Principles of Empirical Analysis: Stata Tutorial

Aapo Stenhammar

January 12, 2022

Introduction

Today we will cover some basic commands in Stata

- The content is specifically planned to help you get started in your first problem set
- Code and (raw) data for today is already posted to course homepage

All Aalto students can download the latest version of Stata for free from download.aalto.fi

Stata is widely used by economist and there exists several online guides that you can refer to

- For Stata tutorial by UCLA see here
- For Stata cheat sheet see here
- For coding guide by Matthew Gentzkow and Jesse M. Shapiro see here

Plan for today

1. Downloading a data set to Stata
2. Cleaning a data set
3. Merging data sets
4. Descriptive statistics and graphs

Today our primary data source is employment statistics 115c downloaded from Statistics Finland's PxWeb databases

- First click Population and then Employment. Download only the total numbers and not by sex. You can also find the data set from course home page
- The data set contains the total number of people under different main activities (employed, unemployed, student etc.) by year and age group

Documentation of your data, sources and analysis

An extremely important and topical issue in science, social science and economics included, is the replication crisis

- For more details see for example this article on Science

Objective: document your empirical research in such a way that acknowledgeable researcher can:

1. go to your original data sources
2. recreate your analysis data
3. redo you analysis to reproduce your results, and
4. critically evaluate the decisions you've made going from the raw data to the data used in the analysis and the choices you've made in the analysis

Getting started

Create a new folder on your desktop

Under the main folder create two sub-folders: 1) data 2) output

Download the dataset in csv format and save it to the data folder

Open Stata and open a new do-file by clicking New Do-file Editor on the toolbar

- To execute your code from your do-file: highlight the part of code you want to run and click Execute (do) from the toolbar or Ctrl+D keystroke

Downloading a data set to Stata

Start your do file by creating paths to folders where you have your data and where you want to save your results using command *global*

- After this you can call these folders by adding a dollar sign in front of the folder names you specified

We can bring in the raw data to Stata using command *import delimited $datafoldername\dataname.csv*

- Remember to specify which row contains the variable names *varnames()* and the range of data you want to import *rowrange()*

You can check how the data looks by clicking open the data editor

- First click Data from the top of the window and then choose Data editor

Cleaning a data set (1/2)

You can check what variables your data set contains by typing *describe*. If you want to have a closer look of a particular variable type *codebook varname*

To rename a variable type *rename oldname newname*, to drop a redundant variable type *drop varname* and to drop observations with values over (under) certain limit type *drop if varname > (<) limit*

To change a variable from string format to a numeric format type *destring varname, replace*

Using these commands we convert the variable age from string to numeric and then drop age groups under 15 and over 74 from the data

Cleaning a data set (2/2)

Our aim is to transform the data so that the observational unit (one row) is year - activity type. We do this using *wide to long* and *long to wide* commands

- For a tutorial of these two commands see here and here

Next we aggregate the age groups to 10-year bins (i.e. 15-24, 25-34, ...). To do this we need to use a loop and the egen command

- For a guide on how to loop over a group of values (*forvalues*) see here and for a guide on how to loop over a group of varnames (*foreach*) see here
- For a guide how to use egen command you can type *help egen* to Stata

Merging a data set

Often you want to combine data from multiple sources. Today we want to merge GDP information to our employment statistics. We do this using the command *merge*

For the merge command one needs to specify:

- A common identifier, which in our case is year
- The name and location of the data set we want to merge
- Whether we want to merge many observation to one (m:1), one observation to many (1:m) or one observation to one observation (1:1)

For a guide on how to use the merge command see here

Creating new variables and summary statistics

Now our data set is ready! First let's save it in .dta format using the *save* command and then open a separate do-file in which we do our data analysis

Often times one needs to create some additional variables that combine information from several observations or variables

- We create a variable, which sums over all age groups using the *egen* command and using command *gen* we calculate the yearly change in GDP and in number of employed people

To find out the mean, min and max of yearly changes in GDP and number of employed people we type *sum change_in_GDP change_in_employed*

To find out the mean number of population under different types of main activities we type *tabstat population, by(main_activity)*

Graphs

Visualizing data is often a key part of empirical analysis

- For a guide on how to visualize data see this paper in Journal of Economic Perspectives by Jonathan Schwabish

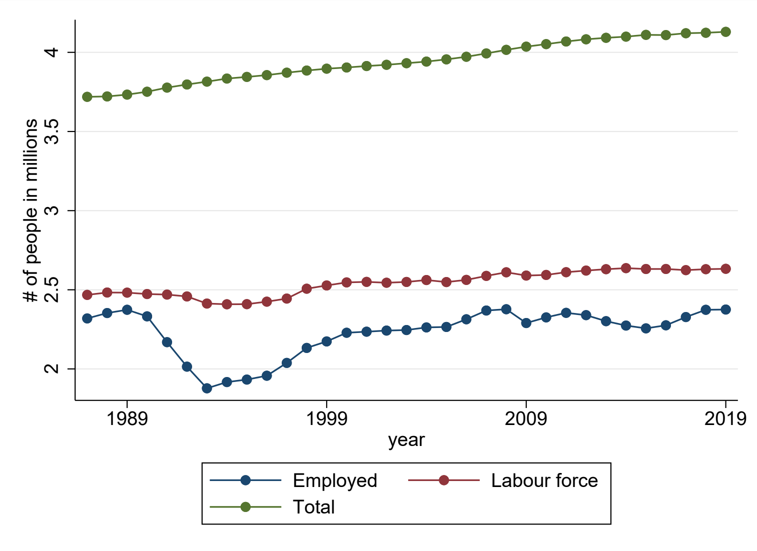
We can create different kinds of graphs in Stata using the *twoway* command

We plot the development of labour force and employed population over time using *twoway connected*

We plot the statistical relationship between change in employment and change in GDP using *twoway scatter* and fit a linear regression line in the same graph using *twoway lfit*

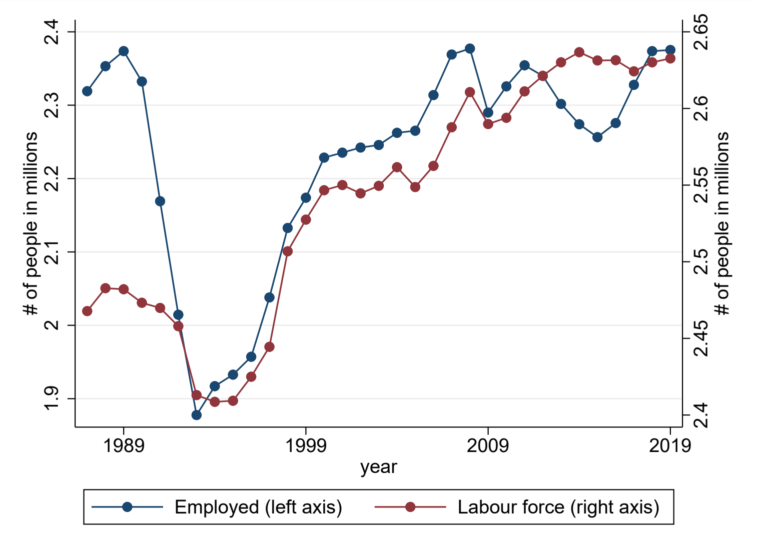
You can add a second yaxis using command *, yaxis(2)*

Graph 1

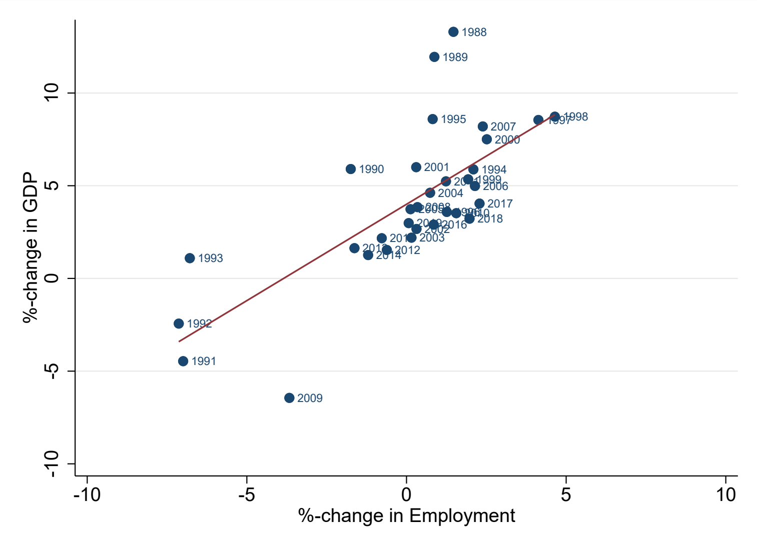


# of people in millions
2
2.5
3
3.5
4
1989
1999
2009
2019
year
Employed
Labour force
Total

Graph 2


2.4
2.3
2.2
2.1
2
1.9
# of people in millions
2.65
2.6
2.55
2.5
2.45
2.4
# of people in millions
1989
1999
2009
2019
year
Employed (left axis)
Labour force (right axis)

Graph 3


%-change in GDP
%-change in Employment
-10
-5
0
5
10
-10
-5
0
5
10
1988
1989
1995
2007
1997
1998
2000
1990
2001
1994
1999
2006
2004
2017
2008
2005
2010
2018
2016
2002
2003
2013
2012
2014
1993
1992
1991
2009