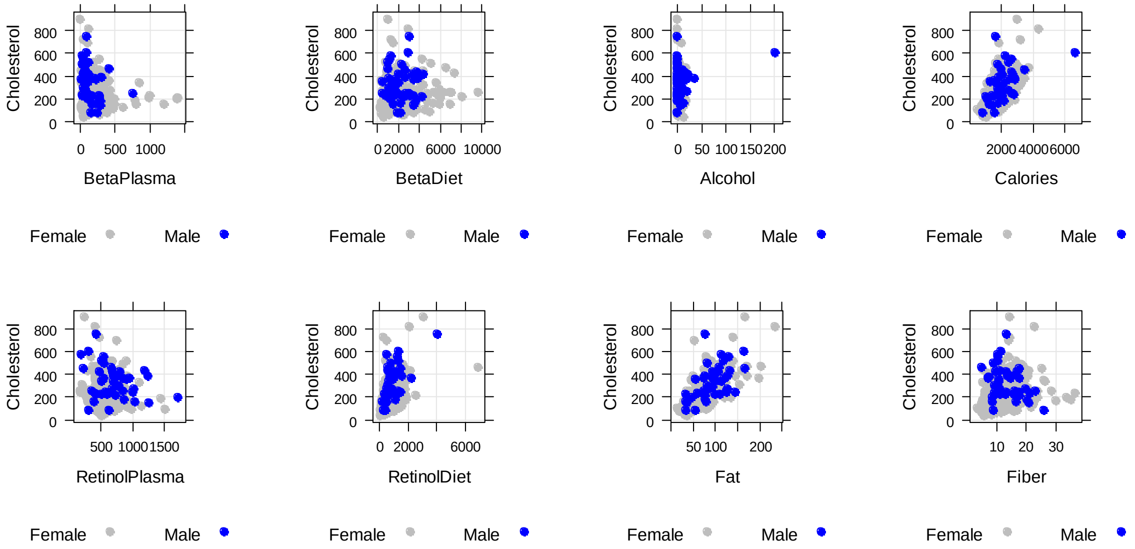
Assignment #6: Fitting and Interpreting ANCOVA Models

Points: 100 points

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

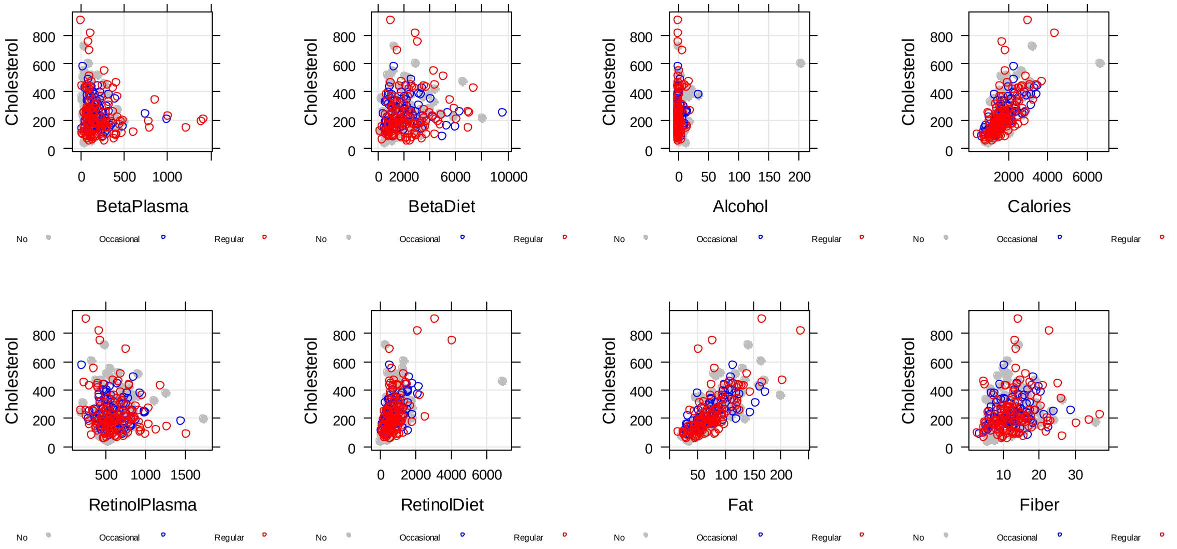
Assignment Instructions:

In this assignment we are going to look at a variety of statistical graphics that we may use in the Analysis of Covariance, and then fit some basic ANCOVA 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 #6 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.



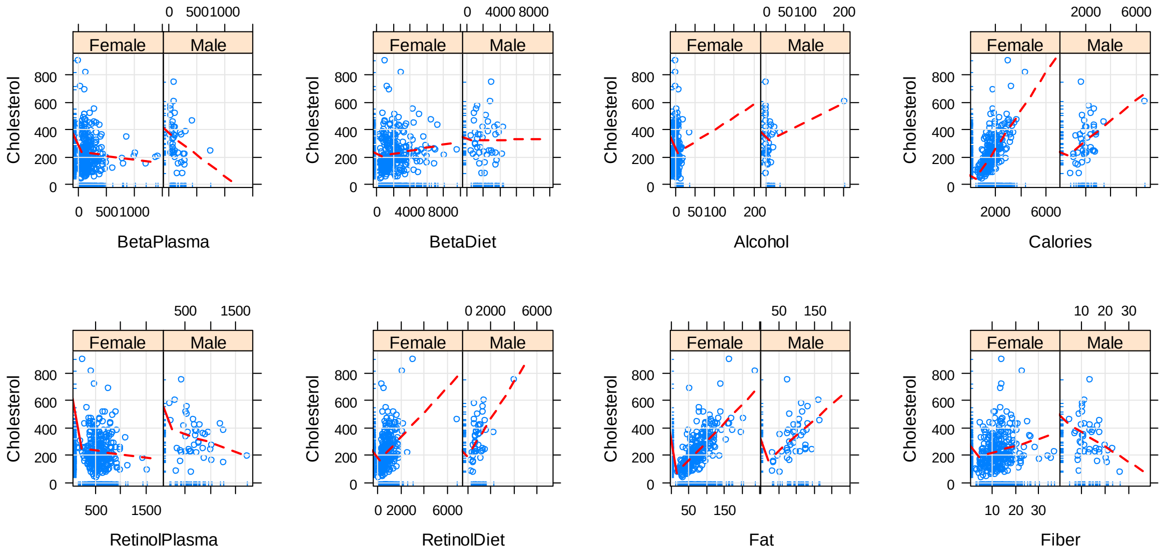
(1) (5 points) If we had to pick one point in all of these plots conditioned on Gender to be an outlier, then which point would we pick? Find the observation in the data set and provide the ID number?

I would pick the Male outlier point on the Alcohol vs Cholesterol chart. This corresponds to ID 62.



(2) (5 points) Do these plots conditioned on vitamin use suggest that vitamin use affects cholesterol? Why or why not?

There doesn't seem to be definite trends that can be identified based on frequency of vitamin use. The three types of points are clustered together and don't indicate a difference based on purely visual analysis.



(3) (10 points) Which two variables appear to have the best regression lines?

Calories and fat have the best regression lines. Outliers heavily influence the regression lines of the other variables

(4) (10 points) The variable retinol diet appear to have two regression lines with nice slopes to them. Should we trust those regression lines? Are there reasons to consider those regression lines to be bad regression lines?

The regression lines for retinol diet seem to follow the basic trend of the variable, but they might not be completely accurate. The female regression line is pulled to a lower slope because of one observation with a high retinol diet value but a mid-cholesterol value. The male regression line is also pulled to a smaller slope because of one of these observations.

(5) (10 points) Are there any plots that do not make any sense? Hint: Sometimes we condition on a variable and then the relationship between the predictor variable and the response variable becomes nonsense. Do we see that behavior in any of the plots?

Some of the variables (BetaDiet, RetinolPlasma) have data points which are spread in a way where a definite pattern is not identifiable. For other plots like Alcohol, the scale makes the relationship between predictor and response very difficult to tell.

We fit ANCOVA models in order to specify regression models with segmented effects on a pooled sample, that is, instead of fitting many separate linear regression models on subsets of the data, we fit a single complex model on the pooled sample. These specifications allow shared effects and separate effects across the segments. For the interpretation of ANCOVA models we typically decompose the ANCOVA models into separate models. Let’s fit some ANCOVA models, and see if we can understand what they are.

(6) (10 points) Fit the following ANCOVA model. Write out the regression equations for Male and Female.

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

Male: Cholesterol = 76.7355 + 2.6775*Fat

Female: Cholesterol = 29.9715 + 2.6775*Fat

(7) (10 points) Fit the following ANCOVA model. Write out the regression equations for Male and Female.

```
model.2 <- lm(Cholesterol ~ Gender:Fat, data=my.df);
```

Male: Cholesterol = 33.6884 + 3.0383*Fat

Female: Cholesterol = 33.6884 + 2.6465*Fat

(8) (10 points) These two ANCOVA models have different model specifications. Compute the Mean Absolute Error (MAE) and the Mean Square Error (MSE) for both model.1 and model.2. Which model should we prefer?

MAE model 1 = 62.11234

MSE model 1 = 8372.765

MAE model 2 = 62.3326

MSE model 2 = 8454.032

Although very close, model 1 has slightly lower error, which would make it the preferred model

(9) (10 points) Fit the following ANCOVA model. Write out the regression equations for Male and Female.

```
my.df$Male <- ifelse(my.df$Gender=='Male',1,0);

model.3 <- lm(Cholesterol ~ Male + Fat + Male:Fat, data=my.df);
```

Male: Cholesterol = 115.9322 + 2.26*Fat
Female: Cholesterol = 25.1472 + 2.7423*Fat

(10) (10 points) Compute the Mean Absolute Error (MAE) and the Mean Square Error (MSE) for model.3. Which of the three models should we prefer?

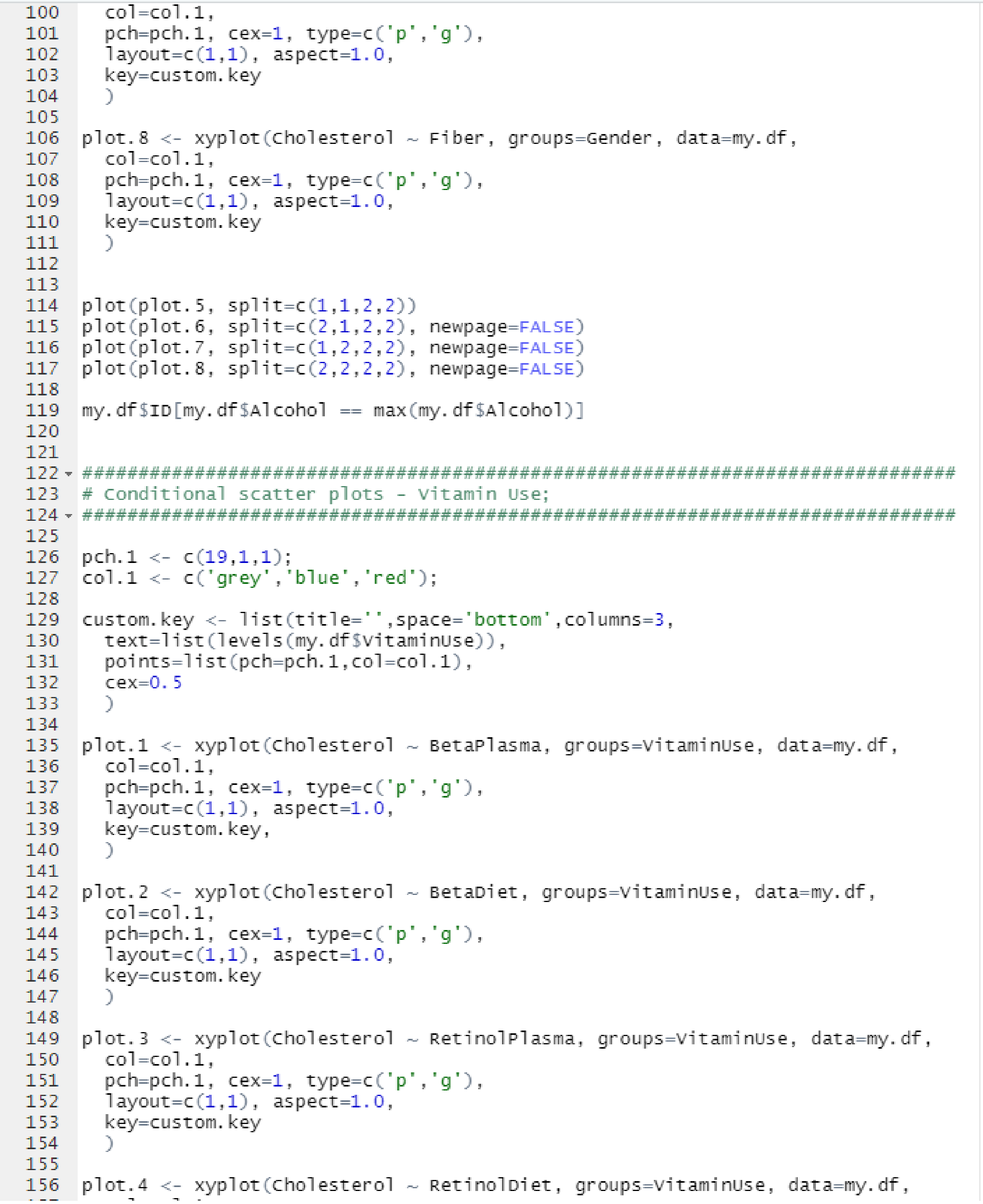
MAE model 3 = 62.34995
MSE model 3 = 8343.083

This model has about the same MAE as model 1 and 2, but it has a lower MSE than both. This might indicate that model 3 is better.

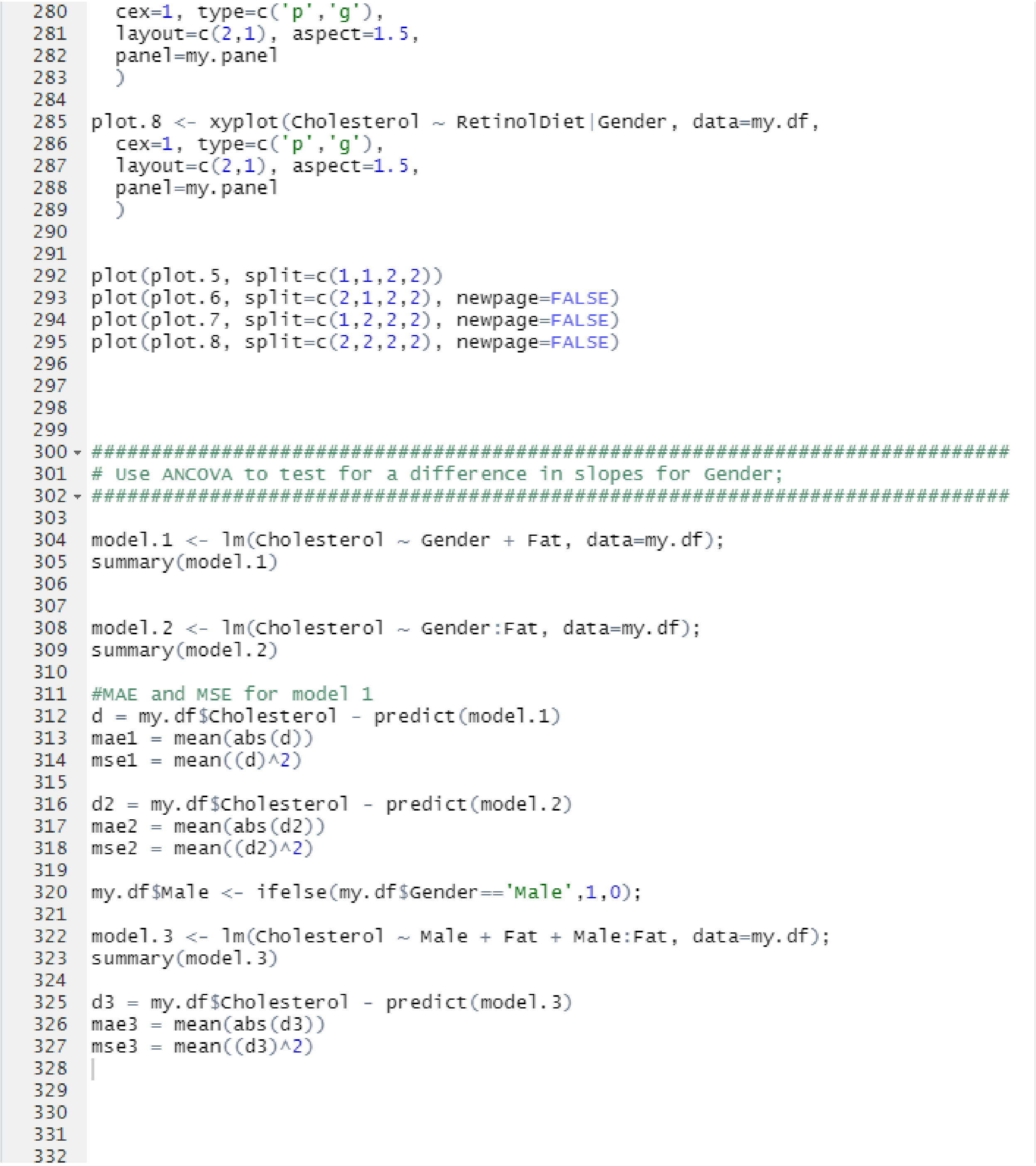
(11) (10 points) How should we interpret the regression model output for model.3? Does model.3 suggest that each Gender should have its own intercept and its own slope?

The model is not suggesting each gender gets its own intercept and slope. The default intercept/fat values are for females since we are building our model based on gender. The male estimates are added to the initial coefficients to signify the male component of the regression equation.

Appendix



```
  col=col.1,
  pch=pch.1, cex=1, type=c('p','g'),
  layout=c(1,1), aspect=1.0,
  key=custom.key
  )

plot.8 <- xyplot(Cholesterol ~ Fiber, groups=Gender, data=my.df,
  col=col.1,
  pch=pch.1, cex=1, type=c('p','g'),
  layout=c(1,1), aspect=1.0,
  key=custom.key
  )


plot(plot.5, split=c(1,1,2,2))
plot(plot.6, split=c(2,1,2,2), newpage=FALSE)
plot(plot.7, split=c(1,2,2,2), newpage=FALSE)
plot(plot.8, split=c(2,2,2,2), newpage=FALSE)

my.df$ID[my.df$Alcohol == max(my.df$Alcohol)]


##############################################################################
# Conditional scatter plots - Vitamin Use;
##############################################################################

pch.1 <- c(19,1,1);
col.1 <- c('grey','blue','red');

custom.key <- list(title='',space='bottom',columns=3,
  text=list(levels(my.df$VitaminUse)),
  points=list(pch=pch.1,col=col.1),
  cex=0.5
  )

plot.1 <- xyplot(Cholesterol ~ BetaPlasma, groups=VitaminUse, data=my.df,
  col=col.1,
  pch=pch.1, cex=1, type=c('p','g'),
  layout=c(1,1), aspect=1.0,
  key=custom.key,
  )

plot.2 <- xyplot(Cholesterol ~ BetaDiet, groups=VitaminUse, data=my.df,
  col=col.1,
  pch=pch.1, cex=1, type=c('p','g'),
  layout=c(1,1), aspect=1.0,
  key=custom.key
  )

plot.3 <- xyplot(Cholesterol ~ RetinolPlasma, groups=VitaminUse, data=my.df,
  col=col.1,
  pch=pch.1, cex=1, type=c('p','g'),
  layout=c(1,1), aspect=1.0,
  key=custom.key
  )

plot.4 <- xyplot(Cholesterol ~ RetinolDiet, groups=VitaminUse, data=my.df,
```



```
  cex=1, type=c('p','g'),
  layout=c(2,1), aspect=1.5,
  panel=my.panel
  )

plot.8 <- xyplot(Cholesterol ~ RetinolDiet|Gender, data=my.df,
  cex=1, type=c('p','g'),
  layout=c(2,1), aspect=1.5,
  panel=my.panel
  )


plot(plot.5, split=c(1,1,2,2))
plot(plot.6, split=c(2,1,2,2), newpage=FALSE)
plot(plot.7, split=c(1,2,2,2), newpage=FALSE)
plot(plot.8, split=c(2,2,2,2), newpage=FALSE)




################################################################################
# Use ANCOVA to test for a difference in slopes for Gender;
################################################################################

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


model.2 <- lm(Cholesterol ~ Gender:Fat, data=my.df);
summary(model.2)

#MAE and MSE for model 1
d = my.df$Cholesterol - predict(model.1)
mae1 = mean(abs(d))
mse1 = mean((d)^2)

d2 = my.df$Cholesterol - predict(model.2)
mae2 = mean(abs(d2))
mse2 = mean((d2)^2)

my.df$Male <- ifelse(my.df$Gender=='Male',1,0);

model.3 <- lm(Cholesterol ~ Male + Fat + Male:Fat, data=my.df);
summary(model.3)

d3 = my.df$Cholesterol - predict(model.3)
mae3 = mean(abs(d3))
mse3 = mean((d3)^2)
```