Lab 10

Statistical Computing, 36-350

Friday November 13, 2015

Agenda: Practicing split-apply-combine!

(This lab shares a dataset with homework 9, and the description of the data is reproduced here for your convenience.)

Gross domestic product (GDP) is a measure of the total market value of all goods and services produced in a given country in a given year. The percentage growth rate of GDP in year t is

$$100 \times \left(\frac{GDP_{t+1} - GDP_t}{GDP_t} \right) - 100$$

An important claim in economics is that the rate of GDP growth is closely related to the level of government debt, specifically with the ratio of the government's debt to the GDP. The file http://stat.cmu.edu/~ryantibs/statcomp/labs/debt.csv on the class website contains measurements of the GDP growth rate (column name `growth`) and of the debt-to-GDP ratio (column name `ratio`) for twenty countries around the world, from the 1940s to 2010. Note that not every country has data for the same years, and some years in the middle of the period are missing data for some countries but not others. **Throughout, use 3 significant digits for numerical answers!!** (That is, `signif(mydat,3)` is your friend). Also, make sure the package `plyr` is installed and loaded.

```
library(plyr)
debt = read.csv("http://stat.cmu.edu/~ryantibs/statcomp/labs/debt.csv")
```

1. Calculate the average GDP growth rate for each country (averaging over years). This is a classic split/apply/combine problem, and you will use `daply()` to solve it.

    a. Begin by writing a function, `mean.growth()`, that takes a data frame as its argument and returns the mean of the `growth` column of that data frame.

    b. Use `daply()` to apply `mean.growth()` to each country in `debt`. You should not need to use a loop to do this. Don't use something like `mean(debt$growth[debt$Country=="Australia"])`, except to check your work. (The average growth rates for Australia and the Netherlands should be 3.72 and 3.03.) Report the average GDP growth rates clearly.

2. Using the same instructions as problem 1, calculate the average GDP growth rate for each year (now averaging over countries). (The average growth rates for 1972 and 1989 should be 5.63 and 3.19, respectively.) Make a plot of the growth rates (y-axis) versus the year (x-axis). Make sure the axes are labeled appropriately.

3. Recall that the function `cor(x,y)` calculates the correlation coefficient between two vectors `x` and `y`.

    a. First we want to see the correlation between GDP Growth and debt ratio *overall*, in the entire dataset. To this end, compute the correlation coefficient of these two variables over all countries and all years (in other words, no grouping). Your answer should be -0.1995.

    b. Compute the correlation coefficient separately for each country, and plot a histogram of these coefficients (with 10 breaks). The mean of these correlations should be -0.1778. Do not use a loop. (Hint: consider writing a function and then making it an argument to `daply()`). Are there any countries or years where the correlation goes against the general trend?

c. Calculate the correlation coefficient separately for each year, and plot a histogram of these coefficients. The mean of these correlations should be -0.1906. Are there any countries or years where the correlation goes against the general trend?

4. Make a scatter-plot of GDP growth (vertical) against the debt ratio (horizontal). Describe the over-all shape of the point cloud in words. Does its shape match what you expect from problem 3? There should be four countries with a correlation smaller than -0.5. Separately, plot GDP growth versus debt ratio from each of these four countries and put the country names in the titles. This should be four plots. Call `par(mfrow=c(2,2))` before plotting so all four plots will appear in the same figure.

   (Think about what this shows: individual relationships at the country level are sometimes concealed or "smudged out" when data is aggregated over *all* groups (countries). This conveys the importance of careful analysis at a more granular group level, when such groupings are available!)

5. **Bonus (2pt)**: In question 2, a trend certainly seems to exist, but how certain are we of each average we are plotting? In order to portray this, it is useful to plot **error bars** representing how certain/uncertain we are about the quantities we're estimating. Towards this end, also calculate the sample standard deviation ($\hat{\sigma}$) of the growth rates each year, as in problem 3. The standard error of the *average* growth rate should be $\hat{\sigma}/\sqrt{n}$, where n is the number of data points that went into calculating that average. Plot the +-2 standard error bars for each point in the plot from problem 3, using the following code. Make sure none of the error bars are cut off by the margins (i.e. adjust the `ylim` and `xlim` argument in `plot()`).

```
plot.errbar = function(year, mn, std, nsample){
  # Input: takes as input the year, mean (mn), sample standard deviation (std),
  #        and number of data rows in a given year (nsample)
  # Plots a single line that is +-2 standard error bar
  lines(x=rep(year,2), y = mn + 2*std/sqrt(nsample)*c(-1,1), col='blue', lwd=.5)
}
```