ECON 282, Professor Hogendorn

Assignment 3

For this assignment, you can pick your own data set. Try to find one with more than 100 rows and several columns. If you prefer, you can use the sample dataset linked on the syllabus.

Some generic R Instructions for importing your dataset: If your dataset is in .txt use DataSetName <- read.delim("DataSet.txt", header=TRUE, stringsAsFactors=FALSE). If your dataset is in .csv use DataSetName <- read.csv("DataSet.csv", stringsAsFactors=FALSE). If your dataset is in another format like from Stata, SAS, etc, you may need to download the "foreign" package to handle it.

1. Describe the dataset in words. Explain a hypothesis about why a correlation might exist between some variables and others.

2. Pick two variables (e.g. y and x) from your hypothesis, and run a t-test to see if there is a difference in the conditional means $(\overline{y}|x < \overline{x})$ and $(\overline{y}|x > \overline{x})$. (Depending on what variable you choose for x, other conditions might be more appropriate, such as sex = men versus women or age = greater versus less than 18.) What do you conclude from your t-test?

3. Using your hypothesis, run the following commands using just one dependent variable and one independent variable:

   plot($y \sim x$, xlab = "name", ylab = "name")

   model1 <- lm($y \sim x$)

   summary(model1)

   abline(model1)

4. Now try a multiple regression where you use more than one independent variable.

5. In your multiple regression from part (4), do you have significant p-values (p-value less than .05) on the coefficients and the regression as a whole? Does this mean your hypothesis is definitely correct? Can you think of possible confounding variables? Is your r-squared value low or high? Why do you think this is?