

Economics 253: Introduction to Econometrics
Sebastian Anti, Assistant Professor of Economics
PROBLEM SET 1

GENERAL INSTRUCTIONS:

You may only work with members of your group in completing this assignment, although you are free to consult online and print resources as much as you wish.

As you work through the questions continually check your work. If something doesn't seem to make sense try your best to work out if something went wrong and correct it. Give yourself a lot of time to work through these problem sets in case you get stuck and need to see me in office hours to work through sticking points.

When you submit your work on Moodle, you must submit your write-up in .docx or .pdf format, as well as your .do file.

See the Moodle posting for the due date for this assignment.

PROBLEM SET 1: STATA, MANAGING DATA, CLEANING DATA, DATA VISUALIZATION, AND DESCRIPTIVE STATISTICS

SECTION 1: GETTING STARTED WITH STATA

1.1 Accessing Stata

1. Access Stata remotely on a Bryn Mawr College computer lab computer. You can do this using LabStats. Do this click on the following URL and follow the instructions: <https://techdocs.blogs.brynmawr.edu/9708>
2. Open Stata

1.2 Importing Data into Stata

3. Create a folder on your computer called "Problem Set 1"
4. Download the data file PS1_Data.dta from Moodle Site to the "Problem Set 1" folder
5. Change Stata's working directory to the "Problem Set 1" folder. You can do this by going to **File >> Change working directory... >> Select Problem Set 1 Folder**.
6. Import PS1_Data.dta into Stata. This is easy. You can just double click on the data file. Or, just go to **File >> Import >> PS1_Data.dta**

SECTION 2: ABOUT THE DATA

2.1 Context of the Dataset

Northern Uganda has experienced civil war since 1986 when the Lord's Resistance Army (LRA) rebelled against the government in Kampala. A distinctive phenomenon of the civil war in northern Uganda was

the use of child soldiers in the LRA. Researchers estimate that the LRA abducted between 60,000 and 80,000 young people and forced them to fight for them in the civil war. These young people were typically abducted by roving groups of rebels during night raids on rural homes (Blattman and Annan, 2010).

Northern Uganda is a highly rural area of the country. The most common economic activity for people in Uganda is smallholder, often subsistence, farming. Keep this in mind as you work with the data.

2.2 Dataset Characteristics

The dataset you are using for this problem set is a paired-down version of the first round of the Survey of War-Affected Youth (SWAY) used by Jeannie Annan of the International Rescue Committee, Chris Blattman at the University of Chicago, and Dyan Mazurana and Khristopher Carlson both at Tufts University, as the basis for a number of papers published between 2008 and 2013 studying the effects of child soldiering and experience of civil war on various economic outcomes.

This component of the SWAY data was collected in 2005 and 2006 of men (the more common target of LRA abduction) born between 1975 and 1991, and therefore between 20 and 5 years old when child soldiering became a prominent tactic of the LRA. The sample is representative from eight counties in the districts of Kitgum and Pader (see map). The respondents in the survey provided information on their war experiences as well as current well-being and life circumstances.

FIGURE 1.—MAP OF UGANDA AND SURVEY FIELD SITE



2.3 Citation and Other Information

For more information on the dataset as well as historical context consult the following source:

Blattman, Christopher, and Jeannie Annan, 2010. "The Consequences of Child Soldiering." *The Review of Economics and Statistics* 92 (4): 882-898.

SECTION 3: EXPLORING THE DATA, DATA CLEANING AND ORGANIZATION

Use the `desc` command to see all the variables and their definitions.

7. Which variable contains the Individual ID number? (2 points)
8. Which variable identifies the number of years of education for each individual? (2 points)
9. How many total individuals are in the dataset? Cite the part of the output that tells you this. (2 points)
10. How many variables does the dataset contain? (2 points)
11. Which variable identifies whether the individual was ever abducted by the LRA or not? What value for this variable identifies that the individual was abducted? What value for this variable identifies that the individual was not abducted? (3 points)
12. What is the `distress` variable measuring? What is its scale? (2 points)

To see the data in spreadsheet form, use the "browse" command. It is helpful because often you need to inspect the data visually to get a sense of how it is organized.

13. What is the value that Stata uses when there is no data in a cell? (2 points)
 - Hint: Look at the value for one's wage when that person is not employed. Since they aren't employed, the entry for their wage has to take on a missing value.
 - Note: Stata drops observations with this value from any calculations you ask it to do if its missing value is involved in the calculation.

You can use Stata to calculate new variables based on the other variables in the dataset. The command for this is `gen`. The syntax for `gen` is as follows: `gen newvar = exp`, where `newvar` is the name of the new variable you want to calculate. And `exp` is the mathematical expression for this new variable.

14. Often children were abducted by the LRA multiple times (up to three in the data). The SWAY data collects the total amount of time for each abduction (in months). Use the `gen` command to calculate the total amount of time (in months) each observation experienced abduction by the LRA. Name this new variable `total_time_abd`. (5 points)

SECTION 4: CALCULATING DESCRIPTIVE STATISTICS

Let's get a sense of the shape of the SWAY data. Of course a dataset like this is way too big to get a good sense of it by visually inspecting it with the `browse` command. So we need to summarize it using statistics.

15. Use the `sum` command to calculate the means, standard deviations and ranges for the variables `total_time_abd`, `abd`, `educ`, `distress`, `age`, `abd_age`, and `wage`. Include this table in your write up. (5 points)
16. Of the information provided by the `sum` command, which is/are (a) measure(s) of central tendency and which is/are (a) measure(s) of dispersion? (3 points)
17. How many missing values are there in the `wage` variable? Explain how you figured out the answer to this. (5 points)
 - Note: Stata drops observations with missing values in the variable that it is using in its calculations.
18. Why are there so many missing values in the `wage` variable? (3 points)

You can use restrictions to ask Stata to do calculations for subsamples within the sample. For example, you might want to calculate summary statistics for just the women in your sample, or just those living in a particular region. Restrictions are added at the end of a command in the form of an `if` statement. For example, `sum var1 var2 if N==T` will produce a summary table for just those in the sample where `N` is exactly equal to `T`.

19. Use this formulation to produce the summary statistics for the variables `total_time_abd`, `abd`, `educ`, `distress`, `age`, `abd_age`, and `wage` for only those in the sample that were abducted by the LRA. Include the resulting table in your write up. (5 points)

If you want more detail on the distribution than what `sum` offers, you can add the option `, detail` to the end of the `sum` command. This provides you with the values in the data that correspond with various percentile cutoffs in the variable.

- Note that if you want to use a restriction with the `, detail` option, the `if` statement comes before the comma in the syntax.
20. Produce the output of this command for the `total_time_abd` variable. (5 points)
 21. What is the median for the `total_time_abd` variable? (3 points)

The `tab` command allows you to see the number of instances that observations take on certain values within a variable. It is particularly useful to get a sense of the representations of different groups in the dataset. For instance, `camp` is a categorical variable. It tracks the refugee camp in which the respondent was surveyed.

22. Enter `tab camp` and examine the table that Stata returns. Include this table in your write up. The "Freq." column shows how many respondents are members of a particular category. Which camp is the most represented in the data? Which is the least represented? (3 points)
 - Tip: Avoid the mistake of using `tab` with a continuous variable. The output will be completely uninformative and may take Stata a long time to provide it.

Histograms provide a useful way to get a sense of a random variable's probability density function. You can do this easily in Stata with the `hist` command. The syntax for the `hist` command is `hist varX`, where `varX` is the variable whose distributions you want to look at.

23. Use the `hist` command to generate the histogram for `abd_age`. Include the histogram in your write up. (5 points)

You can adjust the presentation of the histogram by adjusting the “bin”. A higher bin will make the histogram more jagged while a smaller bin will make the histogram smoother and chunkier. If left unspecified like in question 23, Stata selects a bin for you based on what it thinks is optimal.

24. The command `hist abd_age, bin(10)` will tell Stata to display a histogram of the variable `abd_age` with a bin of 10. Enter this in and report it in your write up. Also, try a bin width of 50 and report it in your write up. Compare the two. Which do you prefer? Why? (10 points)

Stata can provide a visual estimate of the variable’s PDF by adding to the end of the `hist` command the following option: “, `frequency kdensity`”. You can specify the bin size as you wish or let Stata do this for you. It’s up to you.

25. Use this to attain the PDF for the variable `abd_age`. Include it in your write up. Describe the PDF. Is it normal? Right skewed? Left skewed? If there is a skew to the data explain what this means and provide a reason for it. (10 points)
26. Most commands in Stata accept the `if` statement we practiced earlier. Use an `if` restriction to attain the histogram for variable `abd_age` and its PDF for just the respondents in the sample from the “Pader” camp. Include it in your write up. (5 points)
- Note: While the values for the `camp` variable display as text, the blue color indicates that Stata contains them as a numeric value. The numeric value for the Pader camp is 5.

Now, let’s start to get a sense of how variables in the data are related to one another. For this we’re going to use the `scatter` command. The core syntax for the `scatter` command is simple. It is just `scatter varY varX`. `varY` is always variable you want on the y axis, `varX` is always the variable you want on the x axis.

27. Blattman and Annan (2010) are interested in the relationships between experiences being abducted by the LRA and certain economic and health outcomes. Let’s think about the relationship between time abducted by the LRA and current emotional distress. Before looking at the data, what does your intuition tell you the relationship would be? Explain. (10 points)
28. Use the `scatter` command to produce the scatter plot of `total_time_abd` and `distress` for those in the sample that were abducted (where `abd==1`). Include this chart in your write up. Be sure to put the `distress` variable on the y axis and `total_time_abd` on the x axis. Describe this relationship. Is it positive or negative, or is it hard to tell? Is the correlation “tight” or “loose”? (10 points)

We can use the `twoway` command, the `scatter` command and the `lfit` command together to project the line of best fit over this scatter plot to see if the data actually do indicate a bivariate relationship that might be hard to discern from the simple scatter plot. The `twoway` command allows you to plot both the line of best fit and the scatter plot in the same chart. The `lfit` command allows you to plot the line of best fit and its syntax is the same as `scatter`. The syntax for using this all together is `twoway (scatter varY varX) (lfit varY varX)`.

29. Use `twoway`, `scatter` and `lfit` to add the line of best fit to the scatter plot you made for question 28. Include this chart in your write up. (5 points)

- Note: Don't forget the restriction to just those who were abducted. The `if` statement goes inside the parentheses for both the `scatter` command and the `lfit` command.

30. Does this chart support or call into question your hypothesized relationship between `distress` and `total_time_abd`? What does this scatter plot have to say, if anything, about the causal relationship between `total_time_abd` and `distress`? If this is not evidence of a causal relationship, Explain. (10 points)

A great thing about Stata is that it allows the general public to write useful commands that the Stata programmers in College Station, TX haven't thought of or (more likely) haven't had the time to do. These user-written commands are hosted in a database easily accessible to any Stata user. We're going to use one of these to produce a joint probability density function for the variables `total_time_abd` and `distress`. The way to install a user-written command is by entering in `ssc install ProgramName` where `ProgramName` is the name of the program we want to use. In this case we want to work with the program called "tddens".

31. Install `tddens` into Stata.

32. The syntax for `tddens` to produce a joint PDF is simple. It is just `tddens var1 var2, sgraph`. Use this to attain and report the joint probability density function for the variables `total_time_abd` and `distress` for respondents that were abducted (the `if` statement comes before the comma). Include the resulting joint PDFs in your write up. (10 points for 31 and 32 combined)

- Sometimes `tddens` produces an error on this question due to a bug. The fix is to define a new variable that is identical to `distress`. You can call it `distress_2`. You can then just run the `tddens` command using this new variable.

33. What information does the joint PDF provide that the scatter plot does not? (5 points)