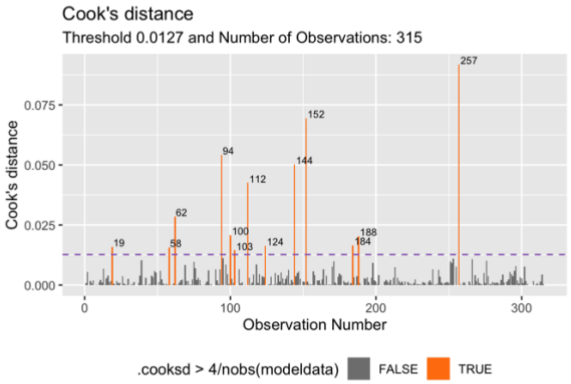
The Omnibus Overall F-statistic for Model 2:

c. Null Hypothesis (H_o): ß$_1$ = 0
d. Alternative Hypothesis (H_a): ß$_i$ ≠ 0 for at least one value of i (e.g.: 1)

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







Similar to Model 1, our dependent variable VitaminUse_Code is negative so for every 1 unit increase then the persons Cholesterol will decrease. Since the F-statistic for Model 2 is 0.3332, which is less than the critical F-statistic for Model 2 at 3.8713 and p-value is 0.5642 then we fail to reject the Null Hypothesis (alpha = 0.05). This means that our model does not contain significant relationship between the explanatory variable and the response variable of Cholesterol. In addition, the R-Squared value is low, which means that our dependent variable is unable to explain the variance in Cholesterol.

In comparing Model 1 and Model 2, there are several outliers that should be considered in these models as they may have an influencing impact.

Task 3:

Create a set of dummy coded (0/1) variables for the VITAMIN categorical variable. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL (Y). Remember, you need to leave one of the dummy coded variables out of the equation. That category becomes the "basis of interpretation." Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Compare the findings here to those in task 2). What has changed?

Model 3: lm(formula = Cholesterol ~ VitaminUse_Occasional + VitaminUse_Regular, data = mydata3)

Note: full summary statistics in Appendix.

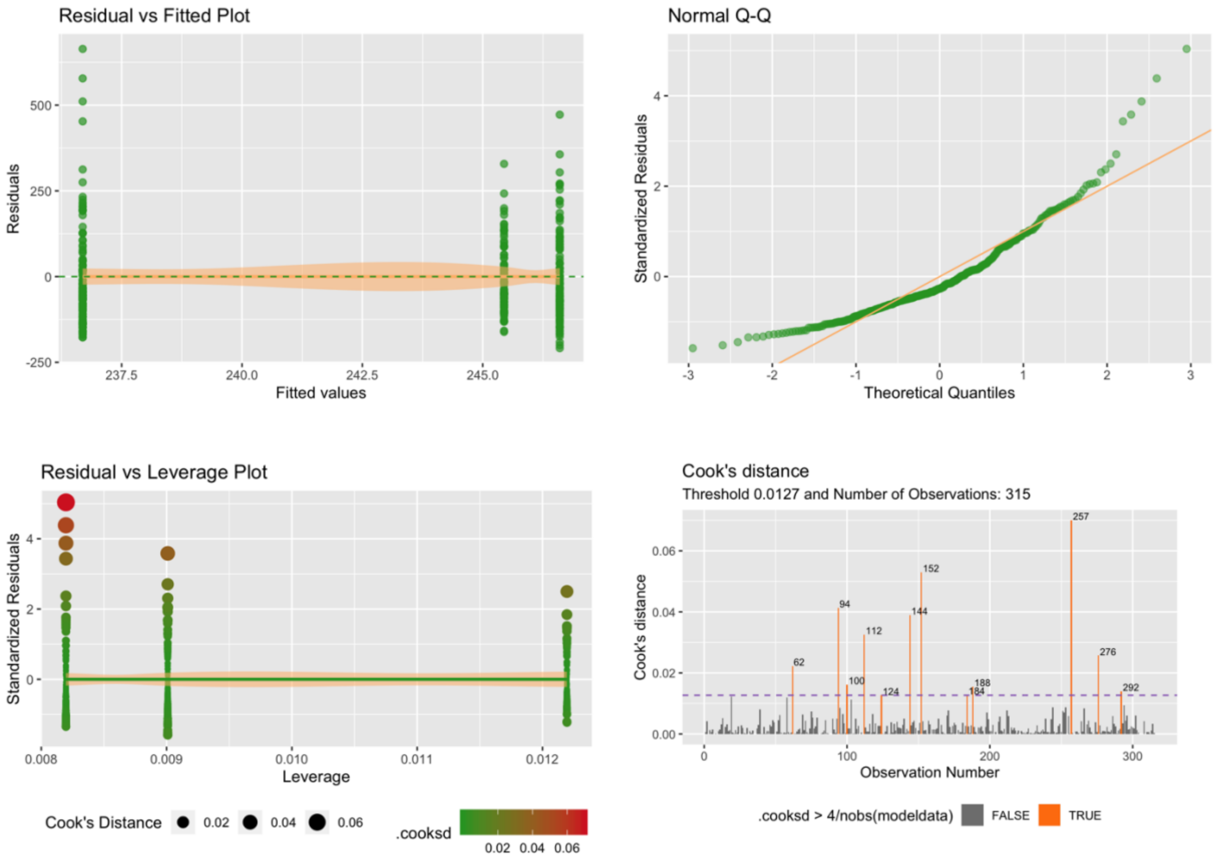
- $\hat{Y} = 246.599 - 1.156\beta_1 \ - 9.908\beta_2$
- $R^2 = -0.005179$

The Omnibus Overall F-statistic for Model 3:

e. Null Hypothesis (H_o): $ß_1 = ß_2 = 0$

f. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1 or 2)

Figure 6: Model 3 - Residual, QQ, Leverage and Cook's Distance Plots



When we compare our Model 3, it looks very similar to Model 1. The reason why it looks similar is that when we created a dummy variable, we had to leave one of them out in order to run our model. It just so happens the dummy variable we left out (VitaminUse = No), is the same one that was used as the intercept for Model 1.

As in Model 1, the Model 3's y-intercept includes users who do not take vitamins so if a person does not take any vitamins then their Cholesterol is 246.599. VitaminUseRegular and VitaminUseOccasional are negative so for every 1 unit increase then the persons Cholesterol will decrease. Since the F-statistic for Model 3 is 0.1911, which is less than the critical F-statistic for Model 3 at 3.0247 and p-value is 0.8262 then we fail to reject the Null Hypothesis (alpha = 0.05). This means that our model does not contain significant relationship between the explanatory variable and the response variable of Cholesterol. In addition, the R-Squared value is low, which means that our dependent variable is unable to explain the variance in Cholesterol.

Task 4:

For the VITAMIN categorical variable, use the NEVER categorical as the control or comparative group, and develop a set of indicator variables using effect coding. Save these to the dataset. Fit a multiple regression model using the dummy coded variables to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. Compare the findings here to those in task 3). What has changed? Which do you prefer? Why?

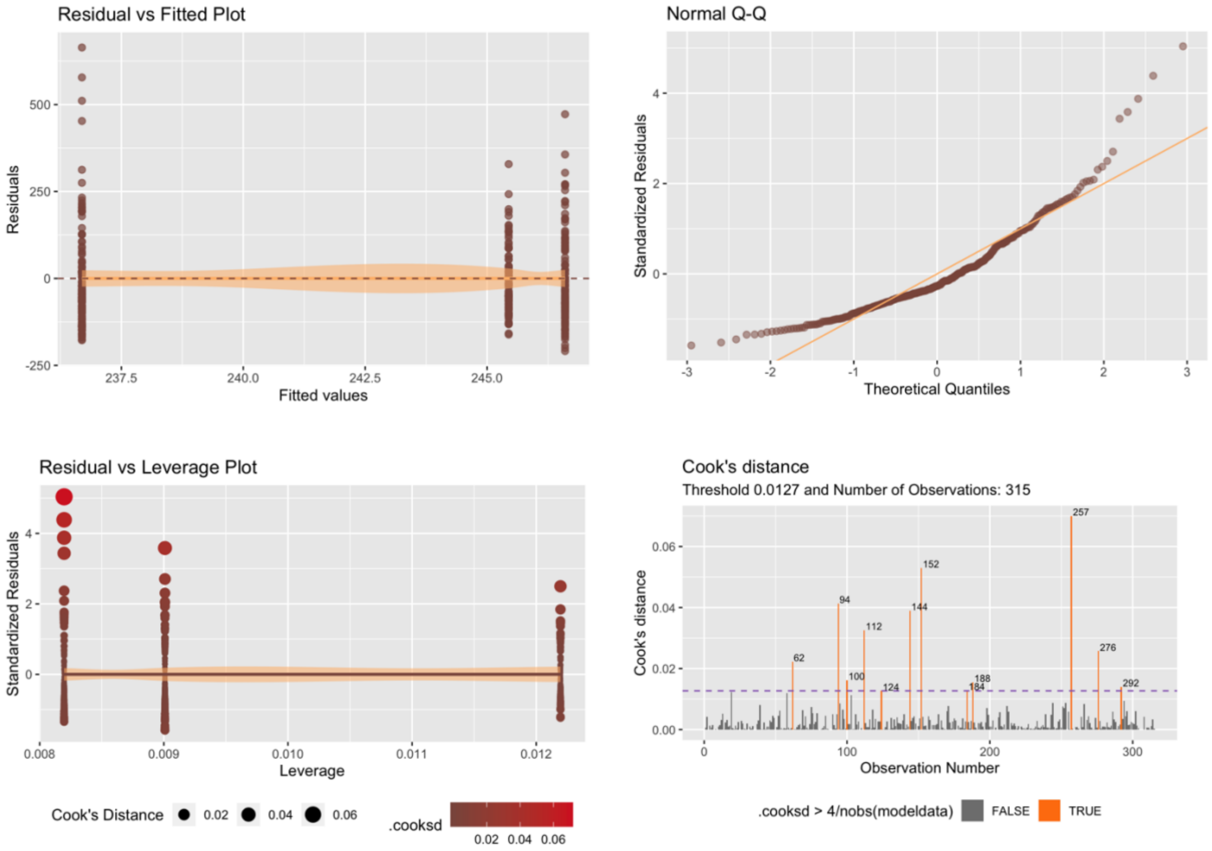
Model 4: lm(formula = Cholesterol ~ VitaminOcc_Eff + VitaminReg_Eff, data = mydata4)
Note: full summary statistics in Appendix.

- $\hat{Y} = 246.599 - 2.532\beta_1 - 6.220\beta_2$
- $R^2 = -0.005179$

The Omnibus Overall F-statistic for Model 4:

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

Figure 7: Model 4 - Residual, QQ, Leverage and Cook's Distance Plots



When we compare our Model 4, it looks very similar to Model 1 and Model 3. The reason why it looks similar is that when we created a dummy variable and indicator variables using effects coding, we had to leave one of them out or make one of them a control group in order to run our model. It just so happens the effect variable (VitaminUse = No), is the same one that was left out in the previous Model 1 and Model 3. However, the coefficients for variables VitaminOcc_Eff and VitaminReg_Eff are different than from Model 1 and Model 3. In addition, the VitaminOcc_Eff coefficient is now positive so for every 1 unit increase of VitmainOcc_Eff, it would on average increase Cholesterol by 2.253.

The F-statistic for Model 4 is 0.1911, which is less than the critical F-statistic for Model 4 at 3.0247 and p-value is 0.8262 then we fail to reject the Null Hypothesis (alpha = 0.05). This means that our model does not contain significant relationship between the explanatory variable and the response variable of Cholesterol. In addition, the R-Squared value is low, which means that our dependent variable is unable to explain the variance in Cholesterol.

In terms of preference, I prefer the dummy coded variables as they seems easier to remember and seem more logical to me.

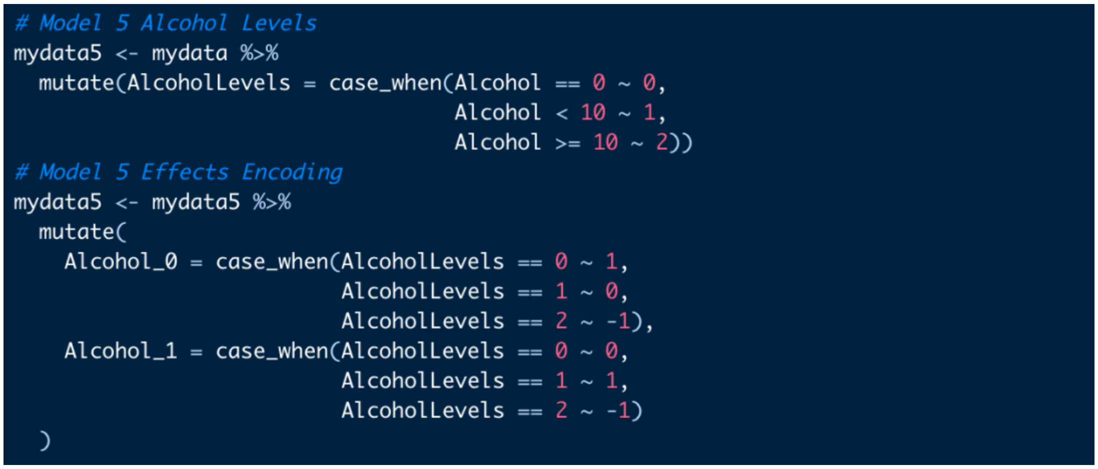
Task 5:

Discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:

0 if ALCOHOL = 0
1 if 0 < ALCOHOL < 10
2 if ALCOHOL >= 10

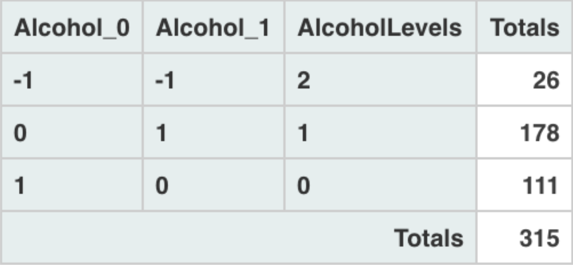
Use these categories to create a set of indicator variables for ALCOHOL that use effect coding. Save these to your dataset.

Figure 8: Alcohol Effects



```
# Model 5 Alcohol Levels
mydata5 <- mydata %>%
  mutate(AlcoholLevels = case_when(Alcohol == 0 ~ 0,
                                   Alcohol < 10 ~ 1,
                                   Alcohol >= 10 ~ 2))
# Model 5 Effects Encoding
mydata5 <- mydata5 %>%
  mutate(
    Alcohol_0 = case_when(AlcoholLevels == 0 ~ 1,
                          AlcoholLevels == 1 ~ 0,
                          AlcoholLevels == 2 ~ -1),
    Alcohol_1 = case_when(AlcoholLevels == 0 ~ 0,
                          AlcoholLevels == 1 ~ 1,
                          AlcoholLevels == 2 ~ -1)
  )
```

Figure 9: Alcohol Levels Table

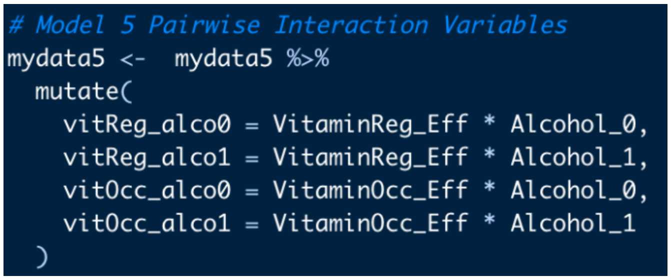


Alcohol_0	Alcohol_1	AlcoholLevels	Totals
-1	-1	2	26
0	1	1	178
1	0	0	111
		Totals	315

Task 6:

At this point, you should have effect coded indicator variables for VITAMIN and 2 effect coded indicator variables for ALCOHOL. Create 4 product variables by multiplying each of the effect coded indicator variables for VITAMIN by the effect coded indicator variables for ALCOHOL. This is all pairwise products of the effect coded variables. Now, we are going to test for interaction. Fit an OLS multiple regression model using the 4 VITAMIN and ALCOHOL effect coded indicator variables plus the 4 product variables to predict CHOLESTEROL. Call this the full model. For the Reduced model, fit an OLS multiple regression model using only the effect coded variables for VITAMIN and ALCOHOL to predict CHOLESTEROL. Conduct a nested model F-test using the Full and Reduced Models described here. Be sure to state the null and alternative hypothesis, make a decision regarding the test, and interpret the result. Obtain a means plot to illustrate any interaction, or lack thereof, to help explain the result.

Figure 10: Pairwise Products (Vitamin and Alcohol)



```
# Model 5 Pairwise Interaction Variables
mydata5 <-  mydata5 %>%
  mutate(
    vitReg_alco0 = VitaminReg_Eff * Alcohol_0,
    vitReg_alco1 = VitaminReg_Eff * Alcohol_1,
    vitOcc_alco0 = VitaminOcc_Eff * Alcohol_0,
    vitOcc_alco1 = VitaminOcc_Eff * Alcohol_1
  )
```

Full Model: lm(formula = Cholesterol ~ vitReg_alco0 + vitReg_alco1 + vitOcc_alco0 + vitOcc_alco1 + VitaminReg_Eff + Alcohol_0 + VitaminOcc_Eff + Alcohol_1, data = mydata5)
Note: full summary statistics in Appendix.

The Omnibus Overall F-statistic for Full Model:

i. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = ß_4 = ß_5 = ß_6 = ß_7 = ß_8 = 0$
j. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2, 3, 4, 5, 6, 7 or 8)

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{5470440.852 - 5342216.257}{8}\right)}{\left(\frac{5342216.257}{315-8-1}\right)} = 0.9181$$

The critical F-statistic for Full Model is:

$$F_{i,n-k-p-1,1-a} = F_{8,315-8-1,0.95} = 1.9687$$

Since the F-statistic for Full Model is 0.9181, which is less than the critical F-statistic at 1.9687 then we fail to reject the Null Hypothesis. This means that our model does not contain significant relationship between the explanatory variables and the response variable of Cholesterol.

Reduced Model: lm(Cholesterol ~ VitaminReg_Eff + VitaminOcc_Eff + Alcohol_0 + Alcohol_1, data = mydata5)
Note: full summary statistics in Appendix.

The Omnibus Overall F-statistic for Reduced Model:

a. Null Hypothesis (H_o): $ß_5 = ß_6 = ß_7 = ß_8 = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 5, 6, 7 or 8)

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

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{5470440.852 - 5426297.463}{4}\right)}{\left(\frac{5426297.463}{315-4-1}\right)} = 0.6305$$

The critical F-statistic value is:

$$F_{i,n-k-p-1,1-a} = F_{4,315-4-1,0.95} = 2.4008$$

Since the F-statistic for Reduced Model is 0.6305, which is less than the critical F-statistic at 2.4008 then we fail to reject the Null Hypothesis. This means that our model does not contain 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$_5$ = ß$_6$ = ß$_7$ = ß$_8$ = 0
b. Alternative Hypothesis (H_a): ß$_i$ ≠ 0 for at least one value of i (e.g.: 5, 6, 7 or 8)

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

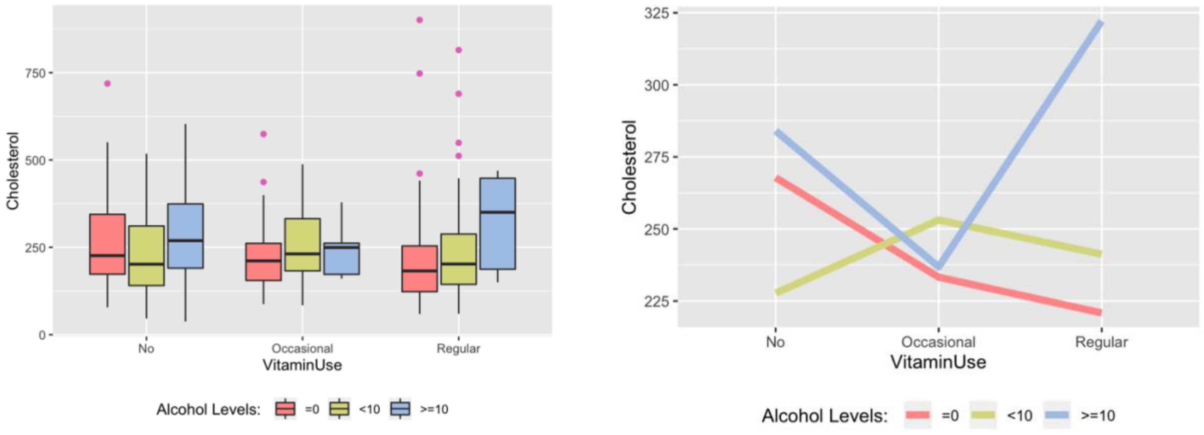
$$F = \frac{\frac{(SSE_R - SSE_C)}{s}}{\left(\frac{SSE_C}{n-k-p-1}\right)} = \frac{\frac{(5426297.463 - 5342216.257)}{4}}{\left(\frac{5342216.257}{315-4-1}\right)} = 1.21$$

The critical F-statistic value is:

$$F_{i,n-k-p-1,1-a} = F_{4,315-4-1,0.95} = 2.402$$

Since the F-statistic value of 1.21 is less than the critical value of 2.402 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.

Figure 11: Boxplot & Means Plot comparing Vitamin and Alcohol, modeling Cholesterol

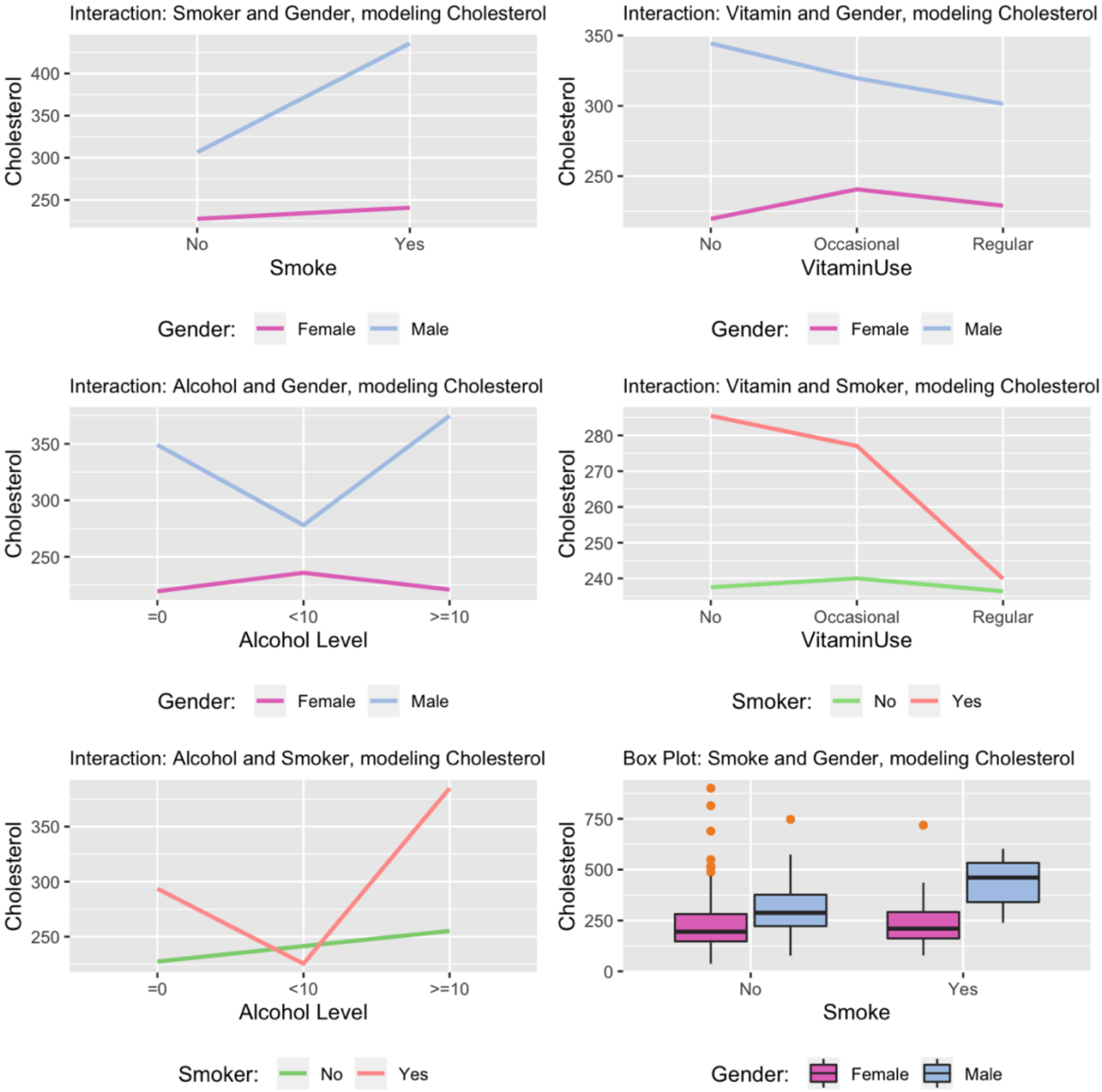


Based on the plots above, we can see the interactions occur between the various variables (VitaminUse, Alcohol Levels) against Cholesterol. The Means Plot shows that high Alcohol Levels increases Cholesterol on average.

Task 7:

There are 2 other categorical variables in this dataset, namely GENDER and SMOKE.   Do these variables interact amongst themselves or with VITAMIN or ALCOHOL when it comes to modeling CHOLESTEROL? Obtain means plots to see if there is interaction. Conduct nested model F-tests to rule out randomness as the explanation for observed patterns. Report your findings.

Figure 12: Means Plots comparing Smoke, Gender, Alcohol & Vitamin modeling Cholesterol



The Omnibus Overall F-statistic for Nested Model:

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

Model: lm(formula = Cholesterol ~ vitReg_alco0 + vitReg_alco1 + vitOcc_alco0 + vitOcc_alco1 + VitaminReg_Eff + Alcohol_0 + VitaminOcc_Eff + Alcohol_1 + Smoke + Gender, data = mydata5)
Note: full summary statistics in Appendix.

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

a. Null Hypothesis (H_o): $ß_9 = ß_{10} = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 9 or 10)

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

$$F = \frac{\frac{(SSE_F - SSE_{F+SG})}{s}}{\left(\frac{SSE_{F+SG}}{n-k-p-1}\right)} = \frac{\frac{(128224.6 - 452516)}{2}}{\left(\frac{452516}{315-10-1}\right)} = 9.8232$$

The critical F-statistic value is:

$$F_{i,n-k-p-1,1-a} = F_{2,315-10-1,0.95} = 3.025$$

Since the F-statistic value of 9.8232 is greater than the critical value of 3.025 at a confidence of 95%, then we can reject the null hypothesis that the Full Model SG is no more robust than the Full Model. This means that the additional variables add significant information in predicting Cholesterol.

CONCLUSION & REFLECTION:

This was a fun assignment though, still takes me many more hours to do than I anticipated. I learned quite a bit about dummy variables and effects coding. In particular, how to transform the variable to these formats and avoid using all of them as it can through our model off and must keep one out when using dummy variables. We also explored using effect coding to control one variable versus others in determining the difference in group means. We generated mean plots to evaluate how the mean varies across different groups of data. I also found box plots to compare variables to be quite useful, especially, when showing outliers.

In addition, we continued to learn about formulating hypothesis for validating individual components, such as, beta coefficients, performing t-tests on individual variables, formulating omnibus overall F-statistic, calculating how to generate statistics for nested models, etc. These tasks were really beneficial for me as it allowed me to get a bit more in the weeds and understand how to assess models and variables within them.

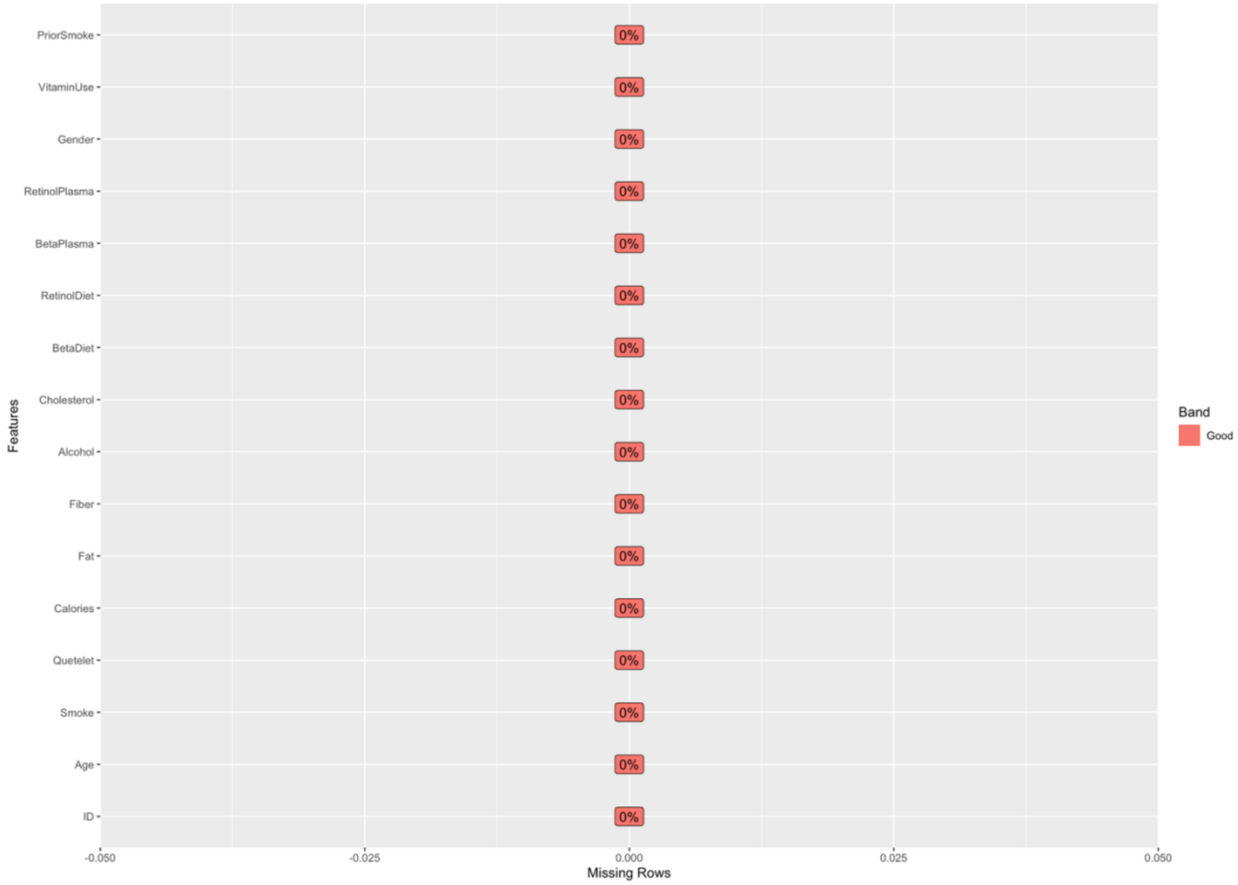
We also further explored getting comfortable with nested models and being able to interpret them with different variables. This allowed me to reinforce calculating statistics of variables and models, so I became more comfortable calculating it.

Overall, this was a good assignment allowing me to reinforce previous concepts and learning new ones.

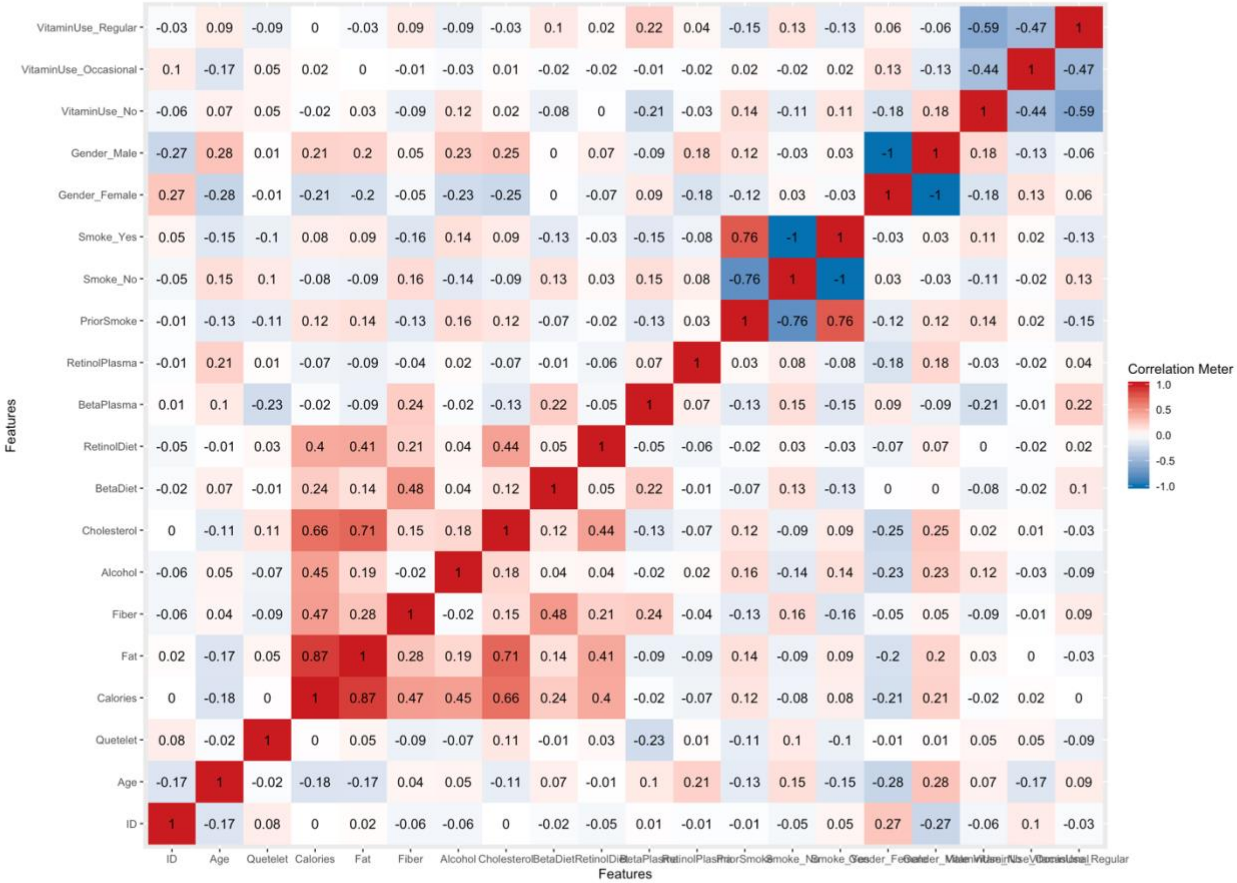
Appendix

A: Exploratory Data Analysis

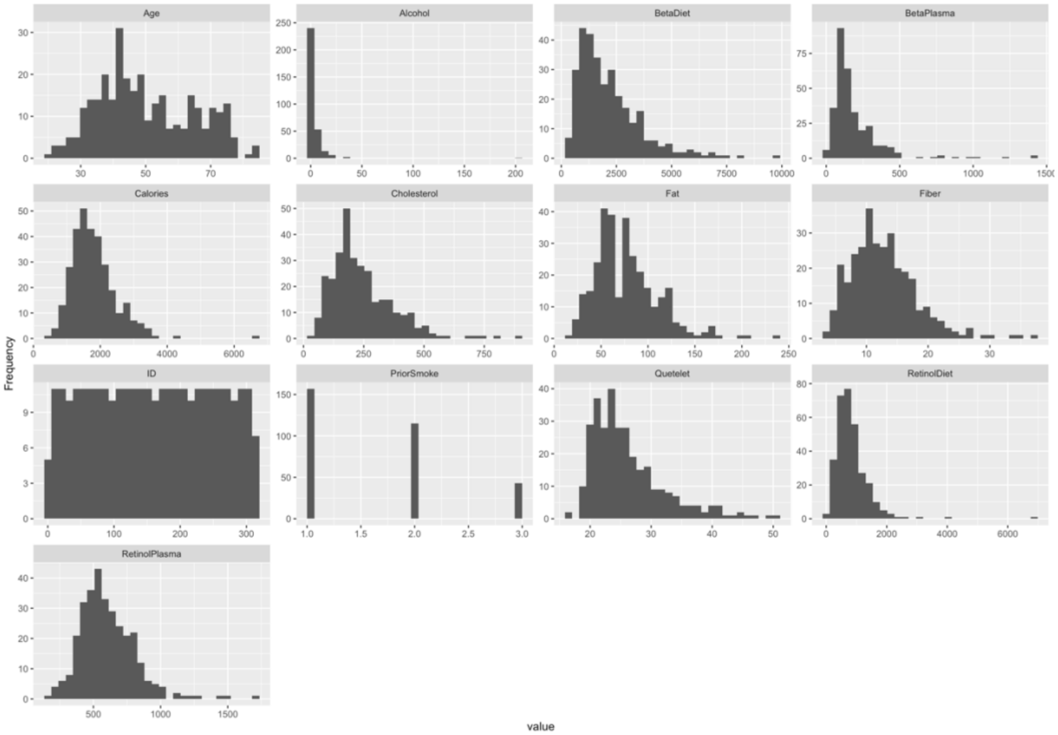
Missing Data



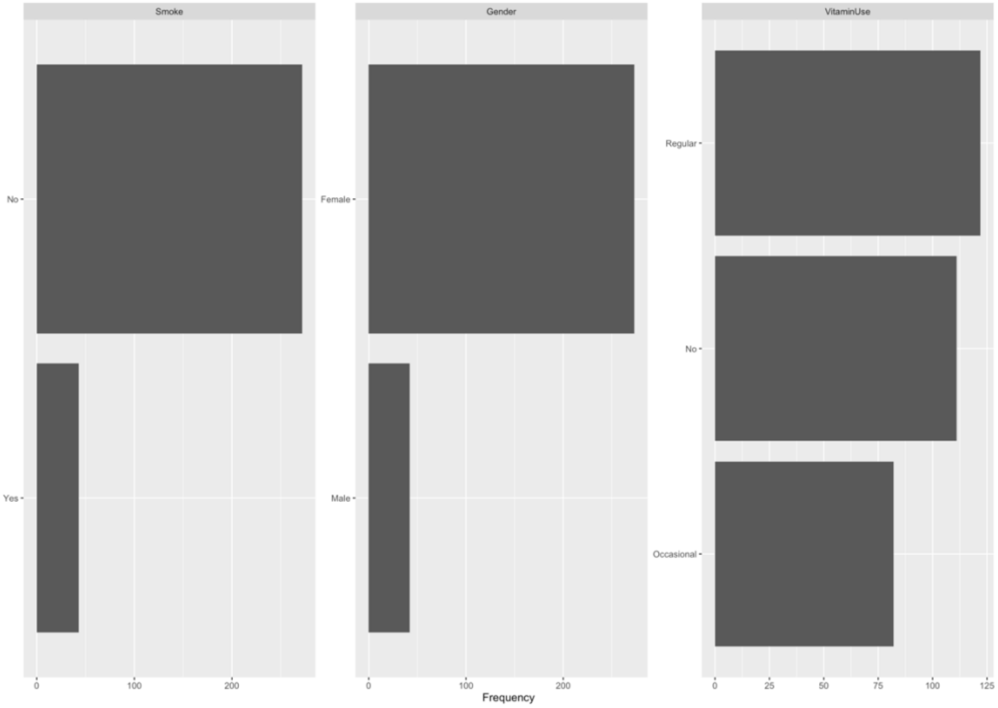
Correlation Analysis



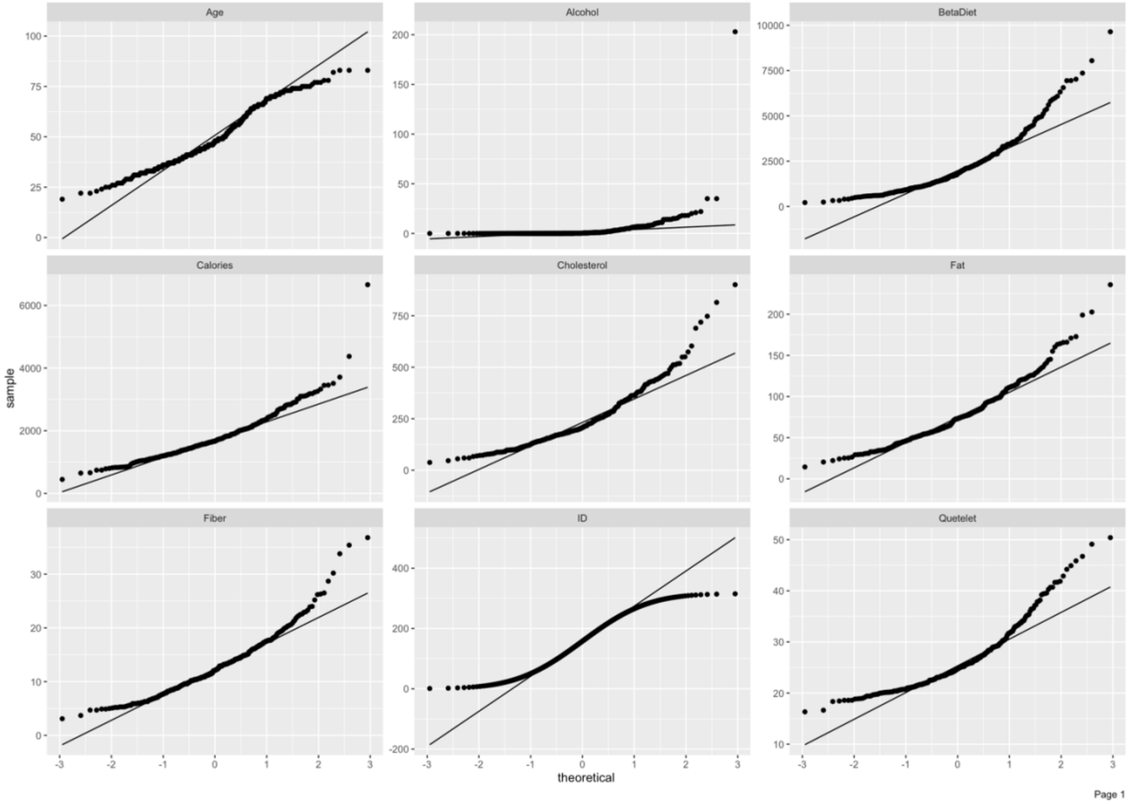
Univariate Distribution (Histograms)

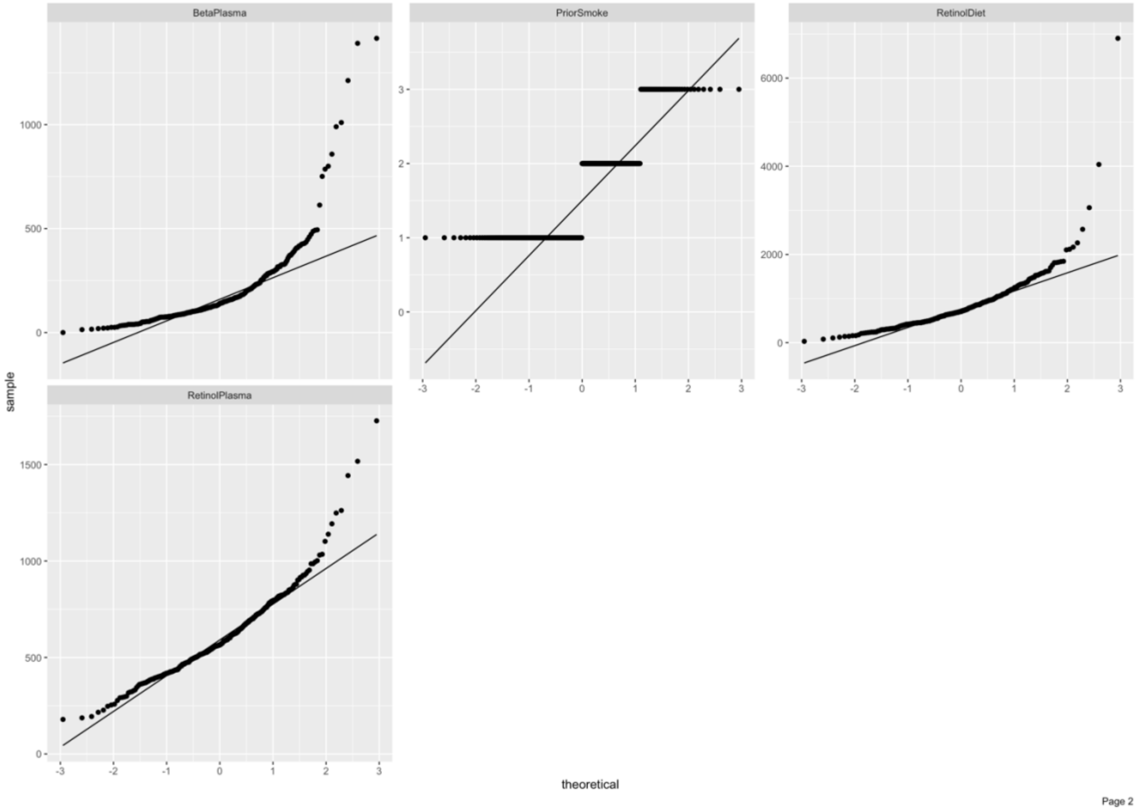


Frequency (Bar Chart)

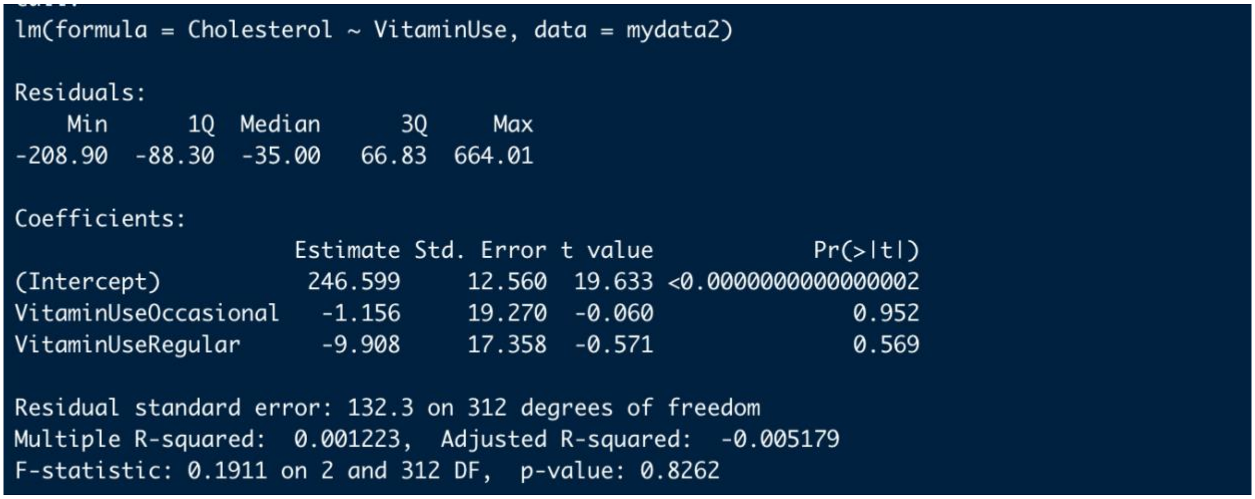


QQ Plot


Age
Alcohol
BetaDiet
Calories
Cholesterol
Fat
Fiber
ID
Quetelet
sample
theoretical
-3
-2
-1
0
1
2
3
Page 1


BetaPlasma
PriorSmoke
RetinolDiet
RetinolPlasma
sample
theoretical
-3
-2
-1
0
1
2
3
Page 2

B: Model 1 Summary Statistics (VitaminUse)



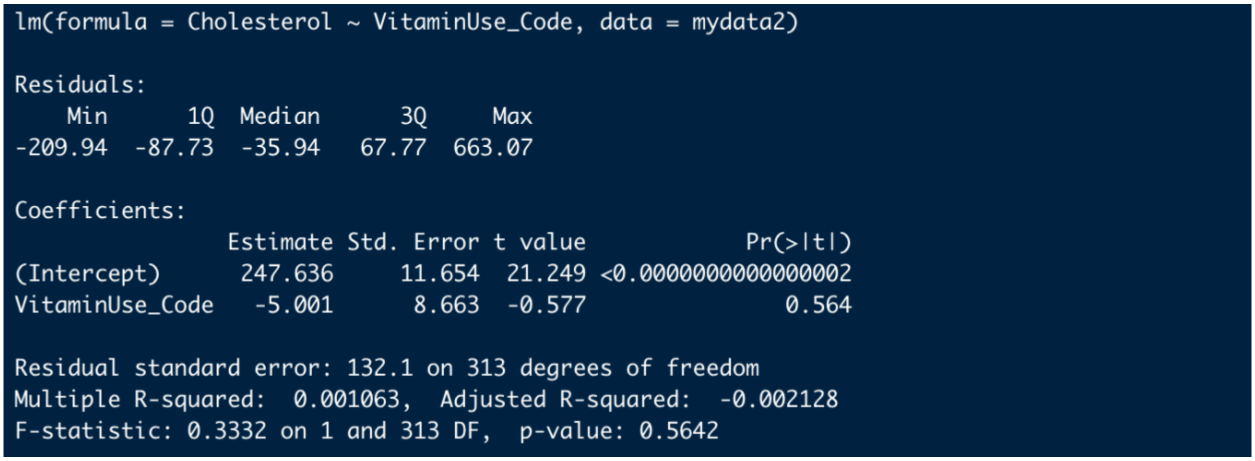
```
lm(formula = Cholesterol ~ VitaminUse, data = mydata2)

Residuals:
    Min      1Q  Median      3Q     Max 
-208.90  -88.30  -35.00   66.83  664.01 

Coefficients:
                     Estimate Std. Error t value            Pr(>|t|)    
(Intercept)           246.599     12.560  19.633 <0.0000000000000002
VitaminUseOccasional   -1.156     19.270  -0.060               0.952    
VitaminUseRegular      -9.908     17.358  -0.571               0.569    

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,	Adjusted R-squared:  -0.005179 
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

C: Model 2 Summary Statistics (VitaminUse_Code)



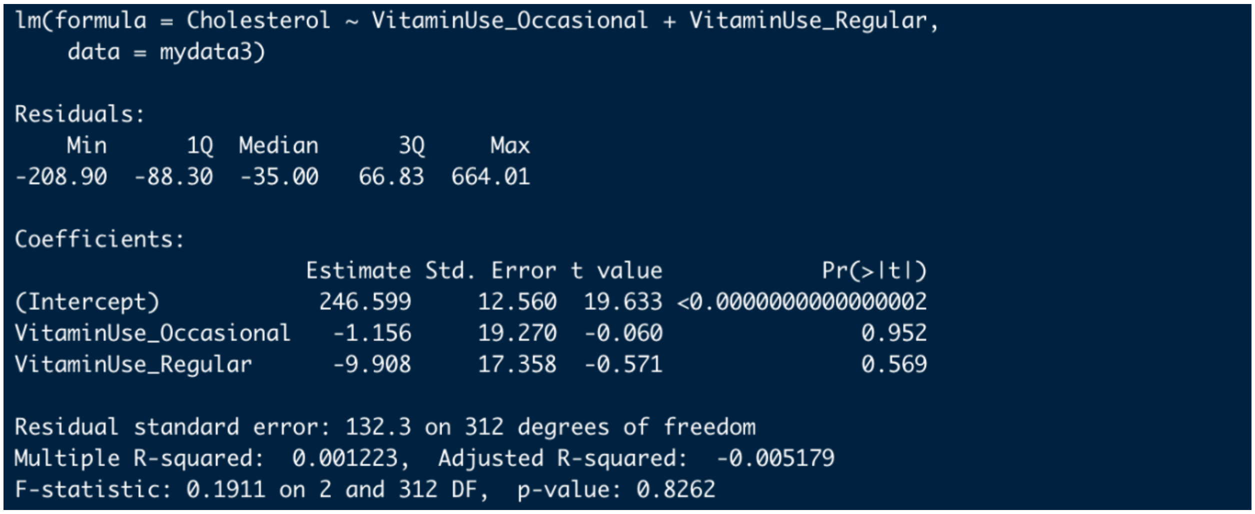
```
lm(formula = Cholesterol ~ VitaminUse_Code, data = mydata2)

Residuals:
    Min      1Q  Median      3Q     Max 
-209.94  -87.73  -35.94   67.77  663.07 

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)    
(Intercept)      247.636     11.654  21.249 <0.0000000000000002
VitaminUse_Code   -5.001      8.663  -0.577               0.564    

Residual standard error: 132.1 on 313 degrees of freedom
Multiple R-squared:  0.001063,	Adjusted R-squared:  -0.002128 
F-statistic: 0.3332 on 1 and 313 DF,  p-value: 0.5642
```

D: Model 3 Summary Statistics (VitaminUse_Occasional + VitaminUse_Regular)



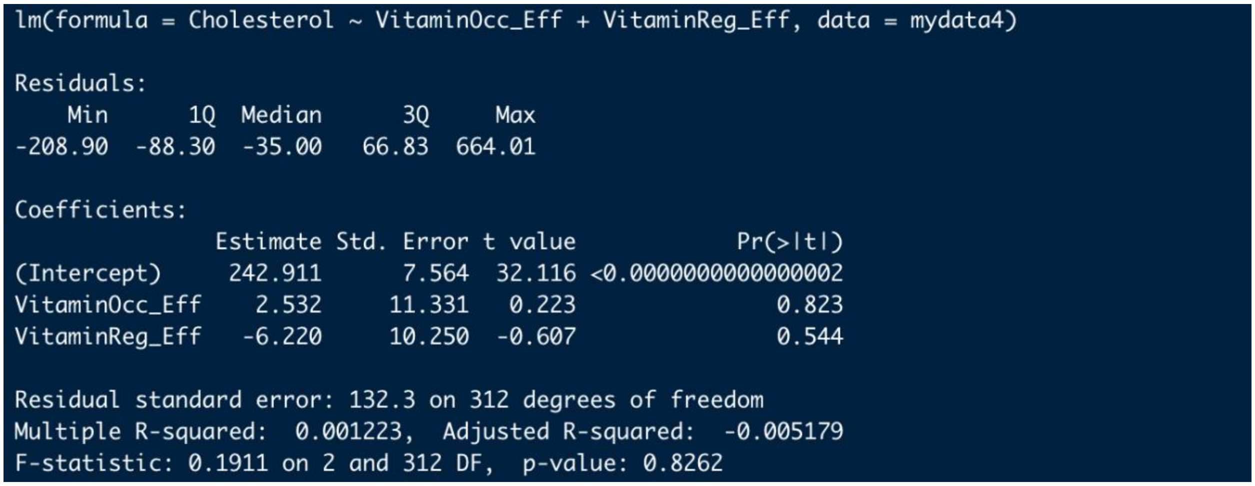
```
lm(formula = Cholesterol ~ VitaminUse_Occasional + VitaminUse_Regular,
    data = mydata3)

Residuals:
    Min      1Q  Median      3Q     Max
-208.90  -88.30  -35.00   66.83  664.01

Coefficients:
                      Estimate Std. Error t value            Pr(>|t|)
(Intercept)            246.599     12.560  19.633 <0.0000000000000002
VitaminUse_Occasional   -1.156     19.270  -0.060               0.952
VitaminUse_Regular      -9.908     17.358  -0.571               0.569

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,  Adjusted R-squared:  -0.005179
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

E: Model 4 Summary Statistics (VitaminOcc_Eff + VitaminReg_Eff)



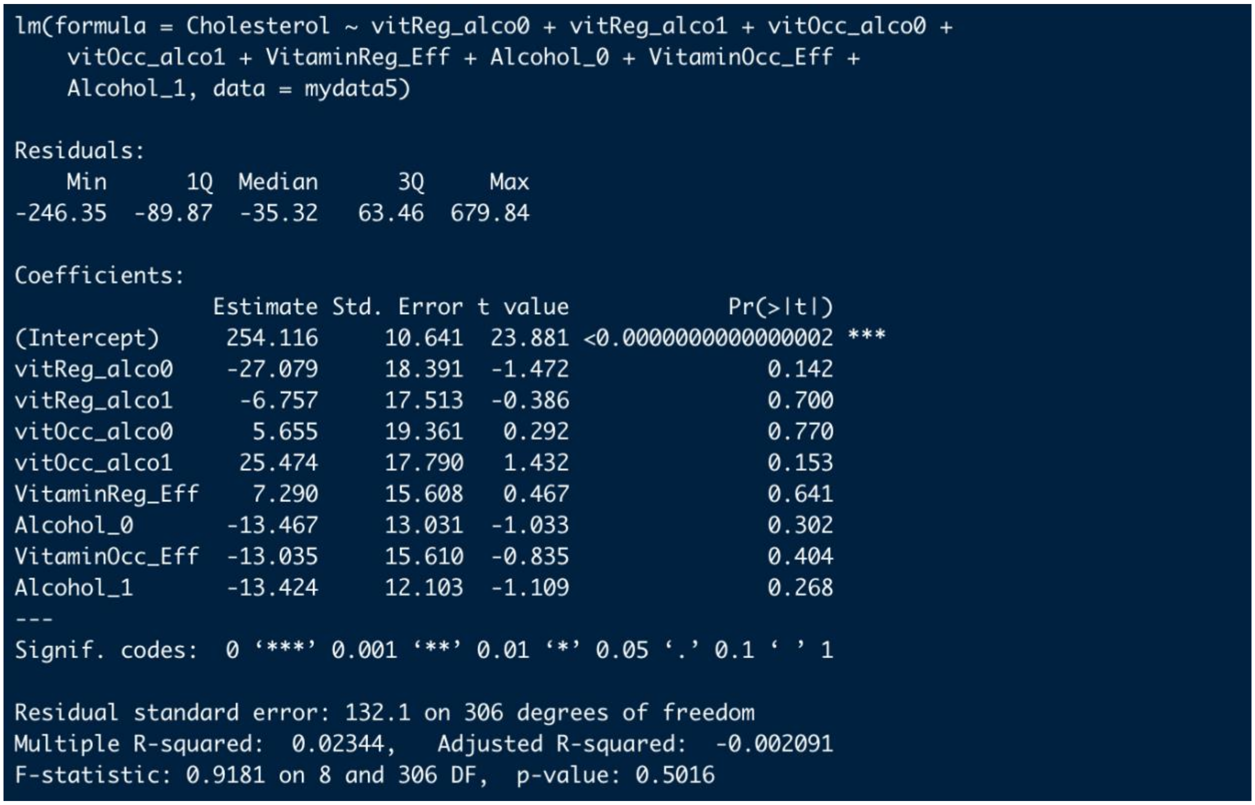
```
lm(formula = Cholesterol ~ VitaminOcc_Eff + VitaminReg_Eff, data = mydata4)

Residuals:
    Min      1Q  Median      3Q     Max
-208.90  -88.30  -35.00   66.83  664.01

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)
(Intercept)     242.911      7.564  32.116 <0.0000000000000002
VitaminOcc_Eff    2.532     11.331   0.223               0.823
VitaminReg_Eff   -6.220     10.250  -0.607               0.544

Residual standard error: 132.3 on 312 degrees of freedom
Multiple R-squared:  0.001223,  Adjusted R-squared:  -0.005179
F-statistic: 0.1911 on 2 and 312 DF,  p-value: 0.8262
```

F: Full Model Summary Statistics



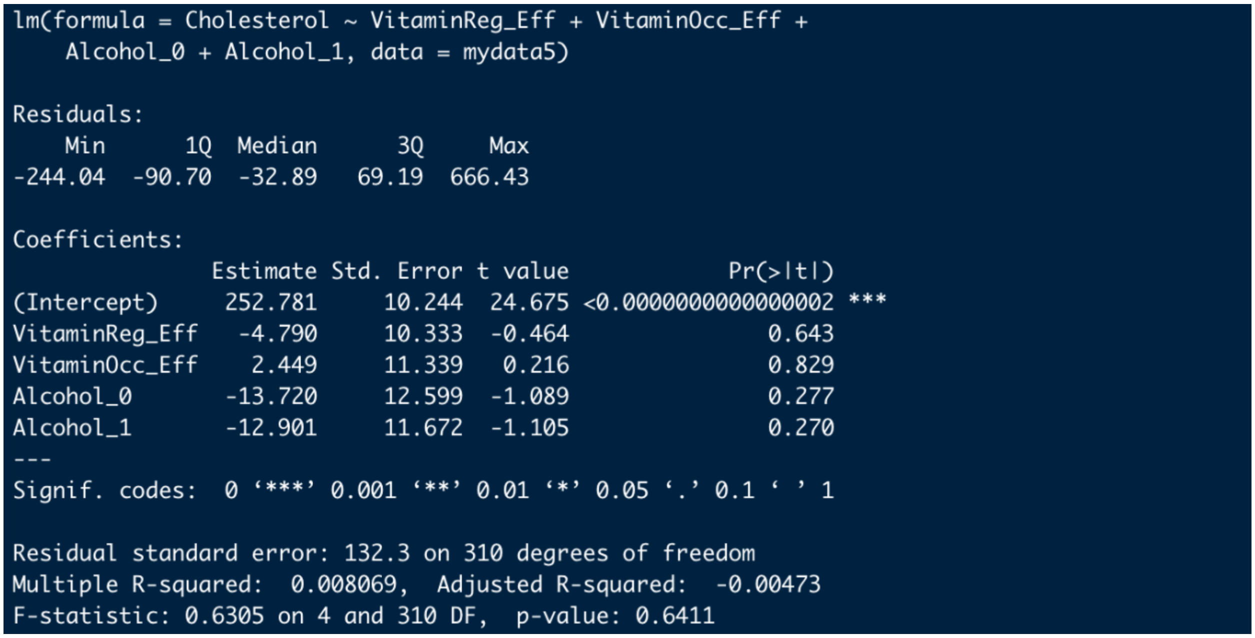
```
lm(formula = Cholesterol ~ vitReg_alco0 + vitReg_alco1 + vitOcc_alco0 +
    vitOcc_alco1 + VitaminReg_Eff + Alcohol_0 + VitaminOcc_Eff +
    Alcohol_1, data = mydata5)

Residuals:
    Min      1Q  Median      3Q     Max
-246.35  -89.87  -35.32   63.46  679.84

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)
(Intercept)     254.116     10.641  23.881 <0.0000000000000002 ***
vitReg_alco0    -27.079     18.391  -1.472               0.142
vitReg_alco1     -6.757     17.513  -0.386               0.700
vitOcc_alco0      5.655     19.361   0.292               0.770
vitOcc_alco1     25.474     17.790   1.432               0.153
VitaminReg_Eff    7.290     15.608   0.467               0.641
Alcohol_0       -13.467     13.031  -1.033               0.302
VitaminOcc_Eff  -13.035     15.610  -0.835               0.404
Alcohol_1       -13.424     12.103  -1.109               0.268
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.1 on 306 degrees of freedom
Multiple R-squared:  0.02344,   Adjusted R-squared:  -0.002091
F-statistic: 0.9181 on 8 and 306 DF,  p-value: 0.5016
```

G: Reduced Model Summary Statistics



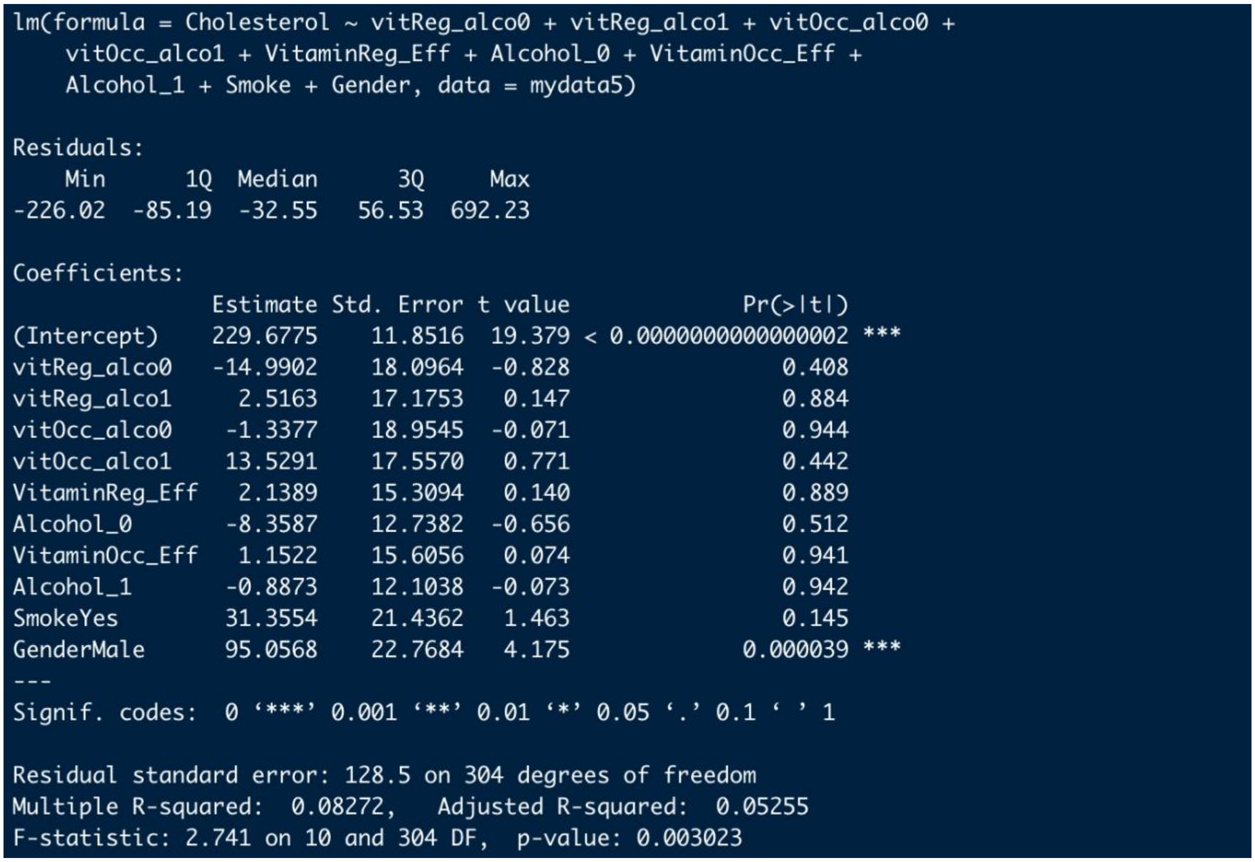
```
lm(formula = Cholesterol ~ VitaminReg_Eff + VitaminOcc_Eff +
    Alcohol_0 + Alcohol_1, data = mydata5)

Residuals:
    Min      1Q  Median      3Q     Max
-244.04  -90.70  -32.89   69.19  666.43

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)
(Intercept)     252.781     10.244  24.675 <0.0000000000000002 ***
VitaminReg_Eff   -4.790     10.333  -0.464               0.643
VitaminOcc_Eff    2.449     11.339   0.216               0.829
Alcohol_0       -13.720     12.599  -1.089               0.277
Alcohol_1       -12.901     11.672  -1.105               0.270
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.3 on 310 degrees of freedom
Multiple R-squared:  0.008069,  Adjusted R-squared:  -0.00473
F-statistic: 0.6305 on 4 and 310 DF,  p-value: 0.6411
```

H: Full Model SG Summary Statistics



```
lm(formula = Cholesterol ~ vitReg_alco0 + vitReg_alco1 + vitOcc_alco0 +
    vitOcc_alco1 + VitaminReg_Eff + Alcohol_0 + VitaminOcc_Eff +
    Alcohol_1 + Smoke + Gender, data = mydata5)

Residuals:
    Min      1Q  Median      3Q     Max
-226.02  -85.19  -32.55   56.53  692.23

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)    229.6775    11.8516  19.379 < 0.0000000000000002 ***
vitReg_alco0   -14.9902    18.0964  -0.828                0.408
vitReg_alco1     2.5163    17.1753   0.147                0.884
vitOcc_alco0    -1.3377    18.9545  -0.071                0.944
vitOcc_alco1    13.5291    17.5570   0.771                0.442
VitaminReg_Eff   2.1389    15.3094   0.140                0.889
Alcohol_0       -8.3587    12.7382  -0.656                0.512
VitaminOcc_Eff   1.1522    15.6056   0.074                0.941
Alcohol_1       -0.8873    12.1038  -0.073                0.942
SmokeYes        31.3554    21.4362   1.463                0.145
GenderMale      95.0568    22.7684   4.175             0.000039 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 128.5 on 304 degrees of freedom
Multiple R-squared:  0.08272,   Adjusted R-squared:  0.05255
F-statistic: 2.741 on 10 and 304 DF,  p-value: 0.003023
```