Pooja Deshpande

Computational Assignment 3
MSDS – 410 Data Modeling for Supervised Learning,
Summer 2020
Northwestern University

Contents:

Q.1

Recoding categorical variables into numerical values

VitaminUse is recoded as “Regular” = 1, “Occasional” = 2, “No” = 3,

Name of new variable: VitaminRecoded

Gender is recoded as “Female” = 1, “Male” = 2,

Name of new variable: GenderRecoded

Smoke is recoded as “No” = 1, “Yes” = 2

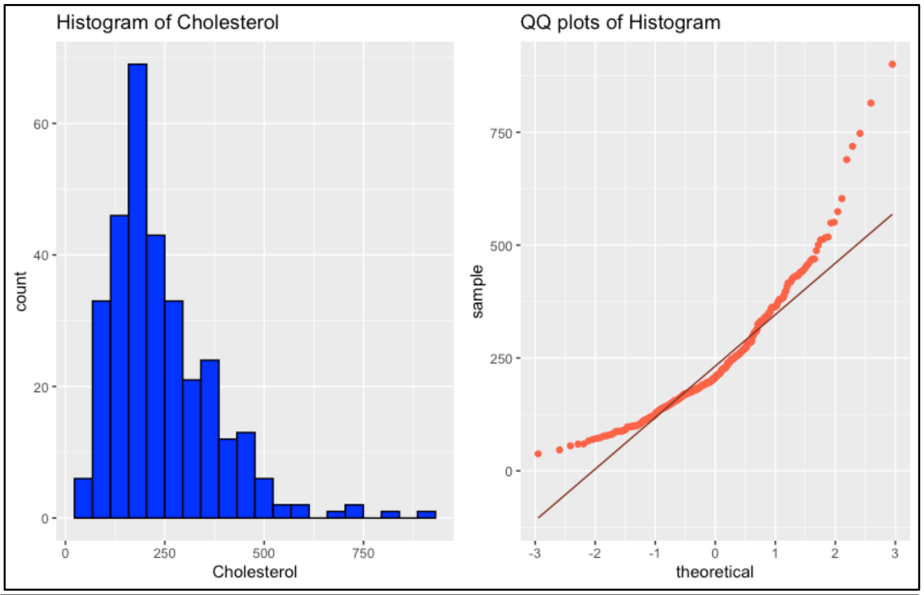
Name of new variable: SmokeRecoded

```
> #Q.1
> #Regular = 1, Occasional = 2, No = 3
> data$VitaminRecoded <- revalue(data$VitaminUse,c("Regular" = 1,"Occasional" = 2, "No" = 3))
> #Female = 1, Male = 2
> data$GenderRecoded <- revalue(data$Gender,c("Female" = 1,"Male" = 2))
> #No = 1, Yes = 2
> data$SmokeRecoded <- revalue(data$Smoke,c("No" =1, "Yes" = 2))
> str(data)
```

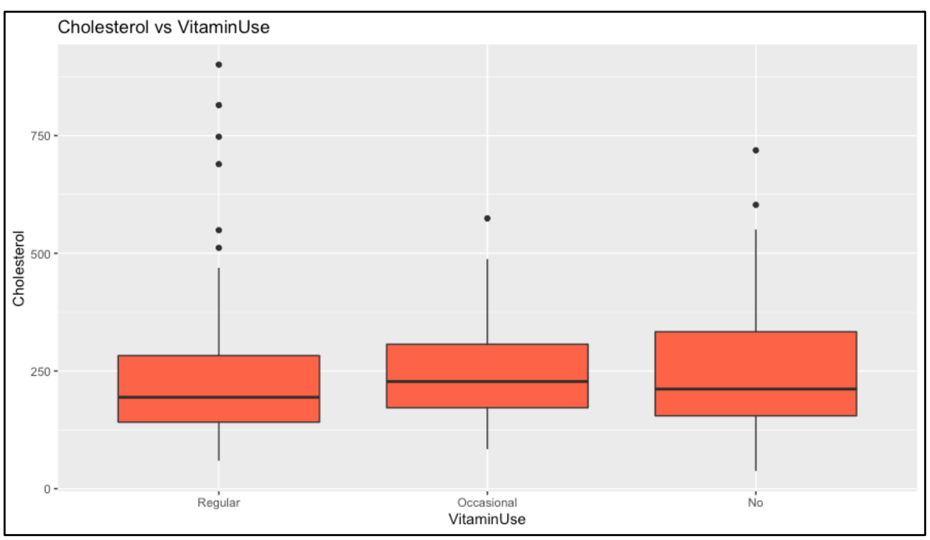
```
> head(data)
  ID Age Smoke Quetelet Calories  Fat Fiber Alcohol Cholesterol BetaDiet RetinolDiet BetaPlasma RetinolPlasma
1  1  64    No  21.4838   1298.8 57.0   6.3     0.0       170.3     1945         890        200           915
2  2  76    No  23.8763   1032.5 50.1  15.8     0.0        75.8     2653         451        124           727
3  3  38    No  20.0108   2372.3 83.6  19.1    14.1       257.9     6321         660        328           721
4  4  40    No  25.1406   2449.5 97.5  26.5     0.5       332.6     1061         864        153           615
5  5  72    No  20.9850   1952.1 82.6  16.2     0.0       170.8     2863        1209         92           799
6  6  40    No  27.5214   1366.9 56.0   9.6     1.3       154.6     1729        1439        148           654
  Gender VitaminUse PriorSmoke VitaminRecoded GenderRecoded SmokeRecoded VitaminOccasional VitaminNo
1 Female    Regular          2              1             1            1                 0         0
2 Female    Regular          1              1             1            1                 0         0
3 Female Occasional          2              2             1            1                 1         0
4 Female         No          2              3             1            1                 0         1
5 Female    Regular          1              1             1            1                 0         0
6 Female         No          2              3             1            1                 0         1
```

Q.2

EDA before building the model:



Cholesterol is not normally distributed. It is positively skewed with most of the values below 500 and a peak at around 200.



There are several outliers in the cholesterol values for people who consume vitamins regularly. There are fewer outliers in the cholesterol values for people who never consume vitamins or who consume them occasionally. From the boxplots it can be observed that there isn't much of a difference in the cholesterol values between the people who take vitamins regularly, occasionally or never. For this reason, it may not be a good predictor of cholesterol.

Model with 'VitaminRecoded' as the predictor:

```
> summary(model1)

Call:
lm(formula = Cholesterol ~ VitaminRecoded, data = data)

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 ***
VitaminRecoded2   -1.156     19.270  -0.060               0.952    
VitaminRecoded1   -9.908     17.358  -0.571               0.569    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

> anova(model1)
Analysis of Variance Table

Response: Cholesterol
                Df  Sum Sq Mean Sq F value Pr(>F)
VitaminRecoded   2    6692  3345.8  0.1911 0.8262
Residuals      312 5463749 17512.0               
```

Model equation:

*Cholesterol = 246.599 - 1.156*VitaminRecoded2 - 9.908*VitaminRecoded1*

Model Interpretation:

Baseline: VitaminRecoded3 (No vitamin)

Category 1: VitaminRecoded2 (Occasional)

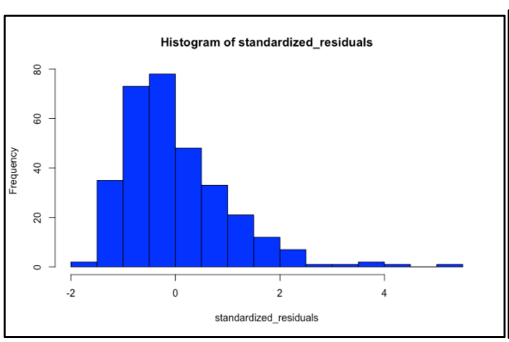
Category 2: VitaminRecoded1 (Regular)

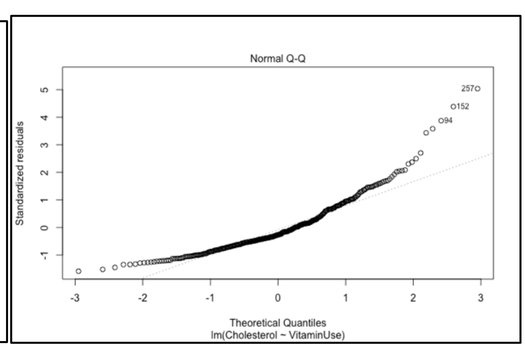
The baseline category is the one where the person doesn't consume any vitamins. The estimated cholesterol level for a person who doesn't take vitamins is 246.599. The estimated cholesterol level for a person who takes vitamins occasionally is 246.599 – 1.156 = 245.443. The estimated cholesterol level for a person who takes vitamins regularly is 246.599 – 9.908 = 236.691. However, the p-values associated with the dummy variables is large which indicates no real difference between the vitamin categories (regular and occasional) and the baseline category.

R-square = 0.00122 which means 0.1% of the variance in cholesterol is explained by VitaminUse.

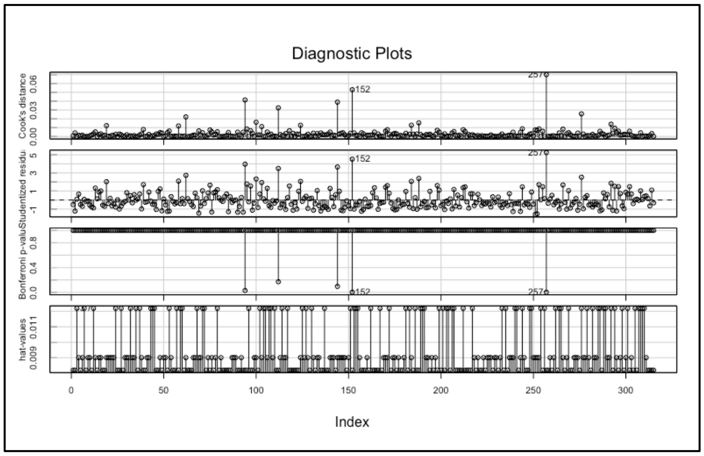
Omnibus F-test has a p-value of 0.82 which suggests that there is no significant relationship between the predictor and the response variable.

__**Residual plots:**__





Residuals are positively skewed which is expected since the response variable is also positively skewed.



From the diagnostic plot it can be observed that there are a few outliers and influential points.

<u>Recoding the 'VitaminUse' variable and rebuilding the model:</u>

```
> #Rebuild the model
> model2 <- lm(Cholesterol~VitaminRecoded,data = data)
> summary(model2)

Call:
lm(formula = Cholesterol ~ VitaminRecoded, data = data)

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 ***
VitaminRecoded2   -1.156     19.270  -0.060               0.952
VitaminRecoded3   -9.908     17.358  -0.571               0.569
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

> anova(model2)
Analysis of Variance Table

Response: Cholesterol
                Df  Sum Sq Mean Sq F value Pr(>F)
VitaminRecoded   2    6692  3345.8  0.1911 0.8262
Residuals      312 5463749 17512.0
```

<u>Model equation:</u>

*Cholesterol = 246.599 - 1.156*VitaminRecoded2 - 9.908*VitaminRecoded3*

<u>Model Interpretation:</u>

Baseline: VitaminRecoded1 (No vitamin)

Category 1: VitaminRecoded2 (Occasional)

Category 2: VitaminRecoded3 (Regular)

The baseline category is the one where the person doesn't consume any vitamins. The estimated cholesterol level for a person who doesn't take vitamins is 246.599. The estimated cholesterol level for a person who takes vitamins occasionally is 246.599 – 1.156 = 245.443. The estimated cholesterol level for a person who takes vitamins regularly is 246.599 – 9.908 = 236.691. However, the p-values associated with the dummy variables is large which indicates no real difference between the vitamin categories (regular and occasional) and the baseline category.

No matter which category is considered as the baseline, the model interpretation always remains the same. The coefficients are adjusted to accurately describe the effect of each dummy variable.

<u>Q.3</u>

Manually created two dummy variables: VitaminOccasional and VitaminNo

	VitaminOcassional	VitaminNo
Regular	0	0

Occasional	1	0
No	0	1

<u>New model fit with the dummy variables:</u>

```
> #Manually create dummy variables with "Regular" as the base class
> data$VitaminOccasional <- ifelse(data$VitaminUse=='Occasional',1,0)
> data$VitaminNo <- ifelse(data$VitaminUse=='No',1,0)
> #Build the model
> model3 <- lm(Cholesterol~VitaminOccasional+VitaminNo,data = data)
> summary(model3)

Call:
lm(formula = Cholesterol ~ VitaminOccasional + VitaminNo, data = data)

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)        236.691     11.981  19.756 <0.0000000000000002 ***
VitaminOccasional    8.752     18.897   0.463               0.644    
VitaminNo            9.908     17.358   0.571               0.569    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

<u>Model equation:</u>

*Cholesterol = 236.691 + 8.752*VitaminOccasional + 9.908*VitaminNo*

<u>Model Interpretation:</u>

Baseline: Regular Vitamin

Category 1: VitaminOccasional

Category 2: VitaminNo

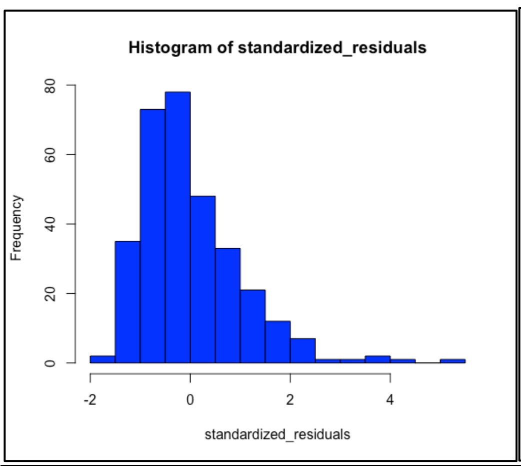
The baseline category is the one where the person consumes vitamins regularly. The estimated cholesterol level for such a person is 236.691. The estimated cholesterol level for a person who takes vitamins occasionally is 236.691 + 8.752 =

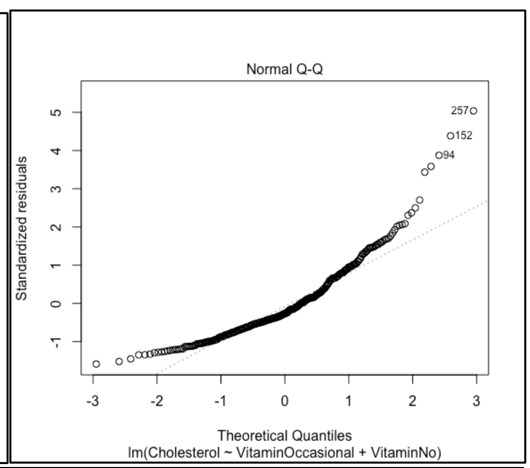
245.443. The estimated cholesterol level for a person who takes vitamins regularly is 236.691+ 9.908 = 246.599. However, the p-values associated with the dummy variables is large which indicates no real difference between the vitamin categories (no and occasional) and the baseline category.

R-square = 0.00122 which means 0.1% of the variance in cholesterol is explained by VitaminUse.

Omnibus F-test has a p-value of 0.82 which suggests that there is no significant relationship between the predictor and the response variable.

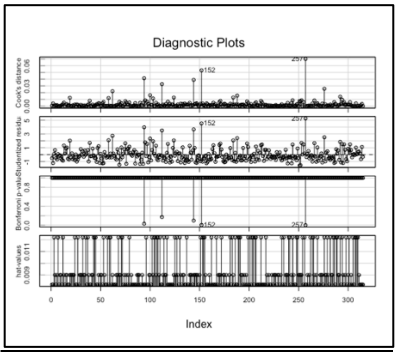
<u>Residual plots:</u>





Residuals are positively skewed. They do not follow a normal distribution.

<u>Diagnostic plots</u>



As observed in task 2, there are a few outliers/ influential points.

There is no difference between the results in Q.2 and Q.3 because when a factor variable is passed as a regressor, R automatically encodes it into dummy variables. The number of dummy variables is equal to 1 minus the number of categories. Although, the coefficients are different in this model, the final interpretation is the same as the model in Q.2.

Q.4

VitaminUse is effect coded with "No" as the comparative group.

	VitaminOcc (effect coded)	VitaminReg (effect coded)
"Regular"	0	1
"Occasional"	1	0
"No"	-1	-1

```
> #Effect coding with "No" as the comparative group
> data$VitaminOcc <- ifelse(data$VitaminUse=='Occasional',1,ifelse(data$VitaminUse=='Regular',0,-1))
> data$VitaminReg <- ifelse(data$VitaminUse=='Regular',1,ifelse(data$VitaminUse=='Occasional',0,-1))
> head(data)
  ID Age Smoke Quetelet Calories  Fat Fiber Alcohol Cholesterol BetaDiet RetinolDiet
1  1  64    No  21.4838   1298.8 57.0   6.3     0.0       170.3     1945         890
2  2  76    No  23.8763   1032.5 50.1  15.8     0.0        75.8     2653         451
3  3  38    No  20.0108   2372.3 83.6  19.1    14.1       257.9     6321         660
4  4  40    No  25.1406   2449.5 97.5  26.5     0.5       332.6     1061         864
5  5  72    No  20.9850   1952.1 82.6  16.2     0.0       170.8     2863        1209
6  6  40    No  27.5214   1366.9 56.0   9.6     1.3       154.6     1729        1439
  BetaPlasma RetinolPlasma Gender VitaminUse PriorSmoke VitaminRecoded GenderRecoded
1        200           915 Female    Regular          2              1             1
2        124           727 Female    Regular          1              1             1
3        328           721 Female Occasional          2              2             1
4        153           615 Female         No          2              3             1
5         92           799 Female    Regular          1              1             1
6        148           654 Female         No          2              3             1
  SmokeRecoded VitaminOccasional VitaminNo VitaminOcc VitaminReg
1            1                 0         0          0          1
2            1                 0         0          0          1
3            1                 1         0          1          0
4            1                 0         1         -1         -1
5            1                 0         0          0          1
6            1                 0         1         -1         -1
```

Model building:

```
> #Build the model
> model4 <- lm(Cholesterol~VitaminOcc+VitaminReg,data = data)
> summary(model4)

Call:
lm(formula = Cholesterol ~ VitaminOcc + VitaminReg, data = data)

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     2.532     11.331   0.223               0.823
VitaminReg    -6.220     10.250  -0.607               0.544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

> anova(model4)
Analysis of Variance Table

Response: Cholesterol
            Df  Sum Sq Mean Sq F value Pr(>F)
VitaminOcc   1     243   242.5  0.0138 0.9064
VitaminReg   1    6449  6449.0  0.3683 0.5444
Residuals  312 5463749 17512.0
```

Model equation and Interpretation:

*Cholesterol = 242.911 +2.532*VitaminOcc -6.220*VitaminReg*

If the person consumes no vitamins:

Cholesterol = 242.911 +2.532*(-1) -6.220*(-1) = 246.599

If the person consumes vitamins regularly:

Cholesterol = 242.911 +2.532*0 - 6.220*(1) = 236.691

If the person consumes vitamins occasionally:

Cholesterol = 242.911 +2.532*(1) - 6.220*(0) = 245.443

No matter which type of encoding is used, the model will return the same results.

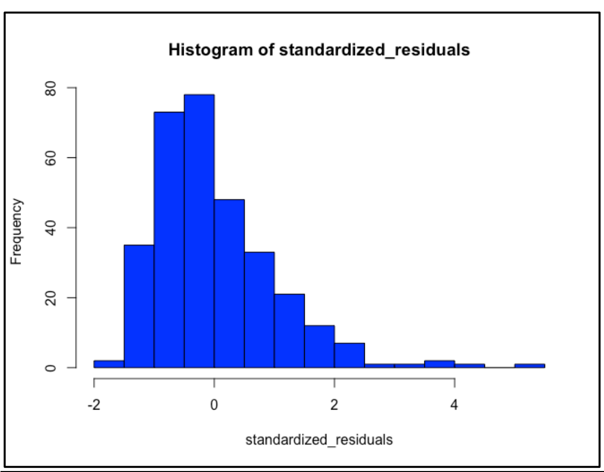
The model interpretation also remains the same.

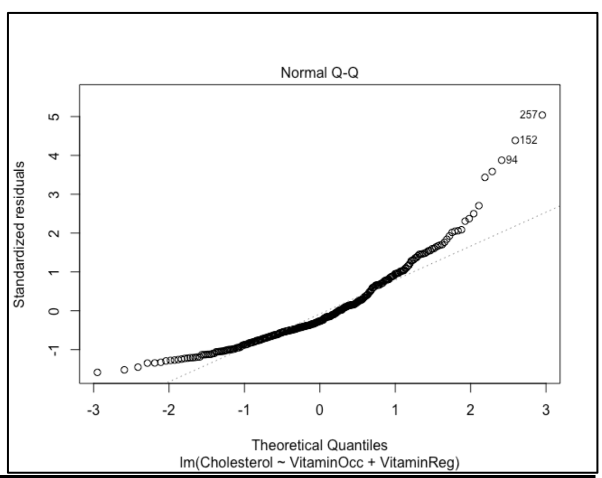
I would prefer using the dummy variables since the indicator variables can only take two values either 0 or 1. The assignment of values to the indicator variables is easier and the calculations too.

R-square = 0.00122 which means 0.1% of the variance in cholesterol is explained by VitaminUse.

Omnibus F-test has a p-value of 0.82 which suggests that there is no significant relationship between the predictor and the response variable.

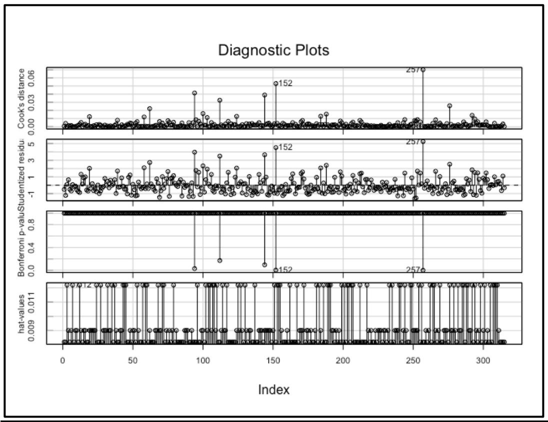
Residual plots:

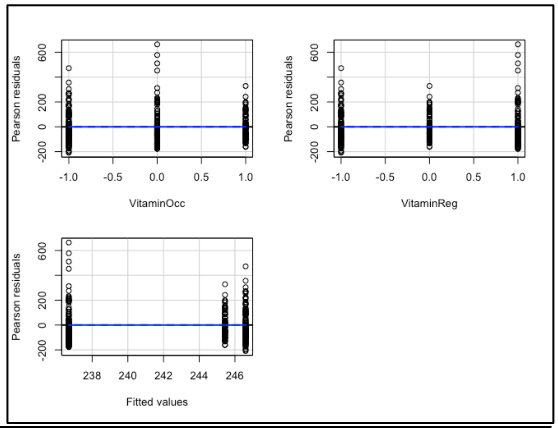




Residuals are positively skewed. Some outliers are also visible in the qq plot.

Influential Index plots:





Observations 257, 152 are clear outliers.

Q.5

Alcohol variable is discretized to form a categorical variable with three categories:

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

From these, two effect coded indicator variables are created and added to the dataset.

```
> #Q.5
> #Discretize 'Alcohol' variable
> data$AlcoholCatg <- ifelse(data$Alcohol==0,0,ifelse((data$Alcohol>0) & (data$Alcohol<10),1,2))
> #Indicator effect coded variables
> data$Alcohol1 <- ifelse(data$AlcoholCatg==1,1,ifelse(data$AlcoholCatg==2,0,-1))
> data$Alcohol2 <- ifelse(data$AlcoholCatg==2,1,ifelse(data$AlcoholCatg==1,0,-1))
> data[1:5,c("Alcohol","AlcoholCatg","Alcohol1","Alcohol2")]
  Alcohol AlcoholCatg Alcohol1 Alcohol2
1     0.0           0       -1       -1
2     0.0           0       -1       -1
3    14.1           2        0        1
4     0.5           1        1        0
5     0.0           0       -1       -1
```

Q.6

Created four product variables using the effect coded Vitamin and Alcohol variables:

```
#Creating interaction variables
data$VitaminOcc_Alcohol1 <- data$VitaminOcc*data$Alcohol1
data$VitaminOcc_Alcohol2 <- data$VitaminOcc*data$Alcohol2
data$VitaminReg_Alcohol1 <- data$VitaminReg*data$Alcohol1
data$VitaminReg_Alcohol2 <- data$VitaminReg*data$Alcohol2
```

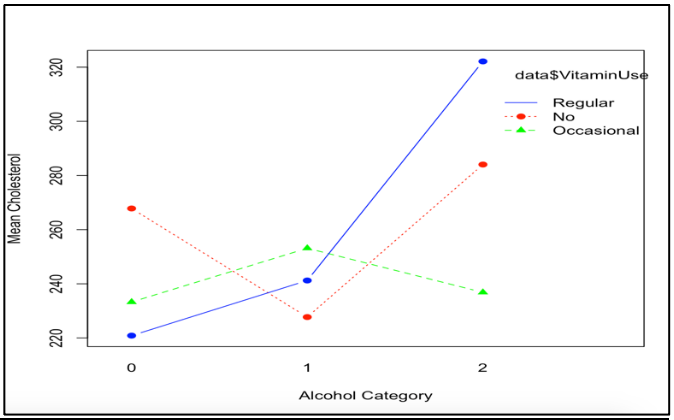
Full model containing interaction variables:

```
> #Full model with interaction variables
> fullmodel <- lm(Cholesterol~VitaminOcc+VitaminReg+Alcohol1+Alcohol2+VitaminOcc_Alcohol1+VitaminOcc_Alcohol2+
+                 VitaminReg_Alcohol1 + VitaminReg_Alcohol2,data =data)
> summary(fullmodel)

Call:
lm(formula = Cholesterol ~ VitaminOcc + VitaminReg + Alcohol1 +
    Alcohol2 + VitaminOcc_Alcohol1 + VitaminOcc_Alcohol2 + VitaminReg_Alcohol1 +
    VitaminReg_Alcohol2, data = data)

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 ***
VitaminOcc           -13.035     15.610  -0.835               0.404
VitaminReg             7.290     15.608   0.467               0.641
Alcohol1             -13.424     12.103  -1.109               0.268
Alcohol2              26.891     19.055   1.411               0.159
VitaminOcc_Alcohol1   25.474     17.790   1.432               0.153
VitaminOcc_Alcohol2  -31.129     27.761  -1.121               0.263
VitaminReg_Alcohol1   -6.757     17.513  -0.386               0.700
VitaminReg_Alcohol2   33.836     28.580   1.184               0.237
---
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
```

Reduced model without the interaction variables:

```
> #Reduced model
> reducedmodel <- lm(Cholesterol~VitaminOcc+VitaminReg+Alcohol1+Alcohol2,data = data)
> summary(reducedmodel)

Call:
lm(formula = Cholesterol ~ VitaminOcc + VitaminReg + Alcohol1 +
    Alcohol2, data = data)

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 ***
VitaminOcc     2.449     11.339   0.216               0.829
VitaminReg    -4.790     10.333  -0.464               0.643
Alcohol1     -12.901     11.672  -1.105               0.270
Alcohol2      26.621     18.216   1.461               0.145
---
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
```

<u>Interaction plot of VitaminUse and AlcoholCatg:</u>



It can be observed from the plot that there is some interaction between VitaminUse and AlcoholCatg.

<u>Comparison of reduced and final models:</u>

```
> #Comparison of reduced and full model
> anova(reducedmodel,fullmodel)
Analysis of Variance Table

Model 1: Cholesterol ~ VitaminOcc + VitaminReg + Alcohol1 + Alcohol2
Model 2: Cholesterol ~ VitaminOcc + VitaminReg + Alcohol1 + Alcohol2 +
    VitaminOcc_Alcohol1 + VitaminOcc_Alcohol2 + VitaminReg_Alcohol1 +
    VitaminReg_Alcohol2
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1    310 5426297
2    306 5342216  4     84081 1.204 0.3091
```

Ho: Reduced model is adequate

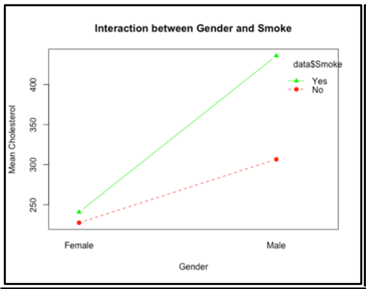
(or)

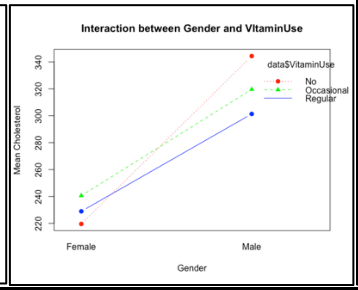
The coefficients of the interaction terms are zero

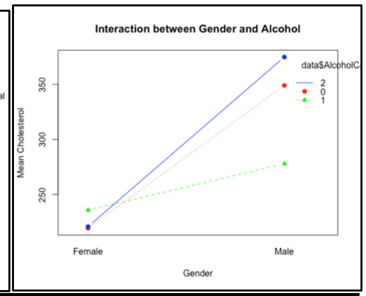
Ha: Full model is adequate

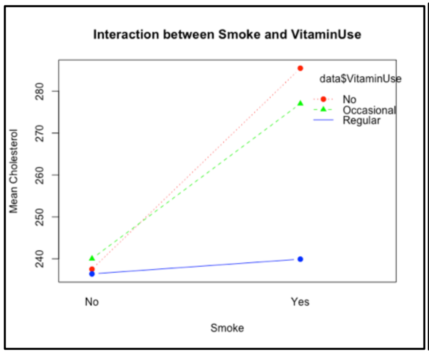
<u>F-test, p-value</u>: 0.3091 which suggests that although interaction is observed in the plots, there isn't enough evidence to suggest that it is significant. The reduced model without the interaction terms is adequate.

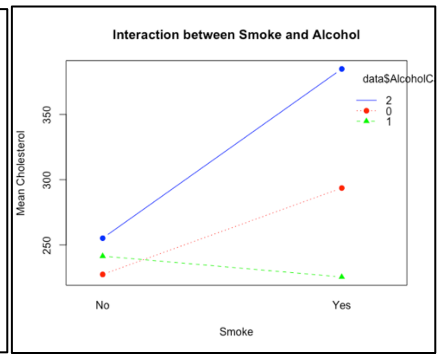
Q.7











ANOVA comparisons:

```
Analysis of Variance Table

Model 1: Cholesterol ~ Gender + Smoke
Model 2: Cholesterol ~ Gender + Smoke + Gender * Smoke
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1    312 5078043
2    311 5011965  1     66078 4.1002 0.04373 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Variance Table

Model 1: Cholesterol ~ Gender + VitaminUse
Model 2: Cholesterol ~ Gender + VitaminUse + Gender * VitaminUse
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    311 5103169
2    309 5080751  2     22419 0.6817 0.5065
```

```
Analysis of Variance Table

Model 1: Cholesterol ~ Gender + AlcoholCatg
Model 2: Cholesterol ~ Gender + AlcoholCatg + Gender * AlcoholCatg
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    312 5107177
2    311 5106891  1    286.03 0.0174 0.8951
```

```
Analysis of Variance Table

Model 1: Cholesterol ~ Smoke + AlcoholCatg
Model 2: Cholesterol ~ Smoke + AlcoholCatg + Smoke * AlcoholCatg
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    312 5405504
2    311 5405197  1    307.16 0.0177 0.8943
```

```
Analysis of Variance Table

Model 1: Cholesterol ~ Smoke + VitaminUse
Model 2: Cholesterol ~ Smoke + VitaminUse + Smoke * VitaminUse
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    311 5422373
2    309 5410426  2     11947 0.3412 0.7112
```

Results:

Variable pairs	Any interaction in the plots?	p-value of ANOVA	Significant interaction?
Gender, Smoke	Yes	0.04	Yes
Gender, VitaminUse	Yes	0.5	No
Gender, AlcoholCatg	Yes	0.89	No
Smoke, AlcoholCatg	Yes	0.89	No
Smoke, VitaminUse	Yes	0.71	No
VitaminUse, alcoholCatg	Yes	0.3	No

Q.8

Through this assignment I have learnt how categorical variables behave in a regression model. I learnt how to interpret the coefficients and the different kinds of encoding such as dummy, effect. It is a good practice to always start with a simple model and then add more predictors to make it complex until the desired results are achieved. Plots can be constructed to check for interaction/dependency between two predictors. However, interaction observed in the plots may not always be significant. Hence, tests should be performed to check if there is enough evidence to suggest that the interaction is not due to chance.