Scott Jue
MSDS 410
2/05/2023

Modeling Assignment #5: Modeling with Categorical Explanatory Variables

Introduction

For this analysis we will be using data collected from a nutrition study to develop several regression models using categorical variables as the explanatory variables to predict the response variable which will be cholesterol. The dataset contains 16 variables and has 315 observations. Four of these variables are categorical which are Smoke, Gender, VitaminUse, and PriorSmoke. I have defined the sample population as individuals who are smoking and non-smoking adults between the ages 19 and 83, who eat 4,000 calories or less and have a Beta Plasma level greater than 0. After these filters have been applied to the dataset, there are a total of 312 observations remaining in the dataset. To prepare the data for regression modeling, I have transformed these categorical variables and also created dummy coded variables to represent the different categorial variables. I recoded the Gender variable to be 0 for female and 1 for male and also the Smoke variable to be 0 for No and 1 for Yes. Furthermore, dummy variables have been created for the different categories/levels for the VitamineUse and PriorSmoke variables. Lastly, the Alcohol variable has been transformed and converted to Alcohol consumption categorical variable by levels of None (Alchol = 0), Some (Alcohol >0 and <10), and A lot (Alcohol >10). These levels have been discretized and coded with dummy variables.

This is observational data from self-reported adults and it is unknown if this is a random sample. Furthermore, from my exploratory data analysis, the data does not appear to show normal distribution patterns. As such, the results and purpose of this analysis cannot be used for statistical inference. This modeling exercise and analysis being performed below is for the purpose finding the best fitting a model using categorical variables on the observed dataset.

Results:

1. The below table displays the descriptive stats for Cholesterol by the PriorSmoke group. We can see that each group is associated with a different mean value. This means that the PriorSmoke group that an individual is associated with can have a significant impact on the individual's cholesterol level. PriorSmoke group 1 has the lowest mean cholesterol level of 220.27. PriorSmoke group 2 had a higher mean cholesterol level of 250.42. And PriorSmoke group 3 had the highest mean cholesterol level of 264.66. There appears to be a relationship between the group number and cholesterol as the higher group number is associated with a higher mean cholesterol level. For this reason, the PriorSmoke variable will be considered as an explanatory variable to predict the response variable Cholesterol.

DESCRIPTIVE STATISTICS

	n	mean	sd	min	max
1	155	220.27	114.09	37.70	689.40
2	115	250.42	121.69	46.30	747.50
3	42	264.66	138.15	78.30	718.80

Grand Mean: 237.361

2. For the first model we will use the dummy coded variables for PriorSmoke as explanatory variables to predict the response variable Cholesterol (Y). Below is the prediction equation for this model.

Model 1:

Y_hat = 220.271 + 30.153*d2_PriorSmoke + 44.393*d3_PriorSmoke

This intercept for this equation can be interpreted as if d2_PriorSmoke and d3_PriorSmoke are both 0, then the predicted cholesterol level is 220.27. For context, d2_PriorSmoke and d3_PriorSmoke will only both be 0 when the individual falls in to the PriorSmoke group 1. As such, PriorSmoke group 1 is the basis of interpretation for this model. The coefficient for d2_PriorSmoke can be interpreted as when an individual is associated with PriorSmoke group 2, then the predicted cholesterol increases by 30.15 units from the 220.27 intercept. The coefficient for d3_PriorSmoke can be interpreted as when an individual is associated with PriorSmoke group 3, then the predicted cholesterol increases by 44.39 units from the 220.27 intercept. It is important to note that since these dummy variables reflect the category of the PriorSmoke group, only one of these variables can have a value of 1 in the prediction equation. Therefore, the predicted cholesterol level can take on one of three values. The first value is the intercept value (220.27) when both dummy variables are 0, meaning the individual is in PriorSmoke group 1. The second value is when the individual is in PriorSmoke group 2 and the predict value is the intercept value plus the 30.15 increase which id 250.42. The third value that is when the individual is in the PriorSmoke group 3 category and the predicted value is the intercept value plus the 44.39 increase which is 264.66.

Model 1 Coefficient Summary Table:

```
Call:
lm(formula = Cholesterol ~ d2_PriorSmoke + d3_PriorSmoke, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-204.12  -86.45  -25.56   64.15  497.08

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)    220.271      9.667  22.787 <0.0000000000000002 ***
d2_PriorSmoke   30.153     14.812   2.036              0.0426 *
d3_PriorSmoke   44.393     20.935   2.121              0.0348 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120.3 on 309 degrees of freedom
Multiple R-squared:  0.02104,   Adjusted R-squared:  0.01471
F-statistic: 3.321 on 2 and 309 DF,  p-value: 0.0374
```

Model 1 ANOVA Table:

```
Analysis of Variance Table

Response: Cholesterol
               Df  Sum Sq Mean Sq F value  Pr(>F)
d2_PriorSmoke   1   31080   31080  2.1459 0.14397
d3_PriorSmoke   1   65125   65125  4.4965 0.03476 *
Residuals     309 4475397   14483
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results from the regression model are very much related to the results from the ANOVA model in Task 1. As mentioned previously, PriorSmoke group 1 is the basis of interpretation for the model; therefore, the intercept value is equal to the mean value of PriorSmoke group 1 that is shown in the ANOVA table from Task 1. Additionally, the coefficient for d2_PriorSmoke is the difference between the mean values of group 1 and group 2. Therefore, when the individual is in group 2 the predict cholesterol value is equal the mean of group 2 and this is achieved by adding the difference in means to the group 1 mean value which is also the intercept value. Similarly, the coefficient for d3_PriorSmoke is the difference between the mean values of group 1 and group 3. Therefore, when the individual is in group 3 the predict cholesterol value is equal the mean of group 3 seen in the Task 1 ANOVA table and this is achieved by adding the difference in means to the group 1 mean value.

3. For Model 2, we will start with Model 1 and add in the continuous variable Fat to fit a multiple linear model to predict the response variable, Cholesterol. Below is the model equation, coefficient table, and ANOVA table.

   Model 2:

   Y_hat = 41.4939 + 3.7929*d2_PriorSmoke + 14.2748*d3_PriorSmoke + 2.5348*Fat

   For this model, PriorSmoke group 1 will be used as the basis of interpretation again. This model can be interpreted as if the individual is in PriorSmoke group 1 then their predicted cholesterol value is 41.49 which is the intercept value plus 2.54 units for every unit of fat. If an individual is in group 2, then the predict value increases by 3.79 units from group plus 2.54 units for every unit of fat. If an individual is in group 3, then the predict value increases by 14.27 units from group 1 plus 2.54 units for every unit of fat. For all groups, the predicted cholesterol increases by 2.54 for every unit of fat increase, so the only change attributable to going from PriorSmoke group 1 to group 2 or group 3 can be observed in the dummy variable parameters.

Model 2 Coefficient Summary Table:

```
Call:
lm(formula = Cholesterol ~ d2_PriorSmoke + d3_PriorSmoke + Fat,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-190.43  -51.60  -11.87   27.10  514.07

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   41.4938    13.4977   3.074              0.0023 **
d2_PriorSmoke  3.7929    11.1912   0.339              0.7349
d3_PriorSmoke 14.2748    15.7561   0.906              0.3657
Fat            2.5348     0.1617  15.677 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.9 on 308 degrees of freedom
Multiple R-squared:  0.4555,    Adjusted R-squared:  0.4502
F-statistic: 85.89 on 3 and 308 DF,  p-value: < 0.00000000000000022
```

Model 2 ANOVA Table:

```
Analysis of Variance Table

Response: Cholesterol
               Df  Sum Sq Mean Sq  F value               Pr(>F)
d2_PriorSmoke   1   31080   31080   3.8457              0.05077 .
d3_PriorSmoke   1   65125   65125   8.0583              0.00483 **
Fat             1 1986225 1986225 245.7674 < 0.0000000000000002 ***
Residuals     308 2489172    8082
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model has an R-squared value of 0.46 which means that these explanatory variables are only able to account for approximately 46% of the variance observed in cholesterol.

Hypothesis Tests:

For the below Omnibus Overall F-test, we will use a critical of 2.6339 which is the value associated with degrees of freedom of 3 and 308, at the 0.05 significance level.

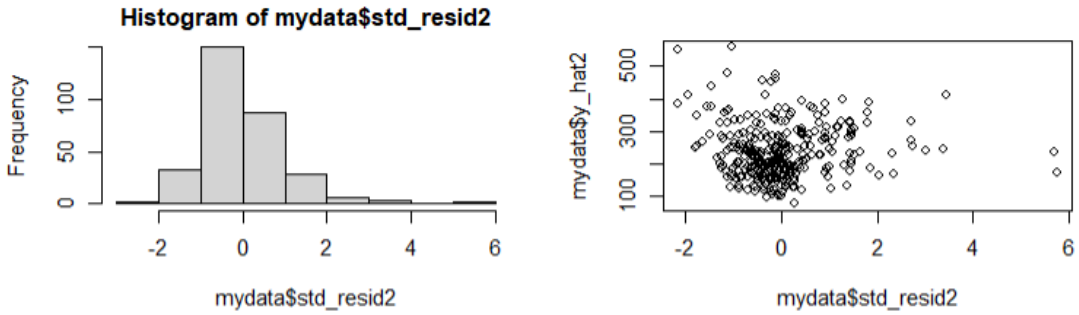
Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

From the Coefficient Summary Table, we can see that the F-statistic of 85.89 is greater than the critical F-value of 2.6339. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Hypothesis Test Underlying Assumptions:

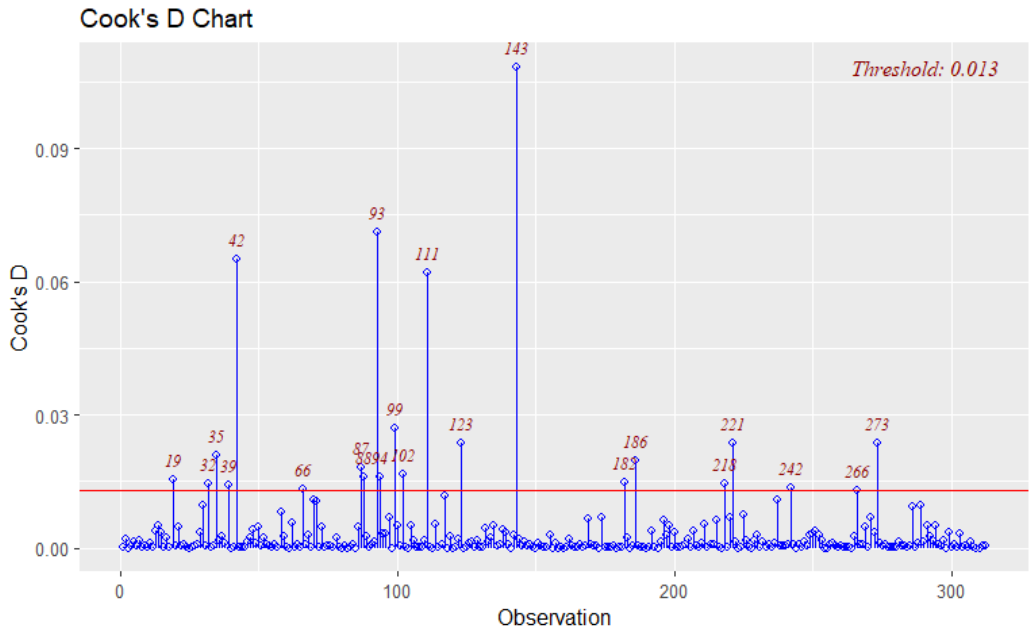
We can see in the below histogram that the distribution of the standardized residuals is right-skewed. Additionally, the scatterplot shows that there is some heteroscedasticity seen in the residuals as the lower predicted cholesterol levels have higher standardize residuals compared to the higher predicted cholesterol levels which show to have slightly negative standardize residuals. Therefore, the underlying assumptions have been violated and the hypothesis tests do not provide much value.

Histogram and Scatterplot of Standardized Residuals:

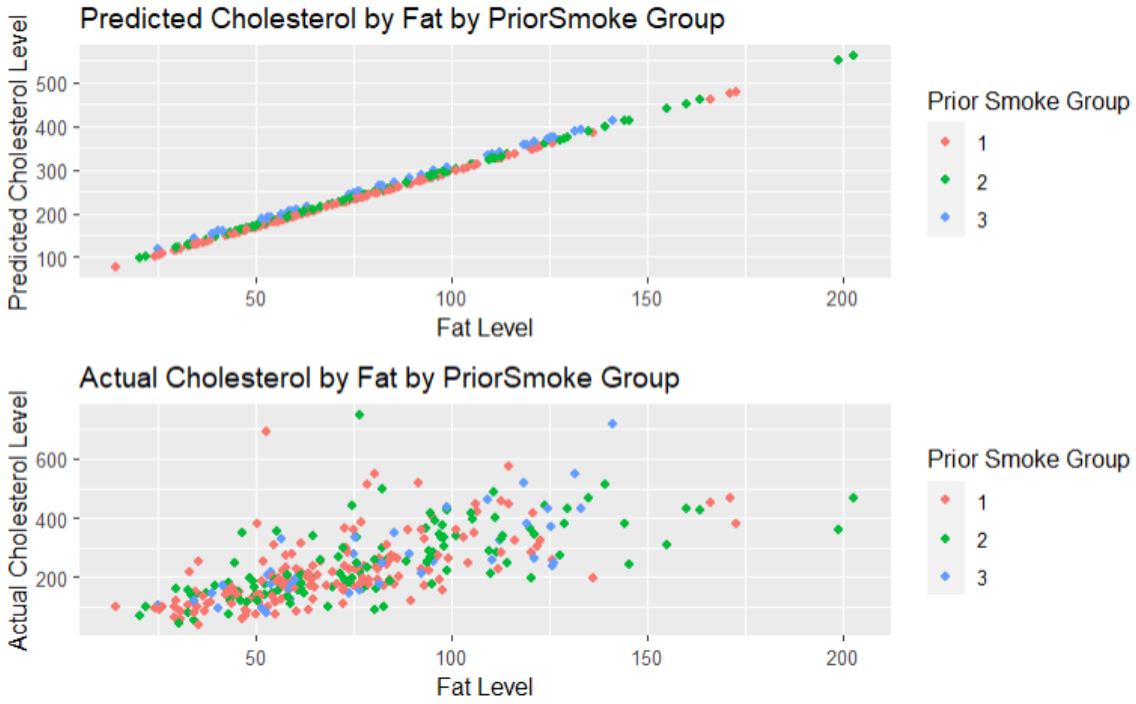


Influential Observations:

The below Cook's Distance chart shows several observations that are greater than threshold limit of 0.013 indicating that these are influential observations. We should be concerned that these observations may pull the regression model too far in one direction, thereby decreasing the overall fitness of the model. We could potentially remove these observations to increase the model performance and fit. However, these observations could also be valid representations of the sample population. So, further analysis should be done on these influential observations before removing them and refitting the model.



4. In the below scatterplots using the predict cholesterol level and the actual cholesterol level by fat level and colorized by group, we can see how this model has all the predicted values in a straight line. We can see in the plot with the actual cholesterol level that a similar linear pattern is present; however, it could be possible that there is a slight difference in slopes between groups. When looking at the higher levels of fat, we can see that the Prior Smoke Group 3 points (blue) are slightly higher up on the y-axis. Groups 1 and 2 points (red and green) are lower on the y-axis. This suggests that these groups may potentially have uneven slopes. In which case, a more complex model could be considered in order to better fit the observed data.



5. For Model 3 we will create an Unequal Slopes Model to look at the interaction between the explanatory variables PriorSmoke and Fat. To do this I have created interaction variables (fat_smoke1, fat_smoke2, and fat_smoke3) by multiplying the dummy coded variables for PriorSmoke and the Fat variable value. Starting with the ANCOVA model from Model 2, we will add in these newly created interaction variables to the multiple regression model to predict the response variable, Cholesterol. To remain consistent with the prior models, I will leave out PriorSmoke group 1's dummy variable and interaction variable from the model, so that group 1 can be the basis of interpretation again. Below is the model equation, coefficient table, and ANOVA table.

   Model 3:

   Y_hat = 41.2814 + 2.5379*Fat + 23.8104*d2_PriorSmoke - 53.1546*d3_PriorSmoke – 0.2477*fat_smoke2 + 0.8178*fat_smoke3

Since PriorSmoke group 1 will be used as the basis of interpretation again, the interpretation of this model is when an individual is in group 1, meaning all dummy variables and interaction variables equal 0, then the predicted cholesterol value is the intercept value of 41.28 plus 2.54 units for every unit of Fat. If an individual is in group 2, then d2_PriorSmoke will have a value of 1, meaning that predicted cholesterol value increases by 23.81 units from group 1 and plus 2.54 units for every unit of Fat. Then, we also need to take in to account the interaction variable for PriorSmoke group 2. The interpretation of this coefficient is that cholesterol decreases by -0.25 for every unit increase in this interaction variable. If an individual is in group 3, then d3_PriorSmoke will have a value of 1, meaning that predicted cholesterol value decreases by 53.15 units from group 1 and plus 2.54 units for every unit of Fat. Then we also need to take in to account the interaction variable for PriorSmoke group 3. The interpretation of this coefficient is that cholesterol increases by 0.82 for every unit increase in this interaction variable. These interaction variables effectively will produce a different slope for each PriorSmoke group. For all groups, the predicted cholesterol increases by 2.54 for every unit of fat increase. Therefore, the only change in predicted cholesterol from PriorSmoke group 1 to group 2 or group 3 is observed in the dummy variables and interaction variable parameters.

Model 3 Coefficient Summary Table:

```
Call:
lm(formula = Cholesterol ~ Fat + d2_PriorSmoke + d3_PriorSmoke +
    fat_smoke2 + fat_smoke3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-190.63  -52.57   -9.95   28.25  514.12

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   41.2814    19.3370   2.135              0.0336 *
Fat            2.5379     0.2545   9.971 <0.0000000000000002 ***
d2_PriorSmoke 23.8104    28.4633   0.837              0.4035
d3_PriorSmoke -53.1546   43.0497  -1.235              0.2179
fat_smoke2     -0.2477     0.3475  -0.713              0.4764
fat_smoke3      0.8178     0.5045   1.621              0.1060
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.52 on 306 degrees of freedom
Multiple R-squared:  0.4636,    Adjusted R-squared:  0.4549
F-statistic:  52.9 on 5 and 306 DF,  p-value: < 0.00000000000000022
```

Model 3 ANOVA Table:

```
Analysis of Variance Table

Response: Cholesterol
               Df  Sum Sq Mean Sq  F value               Pr(>F)
Fat             1 2075770 2075770 259.0331 <0.0000000000000002 ***
d2_PriorSmoke   1      27      27   0.0033               0.9541
d3_PriorSmoke   1    6634    6634   0.8278               0.3636
fat_smoke2      1   15974   15974   1.9934               0.1590
fat_smoke3      1   21057   21057   2.6277               0.1060
Residuals     306 2452141    8014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This model has an R-squared value of 0.46 which means that these explanatory variables are only able to account for approximately 46% of the variance observed in cholesterol. This is only a slight increase from Model 2.

Hypothesis Tests:

For the below Omnibus Overall F-test, we will use a critical value of 2.2435 which is the value associated with degrees of freedom of 5 and 306, at the 0.05 significance level.

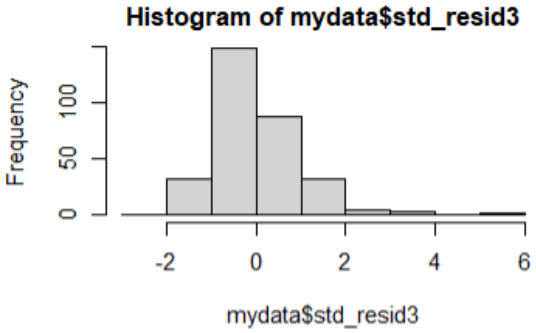
Null hypothesis → H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
Alternative hypotheses → H_A: at least one $\beta_i \neq 0$

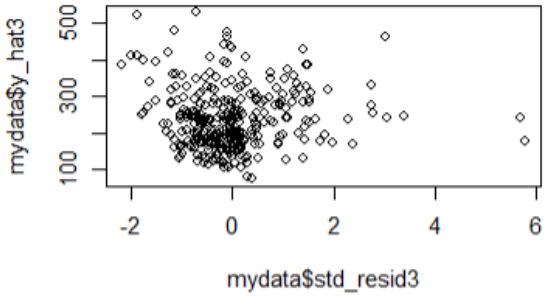
From the Coefficient Summary Table, we can see that the F-statistic of 52.9 is greater than the critical F-value of 2.2435. Therefore, we should reject the null hypothesis and conclude that at least one $\beta_i \neq 0$. This indicates that there is a significant relationship between the independent variables and the response variable.

Hypothesis Test Underlying Assumptions:

We can see similar standardized residual patterns as Model 2 in the below histogram and scatterplot. The histogram shows that the distribution of the standardized residuals is right-skewed. Additionally, the scatterplot shows that there is some heteroscedasticity seen in the residuals as the lower predicted cholesterol levels have higher standardize residuals compared to the higher predicted cholesterol levels which show to have slightly negative standardize residuals. Therefore, the underlying assumptions have been violated and the hypothesis tests do not provide much value.

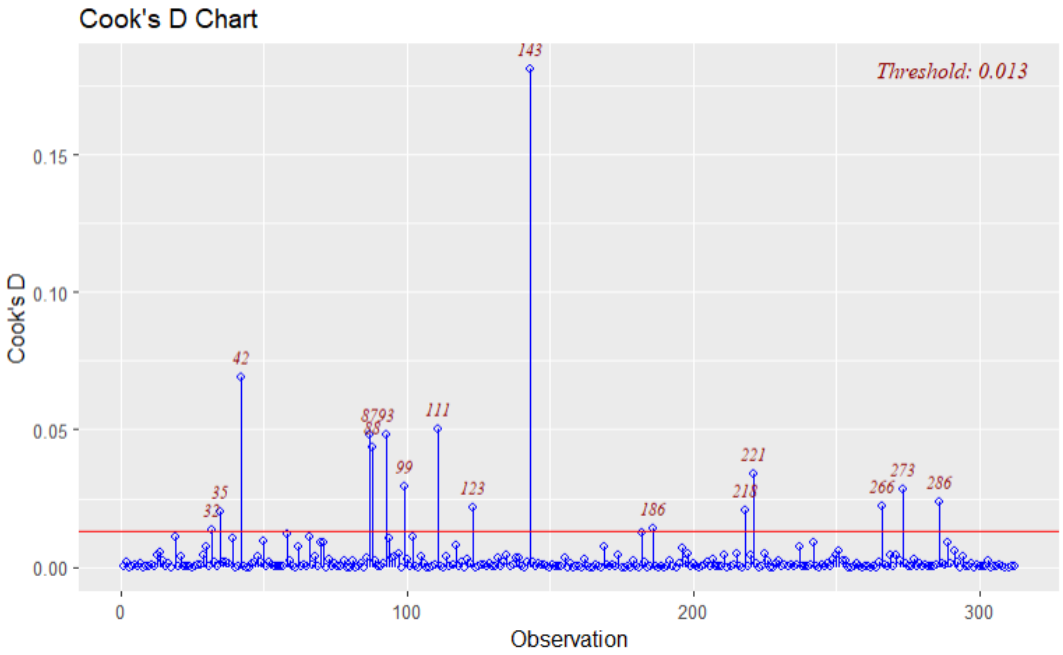
Histogram and Scatterplot of Standardized Residuals:





Influential Observations:

The below Cook's Distance chart shows several observations that are greater than threshold limit of 0.013 indicating that these are influential observations. However, there is not as many influential observations as compared to model 2, but it appears these observations are slightly more influencing since their Cook's Distance is greater. Nonetheless, we should be concerned that these observations may pull the regression model too far in one direction, thereby decreasing the overall fitness of the model. We could potentially remove these observations to increase the model performance and fit. However, these observations could also be valid representations of the sample population. So, further analysis should be done on these influential observations before removing them and refitting the model.



6. In the below scatterplots using the predict cholesterol level and the actual cholesterol level by fat level and colorized by group, we can see how Model 3 has produces predicted values in along separate lines of unequal slopes. The red and green points have a similar slope; however, the blue points have a greater slope. This resembles more of the scatterplot with the actual cholesterol level.

The unequal slopes approach used in Model 3 appears to be a better fit than the previous ANCOVA model that did not include the interaction variables. Therefore, we can see there is some interaction between the PriorSmoke group and fat variable that leads to Group 3 behaving different than Group 1 and Group 2.

7. Model 3 is the full model because Model 2 is nested in Model 3. For this reason, Model 2 is can be considered a reduced model of Model 3. Next, we will use the nested F-test using Model 2 (reduced model) and Model 3 (full model) to determine if the additional intercation variables are useful for predicting cholesterol. The critical F-value for at the 0.05 significance level for a reduced model degrees of freedom equal to 2 in the numerator and a full model degrees of freedom equal to 306 in the denominator is 3.0253. Below are the null and alternate hypotheses for the nested F-test.

   Null hypothesis → H_0: $\beta_4 = \beta_5 = 0$
   Alternative hypotheses → H_A: $\beta_4 \neq 0$ or $\beta_5 \neq 0$

   The computed F-value of 2.3105 is less than the critical F-value of 3.0016. Therefore, we can fail to reject the null hypothesis that the unequal slopes paraments are equal to 0 and conclude that the reduced model is a better model for fitting the observed data.

8. Next, I will explore other categorical variables in the dataset using the above modeling approaches to determine if the variables Smoke, Alcohol Consumption, or Gender along with the Fat variable are most predictive of Cholesterol.

Smoke + Fat:

First, I fitted a reduced model that used the Smoke dummy variables and the fat variable, then I fitted a full model which adds interaction variables and compared using the models using R-squared values and a nested f-test results.

Null hypothesis → H_0: $\beta_3 = 0$
Alternative hypotheses → H_A: $\beta_3 \neq 0$

The computed F-value of 4.033 is greater than the critical F-value of 3.8718. Therefore, we can reject the null hypothesis that the unequal slopes paraments are equal to 0 and conclude that the full model is a better model for fitting the observed data. Additionally, the R-squared is 0.4624 for the full model and 0.4553 for the reduced model. Therefore, the full model with the added interaction variables is able to account for slightly more variance in cholesterol level.

Alcohol Consumption + Fat:

Next, I fitted a reduced model that used the Alcohol consumption dummy variables and the fat variable, then I fitted a full model which adds interaction variables and compared using the models using R-squared values and a nested f-test results.

Null hypothesis → H_0: $\beta_3 = 0$
Alternative hypotheses → H_A: $\beta_3 \neq 0$

The computed F-value of 1.141 is less than the critical F-value of 3.0253. Therefore, we can fail to reject the null hypothesis that the unequal slopes paraments related to the interaction variables are equal to 0 and conclude that the reduced model is a better model for fitting the observed data. Additionally, the R-squared is 0.4587 for the full model and 0.4547 for the reduced model so there is no material improvement by including the interaction variables between alcohol consumption and fat variables.

Gender + Fat:

Lastly, I fitted a reduced model that used the Gender dummy variables and the fat variable, then I fitted a full model which adds interaction variables and compared using the models using R-squared values and a nested f-test results.

Null hypothesis → H_0: $\beta_4 = \beta_5 = 0$
Alternative hypotheses → H_A: $\beta_4 \neq 0$ or $\beta_5 \neq 0$

The computed F-value of 0.9288 is less than the critical F-value of 3.8718. Therefore, we can fail to reject the null hypothesis that the unequal slopes paraments related to the interaction variables are equal to 0 and conclude that the reduced model is a better model for fitting the observed data.

Additionally, the R-squared is 0.4752 for the full model and 0.4736 for the reduced model so there is no material improvement by including the interaction variables between gender and fat variables.

Out of all the categorical variables that were looked at for this analysis, the Gender variable when combined with the Fat variable produced the most predictive model having the highest R-squared and adjusted R-squared values of 0.47. Additionally. a nested F-test determined that adding interaction variables between Gender and Fat do not add any predictive value to the model.

Conclusion:

From this analysis we were able to develop various multiple linear regression models using the categorical variables contained in the dataset by creating dummy variables. We were able to determine if different categories had different mean values in the response variable, Cholesterol. Additionally, we were able to test if there was significant interaction between the categorical variable and another continuous explanatory variable, Fat that provided any information for predicting cholesterol levels. It was determined that the Gender variable had provided the most predictive value when combined with Fat. For further analysis in the future, I would test other continuous variables to see if they could also provide useful information for predicting cholesterol in combination with these categorical variables. One categorical variable not included in this analysis was VitaminUse, so that would be another categorical variable I would consider looking at in future work. Additionally, I am not familiar some of these variables such as BetaDiet, RetinolDiet, BetaPlasma, or RetinolPlasma, so I am unsure what values are in the acceptable range. Therefore, in my analysis above I could have included observations with obscure values that might not be feasibly represent the sample population. This could potentially lead to poor fitting models; however, I did not want to make an uneducated guess at what values were outside the normal range as I did not want to introduce this kind of subjectivity or bias. But perhaps, this is why I observed skewed distributions during the EDA process.

Overall, I learned a lot in this assignment and got to understand how to encode dummy variables. Coming from a finance background, I work with primarily continuous variables. So, I found this assignment particularly interesting as I was able to learn how to transform categorical data to be used in regression models. Additionally, I am getting more comfortable with the concepts from previous modules such as model interpretation, hypothesis testing, and evaluating goodness of fit and model diagnostics. These will become valuable tools in my analytic tool kit going forward.