Psych 240 Lab 3

The purpose of this assignment is to show you how you can use R to view scatterplots, find least-squares regression lines, and evaluate correlations.

Setup Instructions (please complete these steps before the TA begins their presentation for the day)

1) Log on to your lab computer with your OIT username and password. Alert the TA if you have any trouble with this.

2) Go to the course Moodle page. In the “Labs” section, click on the file called “Lab3”. In the dialog box that opens, select “Save File” and hit OK. If you are asked to select a directory, choose the “Downloads” folder. If you are not asked to select a directory, the file should automatically be saved in “Downloads”.

3) On the Start Menu, select “All Programs” – “Statistics” – “R” – “R x64 3.1.0”. You should see R launch on your computer screen.

4) In the R console, type the following text: setwd(“C:/Users/*your-OIT-username*/Downloads”). For example, if your OIT username is “astudent”, you would type in setwd(“C:/Users/astudent/Downloads”).

5) Next, enter this at the R console: source(“Lab3.txt”). If all goes correctly, you should be asked to enter your student ID number. If you get a “file not found” message, then you either didn’t save the Lab2 file from Moodle in the Downloads directory or didn’t set R to your Downloads directory. You can get help from the TA or another student.

6) If you see a prompt that says “Student ID not found”, check your ID number and try to enter it again. If it won’t work after several tries, you can just enter 0 instead of your ID and the program will let you continue. If you do this, you will not get a completion code at the end of the assignment, so you have to show the TA your R screen after you complete the assignment and before you leave.

7) Wait for further instruction from the TA.

Lab Demonstration

The TA will guide you through this section. Please wait until you are asked to begin and follow along with the TA.

1) R has very convenient functions to explore the relationship between two variables, and we will learn about some of them today.

2) Let’s say an insurance company wants to see if the number of insurance claims made by residents of a city in the winter is related to the amount of snowfall for that city. MAclaims in the Lab3 file is the number of claims made by residents of each of 50 Massachusetts cities, and MAinches is the amount of snowfall for each of the cities in inches. Type in “MAclaims” and hit ENTER, and you will see the number of claims for each of the 50 cities. Type in “MAinches” and hit ENTER, and you will see the inches of snowfall for each of the 50 cities.

3) The first step in determining if these two variables are related is looking at a scatterplot. In R, you can do this with the “plot” function. Type in “plot(MAinches,MAclaims)” and hit ENTER. You should see a scatterplot of the two variables come up on your screen. Based on what you see, do the two variables appear to be related? How would you describe that relationship?

4) We can also evaluate the relationship between the two variables by computing a correlation coefficient (*r*). In R, you can do this with the function “cor”. Type in “cor(MAinches,MAclaims)” and hit ENTER. You will see the *r* value for these two variables. Does the value that you got seem consistent with your scatterplot?

5) We might also be interested in using one variable to predict the value of the other, and to do this we need to define a regression line. The “lsfit” function in R can find the least-squares regression line that relates two variables. Type in “reg=lsfit(MAinches,MAclaims)” and hit ENTER. This code finds the least-squares regression line and saves the results in a new R object called “reg” (for “regression”). Now type in “reg” and hit ENTER, and you will see all of the saved output from finding the regression line. The most important output is right at the top under “$coefficients”. This gives you the intercept and slope value for the least-squares regression line. The slope value is labeled “X” in R – this notation isn’t important, I just mentioned it so you would know where to look. The next section under “$residuals” shows the deviation between the predicted Y value and the actual Y value for each city in our data set. “Residual” is another word for this deviation. For example, if we predict that a city would have 3 more claims than it actually did, then the residual is -3. If we predict that it would have 5 fewer claims than is actually did, then the residual is 5. The rest of the output just has details on the process that R used to find the least-squares regression line, and you don’t have to worry about it.

6) R will let you put the least-squares regression line on a plot to help visualize the relationship between two variables. The “abline” function adds a line to a plot with a specified intercept and slope value. Look at “reg$coef” again, and we’ll use these intercept and slope values to see the regression line on our plot. Type in “plot(MAinches,MAclaims)” again if the plot is no longer on

your screen. Now, type in “abline(16.048,.225)” and hit ENTER. You should see the regression line appear on your plot.

Lab Assignment

Complete this section on your own. You can ask the TA for help. Record all of your answers on a sheet of paper with your first and last name on the top. The correct answers are different for each student. R will tell you whether or not your answer is correct, as detailed below. When you enter all of the correct answers, R will tell you that you are finished and give you your completion code to write on your answer sheet. You can either leave when you are finished, but I encourage you to please use any extra time to get help on your homework or the lecture content from the TA.

1) The insurance company also wanted to evaluate how the number of winter insurance claims was related to inches of snowfall for cities in Connecticut and New Hampshire. The claim data are stored in vectors called CTclaims and NHclaims, and the snowfall data are stored in vectors called CTinches and NHinches. “CT” denotes the Connecticut data, and “NH” denotes the New Hampshire data. What is the correlation between snowfall and claims for the Connecticut customers?

[Type in “q1(*your-answer-here*)” and hit ENTER to see if you are correct. For example, if you thought the answer was 12, you would type in “q1(12)”.]

2) What is the correlation between snowfall and claims for the New Hampshire customers?

[Type in “q2(*your-answer-here*)” and hit ENTER to see if you are correct.]

3) If we use snowfall to predict the number of claims for the Connecticut customers, what would the predicted number of claims be for a city that got zero inches of snow that winter?

[Type in “q3(*your-answer-here*)” and hit ENTER to see if you are correct.]

4) If we use snowfall to predict the number of claims for the Connecticut customers, how does the predicted number of claims change when the snowfall total increases by 1 inch? Plot the two variables and check the plot to make sure this value looks plausible.

[Type in “q4(*your-answer-here*)” and hit ENTER to see if you are correct.]

5) If we use snowfall to predict the number of claims for the New Hampshire customers, what would the predicted number of claims be for a city that got zero inches of snow that winter?

[Type in “q5(*your-answer-here*)” and hit ENTER to see if you are correct.]

6) If we use snowfall to predict the number of claims for the New Hampshire customers, how does the predicted number of claims change when the snowfall total increases by 1 inch? Plot the two variables and check the plot to make sure this value looks plausible.

[Type in “q6(*your-answer-here*)” and hit ENTER to see if you are correct.]