Assignment #5: Interpreting ANOVA Models

Points: 50 points

Data: The data for this assignment is the nutrition study data set. This data will be made available by your instructor.

Assignment Instructions:

In this assignment we are going to fit some basic ANOVA models and attempt to understand what they are doing and the output that they produce. We will use a starter script to guide the model fitting so that we all fit the same models correctly, and hence we all get the same output. We will use the output to answer questions in the Assignment #5 template. We will record our answers in green, convert the template to pdf, and submit the pdf as the solution. Hence, this is a question and answer assignment and not a report assignment.

There are three primary take-aways from this assignment:

(1) ANOVA is not predictive modeling. ANOVA is statistical inference.
(2) ANOVA models compute segment or cohort means.
(3) ANOVA models (in their most basic specifications) go hand-in-hand with the concept of indicator variables.

Use the starter script for Assignment #5 to walk you through the questions for this assignment.

(1) ANOVA regression for Gender

Fit this ANOVA model.

```
model.1 <- lm(Cholesterol ~ Gender, data=my.df);
```

(1a) (10 points) What does this ANOVA model estimate? Compute the means for each Gender and demonstrate and interpret how the ANOVA model has computed the mean value for each level of Gender.

This model estimates the mean cholesterol levels for each category of gender. The Anova model has computed the mean cholesterol value for female as the intercept and the additional cholesterol for males as the intercept + GenderMale.

Mean Male Cholesterol: 328.1238

Mean Female Cholesterol: 229.2817

```
> model.1 <- lm(Cholesterol ~ Gender, data=my.df);
> summary(model.1)

Call:
lm(formula = Cholesterol ~ Gender, data = my.df)

Residuals:
    Min      1Q  Median      3Q     Max
-250.62  -85.65  -33.48   54.72  671.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  229.282      7.737  29.635  < 2e-16 ***
GenderMale    98.842     21.188   4.665 4.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 127.8 on 313 degrees of freedom
Multiple R-squared:  0.06501,   Adjusted R-squared:  0.06202
F-statistic: 21.76 on 1 and 313 DF,  p-value: 4.58e-06
```

(1b) (5 points) What does this ANOVA model test? Write out the hypothesis test in verbal form. You should have a null hypothesis H0 and an alternate hypothesis H1. Hint: Take the sqrt() of the F-statistic and see if it matches any other output.

H0: MeanMale = MeanFemale

H1: MeanMale ≠ MeanFemale

The square root of the F-statistic is the t-value of GenderMale, the additional cholesterol for males. This model tests whether the means are equal for each level of gender.

(1c) (5 points) What should we conclude from this ANOVA model?

From this ANOVA model, we can conclude that the mean cholesterol levels are not equal for males and females, thus rejecting the null hypothesis.

(2) (10 points) Regression Model with Indicator for Gender=='Male'

Create an indicator for Male. Fit this regression model.

model.2 <- lm(Cholesterol ~ Male, data=my.df);

Compare the model.2 output to model.1. What did R do under the hood in model.1?

The output for model.2 and model.1 is identical. Since we did not specify a gender level for model.1, R automatically separated the category into each value and computed the model for those values. In model.2 we did specify the level as male, so R essentially performed the same function, but the coefficient is now labeled as Male.

```
> model.2 <- lm(Cholesterol ~ Male, data=my.df);
> summary(model.2)

Call:
lm(formula = Cholesterol ~ Male, data = my.df)

Residuals:
    Min      1Q  Median      3Q     Max
-250.62  -85.65  -33.48   54.72  671.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  229.282      7.737  29.635  < 2e-16 ***
Male          98.842     21.188   4.665 4.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 127.8 on 313 degrees of freedom
Multiple R-squared:  0.06501,   Adjusted R-squared:  0.06202
F-statistic: 21.76 on 1 and 313 DF,  p-value: 4.58e-06
```

(3) (10 points) ANOVA regression model for Smoke

Fit this ANOVA model.

model.3 <- lm(Cholesterol ~ Smoke, data=my.df);

What conclusions should we draw from this model?

This model estimates the mean cholesterol level for non-smokers as 237.707 and for smokers as 237.707+34.826 = 272.533. It tests whether the two means are significantly different from each other. The p-value for the model is .108 which is higher than the standard significance threshold of 0.05. This means we would fail to reject the null hypothesis and conclude there is no difference in the mean cholesterol levels for smokers vs. non-smokers.

```
> model.3 <- lm(Cholesterol ~ Smoke, data=my.df);
> summary(model.3)

Call:
lm(formula = Cholesterol ~ Smoke, data = my.df)

Residuals:
    Min      1Q  Median      3Q     Max
-200.01  -89.22  -35.51   66.18  662.99

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  237.707      7.983  29.777   <2e-16 ***
SmokeYes      34.826     21.606   1.612    0.108
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 131.7 on 313 degrees of freedom
Multiple R-squared:  0.008232,  Adjusted R-squared:  0.005064
F-statistic: 2.598 on 1 and 313 DF,  p-value: 0.108
```

(4) (10 points) Use a 2x2 ANOVA regression model to compute the four means. Verify these means using the model and their separate computations.

Fit this model and verify the four resulting means.

model.4 <- lm(Cholesterol ~ Gender*Smoke, data=my.df);

Means for 2X2 ANOVA:

Male Nonsmoker: 227.534 + 79.058 = 306.5914

Male Smoker: 227.534 + 79.058 + 13.255 + 115.939 = 435.7857

Female Nonsmoker: 227.5338

Female Smoker: 227.534 + 13.255 = 240.7889

These means match the values generated by the manual check code output.

```
> model.4 <- lm(Cholesterol ~ Gender*Smoke, data=my.df);
> summary(model.4)

Call:
lm(formula = Cholesterol ~ Gender * Smoke, data = my.df)

Residuals:
    Min      1Q  Median      3Q     Max
-229.09  -82.96  -31.93   57.64  673.17

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        227.534      8.246  27.593  < 2e-16 ***
GenderMale          79.058     22.988   3.439 0.000663 ***
SmokeYes            13.255     22.708   0.584 0.559832
GenderMale:SmokeYes 115.939    57.257   2.025 0.043732 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 126.9 on 311 degrees of freedom
Multiple R-squared:  0.08381,   Adjusted R-squared:  0.07497
F-statistic: 9.483 on 3 and 311 DF,  p-value: 5.179e-06
```

```
> male.nonsmoker <- subset(my.df$Cholesterol,(my.df$Gender=='Male')&(my.df$Smoke=='No'));
> mean(male.nonsmoker)
[1] 306.5914
> male.smoker <- subset(my.df$Cholesterol,(my.df$Gender=='Male')&(my.df$Smoke=='Yes'));
> mean(male.smoker)
[1] 435.7857
> female.nonsmoker <- subset(my.df$Cholesterol,(my.df$Gender=='Female')&(my.df$Smoke=='No'));
> mean(female.nonsmoker)
[1] 227.5338
> female.smoker <- subset(my.df$Cholesterol,(my.df$Gender=='Female')&(my.df$Smoke=='Yes'));
> mean(female.smoker)
[1] 240.7889
>
```

APPENDIX

```
my.df <- read.csv('NutritionStudy.csv',header=TRUE);
head(my.df)
str(my.df)

# Define some indicator variables;
my.df$SmokeYes <- ifelse(my.df$Smoke=='Yes',1,0);
my.df$Male <- ifelse(my.df$Gender=='Male',1,0);
my.df$RegularVitamin <- ifelse(my.df$VitaminUse=='Regular',1,0);


#####################################################################
# What does this model compute?
#####################################################################
model.1 <- lm(Cholesterol ~ Gender, data=my.df);
summary(model.1)

# Compute the mean cholesterol value for each Gender;
# Use these values to interpret and validate the ANOVA model;
male.cholesterol <- subset(my.df$Cholesterol,my.df$Gender=='Male');
mean(male.cholesterol)

female.cholesterol <- subset(my.df$Cholesterol,my.df$Gender=='Female');
mean(female.cholesterol)

#####################################################################
# What did the ANOVA model test?
#####################################################################
# Write out the test in verbal form.
# Can you write out the test in statistical notation?
# What conclusion should we draw from the ANOVA model?



#####################################################################
# Relationship between ANOVA and indicator variables
#####################################################################
# Now let's use the Male indicator variable that we defined;
# What does this model compute?
model.2 <- lm(Cholesterol ~ Male, data=my.df);
summary(model.2)

# From the output of this model do we understand what R did under the hood?


#####################################################################
# Run an ANOVA regression for the Smoke variable
#####################################################################
# What does this model compute?
model.3 <- lm(Cholesterol ~ Smoke, data=my.df);
summary(model.3)

# What is the statistical test for this model?
# What is the conclusion of this test for this model?



#####################################################################
```

```r
#####################################################################
# Use a 2x2 ANOVA regression to compute the four means
#####################################################################

# As ANOVA models get more complicated they are harder to interpret.
# It becomes better to interpret the model in terms of a factor regression model
# and not a t-test.
# Let's consider this 2x2 model;

model.4 <- lm(Cholesterol ~ Gender*Smoke, data=my.df);
summary(model.4)

# Compute the sample means for each of the four groups and validate the model
# estimates to the sample means;
# Extract the model coefficients using model.4$coef to make your computations;

male.nonsmoker <- subset(my.df$Cholesterol,(my.df$Gender=='Male')&(my.df$Smoke=='No'));
mean(male.nonsmoker)

male.smoker <- subset(my.df$Cholesterol,(my.df$Gender=='Male')&(my.df$Smoke=='Yes'));
mean(male.smoker)

female.nonsmoker <- subset(my.df$Cholesterol,(my.df$Gender=='Female')&(my.df$Smoke=='No'));
mean(female.nonsmoker)

female.smoker <- subset(my.df$Cholesterol,(my.df$Gender=='Female')&(my.df$Smoke=='Yes'));
mean(female.smoker)

```