

Empirical exercise

This empirical exercise must be submitted on Canvas on **Friday 17 December by 2:00pm** (late submission costs you marks). In answering the questions, make sure to include the graphs and tables where appropriate. For each question, write down (at the beginning or at the end of the question) the Stata commands you have used to answer the respective question. Alternatively, you can take a screenshot of the do file that you have created to carry out the empirical exercise and paste it at the end of the document.

Exercise

Use the data set **POLI347_dataset_21_22.dta**.

Load up the data set and open a do-file for you to write the commands in.

1. How many observations does the dataset contain? What are these observations? **(5 points)**
2. The dataset contains a variable called `dpi_mdmh`. What does this variable measure? Does the variable contain missing values? If yes, how many observations are missing? Which observations are missing? **(5 points)**
3. The dataset contains a variable called `e1_gunn2`. What does the variable measure? Does the variable contains missing values? If yes, how many observations are missing? Which observations are missing? What are the maximum and minimum values this variable takes? What is median? And the mean? What is standard deviation? And the variance? **(5 points)**
4. Get summary statistics of the variable GDP (`gdp_log`) and GDP per capita (`gdppc_log`). How can we tell the level of dispersion of these two variables? Which of the two variables is less dispersed? **(5 points)**
5. Tabulate the variables `ciri_wecon` and `fh_cl`. How many observations take the value 1 in both scales? What are those observations? Which observation takes the value 1 in the scale of wome's economic right and 2 in the civil liberties scale? **(5 points)**
6. Looking at the variable `afghanistan`, what does this variable measures? Does the variable contains missing values? What is the the maximum and minimum value this variable takes? What is the mode? What is median? What is the mean? What is standard deviation? Using the same variable, (`afghanistan`), produce a kernel density function. Copy the graph and paste it in the document. Is the variable normally distributed? If not, what are the reasons the distribution does not look like a normal? **(5 points)**
7. The variable `fh_cl` is a categorical variable measuring civil liberties on a scale that goes from 1 (high levels of civil liberties) to 3 (low levels of civil liberties). Generate a bar

chart of this variable. Copy the graph and paste it in the document. Which country or countries had low levels of civil liberties? **(5 points)**

8. Using the variable `gle_rgdp`, first describe what the variable measures then create a histogram with bins sized 2000 each and apply also a smoother (a kernel density). Copy the graph and paste it in the document. How is the variable distributed? Using the same variable (`gle_rgdp`), now create a box plot for EU and non EU member countries. Copy the box plot and paste it in the document. Which category contains outliers? What are these outliers? **(5 points)**
9. The variable `al_ethnic` is an index of ethnic fractionalization that in our sample ranges from .0467574 to .71242. Using this variable, generate a new variable called `al_ethnic2` that takes the following values: 0 for any values of `al_ethnic` below or equal to the first quartile; 1 for values any of `al_ethnic` strictly above the first quatrile and below or equal to the median; 2 for any values of `al_ethnic` that is strictly above the median and below or equal to the third quartile; 3 for any values of `al_ethnic` that is strictly above the third quartile. Assign to the newly created variable the label **Quartiles of ethnic fractionalization**. Assign the labels also to the values of this variable such that: **0** stands for *first quartile*; **1** for *second quartile*; **2** *third quartile*; **3** *fourth quartile*. Tab the variable after adding the labels. Copy the table and paste it in the document.**(5 points)**
10. Produce a box plot of the variable `gdppc` over the newly generated `al_ethnic2`. Is there any category containing outliers? Which observation is the outlier? **(5 points)**
11. Look at the variables listed in the table below. What is the type of each variable and what are their level of measurement, the informative measure(s) of central tendency and the actual value of these measures? **(5 points)**

Variable name	Variable type	Measurement level	Meas. of central tendency	Value of central tendency measure
islamic_tot				
alesina_ethnic_fract				
gd_ptsa				
ciri_physint				
chga_hinst				
uw_gini				

EXAMPLE:

Variable name	Variable type	Measurement level	Meas. of central tendency	Value of central tendency measure
sumimm	Discrete	interval	Mode-Mean-Median	Mode:0 Mean:3.1 Median:3

12. The variable `sumsus` measures the number of counterterrorist regulations that restrict the rights of suspects of terror. What is the mean of this variable in 2000? And in 2007? Which country had the highest number of regulations targeting suspects in 2000? And in 2007? **(5 points)**
13. Using the variable `sumimm` in the year 2000, generate a dummy variable called `sumim2` that takes the value 0 for any value below or equal to 1 and the value 1 for any value strictly above 1. Label the newly generated variable “Dummy of regulations targeting immigrants in 2000”. Assign also the labels to the values of this dummy such that 0 stands for low regulations and 1 stands for high regulations. Generate a bar chart of the newly generated dummy `sumim2` making sure to restrict the graph for the year 2000. Copy the bar chart and paste it into the document. Which countries are classified in the category “high regulations”? **(5 points)**
14. Generate a scatter plot of the percentage of internet users (`internet`) in the y-axis and Average Schooling Years (Total) (`b1_asyt15`) in the x-axis. How does the relationship between these two variables looks like? What is the strength and direction of this relationship? Which observation has a percentage of internet user below 20 and an average schooling years below 7? Now generate a scatter plot with country code labels only for the observations that have a percentage of internet users below 20 or above 40. Copy the graph and paste it into the document. Now compare the means of the percentage of internet users in EU and non EU countries. Which region has a higher percentage of internet users on average? **(5 points)**
15. Using the variable `wbgi_rle`, generate a dummy that takes the value 0 for observations below or equal to the mean and 1 for observations above the mean. Call the newly generated dummy `wbgi_rle2`. Plot the dummy with a bar chart. Copy the graph and paste it into the document. Now generate a kernel density function of the original variable `wbgi_rle`. Copy the graph and paste it into the document. What do the graphs have in common? **(5 points)**
16. Use the `ttest` command with the `unequal` option to test if the mean of life expectancy at birth (`wdi_lidexp`) in EU countries (EU) equals the mean of life expectancy at birth in non EU countries. The alternative hypothesis is that the two means are different. Can we reject the null hypothesis that the difference in the two means is zero at any conventional significance level? Now perform the same analysis controlling for Scandinavian countries (`Scandinavia`). Can we reject the null hypothesis that within Scandinavian countries the life expectancy at birth in EU countries equals the life expectancy at birth in non EU countries at 10%, 5% and 1% significance levels? **(5 points)**
17. A researcher wants to know if the mean of terrorist incidents (measured by the variable `incidents`) in countries with laws that ban support and glorification of terrorism (measured by the variable `suppter`) is the same as the mean of terrorist attacks in countries that do not have such laws. The alternative hypothesis is that the two means are different. Using the appropriate command, perform this analysis. Can we reject the null hypothesis that the two means are the same at 1%, 5% and 10% significance level? Now check if within European Union countries (EU), the means of terrorist incidents in countries that had had laws banning glorification and support of terrorism and those that did not are equal (again, the alternative hypothesis is that the two means are different). Can we reject the null hypothesis that within the EU the mean of terrorist incidents of countries

with and without laws banning glorification and support of terrorism are the same at any conventional statistical level? **(5 points)**

18. Regress the total number of counterterrorist regulations targeting immigrants (`sumimm`) on the percentage of cabinet posts held by right-wing parties (as measured by the variable `gov_right1`).¹ What is the value of the constant? Is it statistically significant? How many anti-terrorist regulations would a country have if the percentage of right wing posts in government were to be zero? How many additional anti-terrorist regulations are associated with an increase of right wing posts by 1? What is the predicted number of anti-terrorist regulations of a country with a mean level of right wing posts in government? With the maximum level of right wing posts? **(5 points)**
19. Now add to the first regression in question 18 a variable measuring religious fractionalization (`al_religion`). Interpret the results of this multivariate regression. Now add to this last regression a variable measuring the rule of law (`fh_rol`). Discuss the results of this last regression. Pay special attention to the difference between the results of this last regression and the second regression (the total number of regulations against immigrants regressed on religious fractionalization). **(5 points)**
20. Now regress the total number of counterterrorist regulations targeting immigrants (`sumimm`) on the total number of people killed in terrorist events (`total`), the percentage of cabinet posts held by right-wing parties (`gov_right1`), the deployment of troops in Afghanistan (`afghanistan`), ethnic fractionalization (as measured by the variable `al_ethnic`) and GDP per capita (`gdppc`). Interpret the effect of `gov_right1` on `sumimm` as causation and as correlations. Explain why it would be misguided to interpret causally the effect of right wing parties in government. **(5 points)**

¹Recall that you need to use the command `reg dependent var independent var` to regress a dependent variable on an independent variable.