R Tasks Instruction

Suanna Oh (sso2118@columbia.edu)

Columbia University

Task 1: Merging data

[Respondent_list.csv] contains ID, Name, and village id of study participants. Another file called [Village_info.csv] contains information about the villages, including district, subcounty, parish and village names as well as an indicator for whether a village is a treatment village or a control village.

1. Read in [Respondent_list.csv] and label all variables
2. Merge it with [Village_info.csv] to see how many participants are in treatment villages, and how many are in control villages
   Hint: check for duplicate IDs. (Fake Name 394)'s ID is actually 1495, not 1491, so use this information to correct the data.
3. In each subcounty, how many people are there and what is the share of people in treatment? Save the relevant information in two new variables, [subcounty_size] and [treatment_share], and add them to the merged data. Save final version of the data as [Respondent_village_data.Rda].

Task 2: Working with time variables

[GDP.csv] [i] contains information on the quarterly US GDP from 1947 to 2016.

1. Import the data into R. Generate a new variable named [time_day] by converting the factor variable [date] into a date variable using as.Date(). Generate another variable named [time_quarter], by extracting the quarter information from [time_day].
2. Create a quarterly time series using ts() with the GDP variable. Split the time series into seasonality, trend, and error using decompose() with additive components, and plot the results.
3. Regress GDP on GDP with 1 quarter lag and GDP with 2 quarter lags.

Task 3: Working with strings and reshaping

[Data_Extract_From_World_Development_Indicators.xlsx] contains data downloaded from World Databank.[ii] It contains three series for years 2010-2014: GDP per capita PPP, Ratio of female to male labor force participation rate, and share of rural population.

1. Create a variable for a shorter version of the [SeriesName]. Specifically, it should only the first part that is outside of the parentheses. In addition, create a variable named [Code] that contains only the third part of the [SeriesCode] (where parts are divided by ".").
2. How many countries contain "St." in their names?

3. Create a long version of the data where each observation is series*country*year. Calculate average yearly growth rate of each series for each country. Save this data as [Dev_indicators.Rda]

Task 4: Using loops and the programming 'if'

Use [Dev_indicators.Rda] from Task 3. We want to calculate the pairwise correlation of yearly growth rates of the three series for each country and print the information.

1. Reshape data so that each observation is country*year and there are three variables for the yearly growth rates of the three series.
2. In a loop that deals with one country at a time, first print the country name. If there are fewer than 4 observations for any of the three growth rates, just print "Missing values" for this country. If all three series have 4 observations, print a pairwise correlation table.

[i] Data downloaded from FRED: https://fred.stlouisfed.org/series/GDP
[ii] Data downloaded from Work Development Indicators: http://databank.worldbank.org/