NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Week 6 Assignment - Computational: OLS Regression Modeling with Continuous and Categorical Variables

MSDS 410

This fourth computational assignment builds on your prior modeling and computing experiences with assignment #3. 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 continuous and 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.

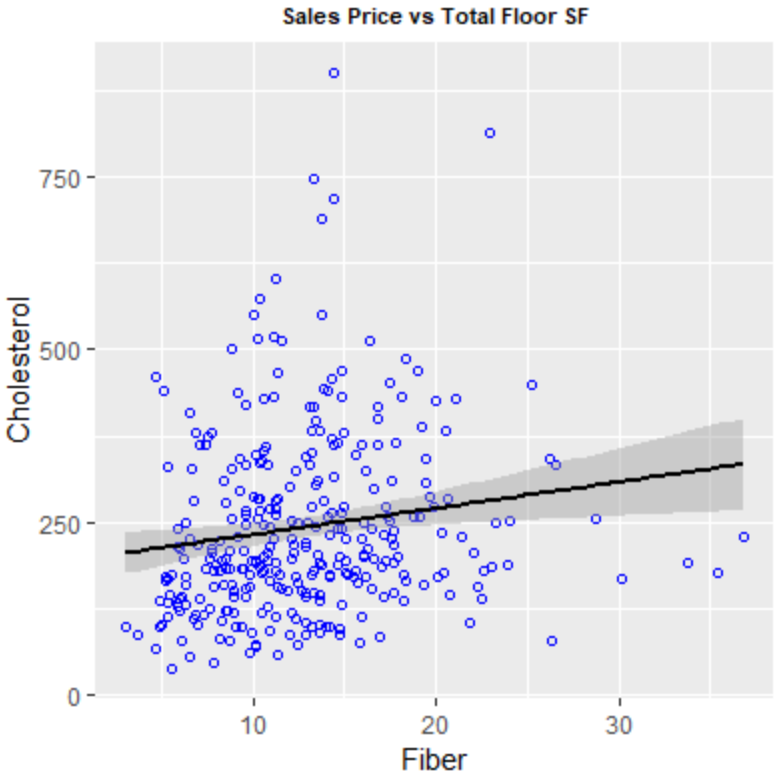
Tasks: To achieve the objective please complete the following tasks enumerated below. You are to use R to obtain any graphs or statistics requested.

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

1) Consider the continuous variable, FIBER. Is this variable correlated with Cholesterol? Obtain a scatterplot and appropriate statistics to address this question.

   Correlation between cholesterol and fiber is 0.15, very weak correlation
   Correlation between log10 (cholesterol) and fiber is 0.19, very weak correlation
   Correlation between log10 (cholesterol) and log10 (fiber) is 0.24, weak correlation

   The scatterplot also confirms that there is a very weak correlation between cholesterol and fiber.



2) Fit a simple linear regression model that uses FIBER to predict CHOLESTEROL(Y). Report the model, interpret the coefficients, discuss the goodness of fit.

Since this task did not ask us the log 10 transformation, we ignored the transformations.

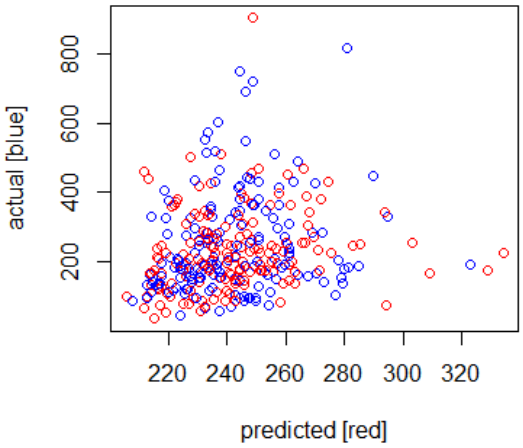
```
Y_hat = 193.701 + 3.813 * Fiber
```

However, we can see that the p-value is very small when we look at the fiber predictor. Therefore, with 95% confidence (two-tail test), we reject the null hypothesis and conclude that the Fiber predictor is significantly helping us in predicting the cholesterol level.

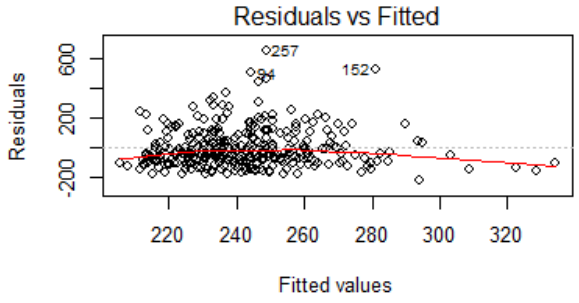
```
Call:
lm(formula = Cholesterol ~ Fiber, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-216.48  -88.58  -34.54   61.18  652.10

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  193.701     19.157  10.111  < 2e-16 ***
Fiber          3.813      1.383   2.757  0.00618 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 130.6 on 313 degrees of freedom
Multiple R-squared:  0.02371,   Adjusted R-squared:  0.02059
F-statistic:   7.6 on 1 and 313 DF,  p-value: 0.006179
```

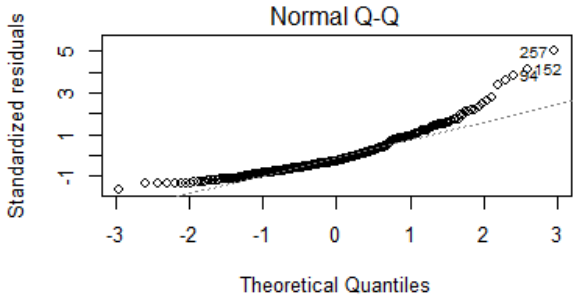
```
> summary(aov(fit_ns1))
             Df  Sum Sq Mean Sq F value  Pr(>F)
Fiber         1  129684  129684     7.6 0.00618 **
Residuals   313 5340757   17063
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

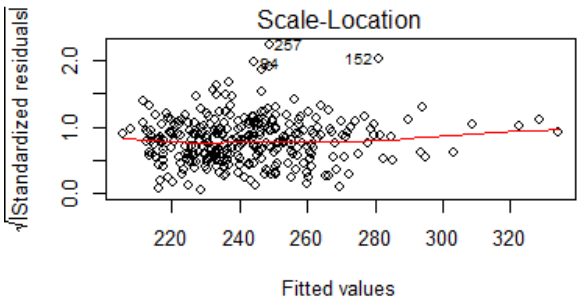
R-squared is 0.024, and we can see how much of the variation in cholesterol level is actually explained by the consumption of fiber. The answer is not much. The consumption of fiber explains only about 2.4% of the variation in cholesterol levels. That means that 97.6 percent of the cholesterol level variation for these users is left unexplained. This lack of explanatory power may not be too surprising because many other variables such as the individuals' lifestyle (smoker, drinker, dieter) influence the cholesterol levels; these factors are necessarily included in the errors in a simple regression analysis.

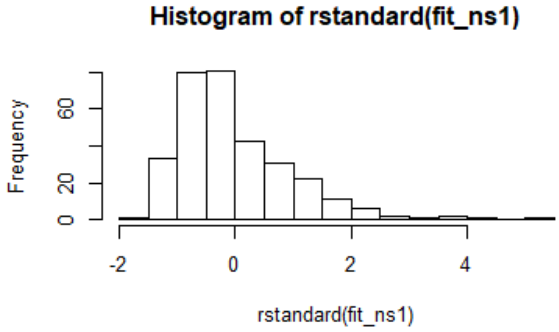


We would like to test if there is a random pattern in the residual plot (essentially the distance of the data points from the fitted regression line). This random pattern indicates that if a linear model provides a decent fit to the data. We can see that residuals are not distributed normally in the histogram, QQ plot, and scatterplots. The plot of the fitted vs. residuals seem to have more variation at lower mid-level values compared with the high fitted values due to the outliers. So the residuals clustered to the left and there are outliers in the right tail. There are negative values for the residual (on the y-axis) which means that the prediction was too high. We would like to see residuals cluster close to or around 0 mean line.









3) For the ALCOHOL categorical variable, create a set of dummy coded (0/1) indicator variables. Fit a multiple linear model that uses the FIBER continuous variable and the ALCOHOL dummy coded variables to predict the response variable Y=CHOLESTEROL. 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 model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics. This is called an Analysis of Covariance Model (ANCOVA)

We dummy coded alcohol variables; there are no alcohol drinkers if it is zero, and there are alcohol drinkers if it is one. We took non-alcohol drinkers (alc_dummy = 0) as a basis of interpretation, baseline.

```
mydata$alc_dummy<-ifelse(mydata$Alcohol == 0, 0, 1)
```

We are trying to predict cholesterol levels by using the fiber (continuous) and alcohol (categorical) variables. In the absence of an interaction term (or sometimes it called as unequal slopes), the model takes this form

```
Y_hat = 191.658 + 3.789 * Fiber + 3.617 * alc_dummy
```

For the baseline level, when there are no alcohol drinkers, the intercept is 191.658, the slope for fiber intake is 3.789. For alcohol drinkers, the intercept is 191.658 + 3.617 = 195.275, the slope for fiber intake is 3.789 (the equal slopes). If we have one unit change in the alcohol drinkers group, we are getting an increase in the cholesterol level by 3.617. So group 1 and group 2 are parallel but separated by 3.617 cholesterol levels. We can also see that the above equation can be rewritten into two equations.

```
Non-alcoholic (alc_dummy = 0): Y_hat = 191.658 + 3.789 * Fiber
Alcoholic (alc_dummy = 1):     Y_hat = 195.275 + 3.789 * Fiber
```

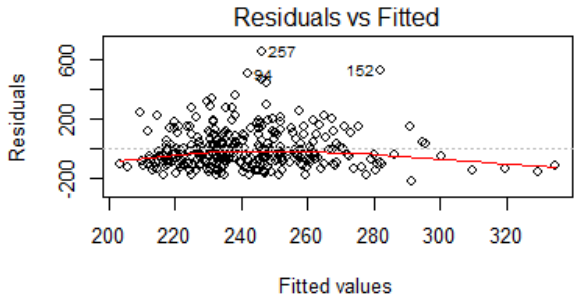
Notice that this amounts to fitting two parallel lines to the data, one for alcohol users and one for non-alcohol users. The lines for alcohol users and non-users have different intercepts, $\beta_0 + \beta_2$ (195.275) versus β_0 (191.658), but the same slope, β_1 (3.789). This allows for the possibility that changes in fiber intake may affect the cholesterol levels of alcohol users and non-alcohol users differently. The fact that the lines are parallel means that the average effect on cholesterol level of a one-unit increase in fiber intake does not depend on whether or not the individual is an alcoholic. Since in fact, a change in fiber intake may have a very different effect on the cholesterol level of alcohol users versus non-users, and this can be addressed by adding an interaction variable, created by multiplying fiber intake with the dummy variable of alcohol users and non-users.

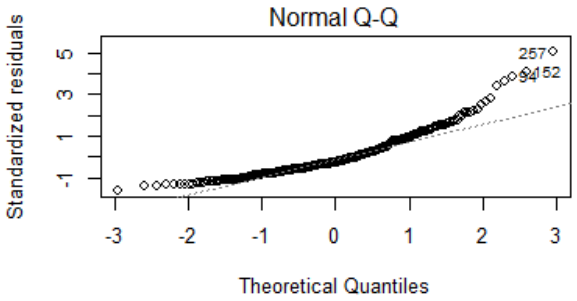
The R-squared shows that the addition of alcohol use to fiber explains only 2.388% of the variation in cholesterol level. The R-squared improvement is very small; so, the R-square value went up by 0.017% only, but the Adjusted R-squared value decreased by 0.297%. We can say that the addition of alcohol users to the existing model (Cholesterol ~ Fiber) without an interaction term did not add any value to improve the model.

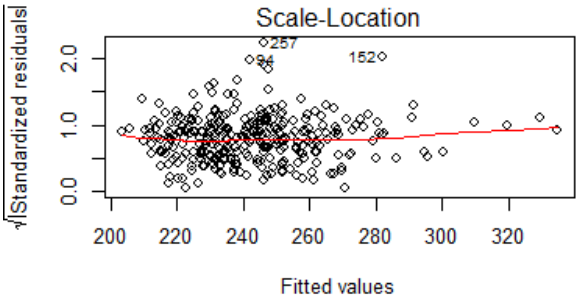
We can use an F-test to test the whole model: H0: $\beta_1 = \beta_2 = 0$; This F-test has a p-value of 0.02305 (less than 0.025; significant), indicating that we reject the null hypothesis and conclude that this model is a significant and there is at least one predictor correlates to response variable Y. We can see that the fiber variable is significantly helping us in predicting the cholesterol level as

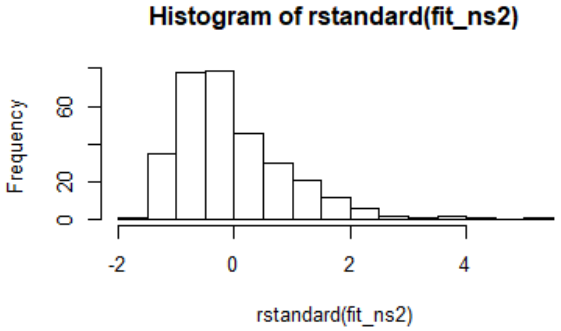
we have seen in tasks 2. However, the p-values associated with the coefficient estimates for the alcohol dummy variable is very large, suggesting no statistical evidence of a real difference in cholesterol levels between the alcohol user groups.

```
Call:
lm(formula = Cholesterol ~ Fiber + alc_dummy, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-213.82  -88.05  -33.07   61.31  654.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  191.658     21.083   9.091  < 2e-16 ***
Fiber          3.789      1.389   2.729  0.00672 **
alc_dummy      3.617     15.470   0.234  0.81526
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.8 on 312 degrees of freedom
Multiple R-squared:  0.02388,   Adjusted R-squared:  0.01762
F-statistic: 3.816 on 2 and 312 DF,  p-value: 0.02305
> summary(aov(fit_ns2))
             Df  Sum Sq Mean Sq F value  Pr(>F)
Fiber         1  129684  129684   7.577 0.00626 **
alc_dummy     1     936     936   0.055 0.81526
Residuals   312 5339821   17115
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```





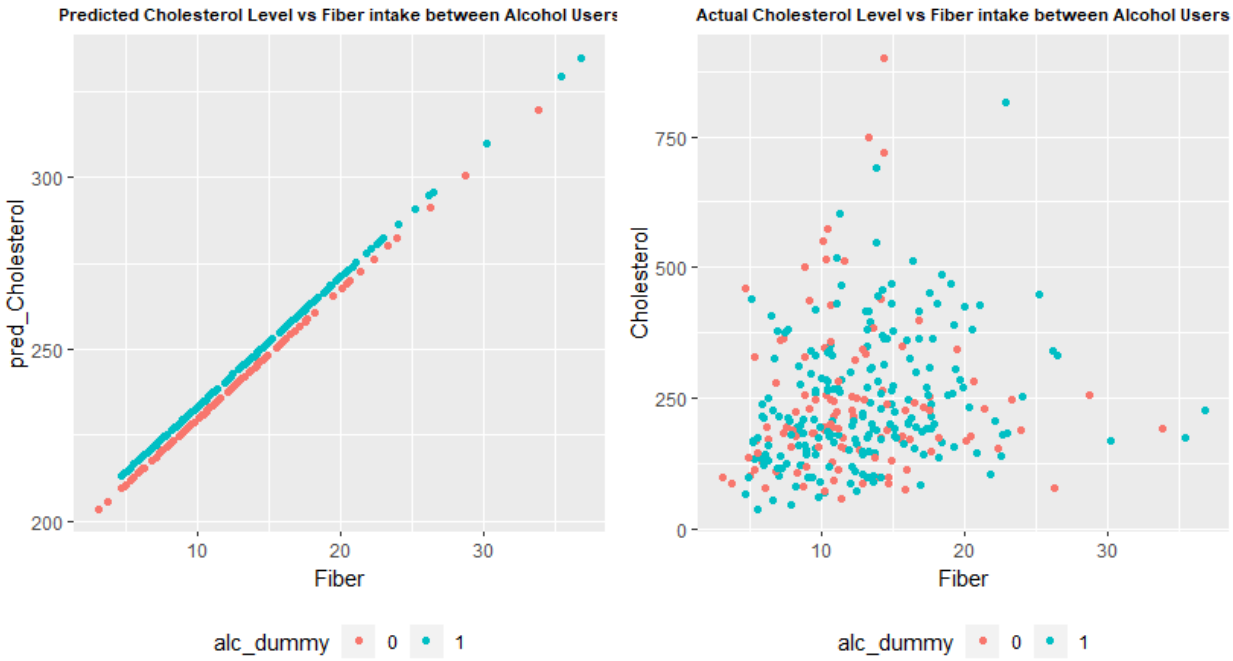




We can also see that the residual distributions are not improved by adding the categorical variable – alcohol usage. Residuals are not distributed normally in the histogram, QQ plot, and scatterplots. The plot of the fitted vs. residuals seem to have more variation at lower mid-level values compared with the high fitted values due to the outliers. So the residuals clustered to the left, and there are outliers in the right tail. There are negative values for the residual (on the y-axis), which means that the prediction was too high. We would like to see residuals cluster close to or around 0 mean line.

We can test parallelism in the following way. Since interaction (B_3) is absent, we know the null hypothesis that the two regression lines are parallel is equivalent to H0: $B_3 = 0$. Since $B_3 = 0$, then the slope for alcohol drinkers, $B_1 = B_1 + B_3$, simplifies to B_1, which is the slope for alcohol non-drinkers. Therefore, two lines are parallel. We won't be able to calculate the F statistics because we are missing interaction terms associated with Regression SS and MSE.

4) Use the ANCOVA model from task 3) to obtain predicted values for CHOLESTEROL(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by FIBER (X), but color code the records for the different groups of ALCOHOL. What do you notice about the patterns in the predicted values of Y? Now, make a scatterplot of the actual values of CHOLESTEROL(Y) by FIBER (X), but color code by the different groups of the ALCOHOL 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?



When I made a scatterplot of the **predicted** values for CHOLESTEROL (y-axis) vs. FIBER (x-axis) and color-coded the records for the different groups of ALCOHOL, I can see that there are two distinct parallel correlated lines between alcohol groups. The correlation between predicted cholesterol level and fiber intake without alcohol groups is 0.996 which has a very strong correlation. Since we had equal slopes β_1 (3.789) for both groups of alcohol drinkers but unequal intercepts, we got parallel lines. Alcohol drinkers have consistently higher (about 3.617) cholesterol level than the non-alcohol drinkers at all fiber intakes.

However, when I made a scatterplot of the **actual** values for CHOLESTEROL (y-axis) vs. FIBER (x-axis) and color-coded the records for the different groups of ALCOHOL, I can see that the values are distributed like an upward cone or funnel shape. It does not have parallel correlated lines. The correlation between actual cholesterol level and fiber intake without alcohol groups is 0.154 which has a very weak correlation.

When I compared the two scatterplots, the ANCOVA model did not appear to fit the observed data very well because we were missing interaction terms. Without interaction terms, we got the same or equal slope, β_1 (3.789), for alcohol and non-alcohol drinkers. The equal slope and unequal intercept suggest that it is a parallel line. So, we need to more complex model; we should test further whether or not that interaction needs to be part of our model.

5) Create new interaction variables by multiplying the dummy coded variables for ALCOHOL by the continuous FIBER(X) variable. Save these product variables to your dataset. Now, to build the model, start with variables in your ANCOVA model from task 4) and add the interaction variables you just created into the multiple regression model. Don't forget, there is one category that is the basis of interpretation. DO NOT include any interaction term that is associated with that category. This is called an Unequal Slopes Model. Fit this model, and save the predicted values. Plot the predicted values for CHOLESTEROL (Y) by FIBER(X). Discuss what you see in this graph. In addition, report the model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics.

```
Call:
lm(formula = Cholesterol ~ Fiber * alc_dummy, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-190.24  -83.41  -32.02   60.80  661.19

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     230.3434    31.6272   7.283  2.7e-12 ***
Fiber             0.6363     2.3719   0.268    0.789
alc_dummy1      -56.3889    39.7615  -1.418    0.157
Fiber:alc_dummy1   4.7842     2.9217   1.637    0.103
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 130.5 on 311 degrees of freedom
Multiple R-squared:  0.03222,   Adjusted R-squared:  0.02289
F-statistic: 3.451 on 3 and 311 DF,  p-value: 0.01692

> summary(aov(fit_ns3))
                Df  Sum Sq Mean Sq F value  Pr(>F)
Fiber            1  129684  129684   7.618 0.00612 **
alc_dummy        1     936     936   0.055 0.81478
Fiber:alc_dummy  1   45643   45643   2.681 0.10255
Residuals      311 5294178   17023
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We took non-alcohol drinkers (alc_dummy = 0) as a basis of interpretation. Instead of creating a sperate interaction calculated variables, I added an interaction term to the model directly. Since

the alc_dummy variable contains 0 or 1 binary codes, I multiplied it to the Fiber variable to get an interaction term for alcohol users only. So, the interaction term automatically excludes the alcohol non-drinkers that is associated with the basis of interpretation. Please note that the following equations produce the same outputs and the second bullet point is the appropriate way of running the interaction terms in R.

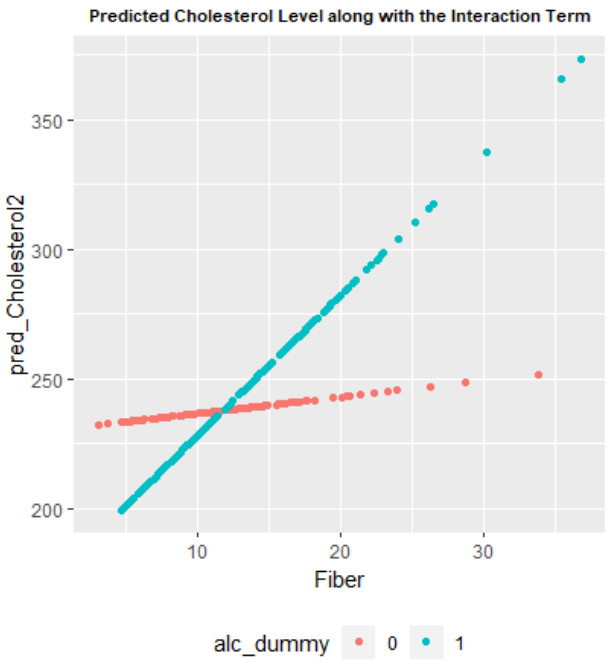
- `Cholesterol ~ Fiber + alc_dummy + Fiber * alc_dummy`
- `Cholesterol ~ Fiber * alc_dummy`

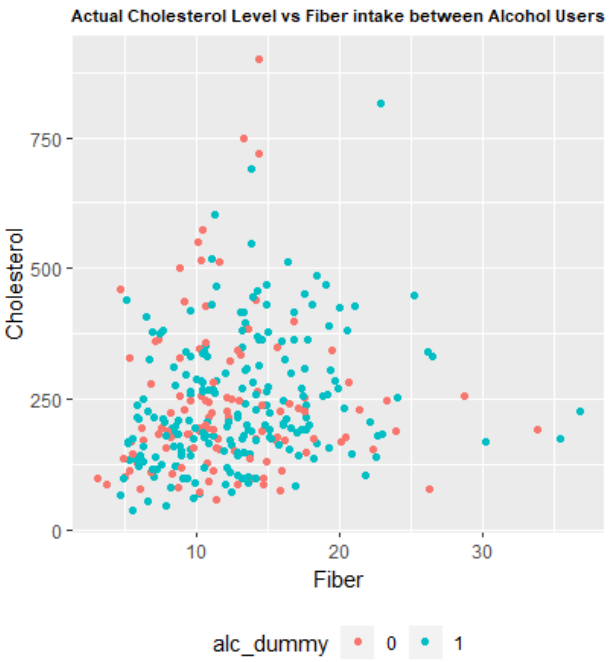
Here is our model with an interaction term:
`Y_hat = 230.343 + 0.636 * Fiber – 56.389 * alc_dummy + 4.784 * (Fiber * alc_dummy)`

For the baseline level when there are no alcohol drinkers, the intercept is 230.343, the slope for fiber intake is 0.636. However, the slope changes for the second group due to the interaction term. So, for alcohol drinkers, the intercept is 230.343 – 56.389 = 173.954, and the slope for fiber intake is 0.636 + 4.784 = 5.420 (now the slopes are unequal). We can also see that the above equation can be rewritten into two equations.

```
Alcohol Non-drinker (alc_dummy = 0): Y_hat = 230.343 + 0.636 * Fiber
Alcohol Drinker (alc_dummy = 1):     Y_hat = 173.954 + 5.420 * Fiber
```

The relationship between fiber intake and mean cholesterol level differs for the alcohol users and nonusers with regard to both the origins and the rates of change. We note that the intercept of alcohol drinkers is lower than alcohol non-drinkers, but it has a steeper slope. This means that there are lower cholesterol levels for alcohol drinkers if they consume less fiber but as they consume more fiber their cholesterol level spikes. This suggests that increases in fiber intake are associated with larger increases in cholesterol levels among alcohol drinkers as compared to alcohol non-drinkers. However, this interpretation does not make sense. We should look at the goodness of fit statistics.



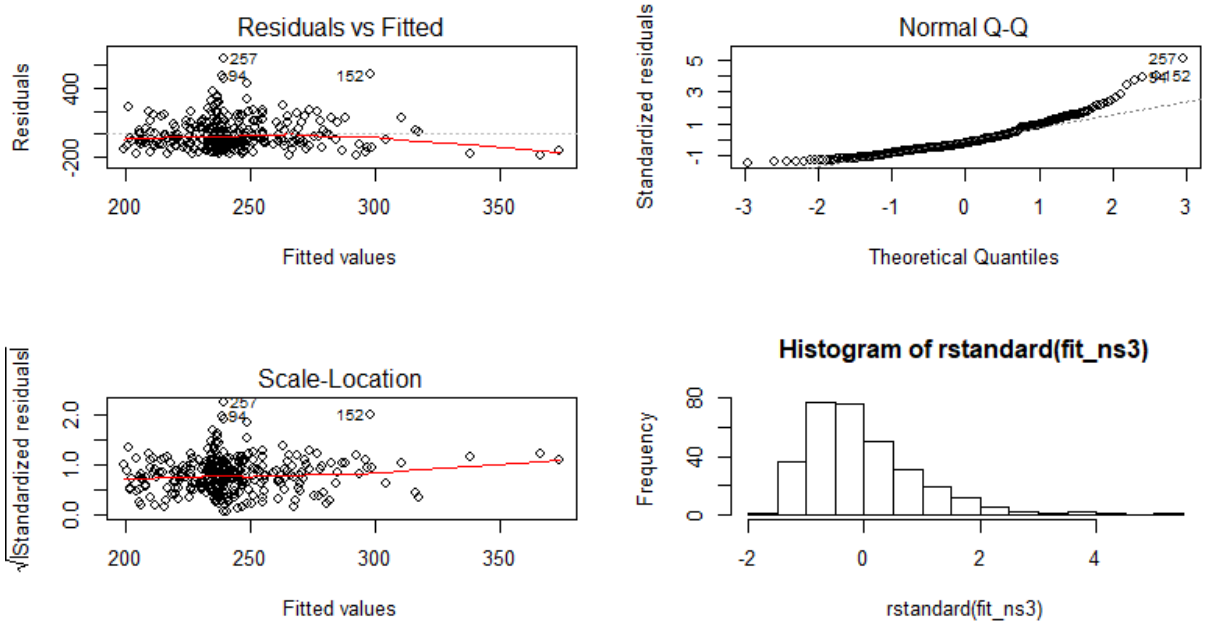


The upper left graph confirms that when we add the interaction term, the model produces different slopes and intercepts between two lines (or drinkers vs. non-drinkers). When I compared the two scatterplots, the ANCOVA model did not appear to fit the observed data very well, even after including the interaction term.

Both of the R-squared and adjusted R-squared values improved by a little bit, 0.832% and 0.527%, respectively. However, those improvements did not move the needle.

We can use an F-test to test the whole model: H0: $\beta_1 = \beta_2 = \beta_3 = 0$; This F-test has a p-value of 0.01692 (insignificant), indicating that we cannot reject the null hypothesis and conclude that this model is a not significant and none of the predictors correlates to response variable Y. We can see that the Fiber variable stopped helping us in predicting the cholesterol level and interpretations do not make sense. Also, p-values associated with the coefficient estimates all variables, including interaction term, are very large, suggesting there is no relationship between cholesterol levels and predictors (fiber, alcohol and their interaction term).

We can also see that the residual distributions are not improved by adding the interaction term. Residuals are not distributed normally in the histogram, QQ plot, and scatterplots. We can see that the residuals clustered to the left. We can also observe the outliers in the right tail.



6) You should be aware that the models of Task 4) and Task 5) 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 in this situation to determine if the slopes are unequal. Use the ANOVA tables from those two models 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? Discuss the findings.

```
Full model:    Y_hat =  β0 + β1 (Fiber) + β2 (alc_dummy) + β3 (Fiber *
alc_dummy)
Reduced model: Y_hat =  β0 + β1 (Fiber) + β2 (alc_dummy)
```

Ho: all slopes are equal

Ha: there is at least one of the slopes are not equal to zero

```
> anova(fit_ns3) # full model
Analysis of Variance Table

Response: Cholesterol
                 Df  Sum Sq Mean Sq F value   Pr(>F)
Fiber             1  129684  129684  7.6181 0.006121
alc_dummy         1     936     936  0.0550 0.814777
Fiber:alc_dummy   1   45643   45643  2.6813 0.102547
Residuals       311 5294178   17023
> anova(fit_ns2) # reduced model
Analysis of Variance Table

Response: Cholesterol
           Df  Sum Sq Mean Sq F value   Pr(>F)
Fiber       1  129684  129684  7.5773 0.006257
alc_dummy   1     936     936  0.0547 0.815265
Residuals 312 5339821   17115
```

Full Model ANOVA

SS(REG) = 129684 + 936 + 45643 = 176263 (degree of freedom = 3)

SS(ERR) = 5294178 (degree of freedom = 311)

Reduced Model ANOVA

SS(REG) = 129684 + 936 = 130620 (degree of freedom = 2)

SS(ERR) = 5339821 (degree of freedom = 312)

$$F = \frac{[SS(Reg\text{-}FULL) - SS(Reg\text{-}Reduced)] / (diff\ of\ df)}{SS(ERR - Full) / df}$$

$$\frac{(176263 - 130620)/(3 - 2) = 45643}{5294178 /311 = 17023.0804} = 45643 / 17023 = 2.6813$$

We can also test it by including full and reduced models into anova function.

```
Analysis of Variance Table

Model 1: Cholesterol ~ Fiber + alc_dummy
Model 2: Cholesterol ~ Fiber * alc_dummy
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    312 5339821
2    311 5294178  1     45643 2.6813 0.1025
```

According to the low F- statistics 2.6813 and high p-value 0.10, we cannot reject the null hypothesis and conclude that the model without interaction term is good enough; do not need the full model with interaction term and it is insignificant. There are not unequal slopes.

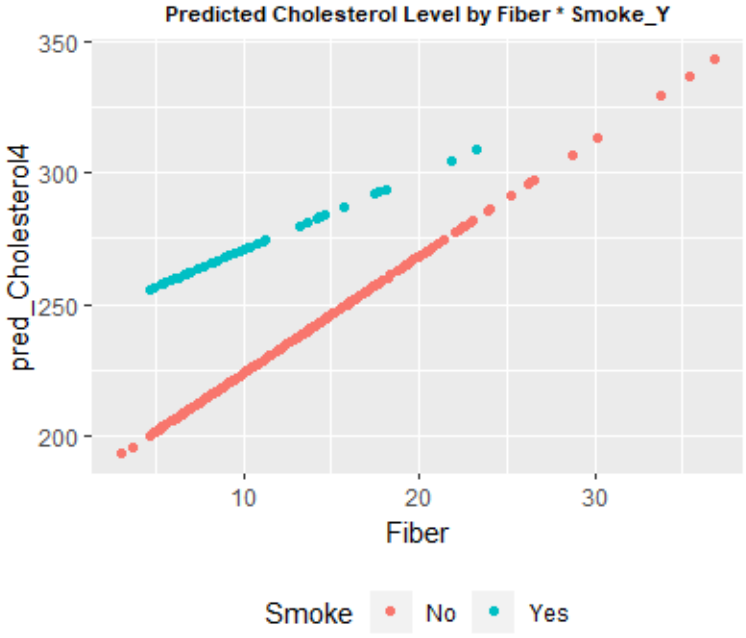
7) Now that you've been exposed to these modeling techniques, it is time for you to use them in practice. Let's examine more of the NutritionStudy data. Use the above practiced techniques to determine if SMOKE, VITAMINS, or GENDER interacts with the FIBER variable and influences the amount of CHOLESTEROL. Formulate hypotheses, construct essential variables (as necessary),

conduct the analysis and report on the results. Which categorical variables are most predictive of CHOLESTEROL, in conjunction with FIBER.

```
Y_hat = 179.184 + 4.455 *(Fiber) + 63.059 * (smoker_Y) – 1.597 * (Fiber *
smoker_Y)
```

- `Smoker = no / 0:  Y_hat = 179.184 + 4.455 * Fiber`
- `Smoker = yes / 1: Y_hat = 242.243 + 2.858 * Fiber`

According to the above equation, smokers have a higher intercept but flatter slope than nonsmokers. This means that smokers have a higher cholesterol level. However, as smokers consume more fiber, their cholesterol increases a little slower than nonsmokers, holding other variables constant.



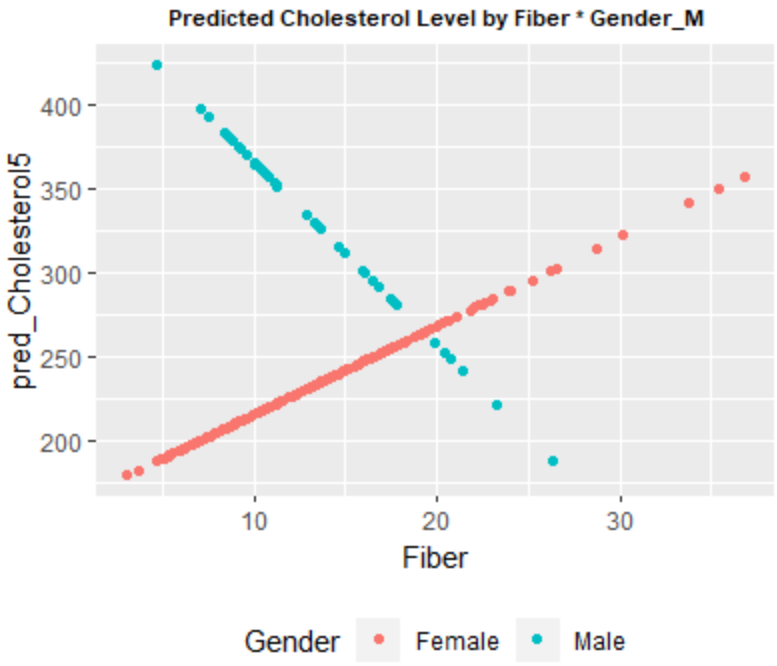
The R-squared shows that the addition of smokers to fiber explains only 3.789% of the variation in cholesterol level. Both R-squared and Adjusted R-squared values are improved slightly. We can use an F-test to test the whole model: H0: $\beta_1 = \beta_2 = 0$; The p-value associated to F-test is 0.007277 (significant), indicating that we reject the null hypothesis and conclude that this model is a significant and there is at least one predictor correlates to response variable Y. We can see that the fiber variable (excluding smokers and interactions) is significantly helping us in predicting the cholesterol level. However, the p-values associated with the coefficient estimates for the smoker predictor is very large, suggesting no statistical evidence of a real difference in cholesterol levels between the smokers and nonsmokers.

```
Call:
lm(formula = Cholesterol ~ Fiber * Smoke_Y, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-218.86  -87.71  -35.15   65.11  657.36

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)   179.184     20.875   8.583 0.000000000000000447
Fiber           4.455      1.471   3.028              0.00267
Smoke_Y        63.059     55.002   1.146              0.25248
Fiber:Smoke_Y  -1.597      4.661  -0.343              0.73218

Residual standard error: 130.1 on 311 degrees of freedom
Multiple R-squared:  0.03789,   Adjusted R-squared:  0.02861
F-statistic: 4.082 on 3 and 311 DF,  p-value: 0.007277
```

Y_hat = 162.359 + 5.273 *(Fiber) + 311.514 * (gender_M) – 16.138* (Fiber * gender_M)

- Gender = female / 0: Y_hat = 162.359 + 5.273 * Fiber
- Gender = male / 1: Y_hat = 473.873 – 10.865 * Fiber

According to the above equation, males have a higher intercept than females. However, males have a negative slope and females have a positive slope. This means that males' cholesterol level drop when they eat more fiber but females' cholesterol level goes up as they eat more fiber, holding other variables constant.



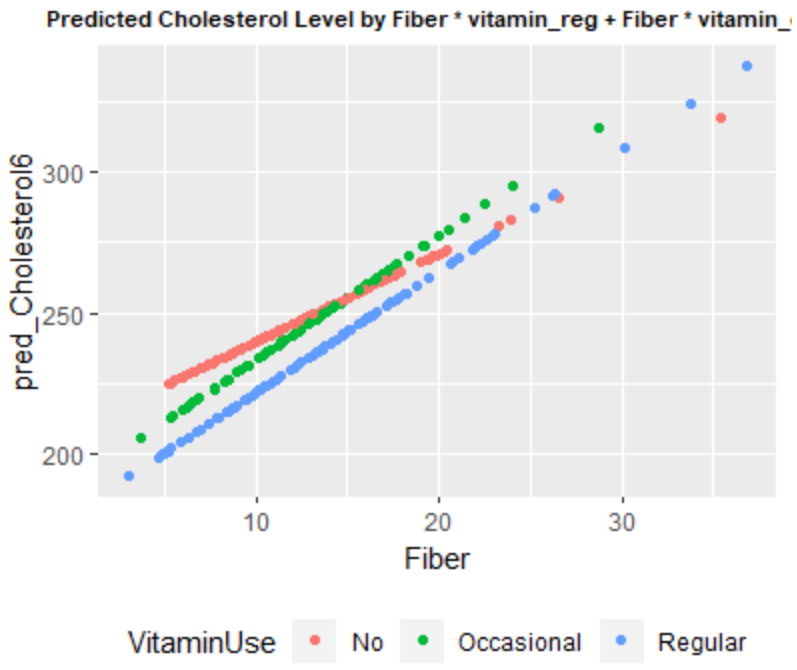
The R-squared shows that the addition of gender to fiber explains 12.61% of the variation in cholesterol level. Both R-squared and Adjusted R-squared values are improved significantly than previous models. We can use an F-test to test the whole model: H0: $\beta_1 = \beta_2 = 0$; The p-value associated to F-test is almost zero (significant), indicating that we reject the null hypothesis and conclude that this model is a significant and there is at least one predictor correlates to response variable Y. We can see that all predictors, including the interaction term, are significantly helping us in predicting the cholesterol level.

```
Call:
lm(formula = Cholesterol ~ Fiber * Gender_M, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-299.55  -80.27  -25.28   53.23  662.41

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)    162.359     19.188   8.462 0.00000000000000105
Fiber            5.273      1.391   3.790            0.000181
Gender_M       311.514     60.083   5.185 0.00000039029294889
Fiber:Gender_M -16.138      4.233  -3.812            0.000166

Residual standard error: 124 on 311 degrees of freedom
Multiple R-squared:  0.1261,    Adjusted R-squared:  0.1177
F-statistic: 14.96 on 3 and 311 DF,  p-value: 0.000000004028
```

Let the "vitamin use regular" be Z1
Let the "vitamin use occasional" be Z2
Let the "vitamin use no" be Z3

```
Y_hat = 192.356 + 3.943 * (Fiber) - 13.477 * (Z1) – 2.988 * (Z2)
+ 0.364 * (Fiber * Z1) +  0.468 * (Fiber * Z2)
```

- vitamin use regular (Z1 =1, Z2 = 0): Y_hat = 178.879 + 4.307 * Fiber
- vitamin use occasional (Z2 =1, Z1 = 0): Y_hat = 189.368 + 4.411 * Fiber
- vitamin use no (Z3 = 1, Z1 = Z2 = 0): Y_hat = 192.356 + 3.943 * Fiber



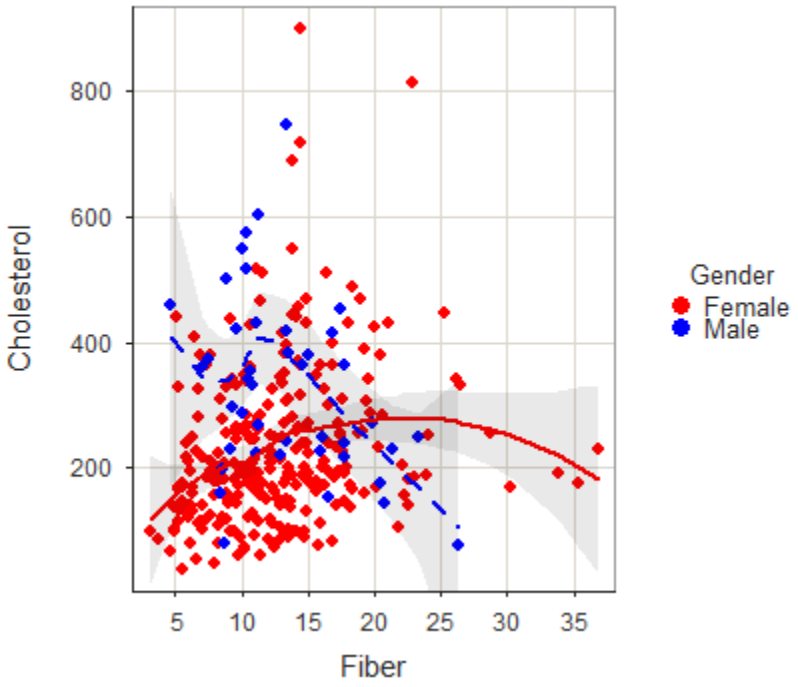
According to the equation above, all vitamin user groups have a positive slope and have somewhat similar intercepts. However, when we look at the graph, there are somewhat parallel lines between regular and occasional vitamin users. However, non-vitamin users start with the highest cholesterol levels. As non-vitamin users eat more fiber, their cholesterol level increases slower and catches up to regular vitamin users' cholesterol levels.

The R-squared shows that the addition of vitamin usage to fiber explains 2.681% of the variation in cholesterol level. Both R-squared and Adjusted R-squared values are declined significantly than previous models. We can use an F-test to test the whole model: H0: $\beta_1 = \beta_2 = 0$; The p-value associated to F-test is non zero (insignificant), indicating that we cannot reject the null hypothesis and conclude that this model is a not significant and none of the predictors correlates to response variable Y well. We can see that the Fiber variable stopped helping us in predicting the cholesterol level when we added the vitamin use predictor. Also, p-values associated with the coefficient estimates

all variables, including interaction terms, are very large, suggesting there is no relationship between cholesterol level and predictors (fiber, vitamin users, and their interaction terms).

```
Call:
lm(formula = Cholesterol ~ Fiber * vitamin_reg + Fiber * vitamin_occ,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-214.64  -91.71  -33.55   63.36  659.80

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)       192.3564    20.2323   9.507 < 0.0000000000000002
Fiber               3.9425     1.4798   2.664              0.00812
vitamin_reg       -13.4772    26.5554  -0.508              0.61216
vitamin_occ        -2.9879    31.5212  -0.095              0.92454
Fiber:vitamin_reg   0.3642     1.8895   0.193              0.84729
Fiber:vitamin_occ   0.4677     2.3174   0.202              0.84018

Residual standard error: 131.3 on 309 degrees of freedom
Multiple R-squared:  0.02681,   Adjusted R-squared:  0.01106
F-statistic: 1.702 on 5 and 309 DF,  p-value: 0.1338
```

Gender categorical variables are most predictive of cholesterol levels in conjunction with fiber because gender significantly improved the R-squared and Adjusted R-squared values. We can see that all predictors, including the interaction term, are significantly helping us in predicting the cholesterol level. We can see that the predicted cholesterol levels by gender group are similar to the trend of the actual data. Here is actual data with fitted non-straight lines.



8) Please write a reflection on your experiences.

I've learned that when we mix continuous and categorical variables, we should also check the unequal slopes (interaction terms). I learned how to compute the intercept and slope values (if

interactions added) for a categorical variable when it mixes with continuous variables. I also learned how to interpret the interaction of continuous and categorical variables.

9) Extra Credit: Feel free to explore models that have other continuous variables, as well as interactions of categorical variables. The more you do, the more extra credit you can accumulate.