Modeling Assignment 5:
Modeling with Categorical Explanatory Variables – ANOVA, ANCOVA, and Unequal Slopes Models

Assignment Overview

In this assignment we will use multiple regression to predict CHOLESTEROL.
The data for this assignment is from a study of nutrition. The Nutrition Study Data 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. Higher scores for the variables translate into having more of that quality.
There is one variable, called QUETELET, that 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 so that the numbers become meaningful. Specifically, QUETELET above 25 is considered overweight, while a QUETELET above 30 is considered obese.

Preparatory Work

There are 4 categorical variables. These are: SMOKE, GENDER, VITAMINUSE, and PRIORSMOKE. Some of these variables use numbers to indicate the levels of the categorical variables, others use text. For regression modeling purposes, you will most likely need to transform these variables, or construct new dummy coded variables. How you do this is as follows:

a) For any dichotomous categorical variable (i.e. a categorical variable with 2 levels), you want to recode such a variable so that the values (or numbers) that indicate the level are set to 0 and 1. The GENDER and SMOKE variables are like this. Often, an analyst will just create a new variable, like d_GENDER, that is the coded version of GENDER.
b) For categorical variables with 3 or more levels, you will need to construct a set of dummy coded (0/1) variables to indicate the levels. The VITAMINUSE and PRIORSMOKE variables are like this. Please see the Module 5 Classroom for directions on how to construct dummy coded variables. Each level must have its own dummy coded variable. As such, there should be 3 dummy coded variables for VITAMINUSE. Similarly, there will be 3 dummy coded variables for PRIORSMOKE.
c) Some analysts like to take continuous variables and discretize or convert them into categorical. For example, the ALCOHOL variable may be easier to work with or interpret results if it were converted into a variable called ALCOHOL CONSUMPTION with levels like: None, Some, A lot. In doing this, you could discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:
    1 if ALCOHOL = 0
    2 if 0 < ALCOHOL < 10
    3 if ALCOHOL >= 10

    Once you have the levels for the new ALCOHOL CONSUMPTION categorical variable, you would then dummy code these levels.

In preparation for modeling, you need to create dummy coded variables for the categorical variables in the Nutrition Study data set. Construct the ALCOHOL CONSUMPTION categorical variable and create dummy coded variables for it.

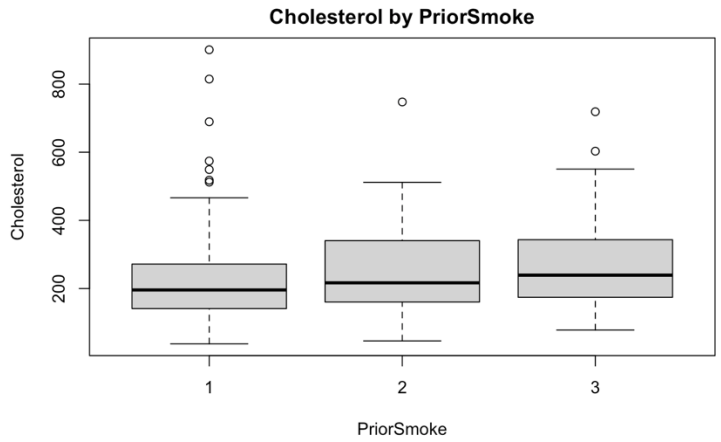
Assignment Tasks

For the tasks in this assignment, the response variable will be: CHOLESTEROL (Y). The remaining variables will be considered explanatory variables (X's).

1. Obtain descriptive statistics (n, mean, s, and any others you want) for Y by the PRIORSMOKE variable. Use the PRIORSMOKE variable as a factor in an ANOVA to test for mean differences in Cholesterol between PRIORSMOKE groups. **Report and interpret these results.**

Here are descriptive statistics for Cholesterol by PriorSmoke

	PriorSmoke=1	**PriorSmoke=2**	**PriorSmoke=3**
5 number summary of Cholesterol	Min: 37.7 1st quartile: 141.2 Median: 195.8 3rd quartile: 271.8 Max: 900.7	Min: 46.3 1st quartile: 160.4 Median: 216.7 3rd quartile: 340.6 Max: 747.5	Min: 78.3 1st quartile: 174.3 Median: 239.2 3rd quartile: 343.2 Max: 718.8
N of Cholesterol	157	115	43
Mean of Cholesterol	228.3911	250.4243	272.5326
SD of Cholesterol	134.2284	121.6916	145.9191



As the value of PriorSmoke increases from 1 to 3, the mean and median value of Cholesterol increases, and the number of datapoints (N) and distribution (SD) of Cholesterol decreases.

Model0: Cholesterol ~ as.factor(PriorSmoke)

```
Analysis of Variance Table

Response: Cholesterol
                       Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(PriorSmoke)   2   77258   38629  2.2347 0.1087
Residuals             312 5393183   17286
```

H_0: mean(Cholesterol | PriorSmoke=1) = mean(Cholesterol | PriorSmoke=3) = mean(Cholesterol | PriorSmoke=3)
H_a: mean(Cholesterol | PriorSmoke=1) != mean(Cholesterol | PriorSmoke=3) != mean(Cholesterol | PriorSmoke=3)

The F test with a p-value of 0.1087 indicates we fail to reject H_0, and conclude that the mean values of Cholesterol for each different group of PriorSmoke are *not* all the same.

2. Fit a linear regression model that uses the dummy coded variables for PRIORSMOKE to predict Cholesterol (Y). Call this Model 1. Remember: you need to leave one of the dummy coded variables out of the equation. That category becomes the "basis of interpretation." **Report the prediction equation and interpret each coefficient in the context of this problem. Report the coefficient and ANOVA tables from this regression model. Discuss how the results from the regression model compare and contrast to the results from the ANOVA model in Task 1.**

Model1: predicted(Cholesterol) = 228.39 + 22.03* d_PriorSmoke_2 + 44.14* d_PriorSmoke_3

The predicted value of Cholesterol for PriorSmoke=1 is **228.39**.
The predicted value of Cholesterol for PriorSmoke=2 is 228.39 + **22.03** = 250.42.
The predicted value of Cholesterol for PriorSmoke=3 is 228.39 + **44.14**= 272.53.

Coefficient Tables

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      228.39      10.49  21.766   <2e-16 ***
d_PriorSmoke_2    22.03      16.14   1.365    0.173
d_PriorSmoke_3    44.14      22.63   1.951    0.052 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131.5 on 312 degrees of freedom
Multiple R-squared:  0.01412,   Adjusted R-squared:  0.007803
F-statistic: 2.235 on 2 and 312 DF,  p-value: 0.1087
```

ANOVA Tables

```
Analysis of Variance Table

Response: Cholesterol
                Df  Sum Sq Mean Sq F value Pr(>F)
d_PriorSmoke_2   1   11487   11487  0.6645 0.4156
d_PriorSmoke_3   1   65771   65771  3.8049 0.0520 .
Residuals      312 5393183   17286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA Table in task 1 indicates that the mean value of Cholesterol is not the same for all 3 groups of PriorSmoke. The ANOVA Table in task 2 agrees and indicates that the mean value of Cholesterol is not the same between group 1 and group 2 of PriorSmoke and between group 1 and group 3 of PriorSmoke.

3. Model 1 illustrates the ANOVA model as a Linear Regression Model. Let's go a step further. Start with Model 1 and add in the continuous variable FAT. In other words, you are using FAT and PRIORSMOKE to predict Cholesterol, but you are using dummy coded variables for the PRIORSMOKE categorical variable. More specifically, fit a multiple linear model that uses the FAT continuous variable and the PRIORSMOKE dummy coded variables to predict the response variable CHOLESTEROL (Y). Remember to leave one of the dummy coded variables out of the model so that you have a basis of interpretation for the constant term. **Report the prediction model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics, if it is relevant. This is called an Analysis of Covariance Model (ANCOVA). Call this Model 2.**

Model2: predicted(Cholesterol) = 28.9401 + 2.7630*Fat - 2.1142* d_PriorSmoke_2 + 10.6358* d_PriorSmoke_3

The predicted value of Cholesterol for PriorSmoke=1 is **28.9401** (given Fat=0).
The predicted value of Cholesterol for PriorSmoke=2 is 28.9401 + **2.1142** = 31.0543 (given Fat=0).
The predicted value of Cholesterol for PriorSmoke=3 is 28.9401 + **10.6358** = 39.5759 (given Fat=0).
The predicted value of Cholesterol for each additional point of Fat increases by **2.7630**.

Coefficient Tables

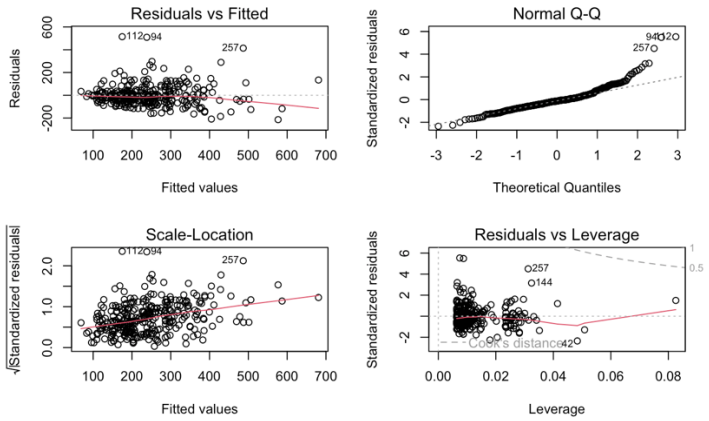
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    28.9401    13.5848   2.130   0.0339 *
Fat             2.7630     0.1574  17.556   <2e-16 ***
d_PriorSmoke_2 -2.1142    11.5372  -0.183   0.8547
d_PriorSmoke_3 10.6358    16.1763   0.657   0.5113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 93.33 on 311 degrees of freedom
Multiple R-squared:  0.5048,    Adjusted R-squared:  0.5001
F-statistic: 105.7 on 3 and 311 DF,  p-value: < 2.2e-16
```

R2 = 0.5048 : 50.48% of variation in Cholesterol can be explained by the predictors in model2.

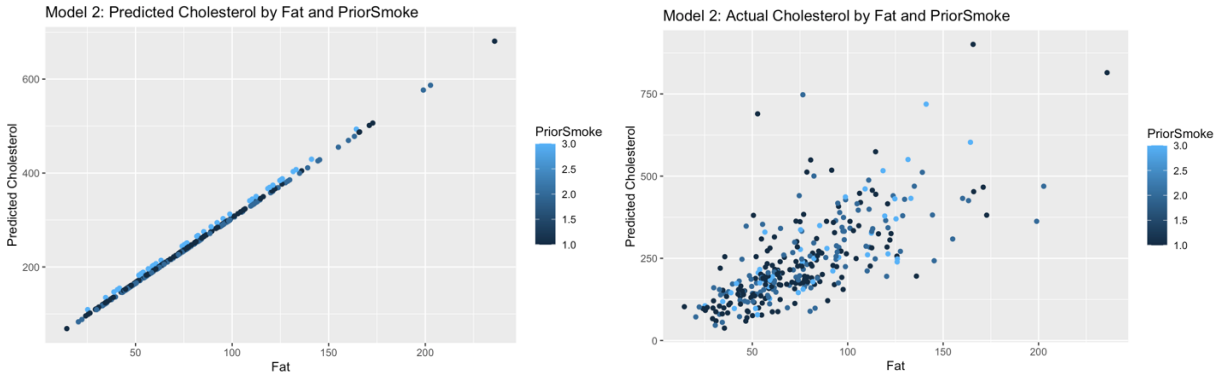
Variable	H_0	H_a	Conclusion
Intercept	$\beta_0 = 0$	$\beta_0 != 0$	Reject H_0
Fat	$\beta_1 = 0$	$\beta_1 != 0$	Reject H_0
d_PriorSmoke_2	$\beta_2 = 0$	$\beta_2 != 0$	Fail to reject H_0
d_PriorSmoke_3	$\beta_3 = 0$	$\beta_3 != 0$	Fail to reject H_0

The T-tests indicate that the intercept and coefficient for Fat are not 0, but that PriorSmoke is 0.

Note in the diagnostic plots below that the residuals vs fitted plot does not appear to violate any assumptions of linear regression. The residuals vs leverage plots also indicate that for CD = 4/N = 4/315 = 0.0127, there are several points that are influential or have high leverage in model2.



4. Use the ANCOVA Model 2 from Task 3) to obtain predicted values for CHOLESTEROL(Y). **Now, make a scatterplot of the Predicted Values for Y (y-axis) by FAT (X), but color code the records for the different groups of PRIORSMOKE. What do you notice about the patterns in the predicted values of Y? Make a second scatterplot of the actual values of CHOLESTEROL(Y) by FAT (X), but color code the data points by the different groups of the PRIORSMOKE variable. If you compare the two scatterplots, does the ANCOVA model appear to fit the observed data very well? Or, is a more complex model needed?**



On the left, model2 indicates that each group of PriorSmoke does not visually differ between each other. On the right, it appears that a more complex model is needed to fit the observed data better – PriorSmoke=3 are typically higher points whereas PriorSmoke=1 are typically lower points which is not accommodated for in model2's predicted values or Fat slopes.

5. Create new product variables by multiplying each of the dummy coded variables for PRIORSMOKE by the continuous FAT(X) variable. Name and save these product variables to your dataset. Now, to build the Unequal Slopes Model, start with the ANCOVA model, Model 2, from Task 3). Add in the interaction variables you just created. You now should have a multiple regression model with the predictor variables of: FAT, two dummy coded PRIORSMOKE variables, and two product variables. This is called an Unequal Slopes Model – call it Model 3. **Fit Model 3 and report the prediction equation, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, leverage, influence, and Outlier statistics, if warranted.**

Model3: predicted(Cholesterol) = 13.7032+ 2.9740*Fat + 51.3886* d_PriorSmoke_2 - 32.8823* d_PriorSmoke_3 - 0.6839*d_Fat_2 + 0.4858* d_Fat_3

The predicted value of Cholesterol for PriorSmoke=1 is **13.7032** (given Fat=0).
The predicted value of Cholesterol for PriorSmoke=2 is 13.7032 + **51.3886** = 65.0918 (given Fat=0).
The predicted value of Cholesterol for PriorSmoke=3 is 13.7032 **- 32.8823** = -19.1791 (given Fat=0).

The predicted value of Cholesterol for each additional point of Fat (for PriorSmoke=1) increases by **2.9740**.
The predicted value of Cholesterol for each additional point of Fat (for PriorSmoke=2) decreases by **0.6839**.
The predicted value of Cholesterol for each additional point of Fat (for PriorSmoke=3) increases by **0.4858**.

Coefficient Tables

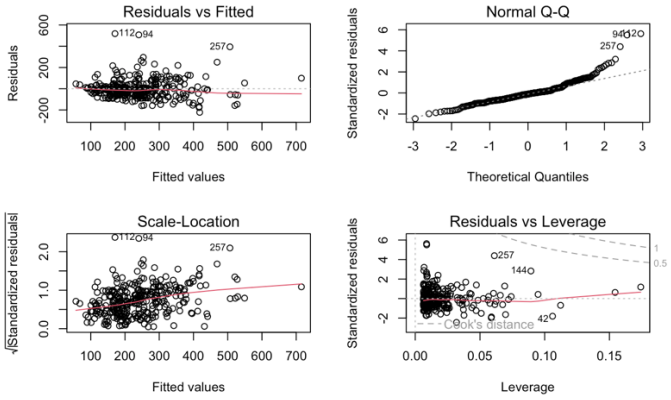
```
(Intercept)     13.7032    18.2752   0.750   0.4539
Fat              2.9740     0.2316  12.843   <2e-16 ***
d_PriorSmoke_2  51.3886    28.2865   1.817   0.0702 .
d_PriorSmoke_3 -32.8823    42.2005  -0.779   0.4365
d_Fat_2         -0.6839     0.3368  -2.031   0.0431 *
d_Fat_3          0.4858     0.4787   1.015   0.3110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.54 on 309 degrees of freedom
Multiple R-squared:  0.5163,    Adjusted R-squared:  0.5085
F-statistic: 65.97 on 5 and 309 DF,  p-value: < 2.2e-16
```

R2 = 0.5163 : 51.63% of variation in Cholesterol can be explained by the predictors in model3.

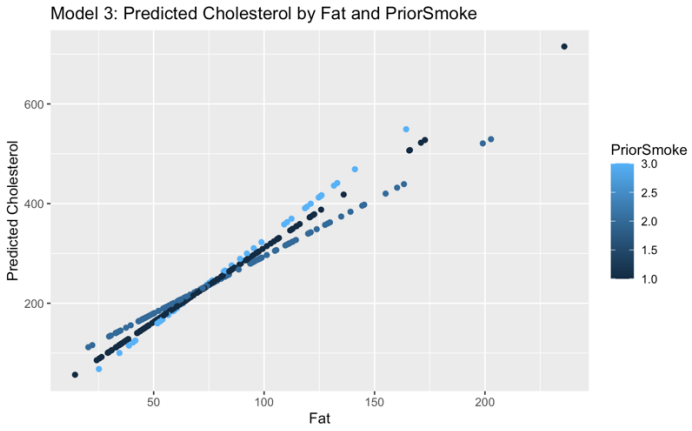
Variable	H_0	H_a	Conclusion
Intercept	$\beta_0 = 0$	$\beta_0 != 0$	Fail to reject H_0
Fat	$\beta_1 = 0$	$\beta_1 != 0$	Reject H_0
d_PriorSmoke_2	$\beta_2 = 0$	$\beta_2 != 0$	Fail to reject H_0
d_PriorSmoke_3	$\beta_3 = 0$	$\beta_3 != 0$	Fail to reject H_0
d_Fat_2	$\beta_4 = 0$	$\beta_4 != 0$	Reject H_0
d_Fat_3	$\beta_5 = 0$	$\beta_5 != 0$	Fail to reject H_0

The T-tests indicate that Fat and the interaction term between Fat and d_PriorSmoke_2 are not 0, but the other coefficients and intercept are 0. This shows that Fat has a non-0 slope influence on predicted Cholesterol and the slope for Fat when PriorSmoke=2 differs from when PriorSmoke != 2.

Note in the diagnostic plots below that the residuals vs fitted plot does not appear to violate any assumptions of linear regression. The residuals vs leverage plots also indicate that for CD = 4/N = 4/315 = 0.0127, there are several points that are influential or have high leverage in model2.



6. Use Model 3 to obtain predicted values. Plot the predicted values for CHOLESTEROL (Y) by FAT(X). **Discuss what you see in this graph.**



This graph more clearly shows that the slope of Fat has different slopes based on the value of PriorSmoke. PriorSmoke=2 has the lowest slope, whereas PriorSmoke=1 or PriorSmoke=3 have higher slopes in comparison.

7. You should be aware that Model 2 and Model 3 are nested. **Which model is the full and which one is the reduced model? Write out the null and alternative hypotheses for the nested F-test to determine if the slopes are unequal. Use the ANOVA tables from Models 2 and 3 you fit previously to compute the F-statistic for a nested F-test using Full and Reduced models. Conduct and interpret the nested hypothesis test. Are there unequal slopes in this situation? Discuss the findings.**

Model 2 is nested in Model 3.
H_0: $\beta_4 = \beta_5 = 0$
H_a: $\beta_4 \neq 0$ or $\beta_5 \neq 0$

```
Analysis of Variance Table

Model 1: Cholesterol ~ Fat + d_PriorSmoke_2 + d_PriorSmoke_3
Model 2: Cholesterol ~ Fat + d_PriorSmoke_2 + d_PriorSmoke_3 + d_Fat_2 +
    d_Fat_3
  Res.Df     RSS Df Sum of Sq     F  Pr(>F)
1    311 2708756
2    309 2645939  2     62817 3.668 0.02665 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With a p-value < 0.05, we reject H_0 and conclude that there are unequal slopes.
As confirmed by the T-tests in the previous task, we may conclude that $\beta_4 \neq 0$.

8. Now that you've been exposed to these modeling techniques, it is time for you to use them in practice. Let's examine more of the NutritionStudy data. Use the above modeling approach to determine if the categorical variables SMOKE, ALCOHOL CONSUMPTION or GENDER, along with the continuous variables FAT variable are predictive of CHOLESTEROL. Formulate hypotheses, construct essential variables (as necessary), conduct the analysis and report on the results. Which categorical variables are most predictive of CHOLESTEROL?

Nested model	Full model (*additional coefficient referenced in H_0 and H_a*)	H_0	H_a	Conclusion
Fat + Smoke	+ Alcohol	Coef = 0	Coef != 0	No
Fat + Smoke	+ Gender	Coef = 0	Coef != 0	**
Fat + Smoke	+ Smoke_interaction	Coef = 0	Coef != 0	.
Fat + Alcohol	+ Smoke	Coef = 0	Coef != 0	No
Fat + Alcohol	+ Gender	Coef = 0	Coef != 0	**
Fat + Alcohol	+ Alcohol_interaction	Coef = 0	Coef != 0	.
Fat + Gender	+ Smoke	Coef = 0	Coef != 0	No
Fat + Gender	+ Alcohol	Coef = 0	Coef != 0	No
Fat + Gender	+ Gender_interaction	Coef = 0	Coef != 0	No
Fat + Smoke + Alcohol	+ Smoke_interaction + Alcohol_interaction	Coef = 0	Coef != 0	.
Fat + Alcohol + Gender	+ Alcohol_interaction + Gender_interaction	Coef = 0	Coef != 0	No
Fat + Smoke + Gender	+ Smoke_interaction + Gender_interaction	Coef = 0	Coef != 0	No
Fat + Smoke + Alcohol	+ Gender	Coef = 0	Coef != 0	**
Fat + Alcohol + Gender	+ Smoke	Coef = 0	Coef != 0	No
Fat + Smoke + Gender	+ Alcohol	Coef = 0	Coef != 0	No
Fat + Smoke + Alcohol + Gender	+ Smoke_interaction + Alcohol_interaction + Gender_interaction	Coef = 0	Coef != 0	No

(*Significance codes are indicated in previous tables*)

By comparing nested models, it appears that only the models that add Gender have a non-0 coefficient. When Gender is added to a nested model, the F-test is significant (yellow rows). But any nested models with Gender that add additional predictors are not significant (orange rows). We conclude that Gender is the most predictive variable for Cholesterol (without interaction with Fat).

9. Please write a conclusion / reflection on your experiences in this assignment.

Conclusion

There is more to linear regression than simple continuous predictors. In this assignment, we were introduced to models that use discrete predictors and models that use a combination of continuous and discrete predictors.

The Nutrition Study Data is a 16 variable dataset with n=315 records. This dataset contained multiple continuous and discrete variables. Our task throughout this assignment was to predict the value of Cholesterol from various datapoints based on other health-related variables such as fat, smoking habits, and gender.

In the beginning of the assignment, we were tasked with creating a model using only a discrete variable which require the implementation of recoded dummy variables. By using an ANOVA test on factored variables, we learned that the mean predicted value of Cholesterol did significantly differ between the different levels of the discrete predictor.

When adding in a continuous predictor to the model with a discrete predictor, we immediately notice from the scatterplots of Cholesterol vs the continuous predictor that a more complex model may be needed to predict the Cholesterol based on different groups from the discrete variable. This problem, also known as unequal slopes, introduces the need for interaction terms. We add these to the model and interpret their significance to the model like any other coefficient, but need to indicate in the interpretation that the interaction terms are the slope of the continuous predictor for a specific group of the discrete predictor.

After implementing both dummy variables and interaction terms in a model that contains continuous and discrete variables, we see that the scatterplot of predicted Cholesterol vs the continuous variable do indeed have visually unequal slopes. We even prove in several of the tasks that one of the slopes is significantly unequal according to both coefficient T-tests and nested F-tests.

An important note in the RStudio analyses is that dummy variables and interaction terms with the dummy variables need a baseline. In the analyses for this assignment, I chose to leave out PriorSmoke=1 in both the dummy variables and interaction terms in order for that group to be the baseline. Interpreting the values of the other coefficients is always relative to the baseline group.

Using everything that we learned, the final task is to create a best fitting model from a combination of 1 continuous and 3 discrete predictors. Using a nested modeling method, I compared the addition of predictors to various different models as well as the presence and absence of interaction terms where appropriate. In the end, the final model used to predict Cholesterol from the given predictors were only Fat and Gender. The other predictors and interaction terms did not add any prediction power to a linear model.

Overall, we learned some great techniques when creating linear regression models with discrete predictors. This assignment taught us that dummy variables, interaction terms, and scatterplots of unequal slopes are all great tools when building models that contain both discrete and continuous variables.