Pol Sci 630: Problem Set 3 - Comparisons and Inference

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Due Date: Tuesday, September 15th, 2015, 10 AM (Beginning of Class)

Note: It is absolutely essential that you show all your work, including intermediary steps, and comment on your R code to earn full credit. Showing all steps and commenting on code them will also be required in future problem sets.

R Programming

Problem 1

Do the following in R:

a) Create a variable x that is a sequence from 1 to 1000 (intervals of 1). Then, create a variable y that is linearly dependent on x. Now evaluate their covariance and correlation. Interpret the results. Then, create a second variable y2 that is linearly dependent on x but additionally has some normally distributed error. Now evaluate their covariance and correlation. Interpret the results. Finally, interpret the difference between the covariances and correlations.

Hint: In order to create the randomly distributed error, you have to first create a vector of length 1000 with random draws from a normal distribution, then add this vector to y2.

b) Write a function that returns the correlation between two vectors in R. For this purpose, do not use the covariance (cov) or the correlation (cor) functions that are built into R. Instead, please refer to the lecture and text book for the mathematical definition of covariance and correlation and emulate those calculations in your function. Demonstrate that your function works by comparing it to the built-in cor function in R.

c) Copy the function you created for problem B. Now integrate two error messages. Message 1 should appear if you plug in two vectors of different lengths. Message 2 should appear if you plug in a non-numeric vector. Demonstrate that your function works by, first, plugging in two vectors of different lengths and, second, plugging in a vector consisting of characters.

Problem 2

Do the following in R:

a) Load the *swiss* dataset in R via the command *data(swiss)*. According to the documentation this is data on "Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888".

b) We are interested in how the level of "education" is correlated with other measures. Use the lm command to run a regression of education on all other variables in the dataset. Note that R is case-sensitive, so pay attention to capital letters.

c) Interpret the results. Be specific about the meaning of t-values and p-values. What do your results indicate? What can you say about causality with respect to the statistical relationships identified?

Covariance and Correlation in Mathematics

Problem 3

Do the following problems. Show every step. For all of the following problems, be aware of the following mathematical definitions:

1. $Cov(X, Y) = \mathbb{E}(X * Y) - \mathbb{E}(X) * \mathbb{E}(Y)$
2. $\mathbb{E}(X) = \sum_{i=1}^{n} x * Pr(X = x)$
3. $\mathbb{E}(X * Y) = \sum_{i=1}^{n} x * y * Pr(X = x, Y = y)$

a) This problem is taken from Pitman (1993) Probability: Let (X, Y) have uniform distribution on the four points $(-1, 0), (0, 1), (0, -1), (1, 0)$. Show that X and Y are uncorrelated. Then prove that they are not independent.

b) This problem is taken from Pitman (1993) Probability: Let X have uniform distribution on -1,0,1 and let $Y = X^2$. Are X and Y uncorrelated? Are X and Y independent? Explain carefully.

c) This problem is taken from Pitman (1993) Probability: Let X_1 and X_2 be the numbers on two independent fair six-sided die rolls, $X = X_1 - X_2$ and $Y = X_1 + X_2$. Show that X and Y are uncorrelated, but not independent.

Hint: For this problem, it might be most convenient to use a for loop in R to calculate the covariance. The solution will be based on an R code. It is also possible to calculate this manually though.