Understanding Data Mining

Lesson 1: Using R for Linear Regression Name___________________________

I. Introduction to R

1. To make sure our results are always the same, type the command **`set.seed(2008).`**

2. We want to use a data set that is already stored in R. To see the data set, simply type the command **`Loblolly`**. If you scroll up to the top of the data set, you should see the column titles "height," "age," and "seed." This data represents the recorded heights (in feet) and ages (in years) of some loblolly pine trees.

3. Next, we will have to load the data into our R workspace using the command **`data(Loblolly).`** No data should show up after you press enter.

4. We now want to create a scatter plot of the data, so use the command **`plot(Loblolly$age, Loblolly$height)`**, which tells R you know age is the independent variable and height is the dependent variable.

5. Notice that R automatically labeled the x- and y-axes, but we also want our scatter plot to have a main title. To add a title, use the command **`title(main = "Loblolly Pine Tree Heights").`**

6. To find a linear model that relates the age and height of the loblolly pine trees, we will use the command **`fit1<-lm(Loblolly$height~Loblolly$age).`** Think of this as slope-intercept form (y=mx+b).

7. To see the model, type **`fit1`**. Record your linear model here:____________________________

8. Now we want to add the graph of this line of best fit to our scatter plot. To do this, use the command **`yfit1<-fit1$fitted.values`**. To see it on your scatter plot, type **`lines(Loblolly$age, yfit1).`**

9. The final piece of information we want about our data is the correlation of the age and height of the Loblolly pine trees. To find the correlation coefficient, use the command **`cor(Loblolly$height, Loblolly$age).`** Record the value of r here: ____________

II. Answer these questions based on our results in 1 – 9 above:

1. What does the value of the correlation coefficient tell us about the relationship between the age and the height of these Loblolly pine trees? Be specific in your explanation.

____

____

____

Understanding Data Mining

2. Interpret the meaning of the slope and the y-intercept of the linear model we recorded in #7.

____

____

3. Use the linear model from #7 above to predict the height of a Loblolly pine tree in my back yard that is 32 years old.

____

4. Use the linear model to predict the age of a Loblolly that is 12 feet tall. _______________

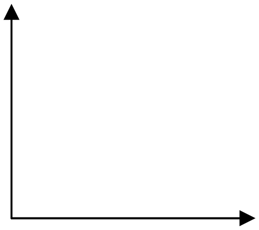
5. Why would using a graphing calculator not be optimal when finding a linear model for a data set like Loblolly?

____

____

III. Now, your job is to practice your R skills using another data set. This time, you are to use **cars**, which gives the speed (in mph) and the braking distance (in feet) for a set of cars in the 1920's.

1. Create a scatter plot of the data. Sketch it here, with its titles!



2. Find a linear model for the data. Add it to your scatter plot, and record the equation below:

____

3. Find the value of the correlation coefficient. Record it below:

____

4. Predict the stopping distance of a car driving 32 miles per hour. ____________

5. Approximately how fast was a car driving if it took 75 feet to stop? ____________

Understanding Data Mining

IV. What if we want to use our own data set? We must put our data in a spreadsheet and load it into R. Let’s use the test score data.

1. Open Microsoft Excel.

2. Name column A **hours** and column B **score**. Then, enter the data.

3. When saving the file, call it **scores** and save it as a text file (.txt). If we don’t, R won’t let us upload the data.

4. In R, change your directory so we can upload the data. To do this, go to “File” and click on “Change dir.” Choose the folder where you saved scores.txt. Now, upload the data using the command **`scores<-read.table(“scores.txt”, header = T).`**

5. Now, let’s draw a scatter plot. Make sure it has a title! Record the code you used below.

code for scatter plot: ____

code for title: ____

6. Let’s generate the linear model for the data, and then let’s add it to the scatter plot. Record the code you used below:

code for linear model: ____

code for adding to scatter plot: ____

7. Calculate the correlation coefficient of the data. Record the code you used below:

____

8. Think about how these results compare to our results from the graphing calculator earlier in our lesson.

(a) Which tool do you think is more useful…the graphing calculator or R? ________________

____

(b) Which tool do you prefer, and why? ____

____

(c) Identify a “real-world” job in which someone may have to utilize statistical software on a regular basis.

____