Stata 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]. Save final version of the data as [Respondent_village_data.dta].

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 Stata. Convert the variable [date] into Stata Internal Forms: generate a variable named [time_day], which gives time in terms of days, and another variable named [time_quarter] that is in terms of quarters.
2. Regress GDP on GDP with 1 quarter lag and GDP with 2 quater lags.
3. De-trend the GDP variable using a Baxter-King filter. Plot the de-trended series starting at 1990q1, and export the graph as a pdf file.

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.dta]

Task 4: Looping and if programming command

Use [Dev_indicators.dta] 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 in a log file.

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 the log file, first print the country name. If there are fewer than 4 observations for any of the three series, print "Missing values". If all series have 4 observations, print a pairwise correlation table.

Task 5: Review

We have data about asset holdings of some farmers in Ghana.[iii] Each farmer is identified with village, hhn, and id. [cash.dta] contains information about the value of their savings, cash, jewelry and cloth. If people have any livestock, information about the type of animal, number of animals, and unit price is contained in a separate data set [animals.dta].

We want to combine these two data sets to create a long data set containing the following variables: village, hhn, id, asset_type, total_value. The unit of observation will be farmer*asset_type, and total_value will contain the total value of the asset. Asset type should include the following: saving (total value saved in bank accounts), cash, jewelry, cloth and animal (at the level of the "animal" variable in [animals.dta]).

Note 1: The final data set should only include the farmers who have some asset holding.

Note 2: Unfortunately many people answered "Don't know" to the value of their assets. We should impute missing values in these cases by calculating the average value for the asset type. For animals, we should calculate the average value for each animal*type. For jewelry, we should impute answers that are "DK" as well as "0".

---

[i] Data downloaded from FRED: https://fred.stlouisfed.org/series/GDP
[ii] Data downloaded from Work Development Indicators: http://databank.worldbank.org/
[iii] Data downloaded from Christopher Udry's website (Round 1 Asset Holdings): http://www.econ.yale.edu/~cru2//ghanadata.html. Data was modified for the purpose of this tutorial.