Robin Singh

NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Computational Assignment #3: OLS Regression Modeling with Categorical Variables

MSDS 410

This third computational assignment builds on your prior modeling and computing experiences. You may begin to work on this assignment anytime you wish.

Data: The data for this assignment is the Nutrition Study data: NutritionStudy.CSV It is a 16 variable dataset with n=315 records. The data was obtained from medical record information and observational self-report of adults. The dataset consists of categorical, continuous, and composite scores of different types. A data dictionary is not available for this dataset, but the qualities measured can easily be inferred from the variable and categorical names for most of the variables. As such, higher scores for the composite variables translate into having more of that quality. The QUETELET variable is essentially a body mass index. It can be googled for more detailed information. It is the ratio of BodyWeight (in lbs) divided by (Height (in inch))^2. Then the ratio is adjusted with an adjustment factor so that the numbers become meaningful. Specifically, QUETELET above 25 is considered overweight, while a QUETELET above 30 is considered obese. There is no other information available about this data.

Objective: Use multiple regression to predict CHOLESTEROL using models with categorical variables. Please note: This assignment is not prescriptive of what you "should do" as an analysis. It is intended to give you experience conducting and reporting on different kinds of multiple regression models.

Tasks: To achieve the objective please complete the following tasks enumerated below. You are to use R to obtain any graphs or statistics requested.

For these analyses, let the response variable be: Y = CHOLESTEROL. The remaining variables will be considered explanatory variables, X's.

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.

```r
mydata$VitaminUseNum[mydata$VitaminUse == "Regular"] <- 1
mydata$VitaminUseNum[mydata$VitaminUse == "Occasional"] <- 2
mydata$VitaminUseNum[mydata$VitaminUse == "No"] <- 3
mydata$VitaminUseNum
```

```r
mydata
```

Alcohol <dbl>	Cholesterol <dbl>	BetaDiet <int>	RetinolDiet <int>	BetaPlasma <int>	RetinolPlasma <int>	Gender <chr>	VitaminUse <chr>	PriorSmoke <int>	VitaminUseNum <dbl>
0.0	170.3	1945	890	200	915	Female	Regular	2	1
0.0	75.8	2653	451	124	727	Female	Regular	1	1
14.1	257.9	6321	660	328	721	Female	Occasional	2	2
0.5	332.6	1061	864	153	615	Female	No	2	3
0.0	170.8	2863	1209	92	799	Female	Regular	1	1
1.3	154.6	1729	1439	148	654	Female	No	2	3
0.0	255.1	5371	802	258	834	Female	Occasional	1	2
0.0	214.1	823	2571	64	825	Female	Regular	1	1
0.6	233.6	2895	944	218	517	Female	No	1	3
0.0	171.9	3307	493	81	562	Female	No	2	3

1–10 of 315 rows | 8–17 of 17 columns

Previous 1 2 3 4 5 6 ... 32 Next

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?

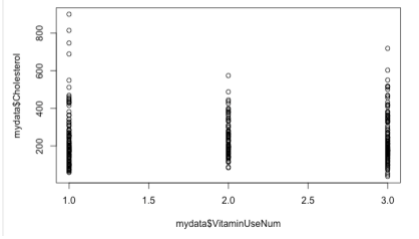
```r
model1=lm(mydata$Cholesterol~mydata$VitaminUseNum)
model1
summary(model1)
```

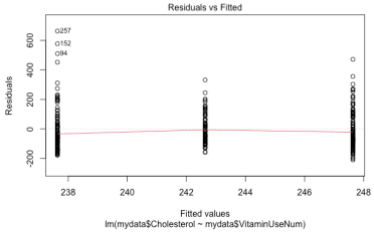
```
Call:
lm(formula = mydata$Cholesterol ~ mydata$VitaminUseNum)

Coefficients:
        (Intercept)  mydata$VitaminUseNum
            232.634                 5.001


Call:
lm(formula = mydata$Cholesterol ~ mydata$VitaminUseNum)

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)          232.634     18.581  12.520   <2e-16 ***
mydata$VitaminUseNum   5.001      8.663   0.577    0.564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

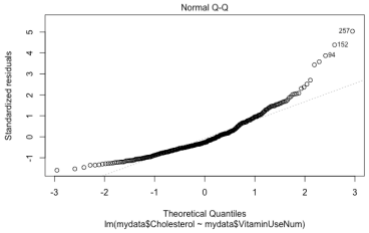
Y = 232.634 + 5.001*B1
Y is the amount of cholesterol, B1 is the Vitamin usage as determined previously (1=never, 2=occasional, 3=regular).

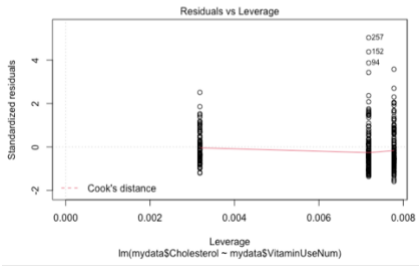
Hypothesis test:
Null: B1= 0
Alt: B1!= 0
t-value=0.577
t-value to analyze against is approximately 2.5.
0.577<2.5, which means the null hypothesis cannot be disproved, and has a significant chance of being true.
For goodness of fit, the p-value is 0.5642, meaning there is a high chance of the null hypothesis being true.
We also have a very low R^2 value of 0.001063, and an adjusted R^2 of -0.002128, indicated a poorly fit model.

```{r}
plot(mydata$VitaminUseNum,mydata$Cholesterol)
```









There is no change, besides in direction of line which is now negative, upon flipping the variables in the model.

```{r }
mydata$VitaminUseNum2[mydata$VitaminUse == "Regular"] <- 3
mydata$VitaminUseNum2[mydata$VitaminUse == "Occasional"] <- 2
mydata$VitaminUseNum2[mydata$VitaminUse == "No"] <- 1

model1.1=lm(mydata$Cholesterol~mydata$VitaminUseNum2)
model1.1
summary(model1.1)
```

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$VitaminUseNum2)

Coefficients:
          (Intercept)  mydata$VitaminUseNum2
              252.637                -5.001


Call:
lm(formula = mydata$Cholesterol ~ mydata$VitaminUseNum2)

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)          252.637     19.137  13.202   <2e-16 ***
mydata$VitaminUseNum2   -5.001      8.663  -0.577    0.564
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

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?

```r
model2=lm(mydata$Cholesterol~mydata$vitReg+mydata$vitOcc)
model2
summary(model2)
```

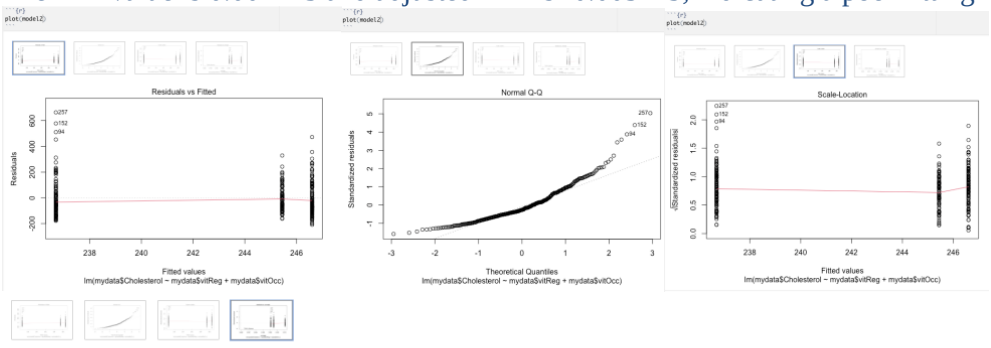
```
Call:
lm(formula = mydata$Cholesterol ~ mydata$vitReg + mydata$vitOcc)

Coefficients:
  (Intercept)  mydata$vitReg  mydata$vitOcc
      246.599         -9.908         -1.156


Call:
lm(formula = mydata$Cholesterol ~ mydata$vitReg + mydata$vitOcc)

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   <2e-16 ***
mydata$vitReg  -9.908     17.358  -0.571    0.569
mydata$vitOcc  -1.156     19.270  -0.060    0.952
---
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
```

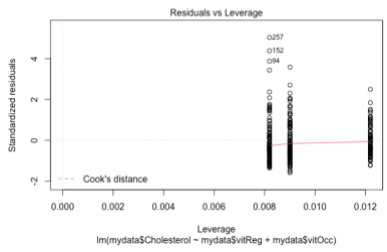
Y=246.599 + -9.908*B1 + -1.156*B2
Y is the amount of cholesterol, -9.908 is the effect of regularly using Vitamins (B1), and -1.156 is the effect of occasionally using vitamins (B2), all compared to never using vitamins, which was the variable that was omitted form the model.

Hypothesis test:
Null: B1=B2=0
Alt: B1 != B2 != 0
The t-values for B1 and B2 are -0.571 and -0.060, which are both less than the threshold of around 2.5, meaning that the null hypothesis cannot be disregarded because it may hold.

The p-value is 0.8262 which is high, indicating the null hypothesis holds.

The R^2 value is 0.001223 and adjusted R^2 is -0.005179, indicating a poor fitting model.





There are not many differences in the model. It still indicates that vitamins alon are a poor predictor of cholesterol.

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?

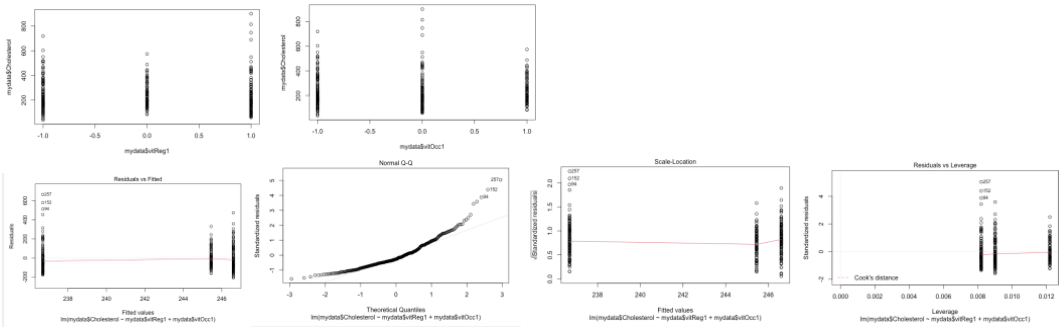
```
Call:
lm(formula = mydata$Cholesterol ~ mydata$vitReg1 + mydata$vitOcc1)

Coefficients:
   (Intercept)  mydata$vitReg1  mydata$vitOcc1
       242.911          -6.220           2.532


Call:
lm(formula = mydata$Cholesterol ~ mydata$vitReg1 + mydata$vitOcc1)

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   <2e-16 ***
mydata$vitReg1   -6.220     10.250  -0.607    0.544
mydata$vitOcc1    2.532     11.331   0.223    0.823
---
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

Analysis of Variance Table

Response: mydata$Cholesterol
               Df  Sum Sq Mean Sq F value Pr(>F)
mydata$vitReg1   1    5817  5817.3  0.3322 0.5648
mydata$vitOcc1   1     874   874.2  0.0499 0.8233
Residuals      312 5463749 17512.0
```

Y=242.911 -6.220*B1 + 2.532*B2
Y is the amount of cholesterol, -6.22 is the effect of regularly using Vitamins (B1), and 2.532 is the effect of occasionally using vitamins (B2), all compared to never using vitamins, which was the variable that was control variable.

Hypothesis test:
Null: B1=B2=0
Alt: B1 != B2 != 0
The t-values for B1 and B2 are -0.607 and -0.223, which are both less than the threshold of around 2.6, meaning that the null hypothesis cannot be disregarded because it may hold.

The p-value is 0.8262 which is high, indicating the null hypothesis holds. This is the same as the dummy coding.

The R^2 value is 0.001223 and adjusted R^2 is -0.005179, indicating a poor fitting model, same as before.



I do like this way better, as it takes into account the control variable in some way, including it in the model.

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.

```r
mydata$AlcoholNumHi[mydata$Alcohol >= 10] <- 1
mydata$AlcoholNumHi[mydata$Alcohol == 0] <- -1
mydata$AlcoholNumHi[mydata$Alcohol > 0 & mydata$Alcohol < 10] <- 0

mydata$AlcoholNumMid[mydata$Alcohol >= 10] <- 0
mydata$AlcoholNumMid[mydata$Alcohol == 0] <- -1
mydata$AlcoholNumMid[mydata$Alcohol > 0 & mydata$Alcohol < 10] <- 1
mydata
```

vitOcc <dbl>	vitNo <dbl>	alcNo <dbl>	alcOcc <dbl>	alcHi <dbl>	AlcoholNum <dbl>	vitReg1 <dbl>	vitOcc1 <dbl>	AlcoholNumHi <dbl>	AlcoholNumMid <dbl>
0	0	1	0	0	0	1	0	-1	-1
0	0	1	0	0	0	1	0	-1	-1
1	0	0	0	1	2	0	1	1	0
0	1	0	1	0	1	-1	-1	0	1
0	0	1	0	0	0	1	0	-1	-1
0	1	0	1	0	1	-1	-1	0	1
1	0	1	0	0	0	0	1	-1	-1
0	0	1	0	0	0	1	0	-1	-1
0	1	0	1	0	1	-1	-1	0	1
0	1	1	0	0	0	-1	-1	-1	-1

1–10 of 315 rows | 20–29 of 29 columns

Previous 1 2 3 4 5 6 ... 32 Next

The 'No Alcohol' is the control variable.

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.

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$regHi + mydata$regMid +
    mydata$occHi + mydata$occMid + mydata$AlcoholNumMid + mydata$AlcoholNumHi +
    mydata$vitReg1 + mydata$vitOcc1)

Coefficients:
        (Intercept)          mydata$regHi         mydata$regMid          mydata$occHi         mydata$occMid  mydata$AlcoholNumMid   mydata$AlcoholNumHi         mydata$vitReg1         mydata$vitOcc1
            254.116                 33.836                 -6.757                -31.129                 25.474                -13.424                 26.891                  7.290                -13.035


Call:
lm(formula = mydata$Cholesterol ~ mydata$regHi + mydata$regMid +
    mydata$occHi + mydata$occMid + mydata$AlcoholNumMid + mydata$AlcoholNumHi +
    mydata$vitReg1 + mydata$vitOcc1)

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   <2e-16 ***
mydata$regHi          33.836     28.580   1.184    0.237
mydata$regMid         -6.757     17.513  -0.386    0.700
mydata$occHi         -31.129     27.761  -1.121    0.263
mydata$occMid         25.474     17.790   1.432    0.153
mydata$AlcoholNumMid -13.424     12.103  -1.109    0.268
mydata$AlcoholNumHi   26.891     19.055   1.411    0.159
mydata$vitReg1         7.290     15.608   0.467    0.641
mydata$vitOcc1       -13.035     15.610  -0.835    0.404
---
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

Analysis of Variance Table

Response: mydata$Cholesterol
                     Df  Sum Sq Mean Sq F value Pr(>F)
mydata$regHi          1   35376   35376  2.0263 0.1556
mydata$regMid         1   11125   11125  0.6372 0.4253
mydata$occHi          1    1721    1721  0.0986 0.7538
mydata$occMid         1   27847   27847  1.5951 0.2076
mydata$AlcoholNumMid  1     327     327  0.0187 0.8912
mydata$AlcoholNumHi   1   39613   39613  2.2690 0.1330
mydata$vitReg1        1      42      42  0.0024 0.9611
mydata$vitOcc1        1   12174   12174  0.6973 0.4043
Residuals           306 5342216   17458
```

For the Reduced model, fit an OLS multiple regression model using only the effect coded variables for VITAMIN and ALCOHOL to predict CHOLESTEROL.

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$regHi + mydata$regMid +
    mydata$occHi + mydata$occMid)

Coefficients:
  (Intercept)   mydata$regHi  mydata$regMid   mydata$occHi  mydata$occMid
     244.6492        22.2208        -0.3898       -19.6190        17.5866


Call:
lm(formula = mydata$Cholesterol ~ mydata$regHi + mydata$regMid +
    mydata$occHi + mydata$occMid)

Residuals:
    Min      1Q  Median      3Q     Max
-204.35  -91.41  -36.34   63.04  677.88

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   244.6492     7.5520  32.395   <2e-16 ***
mydata$regHi   22.2208    19.0795   1.165    0.245
mydata$regMid  -0.3898    13.0868  -0.030    0.976
mydata$occHi  -19.6190    20.6195  -0.951    0.342
mydata$occMid  17.5866    13.9020   1.265    0.207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 131.9 on 310 degrees of freedom
Multiple R-squared:  0.01391,   Adjusted R-squared:  0.001182
F-statistic: 1.093 on 4 and 310 DF,  p-value: 0.3601

Analysis of Variance Table

Response: mydata$Cholesterol
              Df  Sum Sq Mean Sq F value Pr(>F)
mydata$regHi   1   35376   35376  2.0329 0.1549
mydata$regMid  1   11125   11125  0.6393 0.4246
mydata$occHi   1    1721    1721  0.0989 0.7534
mydata$occMid  1   27847   27847  1.6003 0.2068
Residuals    310 5394372   17401
```

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.

Null: B5=B6=B7=B8=B9=0

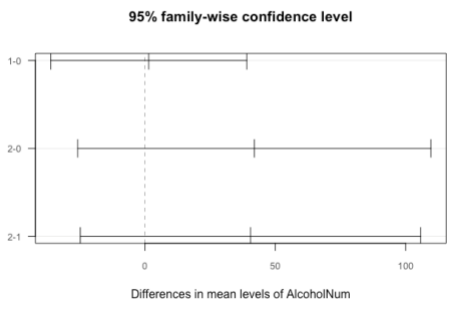
Alt: One does not equal 0

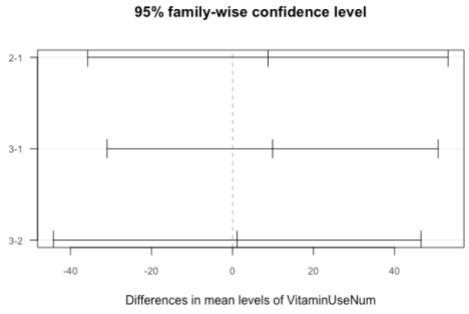
F-test= ((5394372-5342216)/(8-4))/(5394372/(315-8))=0.7421

F-stat to analyze against=2.37

0.7421<2.37

We cannot reject the null hypothesis because the f-stat for our model is lower than the f-stat to analyze against, meaning that the four additional variables in the full model make a significant difference in our model.





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.

```
Call:
lm(formula = mydata$Cholesterol ~ mydata$SmokeBin + mydata$Male)

Coefficients:
    (Intercept)  mydata$SmokeBin      mydata$Male
         225.13            31.49            97.75


Call:
lm(formula = mydata$Cholesterol ~ mydata$SmokeBin + mydata$Male)

Residuals:
    Min      1Q  Median      3Q     Max
-245.38  -84.83  -31.73   58.87  675.57

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)      225.129      8.201  27.453 < 0.0000000000000002
mydata$SmokeBin   31.491     20.949   1.503                0.134
mydata$Male       97.746     21.158   4.620          0.00000563

Residual standard error: 127.6 on 312 degrees of freedom
Multiple R-squared:  0.07173,   Adjusted R-squared:  0.06578
F-statistic: 12.05 on 2 and 312 DF,  p-value: 0.00000906

Analysis of Variance Table

Response: mydata$Cholesterol
                Df  Sum Sq Mean Sq F value      Pr(>F)
mydata$SmokeBin   1   45033   45033  2.7669     0.09724
mydata$Male       1  347365  347365 21.3424 0.000005626
Residuals       312 5078043   16276
```

F-test:
Null:B9=B10=0
Alt: One of them is not zero

```
```{r}
anova(fullModel2,fullModel)
```
```

```
Analysis of Variance Table

Model 1: mydata$Cholesterol ~ mydata$regHi + mydata$regMid + mydata$occHi +
    mydata$occMid + mydata$AlcoholNumMid + mydata$AlcoholNumHi +
    mydata$vitReg1 + mydata$vitOcc1 + mydata$SmokeBin + mydata$Male
Model 2: mydata$Cholesterol ~ mydata$regHi + mydata$regMid + mydata$occHi +
    mydata$occMid + mydata$AlcoholNumMid + mydata$AlcoholNumHi +
    mydata$vitReg1 + mydata$vitOcc1
  Res.Df     RSS Df Sum of Sq      F     Pr(>F)
1    304 5017925
2    306 5342216 -2   -324291 9.8232 0.00007345
```
```

We can reject the null hypothesis that this has no relations combined with the Vitamin and Alcohol variables on the Cholesterol variable because the f-stat value is significantly greater at 9.8232, meaning there is evidence supporting the alternate hypothesis.

Individually, Smoke has a higher impact on the Cholesterol variable with an F-stat of 21.342, while gender significantly less with 2.2597.

```
```{r}
anova(redModel2,smokeModel)
```
```

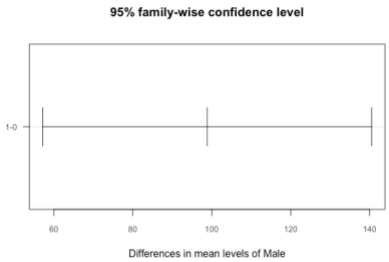
```
Analysis of Variance Table

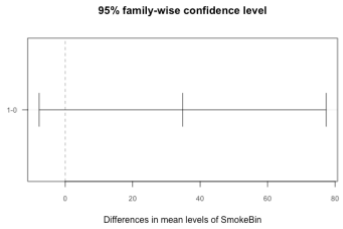
Model 1: mydata$Cholesterol ~ mydata$SmokeBin + mydata$Male
Model 2: mydata$Cholesterol ~ mydata$SmokeBin
  Res.Df   RSS Df Sum of Sq   F   Pr(>F)
1   312 5078043
2   313 5425408 -1   -347365 21.342 0.000005626
```
```

```
```{r}
anova(redModel2,genderModel)
```
```

```
Analysis of Variance Table

Model 1: mydata$Cholesterol ~ mydata$SmokeBin + mydata$Male
Model 2: mydata$Cholesterol ~ mydata$Male
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    312 5078043
2    313 5114821 -1    -36779 2.2597 0.1338
```
```
```





8) Please write a reflection on your experiences from this assignment.

This assignment was initially easy but got a little confusing as things went on. Some of the questions were a little tough to understand in terms of what exactly they wanted, and many were more loaded than others. Still, this taught a lot about how to incorporate categorical features into linear models, and how much of an impact they can truly have on the overall prediction. I did not know how various techniques of implementing them may have different impacts on the overall model.