Reed Ballesteros
MSDS-410-DL, Summer 2022
Dr. Mickelson
7/24/2022

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

Preparatory Work

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.

```
Gender <- mydata$Gender
GenderF <- ifelse(Gender=='Female',1,0)
GenderM <- ifelse(Gender=='Male',1,0)

Smoke <- mydata$Smoke
SmokeNo <- ifelse(Smoke=='No',1,0)
SmokeYes <- ifelse(Smoke=='Yes',1,0)
```

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.

```
VitaminUse <- mydata$VitaminUse
VitaminUseNo <- ifelse(VitaminUse=='No',1,0)
VitaminUseOcc <- ifelse(VitaminUse=='Occasional',1,0)
VitaminUseReg <- ifelse(VitaminUse=='Regular',1,0)

PriorSmoke <- mydata$PriorSmoke
PriorSmoke1 <- ifelse(PriorSmoke==1,1,0)
PriorSmoke2 <- ifelse(PriorSmoke==2,1,0)
PriorSmoke3 <- ifelse(PriorSmoke==3,1,0)
```

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.

```
Alcohol <- mydata$Alcohol
AlcoholConsumption <- vector()

for (i in Alcohol) {
  if (i == 0)
    value <- 'None'
  else if (i>0 & i<10)
    value <- 'Some'
  else
    value <- 'ALot'
  AlcoholConsumption <- append(AlcoholConsumption, value)
}

AlcoholNone <- ifelse(AlcoholConsumption=='None',1,0)
AlcoholSome <- ifelse(AlcoholConsumption=='Some',1,0)
AlcoholALot <- ifelse(AlcoholConsumption=='ALot',1,0)
```

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.

    - We create the following model, Fit 1, where the categorical value PriorSmoke is the predictive explanatory variable for the dependent variable Cholesterol:

```
> fit1

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

Coefficients:
      (Intercept)  mydata$PriorSmoke
           206.32              22.06

> ANOVA(Cholesterol~PriorSmoke, data=mydata)

  BACKGROUND

Data Frame:  mydata

Response Variable: Cholesterol

Factor Variable: PriorSmoke
  Levels: 1 2 3

Number of cases (rows) of data:  315
Number of cases retained for analysis:  315


  DESCRIPTIVE STATISTICS

     n     mean       sd     min      max
1  157   228.39   134.23   37.70   900.70
2  115   250.42   121.69   46.30   747.50
3   43   272.53   145.92   78.30   718.80

Grand Mean: 242.461
```

- The model above shows that every additional category for PriorSmoke increases Cholesterol by just over 22 units, which is also represented by the means for each PriorSmoke factor variable, as shown in the descriptive statistics table above.

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.

- We create the following for Model 1, using PriorSmoke = 1 from our dataframe as our control group/basis of interpretation:

```
> model1

Call:
lm(formula = Cholesterol ~ PriorSmoke2 + PriorSmoke3)

Coefficients:
(Intercept)  PriorSmoke2  PriorSmoke3
     228.39        22.03        44.14
```

- Which results in the following regression model:

Cholesterol = 228.39 + 22.03*PriorSmoke2 + 44.14*PriorSmoke3

- Model 1 is similar to the model we fitted in part 1, such that each category increase in PriorSmoke adds about just over 22 units to Cholesterol. The intercept for Model 1 228.39 is the mean of PriorSmoke1, which matches the mean for factor variable 1 shown in the descriptive statistics from the ANOVA analysis in part 1. Adding the PriorSmoke2 coefficient to the intercept also matches the mean for factor variable 2 from part 1 as well. We can interpret the same for PriorSmoke3 matching the mean for factor variable 3. The grand mean matches for both ANOVA analysis.

```
> ANOVA(Cholesterol ~ PriorSmoke2 + PriorSmoke3, data=mydata)

  BACKGROUND

Data Frame:  mydata

Response Variable: Cholesterol

Factor Variable 1: PriorSmoke2
  Levels: 0 1

Factor Variable 2: PriorSmoke3
  Levels: 0 1

Number of cases (rows) of data:  315
Number of cases retained for analysis:  315

Randomized Blocks ANOVA
  Factor of Interest:  PriorSmoke2
  Blocking Factor:     PriorSmoke3

Note: For the resulting F statistic for PriorSmoke2 to be distributed as F,
      the population covariances of Cholesterol must be spherical.


  DESCRIPTIVE STATISTICS

-- Marginal Means

PriorSmoke2

        X0     X1
  1 237.88 250.42

PriorSmoke3

        X0     X1
  1 237.71 272.53

-- Grand Mean: 242.461
```

```
anova(model1)
# > anova(model1)
# Analysis of Variance Table
#
# Response: Cholesterol
#              Df  Sum Sq Mean Sq F value Pr(>F)
# PriorSmoke2   1   11487   11487  0.6645 0.4156
# PriorSmoke3   1   65771   65771  3.8049 0.0520 .
# Residuals   312 5393183   17286
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model1)
# > summary(model1)
#
# Call:
#   lm(formula = Cholesterol ~ PriorSmoke2 + PriorSmoke3)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -204.12  -90.02  -32.79   61.37  672.31
#
# Coefficients:
#             Estimate Std. Error t value            Pr(>|t|)
# (Intercept)   228.39      10.49  21.766 <0.0000000000000002 ***
# PriorSmoke2    22.03      16.14   1.365               0.173
# PriorSmoke3    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

CholesterolHatM1 <- predict(model1, newdata = mydata)
mydata<-cbind.data.frame(mydata,CholesterolHatM1)
```

- Given the R2 value of 0.0141, the model only represents 1.4% of the variability to Cholesterol. Because of this, the goodness of fit of Model 1 would not be very accurate. Using only PriorSmoke to determine Cholesterol is not recommended and should look to develop a more complex model.

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.

- We create the following for Model 2, still using PriorSmoke = 1 as our basis of interpretation, which results in the following regression model:

Cholesterol = 28.9401 + 2.7630*Fat - 2.1142*PriorSmoke2 + 10.6358*PriorSmoke3

```
> model2

Call:
lm(formula = Cholesterol ~ Fat + PriorSmoke2 + PriorSmoke3, data = mydata)

Coefficients:
(Intercept)          Fat  PriorSmoke2  PriorSmoke3
     28.940        2.763       -2.114       10.636

> anova(model2)
Analysis of Variance Table

Response: Cholesterol
             Df  Sum Sq Mean Sq  F value              Pr(>F)
Fat           1 2756468 2756468 316.4780 <0.0000000000000002 ***
PriorSmoke2   1    1452    1452   0.1667              0.6834
PriorSmoke3   1    3765    3765   0.4323              0.5113
Residuals   311 2708756    8710
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model2)

Call:
lm(formula = Cholesterol ~ Fat + PriorSmoke2 + PriorSmoke3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-214.06  -53.03  -12.01   33.24  514.58

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  28.9401    13.5848   2.130              0.0339 *
Fat           2.7630     0.1574  17.556 <0.0000000000000002 ***
PriorSmoke2  -2.1142    11.5372  -0.183              0.8547
PriorSmoke3  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: < 0.00000000000000022
```

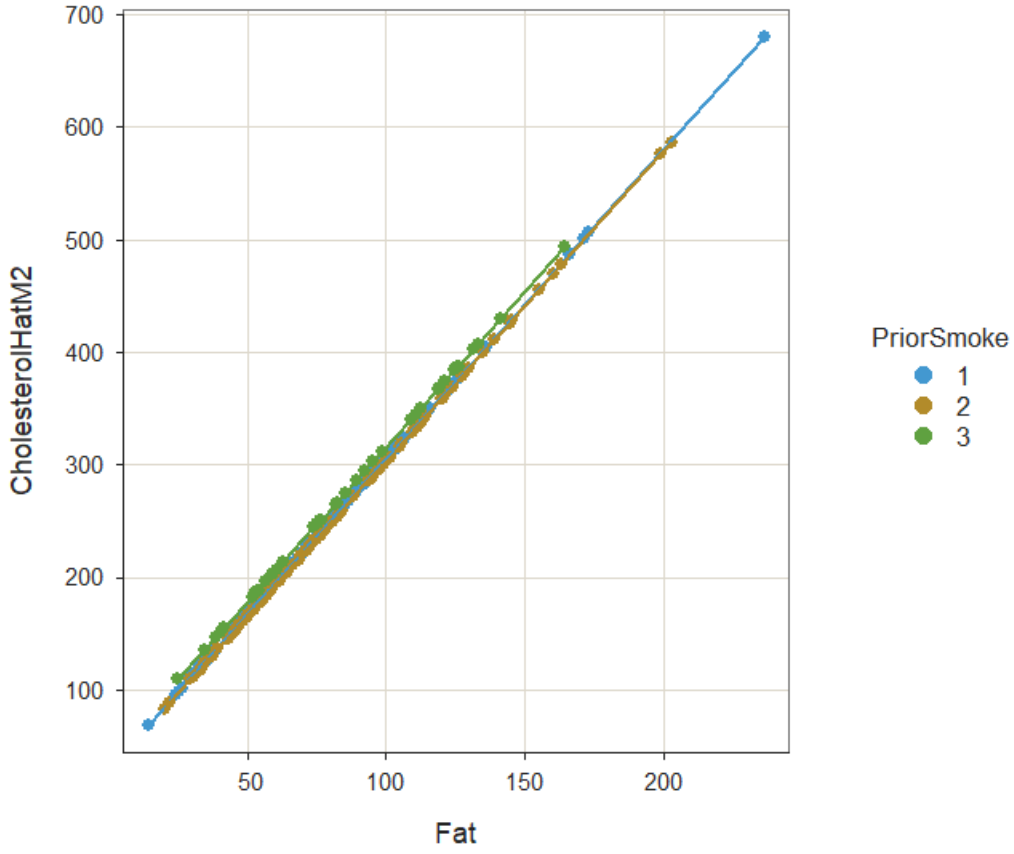
- Model 2 shows that every unit of Fat increases Cholesterol by 2.76, a person classified with PriorSmoke2 can decrease Cholesterol by 2.11 from the mean of PriorSmoke1 (28.94), and people classified with PriorSmoke3 can increase 10.64 Cholesterol from the PriorSmoke1 mean.

- Using Fat along with the PriorSmoke dummy variables for Model 2 is a significant improvement over Model 1, in which the R^2 value goes up to 0.5048, such that the model represents just over 50.5% of the variability to Cholesterol.

4. Use the ANCOVA Model 2 from Task 3) to obtain predicted values for CHOLESTEROL(Y).

```
CholesterolHatM2 <- predict(model2, newdata = mydata)
mydata<-cbind.data.frame(mydata,CholesterolHatM2)
```

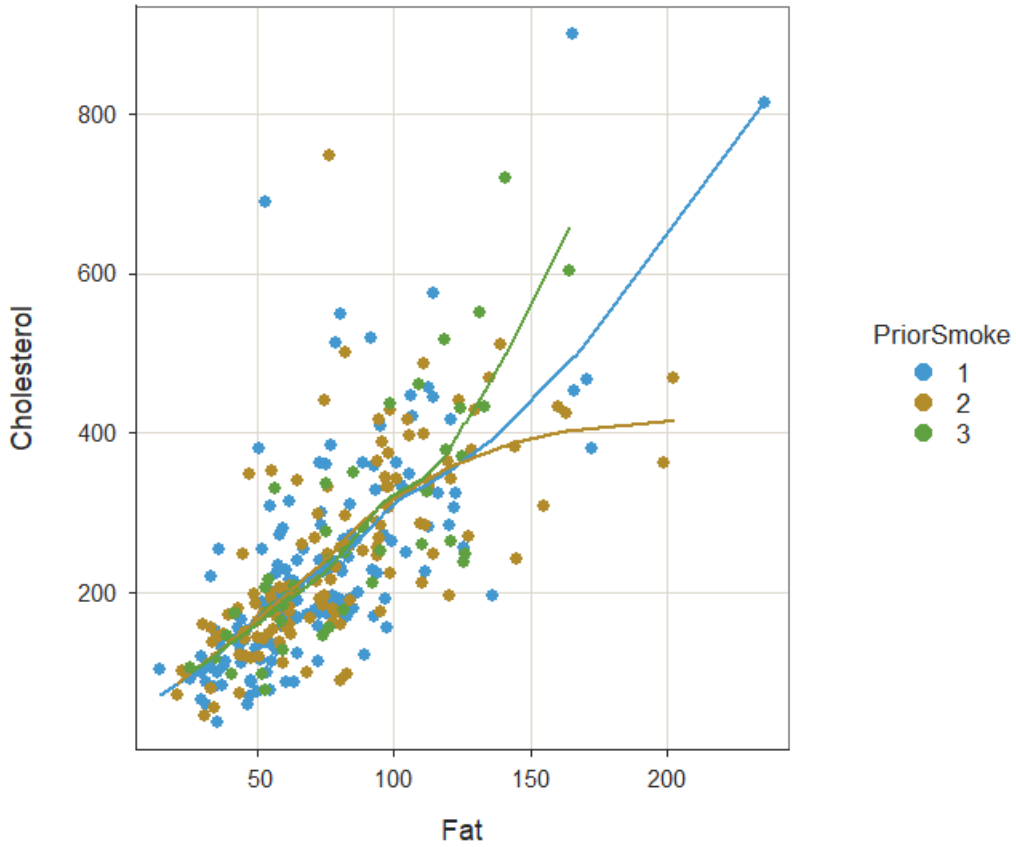
Now, make a scatterplot of the Predicted Values for Y (y-axis) by FAT (X), but color code the records for the different groups of PRIORSMOKE.



What do you notice about the patterns in the predicted values of Y?

- We notice very similar regression lines for each respective PriorSmoke category, such that they have very similar slopes, almost parallel, and do not seem to cross over each other, which indicates no real interaction between the PriorSmoke dummy variables.

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.



- We see somewhat of a linear model for people classified with PriorSmoke1, while we notice a downward curve for those classified with PriorSmoke2, and an upward curve for those classified with PriorSmoke3. We see some interaction between these groups where these models would cross one another.

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?

- Because of the lack of interaction between the regression lines in the predicted model, we'll need to look into a more complex model, despite the improvement in R^2 in this case.

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.

```
FatPS1 <- Fat*PriorSmoke1
FatPS2 <- Fat*PriorSmoke2
FatPS3 <- Fat*PriorSmoke3

mydata<-cbind.data.frame(mydata,
                         FatPS1,FatPS2,FatPS3)
```

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.

- Multiplying the Fat variable to each of the dummy PriorSmoke variables and adding them to the model while still keeping PriorSmoke = 1 as our basis of interpretation results in the following regression model:

```
> model3

Call:
lm(formula = Cholesterol ~ Fat + PriorSmoke2 + PriorSmoke3 +
    FatPS2 + FatPS3, data = mydata)

Coefficients:
(Intercept)          Fat  PriorSmoke2  PriorSmoke3       FatPS2       FatPS3
    13.7032       2.9740      51.3886     -32.8823      -0.6839       0.4858
```

Cholesterol = 13.7032 + 2.974*Fat + 51.3886*PriorSmoke2 - 32.8823*PriorSmoke3 - 0.6839*FatPS2 + 0.4858*FatPS3

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.

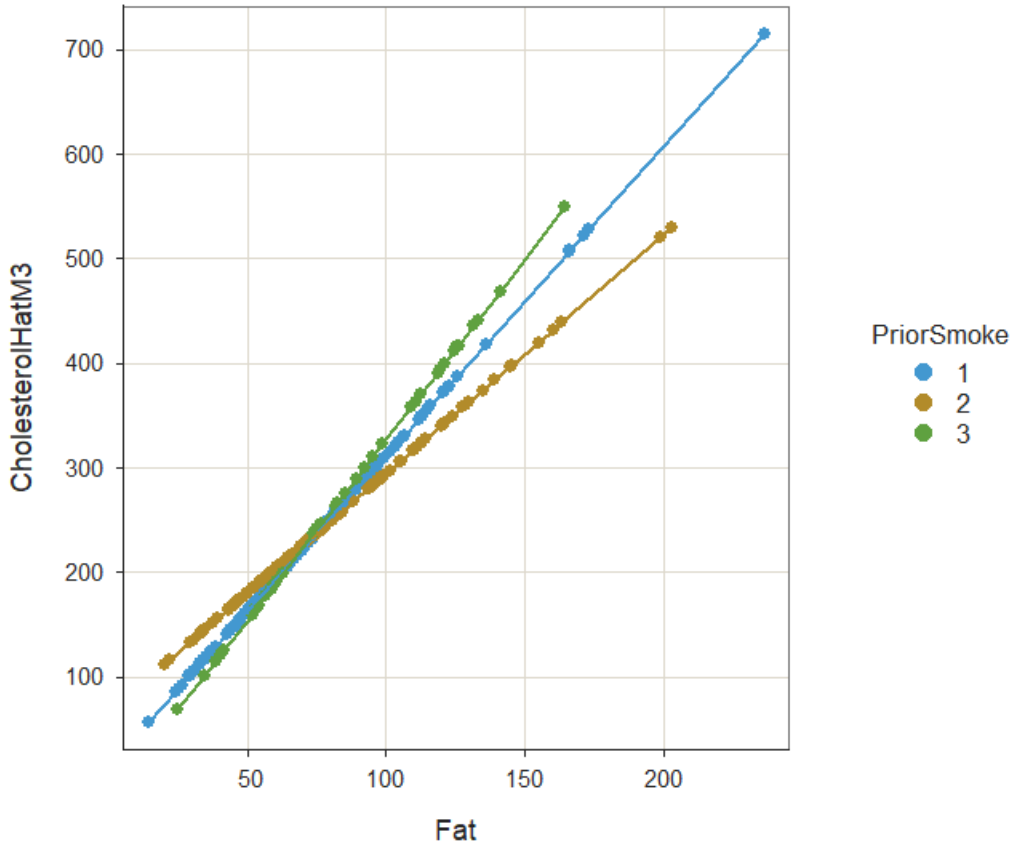
The intercept is 13.7032 and each unit of Fat can increase Cholesterol by 2.974. If a person classified as PriorSmoke2, Cholesterol is increased by 51.39 along with a decrease from their Fat value multiplied by a factor of about 0.69. If a person classified as PriorSmoke3, Cholesterol is decreased by 32.89 but is also increased with their Fat value multiplied by a factor of about 0.49.

```
> anova(model3)
Analysis of Variance Table

Response: Cholesterol
             Df  Sum Sq Mean Sq  F value                Pr(>F)
Fat           1 2756468 2756468 321.9079 < 0.0000000000000002 ***
PriorSmoke2   1    1452    1452   0.1695               0.68083
PriorSmoke3   1    3765    3765   0.4397               0.50775
FatPS2        1   53998   53998   6.3061               0.01254 *
FatPS3        1    8819    8819   1.0298               0.31099
Residuals   309 2645939    8563
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model3)

Call:
lm(formula = Cholesterol ~ Fat + PriorSmoke2 + PriorSmoke3 +
    FatPS2 + FatPS3, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-222.37  -56.18   -9.74   35.48  518.67

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  13.7032    18.2752   0.750               0.4539
Fat           2.9740     0.2316  12.843 <0.0000000000000002 ***
PriorSmoke2  51.3886    28.2865   1.817               0.0702 .
PriorSmoke3 -32.8823    42.2005  -0.779               0.4365
FatPS2       -0.6839     0.3368  -2.031               0.0431 *
FatPS3        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: < 0.00000000000000022
```

The R^2 value for Model 3 is marginally increased by 0.0115 over Model 2, which results to adding 1.15% of the variability to predicting Cholesterol by adding FatPS2 and FatPS3 to the model. We will discuss hypothesis testing in section 7.

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.

```
CholesterolHatM3 <- predict(model3, newdata = mydata)

mydata<-cbind.data.frame(mydata, CholesterolHatM3)
```



Plotting Model 3 grouped by PriorSmoke displays unequal slopes between each PriorSmoke category, as well as showing interaction by crossing over each other.

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?
    - Model 2 is the reduced model and Model 3 is the full 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.

    - H_0: Beta4 = Beta5 = 0 – Model 3 has no interaction
    - H_A: At least one of Beta4 or Beta5 is not zero – Model 3 has interaction

```
# n: 315
# df(RM): 311
# df(FM): 309
# df(RM)-df(FM): 311 - 309 = 2
# df(int): 2
# dim(RM): 3
# dim(FM): 5
# dim(FM)-dim(RM): 5 - 3 = 2

# REDUCED MODEL (RM) RESULTS - Model 2
# SS(reg) = 2756468 + 1452 + 3765 = 2761685 (3)
# SS(err) = 2708756  (311)
#
# FULL MODEL (FM) RESULTS - Model 3
# SS(REG) = 2756468  + 1452 + 3765 + 53998 + 8819 = 2824502  (5)
# SS(ERR) = 2645939  (309)
# SS(INT) = 53998 + 8819 = 62817 (2)
#
# F = SS(interaction) / df(int)
#     ------------------------
#     SS(error - FM) / df(full)
#
# =   62817/2
#   ------------
#   2645939/309
#
# = 31408.5/8562.909
#
# = 3.667971


# F-Critical: df1=2 (df(FM)-df(RM)) and df2=311 (df(RM)): 0.02531987
qf(p=0.05/2, df1=2, df2=311)

# F-Statistic > F-Critical --> 3.667971 > 0.02531987, reject H0
```

Calculating $F_{Statistic}$ 3.668 and $F_{Critical}$ 0.0253 above shows that $F_{Statistic} > F_{Critical}$, such that adding variables FatPS2 and FatPS3 to the regression model is statistically significant and displaying interaction, and that full Model 3 would be preferred over reduced Model 2.

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?

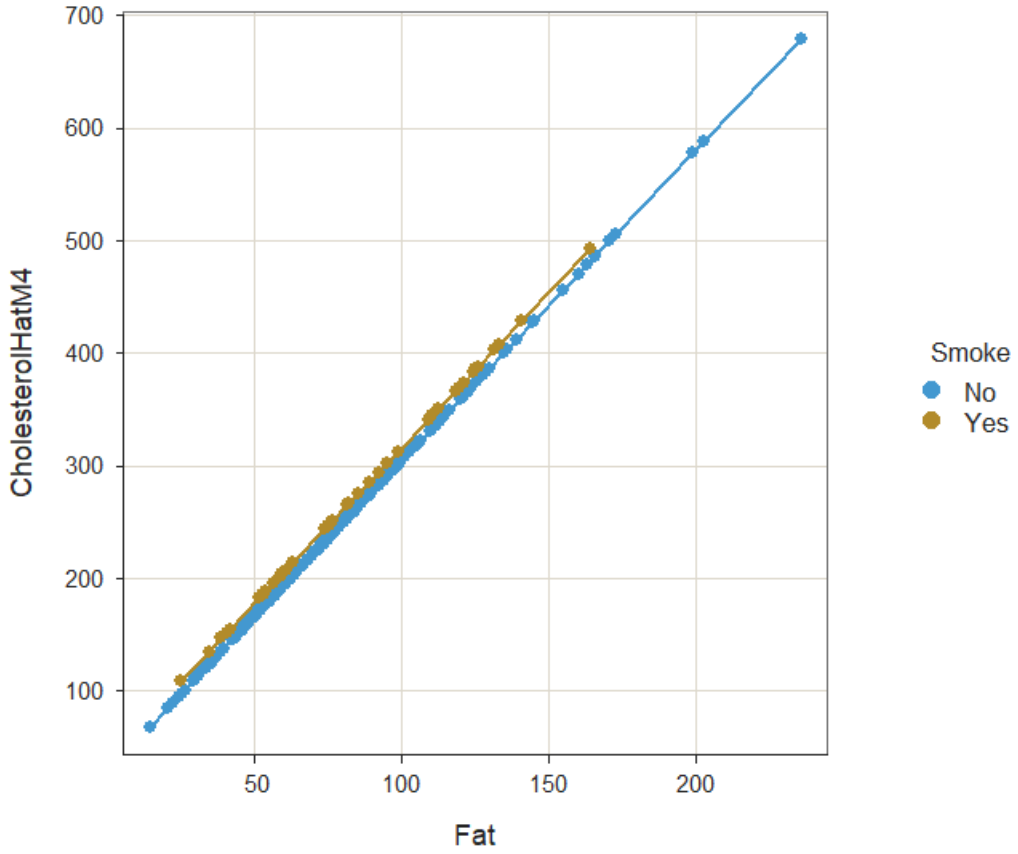
Variable Smoke

We create the following regression Model 4 by adding dummy variables for Smoke and using Smoke = 'No' as our basis for interpretation:

Cholesterol = 28.307+ 2.76*Fat + 11.559*SmokeYes

```
#########################################
# SMOKE - Model 4
#########################################

# control group: Smoke = 'No'
model4 <- lm(Cholesterol ~  Fat + SmokeYes, data= mydata)
model4

anova(model4)
# > anova(model4)
# Analysis of Variance Table
#
# Response: Cholesterol
#            Df  Sum Sq Mean Sq  F value              Pr(>F)
# Fat         1 2756468 2756468 317.4613 <0.0000000000000002 ***
# SmokeYes    1    4924    4924   0.5671               0.452
# Residuals 312 2709048    8683
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model4)
# > summary(model4)
#
# Call:
#   lm(formula = Cholesterol ~ Fat + SmokeYes, data = mydata)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -214.85  -52.98  -12.13   33.23  515.39
#
# Coefficients:
#             Estimate Std. Error t value            Pr(>|t|)
# (Intercept)   28.307     13.118   2.158              0.0317 *
# Fat            2.760      0.156  17.687 <0.0000000000000002 ***
# SmokeYes      11.559     15.349   0.753              0.4520
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 93.18 on 312 degrees of freedom
# Multiple R-squared:  0.5048,  Adjusted R-squared:  0.5016
# F-statistic:   159 on 2 and 312 DF,  p-value: < 0.00000000000000022

CholesterolHatM4 <- predict(model4, newdata = mydata)
mydata<-cbind.data.frame(mydata, CholesterolHatM4)
```

The intercept is 28.307 as the mean value for Smoke='No' and each unit of Fat adds 2.76 to the predicted Cholesterol value, while people categorized as Smoke = 'Yes' adds 11.56 to Cholesterol. Calculating R^2 shows that Model 4 accounts for about 50.5% of the variability for predicting Cholesterol.



Plotting Model 4 shows parallel regression lines between smokers and non-smokers, indicating no interaction between the groups within the model.

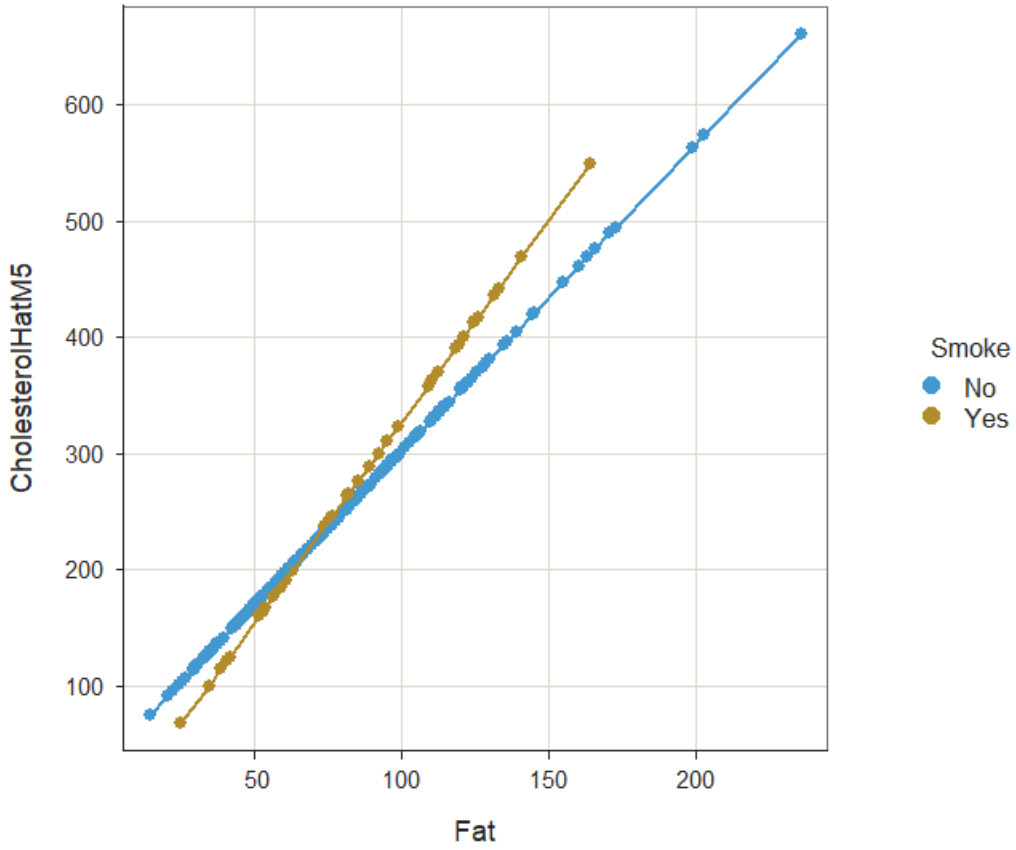
We create the following regression Model 5 by multiplying Fat and the SmokeYes dummy variable and adding it to Model 4:

Cholesterol = 36.7262+ 2.6486*Fat - 55.9053*SmokeYes + 0.8112*FatSmokeYes

```r
#######################################
# add FAT * SMOKE - Model 5
#######################################

FatSmokeNo <- Fat*SmokeNo
FatSmokeYes <- Fat*SmokeYes

mydata<-cbind.data.frame(mydata,FatSmokeNo,FatSmokeYes)

# control group: Smoke = 'No'
model5 <- lm(Cholesterol ~  Fat + SmokeYes + FatSmokeYes, data= mydata)
model5

anova(model5)
# > anova(model5)
# Analysis of Variance Table
#
# Response: Cholesterol
#              Df  Sum Sq Mean Sq  F value                Pr(>F)
# Fat           1 2756468 2756468 319.7147 < 0.0000000000000002 ***
# SmokeYes      1    4924    4924   0.5712               0.45037
# FatSmokeYes   1   27716   27716   3.2147               0.07395 .
# Residuals   311 2681333    8622
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model5)
# > summary(model5)
#
# Call:
#   lm(formula = Cholesterol ~ Fat + SmokeYes + FatSmokeYes, data = mydata)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -201.19  -52.92  -12.12   32.09  512.83
#
# Coefficients:
#             Estimate Std. Error t value            Pr(>|t|)
# (Intercept)  36.7262    13.8895   2.644              0.0086 **
# Fat           2.6486     0.1673  15.829 <0.0000000000000002 ***
# SmokeYes    -55.9053    40.6170  -1.376              0.1697
# FatSmokeYes   0.8112     0.4525   1.793              0.0740 .
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 92.85 on 311 degrees of freedom
# Multiple R-squared:  0.5099,  Adjusted R-squared:  0.5051
# F-statistic: 107.8 on 3 and 311 DF,  p-value: < 0.00000000000000022
#

CholesterolHatM5 <- predict(model5, newdata = mydata)
mydata<-cbind.data.frame(mydata, CholesterolHatM5)
```

The intercept is 36.7262 as the mean value for Smoke='No' and each unit of Fat adds 2.65 to the predicted Cholesterol value. People categorized as Smoke = 'Yes' decreases Cholesterol by 55.9 but

also adds to it with a portion of their Fat value by a factor of 0.8112. R^2 shows that Model 5 shows only a marginal improvement over Model 4, accounting for a 0.5% increase of the variability for predicting Cholesterol.



Plotting Model 5 grouped by Smoke displays unequal slopes between each Smoke category, as well as showing interaction by crossing over each other.

Hypothesis testing:

- H_0: Beta4 = 0 – Model 5 has no interaction
- H_A: Beta4 != 0 – Model 5 has interaction

```
# n: 315
# df(RM): 312
# df(FM): 311
# df(RM)-df(FM): 1
# df(int): 1
# dim(RM): 2
# dim(FM): 3
# dim(FM)-dim(RM): 1

# REDUCED MODEL (RM) RESULTS - Model 4
# SS(reg) = 2756468 + 4924 = 2761392
# SS(err) = 2709048
#
# FULL MODEL (FM) RESULTS - Model 5
# SS(REG) = 2756468 + 4924 + 27716 = 2789108
# SS(ERR) = 2681333
# SS(INT) = 27716
#
# F = SS(interaction) / df(int)
#     -------------------
#     SS(error - FM) / df(full)
#
# =   27716/1
#   ------------
#   2681333/311
#
# = 27716/8621.65
#
# = 3.214698



# F-Critical: df1=1 (df(FM)-df(RM)) and df2=312 (df(RM)): 0.0009836458
qf(p=0.05/2, df1=1, df2=312)

# F-Statistic > F-Critical --> 3.214698 > 0.0009836458, reject H0
```

With Model 4 as the reduced model nested within Model 5, calculating $F_{Statistic}$ 3.215 and $F_{Critical}$ 0.001 above shows that $F_{Statistic} > F_{Critical}$, such that we can reject the null hypothesis and adding variable FatSmokeYes to Model 5 is statistically significant and indicating interaction, and that full Model 5 would be preferred over reduced Model 4.

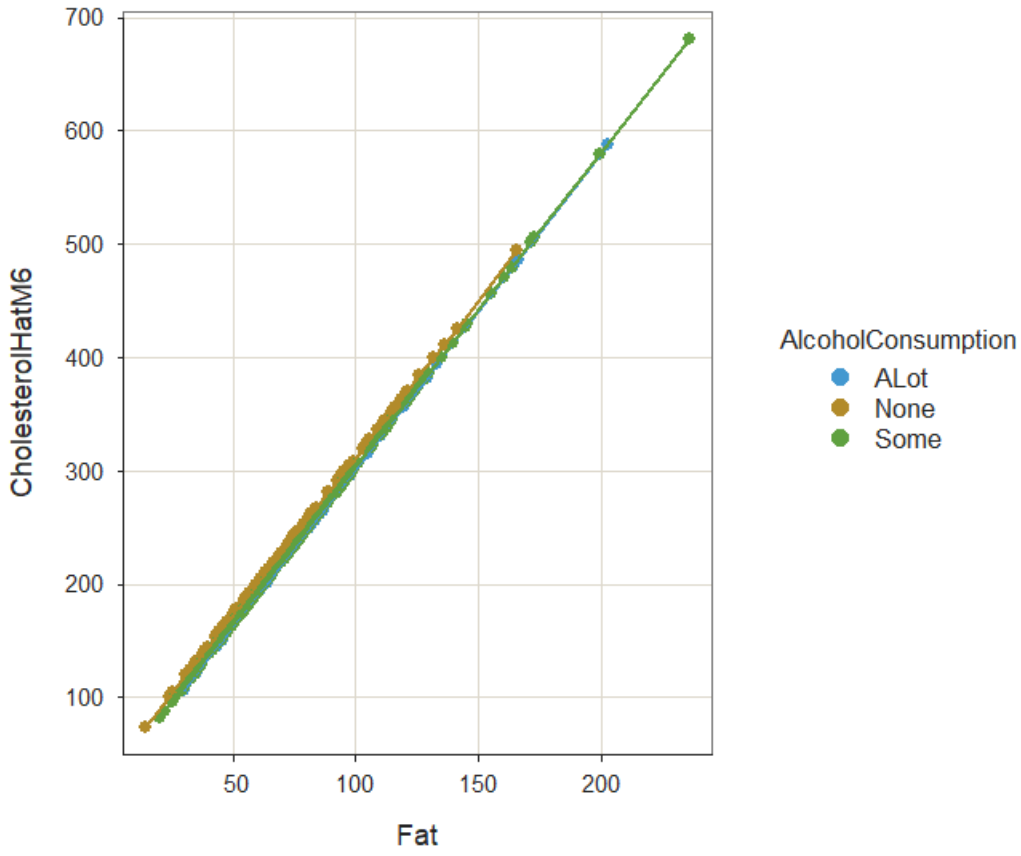
Variable Alcohol Consumption

We create the following regression Model 6 by adding dummy variables for AlcoholConsumption and using AlcoholConsumption = 'None' as our basis for interpretation:

Cholesterol = 33.7425 + 2.78*Fat – 8.2194*AlcoholSome – 9.8299*AlcoholAlot

```
#####################################
# ALCOHOL - Model 6
#####################################

# control group: AlcoholConsumption = 'None'
model6 <- lm(Cholesterol ~  Fat + AlcoholSome + AlcoholALot, data= mydata)
model6

anova(model6)
# > anova(model6)
# Analysis of Variance Table
#
# Response: Cholesterol
#              Df  Sum Sq Mean Sq  F value              Pr(>F)
# Fat           1 2756468 2756468 316.4666 <0.0000000000000002 ***
# AlcoholSome   1    3126    3126   0.3589              0.5496
# AlcoholALot   1    1994    1994   0.2289              0.6327
# Residuals   311 2708853    8710
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model6)
# > summary(model6)
#
# Call:
#   lm(formula = Cholesterol ~ Fat + AlcoholSome + AlcoholALot, data = mydata)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -216.20  -51.92  -10.64   33.34  517.08
#
# Coefficients:
#             Estimate Std. Error t value            Pr(>|t|)
# (Intercept)  33.7425    14.5674   2.316              0.0212 *
# Fat           2.7803     0.1573  17.675 <0.0000000000000002 ***
# AlcoholSome  -8.2194    11.3007  -0.727              0.4676
# AlcoholALot  -9.8288    20.5442  -0.478              0.6327
# ---
#   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:    0.5
# F-statistic: 105.7 on 3 and 311 DF,  p-value: < 0.00000000000000022

CholesterolHatM6 <- predict(model6, newdata = mydata)
mydata<-cbind.data.frame(mydata, CholesterolHatM6)
```

The intercept is 33.7425 as the mean value for AlcoholConsumption='None' and each unit of Fat adds 2.78 to the predicted Cholesterol value, while people categorized as AlcoholConsumption = 'Some' lowers Cholesterol by 8.2194, and people categorized as AlcoholConsumption = 'ALot' lowers Cholesterol by 9.8288. R^2 shows that Model 6 accounts for about 50.5% of the variability for predicting Cholesterol, similar to Model 4.



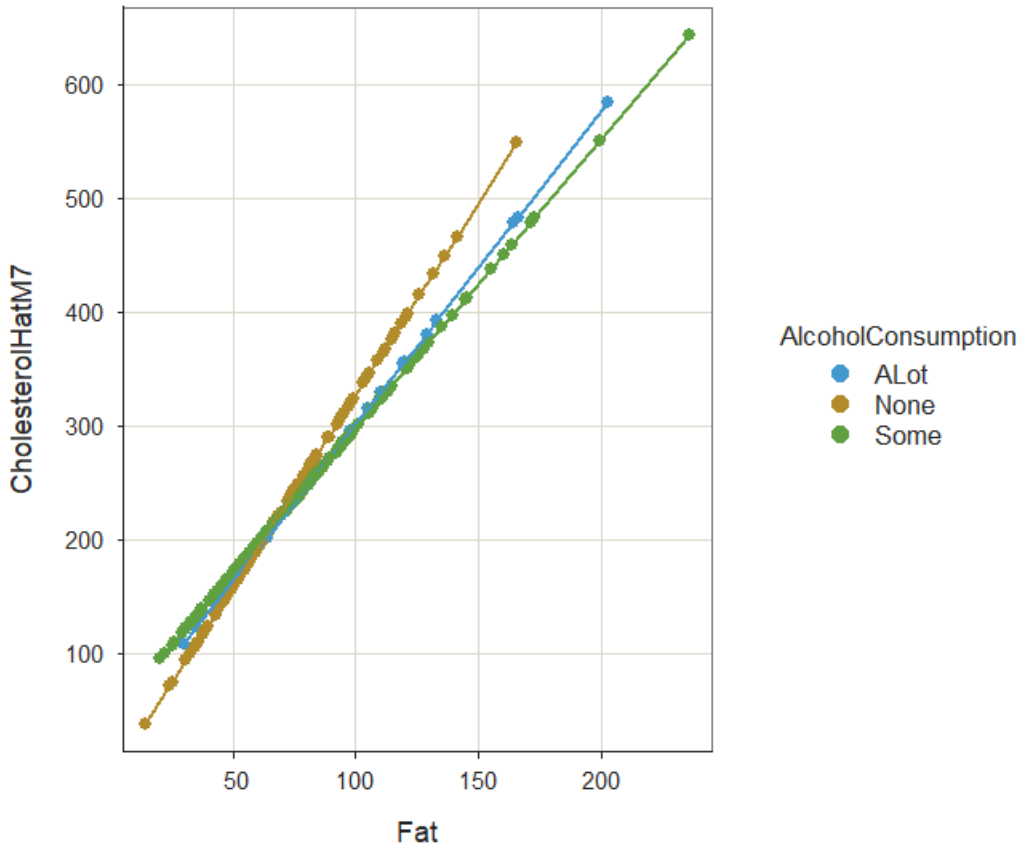
Plotting Model 6 shows close parallel regression lines between the various groups of AlcoholConsumption, indicating no real interaction between the groups.

We create the following regression Model 7 by multiplying Fat and the AlcoholSome and AlcoholALot dummy variables and adding it to Model 6:

Cholesterol = -10.1124 + 3.3768*Fat + 53.7377*FatAlcoholSome + 37.0725*FatAlcoholALot -0.8315*FatAlcoholALot - 0.6296*FatAlcoholALot

```r
#######################################
# add FAT * ALCOHOL - Model 7
#######################################

FatAlcoholNone <- Fat*AlcoholNone
FatAlcoholSome <- Fat*AlcoholSome
FatAlcoholALot <- Fat*AlcoholALot

mydata<-cbind.data.frame(mydata,AlcoholNone,AlcoholSome,AlcoholALot)

# control group: AlcoholConsumption = 'None'
model7 <- lm(Cholesterol ~  Fat + AlcoholSome + AlcoholALot
             + FatAlcoholSome + FatAlcoholALot, data= mydata)

anova(model7)
# > anova(model7)
# Analysis of Variance Table
#
# Response: Cholesterol
#                 Df  Sum Sq Mean Sq  F value                Pr(>F)
# Fat              1 2756468 2756468 319.5620 < 0.0000000000000002 ***
# AlcoholSome      1    3126    3126   0.3624               0.54762
# AlcoholALot      1    1994    1994   0.2311               0.63103
# FatAlcoholSome   1   31479   31479   3.6495               0.05701 .
# FatAlcoholALot   1   12011   12011   1.3924               0.23890
# Residuals      309 2665363    8626
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

summary(model7)
# > summary(model7)
#
# Call:
#   lm(formula = Cholesterol ~ Fat + AlcoholSome + AlcoholALot +
#        FatAlcoholSome + FatAlcoholALot, data = mydata)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -253.33  -53.82   -8.59   34.12  511.39
#
# Coefficients:
#                Estimate Std. Error t value            Pr(>|t|)
# (Intercept)    -10.1124    24.6116  -0.411               0.6814
# Fat              3.3768     0.3125  10.804 <0.0000000000000002 ***
# AlcoholSome     53.7377    29.8214   1.802               0.0725 .
# AlcoholALot     37.0725    50.2510   0.738               0.4612
# FatAlcoholSome  -0.8315     0.3706  -2.244               0.0255 *
# FatAlcoholALot  -0.6296     0.5335  -1.180               0.2389
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 92.88 on 309 degrees of freedom
# Multiple R-squared:  0.5128,  Adjusted R-squared:  0.5049
# F-statistic: 65.04 on 5 and 309 DF,  p-value: < 0.00000000000000022

CholesterolHatM7 <- predict(model7, newdata = mydata)
mydata<-cbind.data.frame(mydata, CholesterolHatM7)
```

The intercept is -10.1124 as the mean value for AlcoholConsumption='None' and each unit of Fat adds 3.38 to the predicted Cholesterol value. People categorized as AlcoholConsumption = 'Some' increases Cholesterol by 55.7377 but also decreases it with a portion of their Fat value by a factor of 0.8315. People categorized as AlcoholConsumption = 'ALot' increases Cholesterol by 37.0725 but also decreases a portion of their Fat value by a factor of 0.6296. Calculating R^2 shows that Model 5 shows only a marginal improvement over Model 4, accounting for 51.3% of the variability for predicting Cholesterol.



Plotting Model 7 grouped by AlcoholConsumption displays unequal slopes between each category, as well as showing interaction by crossing over each other.

Hypothesis testing:

- H_0: Beta4 = Beta5 = 0 – Model 7 has no interaction
- H_A: At least one of Beta4 or Beta5 is not zero – Model 7 has interaction

```
# n: 315
# df(RM): 311
# df(FM): 309
# df(RM)-df(FM): 311 - 309 = 2
# df(int): 2
# dim(RM): 3
# dim(FM): 5
# dim(FM)-dim(RM): 5 - 3 = 2

# REDUCED MODEL (RM) RESULTS - Model 6
# SS(reg) = 2756468 + 3126 + 1994 = 2761588
# SS(err) = 2708853
#
# FULL MODEL (FM) RESULTS - Model 7
# SS(REG) = 2756468 + 3126 + 1994 + 31479 + 12011 = 2805078
# SS(ERR) = 2665363
# SS(INT) = 31479 + 12011 = 43490
#
# F = SS(interaction) / df(int)
#     --------------------
#     SS(error - FM) / df(full)
#
# =   43490/2
#   ------------
#   2665363/309
#
# = 21745/8625.77
#
# = 2.520934

# F-Critical: df1=2 (df(FM)-df(RM)) and df2=311 (df(RM)): 0.02531987
qf(p=0.05/2, df1=2, df2=311)

# F-Statistic > F-Critical --> 2.520934 > 0.02531987, reject H0
```

With Model 6 as the reduced model nested within full Model 7, calculating $F_{Statistic}$ 2.521 and $F_{Critical}$ 0.025 above shows that $F_{Statistic} > F_{Critical}$, such that we can reject the null hypothesis and adding variables FatAlcoholSome and FatAlcoholALot is statistically significant and indicating interaction, and that full Model 7 would be preferred over reduced Model 6.

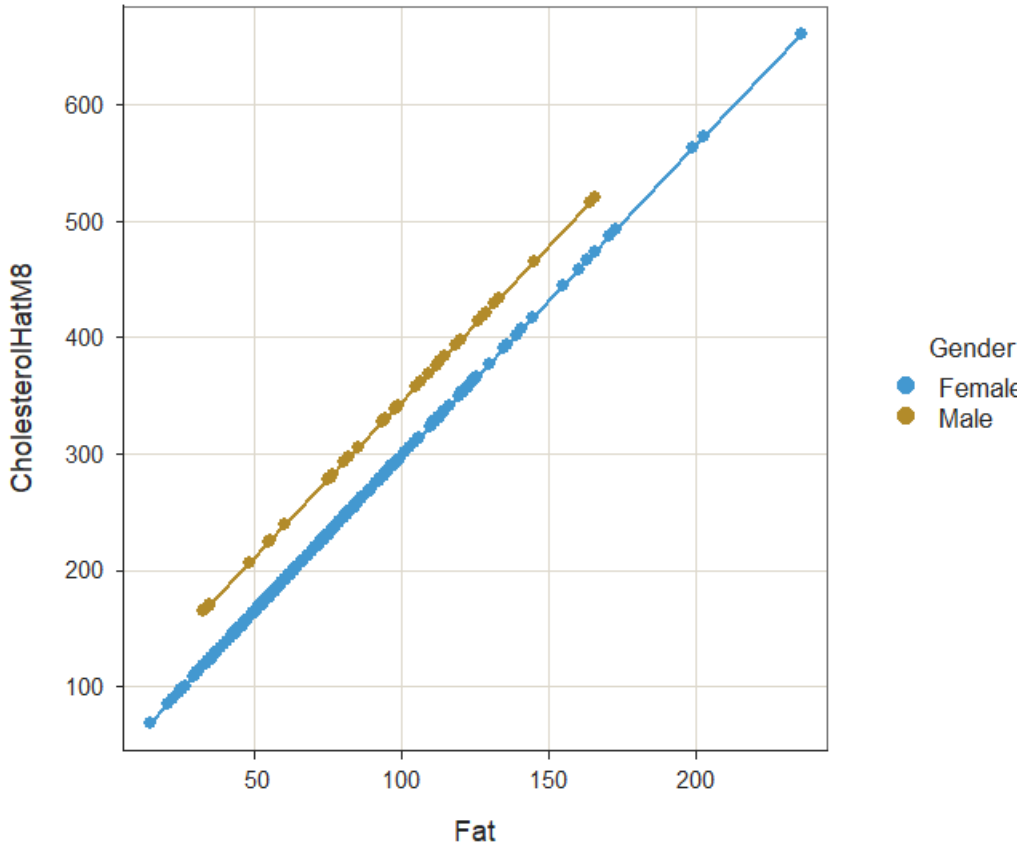
Variable Gender

We create the following regression Model 8 by adding dummy variables for Gender and using Gender = 'Female' as our basis for interpretation:

Cholesterol = 29.9715+ 2.68*Fat + 46.764*GenderM

```
########################################
# GENDER - Model 8
########################################

# control group: Gender = 'Female'
model8 <- lm(Cholesterol ~  Fat + GenderM, data= mydata)
model8

anova(model8)
# > anova(model8)
# Analysis of Variance Table
#
# Response: Cholesterol
#            Df  Sum Sq Mean Sq  F value                Pr(>F)
# Fat         1 2756468 2756468 326.0830 < 0.00000000000000022 ***
# GenderM     1   76552   76552   9.0559              0.002832 **
# Residuals 312 2637421    8453
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model8)
# > summary(model8)
#
# Call:
#   lm(formula = Cholesterol ~ Fat + GenderM, data = mydata)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -223.17  -49.65   -9.89   34.80  518.06
#
# Coefficients:
#             Estimate Std. Error t value             Pr(>|t|)
# (Intercept)  29.9715    12.9039   2.323              0.02084 *
# Fat           2.6775     0.1564  17.119 < 0.00000000000000002 ***
# GenderM      46.7640    15.5399   3.009              0.00283 **
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 91.94 on 312 degrees of freedom
# Multiple R-squared:  0.5179,  Adjusted R-squared:  0.5148
# F-statistic: 167.6 on 2 and 312 DF,  p-value: < 0.00000000000000022

CholesterolHatM8 <- predict(model8, newdata = mydata)
mydata<-cbind.data.frame(mydata, CholesterolHatM8)
```

The intercept is 29.9715 as the mean value for Gender='Female' and each unit of Fat adds 2.68 to the predicted Cholesterol value; Males adds 46.764 to Cholesterol. Calculating R^2 shows that Model 4 accounts for about 51.8% of the variability for predicting Cholesterol.



Plotting Model 8 shows parallel regression lines between genders, indicating no interaction between them within the model.

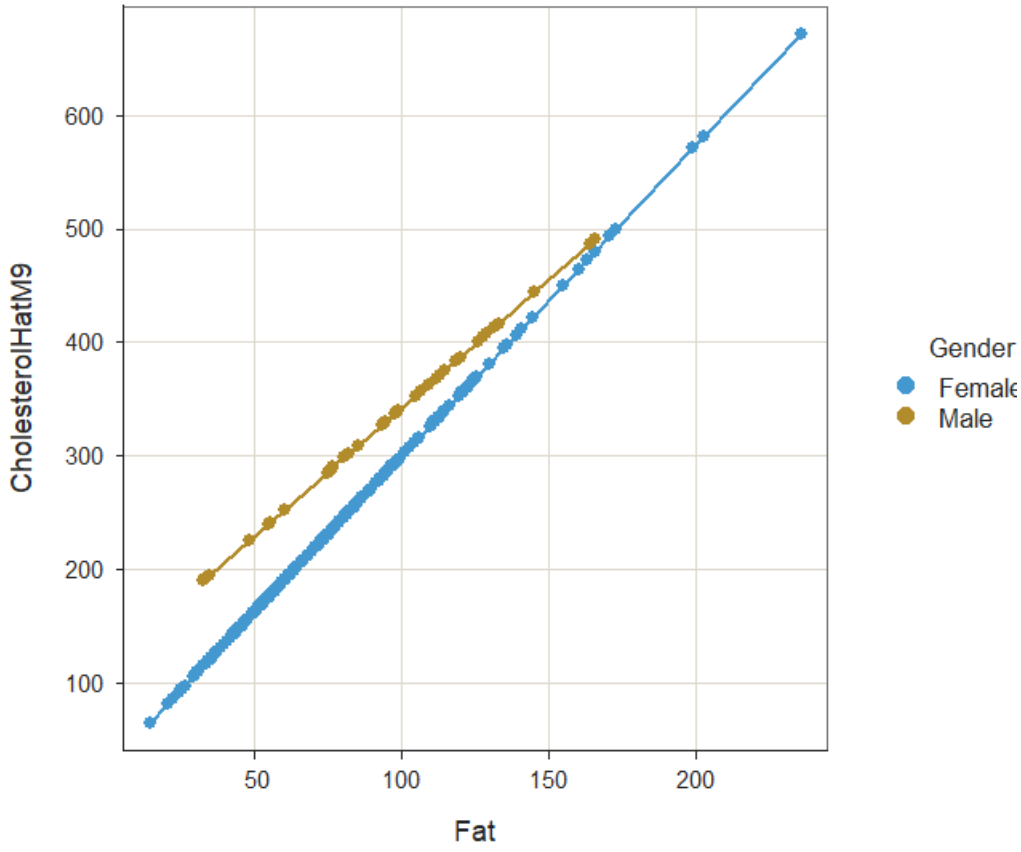
We create the following regression Model 9 by multiplying Fat and the GenderM dummy variable and adding it to Model 9:

Cholesterol = 25.1472 + 2.7423*Fat + 90.785*GenderM - 0.4823*FatGenderM

```r
#######################################
# add FAT * GENDER - Model 9
#######################################

FatGenderF <- Fat*GenderF
FatGenderM <- Fat*GenderM

mydata<-cbind.data.frame(mydata,FatGenderF,FatGenderM)

# control group: Gender = 'Female'
model9 <- lm(Cholesterol ~  Fat + GenderM + FatGenderM, data= mydata)
model9

anova(model9)
# > anova(model9)
# Analysis of Variance Table
#
# Response: Cholesterol
#             Df  Sum Sq Mean Sq  F value                  Pr(>F)
# Fat          1 2756468 2756468 326.1942 < 0.00000000000000022 ***
# GenderM      1   76552   76552   9.0589                0.002828 **
# FatGenderM   1    9350    9350   1.1064                0.293674
# Residuals  311 2628071    8450
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model9)
# > summary(model9)
#
# Call:
#   lm(formula = Cholesterol ~ Fat + GenderM + FatGenderM, data = mydata)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -208.26  -49.71   -9.76   34.62  519.46
#
# Coefficients:
#             Estimate Std. Error t value            Pr(>|t|)
# (Intercept)  25.1472    13.6927   1.837              0.0672 .
# Fat           2.7423     0.1681  16.316 <0.0000000000000002 ***
# GenderM      90.7850    44.6411   2.034              0.0428 *
# FatGenderM   -0.4823     0.4585  -1.052              0.2937
# ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 91.93 on 311 degrees of freedom
# Multiple R-squared:  0.5196,  Adjusted R-squared:  0.515
# F-statistic: 112.1 on 3 and 311 DF,  p-value: < 0.00000000000000022

CholesterolHatM9 <- predict(model9, newdata = mydata)
mydata<-cbind.data.frame(mydata, CholesterolHatM9)
```

The intercept is 25.1472 as the mean value for Gender='Female' and each unit of Fat adds 2.74 to the predicted Cholesterol value. People Listed as Gender = 'Male' increases Cholesterol by 90.8 but

also decreases it by a portion of their Fat value by a factor of 0.4823. Calculating R^2 shows that Model 9 shows only a very marginal improvement over Model 8, just a 0.2% increase of the variability for predicting Cholesterol.



Plotting Model 9 grouped by Gender displays unequal slopes between each category, as well as showing interaction by potentially crossing over each other.

Hypothesis testing:

- H_0: Beta4 = 0 – Model 9 has no interaction
- H_A: Beta4 != 0 – Model 9 has interaction

```
# n: 315
# df(RM): 312
# df(FM): 311
# df(RM)-df(FM): 312 - 311 = 1
# df(int): 1
# dim(RM): 2
# dim(FM): 3
# dim(FM)-dim(RM): 3 - 2 = 1

# REDUCED MODEL (RM) RESULTS - Model 8
# SS(reg) = 2756468 + 76552 = 2833020
# SS(err) = 2637421
#
# FULL MODEL (FM) RESULTS - Model 9
# SS(reg) = 2756468 + 76552 + 9350 = 2842370
# SS(err) = 2628071
# SS(int) = 9350
#
# F = SS(interaction) / df(int)
#     --------------------
#     SS(error - FM) / df(full)
#
# =   9350/1
#   ------------
#   2628071/311
#
# = 9350/8450.389
#
# = 1.106458

# F-Critical: df1=1 (df(FM)-df(RM)) and df2=312 (df(RM)): 0.0009836458
qf(p=0.05/2, df1=1, df2=312)

# F-Statistic > F-Critical --> 1.106458 > 0.0009836458, reject H0
```

With Model 8 as the reduced model nested within Model 9, calculating $F_{Statistic}$ 1.107 and $F_{Critical}$ 0.001 above shows that $F_{Statistic} > F_{Critical}$, such that we can reject the null hypothesis and adding variable FatGenderM to Model 9 is statistically significant and indicating interaction, and that full Model 9 would be preferred over reduced Model 8.

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

   This exercise in understanding categorical variables and integrating it into a regression model was very insightful. While it might not be the case every time, going through a systematic process in using these kinds of variables, from preparatory work to hypothesis testing, gives us a framework to execute for performing an exploratory data analysis (EDA) and expands my toolset for understanding and transforming data in new ways to help create a better-fitting and more appropriate regression model. With more practice and experience we will become more proficient at working with the datasets, and I am looking forward to learning more ways and developing my skillset when it comes to working with various kinds of data.