Introduction to Statistics
homework 1

Johan A. Elkink
jos.elkink@ucd.ie

Due 5 October 2016

You will submit two files: one PDF file[1] *including all plots, tables and interpretations and one command file (SPSS Syntax file (.sps), or Stata do-file, or R-file) with all commands used to answer the exercise and no superfluous commands. Please send both files to* `jos.elkink@ucd.ie`.

(5%) of the grade is used for an overall evaluation of the presentation of your work (the PDF file) and (5%) of the grade for the evaluation of the clarity / presentation of your command file, including the use of comments, clear variable names, and whitespace.

Data

The data set we will use for this homework is based on the replication data for Graham, Gartzke and Fariss (2015*a*), which is available through the Harvard Dataverse Network. The first question will guide you through the process of acquiring the data.

Questions

1. Dataverse is a system, with the most commonly used installation at Harvard, for cataloguing replication data, including tools for basic analysis. We will ignore the latter, but use this to download the data for this homework.
   (a) Find the replication data at `http://thedata.harvard.edu/` (Graham, Gartzke and Fariss 2015*b*).
      i. Under "Search Studies", search for the title of the article, then click on the link to the replication data.
      ii. Read the description and scope on the first page.

[1] Word files will be sent back—note that newer versions of Word can easily save to PDF format.

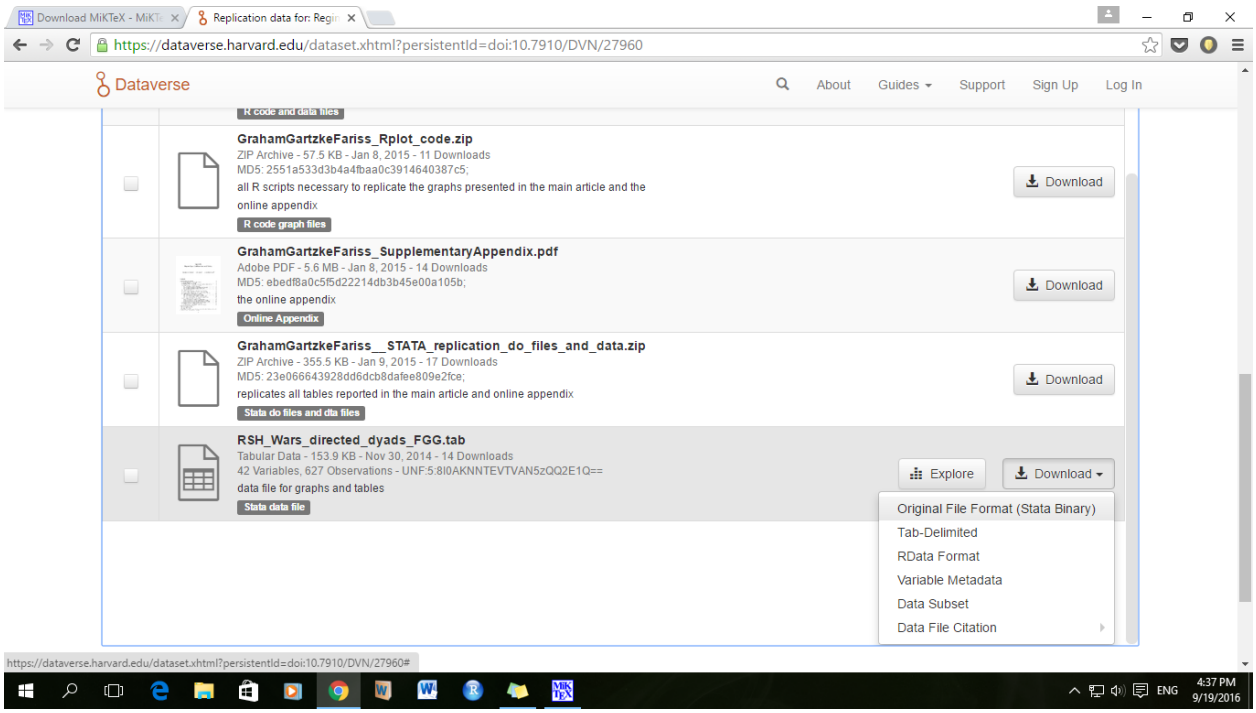

Figure 1: Screen shot of downloading a data set in Stata format from the Harvard Dataverse.

iii. Click on the "Data & Analysis" tab, then "Download as" and select the Stata binary (see also Figure 1).

(b) (5%) Open the data and make sure the command for doing so is in the command file.

Note that in SPSS, the variable labels will be automatically visible, but you might have to adjust measurement levels if they are inappropriately set, for those variables that you use in this homework. Make sure to include this in the SPSS Syntax file when you do so.

(c) (5%) Select only the observations after 1900 from this data set (i.e. the variable **year** should have values greater than 1900)—see "Subsetting data" in Lab 3 for instructions. For the remainder of the homework, use only this data after the year 1900.

2. In this question we will look at the prevalence of democracy in the countries in the dataset and the characteristics of wars waged by these countries.

(a) (5%) Produce a histogram of the variable **larger_war~id**, which is the variable with a distinct code for each war. Explain what you can conclude from the graph. Can you identify the wars with larger coalitions?

(b) (4%) Produce a histogram of the variable **noside1** which is a variable for the size of coalition. Give your comments on what you can learn from this graph?

(c) (4%) Produce a pie chart of the variable **democracy**. Give your comments. Do you

think democracies wage more wars?

(d) (5%) Produce a cross table of whether the country is an initiator of the war (use variable **initiator** or not and prevelance of democracy in them (use variable **democracy** here). Calculate the percentages. Give your comments this.

3. The variable **log_noside1** is the log of the variable for the coalition size **noside1**.

   (a) (4%) Produce a boxplot of the **noside1** variable and interpret the distribution of the variable.

   (b) (4%) Calculate the mean, median, variance and standard deviation of the **noside1** variable.

   (c) The mean is quite different from the median, while both are measures of centrality of the variable. (5%) Discuss why they are different and what this difference implies about the distribution of the variable.

   (d) (4%) Inspect the variable **log_noside1** and recode it to new dummy variable **coalition**, where the value is 1 if observations are above 0 for **log_noside1** or 0 otherwise.

   (e) (4%) Produce a crosstable of the dummy variable **coalition** and the variable **democracy**. What can you conclude? Can you observe any relationship between being a democracy and forming a coalition?

4. We will now look at the relation between military expenditure (variable **milex** and the democracy scale (given by the variable **polity**

   (a) (4%) Produce a scatterplot of **polity** by **milexl**.

   (b) (5%) Compute a new variable **logmil** that will contain the logarithmic transformation of the **milex** variable—see "Transforming the variable" in Lab 2 for an example.

   (c) (4%) Compute the mean and median of the **logmil** variable. What do you conclude about the overall distribution of the variable on the basis of comparing those two numbers?

   (d) (5%) Produce two box-plots, side-by-side, of the **logmil** variable by the two categories of the **democracy** variable.

   (e) (4%) Produce a scatterplot of **polity** by **logmil**.

   (f) (4%) Calculate the covariance and correlation between **polity** and **logmil**.

5. (15%) In approximately 300–400 words, based on the above results (questions 3 and 4, primarily), discuss your interpretation of the relationship between military expenditure and democracy.

References

Graham, Benjamin A.T., Erik Gartzke and Christopher J. Fariss. 2015*a*. "The Bar Fight Theory of International Conflict: Regime Type, Coalition Size, and Victory." *Political Science Research and Methods* pp. 1–27.

Graham, Benjamin A.T., Erik Gartzke and Christopher J. Fariss. 2015*b*. "Replication data for: Regime Type, Coalition Size, and Victory.". *http://dx.doi.org/10.7910/DVN/27960*

Grade conversion scheme

Homeworks	UCD	MDP	Homeworks	UCD	MDP
97-100%	A+	78.33	74-76%	C	51.67
94-96%	A	75.00	71-73%	D+	48.33
91-93%	A-	71.67	68-70%	D	45.00
88-90%	B+	68.33	65-67%	D-	41.67
85-87%	B	65.00	54-64%	E+	38.33
83-84%	B-	61.67	44-53%	E	35.00
80-82%	C+	58.33	33-43%	E-	31.67
77-79%	C	55.00	0-32%	F	25.00

Note that the percentage scores will be translated to UCD grades before entering on the system. Overall module grade will be calculated by the system based on the UCD grades. For MDP students, grades will then be translated to TCD marks. Note that TCD marks are *not* percentages and will therefore reflect the above scale. Thus, a 95% score on all your homeworks will generate a A grade on the UCD system, and a 75 mark on the TCD system.