**POL 454 Data Introduction and Exercises**

The following provides a list of commands that will be useful in exploring and analyzing data relevant to international politics using a statistical software package called Stata. While the written commands are listed in the examples below, Stata also provides a drop-down menu interface where many of these commands can be found, much like a typical Microsoft office program. Before getting started, be sure to familiarize yourself with the program. If you have trouble trying to determine the appropriate command, use Stata’s help feature by typing in the command window:

help *topic*

The help command opens Stata’s help interface which includes commands, explanations, and examples. To begin, open Stata and load the “volgy 454.dta” file for the course. You’ll see a list of 19 variables that are also provided in the excel workbook for the class. These variables are organized in the form of “country-year” data, meaning that each line item represents one country in one year (i.e. USA in 1991, USA in 1992, Canada in 2003, etc.). As a result, use of this data is limited to hypotheses that are either specifically interested in the behaviors and characteristics of states or aggregate changes over time in the international system. The appropriate data are dependent upon the type of research question you seek to explore. While the data provided to you is country-year, the end of this workbook contains a brief explanation on how you can make your own data using an online program called “EUGene”, which allows you to create datasets in a variety of formats.

**Exploring Your Data**

One of the first things you should do when exploring new data is develop an understanding of each of your variables. How is it measured? What years and countries does it cover? What is the average and distribution of the data? Do values of data change over time or across country? To explore the data, we have at our disposal the “*summarize*” command. As an example, let’s take a look at Polity IV scores.

 . summarize polity

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 polity | 4685 1.586126 7.242395 -10 10

This simple command immediately tells us some interesting facts about the polity variable. We have annual observations for each country, which should total to 6169 total observations in the data. However, there are only 4685 observations for polity scores, therefore we have 1484 missing data points. First, let’s check if they’re missing by year, summarizing by year if polity does not equal (~=) missing (.)

. su year if polity~=.

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 year | 4685 1995.555 8.846772 1980 2010

For observations where polity scores are not missing, we cover the entire 1980 to 2010 time span, meaning all years have observations for polity scores. This means that certain countries appear to be missing. We can determine which countries by typing:

tab ccode if polity~=.

[output omitted]

which will give us a table of all countries, and the number of times that country lists a polity score. The table shows us that missing states include those that did not exist at the beginning of the data, like Croatia, Moldova, or Montenegro.

Also important, the summarize function gives us the mean, maximum, minimum, and distribution for the polity scores. The polity score is an aggregate measure of the degree to which a country is autocratic (negative values) and democratic (positive values). A country that is a full democracy would receive a positive 10 value, while a fully autocratic state receives a -10. We see that there exists in the data relative balance between the two extremes, with a mean value of 1.58. The standard deviation represents a measure of dispersion in the data, or the degree to which values are spread out across the range of possibilities. For polity scores, ranging between -10 and 10, we have a very large standard deviation of 7.24. This means that 67% of Polity IV scores will fall within one standard deviation above or one standard deviation below the mean. So most polity scores range from -5 to 8. This large distribution is likely due to a collection of observations at the extremes (10 or-10). To get a better idea, we can graph the data.

. histogram polity, discrete

(start=-10, width=1)



The output graph shows, as we might expect, that full democracies represent the plurality of observations. Note that around 1 and 2 there aren’t many observations available. Given the distribution of the data toward the extremes, the mean may not be very useful beyond telling us that more states are above 0 (more democratic) than below (more autocratic).

We can also generate interesting descriptive graphs to illustrate how a variable might change over time.

graph dot (mean) polity, over(year, descending)

where “graph dot (mean)” means we want to make a graph dot of average values for “polity”. We then specify that we want these averages grouped by year, and want them to be sorted in descending order on the y-axis : “over(year, descending)”



We see from the figure that the average polity score changes dramatically over time, with the greatest shifts occurring between 1989 and 1991 at the end of the Cold War. In fact, prior to 1989, the average polity score for countries was below 0, meaning countries on average were more autocratic than they were democratic.

**Creating Variables**

One thing we typically want to do in testing our theories is create variables to answer potentially interesting questions. For example, we may seek to explain the level of hostility states engage in annually as a function of their militarism. To measure militarism, we would probably be interested in finding a variable that captures their military spending as a percentage of total GDP. We have military expenditures annually, and we have GDP, so we need to combine the two variables so we can get the percentage. First, our military expenditures data is listed in millions, so we need to divide our world bank GDP measure by one million so that the two variables are similarly scaled:

. gen gdp = wbcons/1000000

. gen militarism = milex/gdp

We now have our variable of militarism for our research project.[[1]](#footnote-1) We can then compare countries before running a model using the summarize command. For example, for the United States:

. su militarism if ccode==2

 Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

 militarism | 23 .0526404 .0119258 .0378925 .0793569

**Analyzing the Data**

The simplest form of analysis we can conduct is that of a **basic correlation** to determine whether a statistical relationship exists between two variables. To determine the correlation between two variables we can use the “corr” command in Stata. So, seeking to test the obvious, we decide to explore whether there is a relationship between the size of a country’s economy and the amount of military spending:

. pwcorr milex gdp, sig

 | milex gdp

-------------+------------------

 milex | 1.0000

 |

 |

 gdp | 0.9089 1.0000

 | 0.0000

 |

The relationship between GDP and military expenditures is .9089. So what does this mean? First, the number is positive, meaning that as GDP increases, so too does military expenditures. If the number were negatively signed, we would have an inverse relationship between our two variables. In this case, the correlation serves to ask the question “How good a predictor is GDP of military spending?” A 1.0 would imply a perfect positive relationship, while -1.0 is a perfect negative correlation. Such perfect correlations rarely exist in the social sciences. As a result, any relationship above .6 or below -.6 can be considered strong. To test the significance of the correlation we’ve used “pwcorr” followed by the option “, sig”. he “0.0000” underneath the correlation coefficient “.9089” is the level of significance. Anything less than .05 basically means the relationship is statistically significant 95% of the time, and so we have a meaningful statistical relationship between the two variables. To test this relationship, we’ll want to use **ordinary least squares regression** to determine if the correlation between these two variables is strong enough that we can reject the null hypothesis, or the default position of any testing, that no relationship exists. Likewise, we can determine approximate affect of one variable on another. In this case, we are attempting to explain military expenditures, which is our *dependent* variable. The variable(s) we intend to use to explain our dependent variable are independent variables (in this case, GDP). To run a simple regression we use the regress command:

. regress milex gdp

 Source | SS df MS Number of obs = 3133

-------------+------------------------------ F( 1, 3131) =14878.28

 Model | 4.8848e+12 1 4.8848e+12 Prob > F = 0.0000

 Residual | 1.0280e+12 3131 328319083 R-squared = 0.8261

-------------+------------------------------ Adj R-squared = 0.8261

 Total | 5.9128e+12 3132 1.8879e+09 Root MSE = 18120

------------------------------------------------------------------------------

 milex | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 gdp | .0423701 .0003474 121.98 0.000 .041689 .0430511

 \_cons | -993.5299 332.8934 -2.98 0.003 -1646.241 -340.8184

------------------------------------------------------------------------------

Where “regress” is the regression command in Stata, the first variable listed is our dependent variable, or what we are seeking to explain, and the variables that follow are what we are using to explain the relationship. There are a handful of statistics we can look at to determine whether or not there is a significant relationship between the two variables. First, the t-score (column “t”) represents the level of significance of the relationship, or the degree to which the relationship is significant given the number of observations available. Generally, anything above 1.96 is statistically significant. “P>|t|” is simply the probability that the null hypothesis (that there is no relationship) is upheld. The typical standard in econometrics is a 95% confidence level, so you would like your P<|t| value to be less than or equal to .05. The 95% confidence simply means in 95% of cases, it can be expected that the relationship between military expenditures and gdp will fall in the range of the coefficient (“Coef.” Column), plus or minus the standard error (“Std. Err.” Column). This distribution of possible outcomes is provided to us in the two columns under “[95% Conf. Interval]” column. So long as the range of values in the 95% confidence interval does not include 0, we can reject the null hypothesis that no relationship exists. Rejecting the null means that a relationship does exist, and we have a variable that explains something about the dependent variable. However, we don’t know how much it explains. The last variable of significance asks not just if there is a relationship, but how much of the variance in the dependent variable does our independent variable actually explain. In this case, the value is .8261, which we can interpret as saying that 82% of the variance in military expenditures in our data is explained by variance in GDP.

So, we have a relationship, now we need to properly interpret the specifics. The coefficient provides us the slope of the linear relationship between our two variables. To properly interpret this relationship, we would say “For every 1 million dollar increase in GDP, military expenditures increases by 42,370 dollars, on average.” Generally, proper interpretation entails “For every one unit increase in the independent variable, the dependent variable increases by the coefficient, on average”. The “\_cons” row gives us the intercept of the relationship, or the baseline if of military expenditures if GDP were to equal 0. Obviously, a country with 0 GDP does not and cannot exist, so we get a logically impossible negative value for military expenditures. This is one of the potential problems of ordinary least squares regression: the model assumes the variable is continuous, across both positive and negative values, as well as linear. This assumption may not always, or even often, hold true.

 Almost always testing our theories about international politics involves more than one independent variable. For example, we may be interested to test how multiple factors influence the amount of exports a state engages in as a percentage of their GDP. We hypothesize that the amount of exports is dependent upon the foreign policy behaviors of a state. The more cooperative, the more they export, the more conflictual, the less exports. However, it is well known that the smaller the economy, the more dependent the state’s economy is on trade, so we should control (or include) GDP in our model so that we are properly accounting for this relationship. So, instead of the above model with only two variables, we’re going to run a regression using 5. First, our dependent variable being exports as a percentage of GDP. Second, the independent variables that explain our theory: cooperative events, conflictual events, and number of militarized interstate disputes. Finally, we include the fifth and final variable, GDP to control for its influence on trade dependence. Using the same regress command:

. regress exports coop conf mids gdp

 Source | SS df MS Number of obs = 2022

-------------+------------------------------ F( 4, 2017) = 17.53

 Model | 32868.0406 4 8217.01015 Prob > F = 0.0000

 Residual | 945284.716 2017 468.658759 R-squared = 0.0336

-------------+------------------------------ Adj R-squared = 0.0317

 Total | 978152.757 2021 483.994437 Root MSE = 21.649

------------------------------------------------------------------------------

 exports | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 coop | .0064302 .0036636 1.76 0.079 -.0007546 .0136151

 conflict | -.0196331 .0065523 -3.00 0.003 -.032483 -.0067832

 mids | -.7421322 .319688 -2.32 0.020 -1.369085 -.115179

 gdp | -3.82e-06 1.67e-06 -2.29 0.022 -7.09e-06 -5.46e-07

 \_cons | 37.7071 .5355186 70.41 0.000 36.65687 38.75732

------------------------------------------------------------------------------

We can interpret these results similar to our above model with only two variables. First, note the very low r-squared value of .0336. That means we’re only explaining about 3% of the variance in exports as a percentage of GDP! Don’t be discouraged by this. While it is true that you have a very small substantive effect in explaining the range of values of exports, the complexity of international politics often results in models with very low r-squareds. The relationships between countries are often too complex to fully capture in a simple statistical model (this is why your theories are so important!). Despite the low R-squared values, a number of our variables do exhibit a statistically significant relationship. Conflict, militarized disputes, and GDP all have a significant and negative impact on exports as a percent of GDP. Conflict does appear to negatively impact the level of trade involvement, with each militarized dispute annually decreasing the amount of exports as a percentage of GDP of the country by .74%, on average *all other variables at equal*. Our control, GDP is also significant and negatively signed, demonstrating larger economies are proportionally less dependent on trade. Cooperation, however, is unfortunately not statistically significant, with a low t score (1.76), and a P>|t| value greater than .05. The confidence interval includes 0, signifying that it is possible there is no relationship between cooperation and exports as a percentage of GDP.

Finally, as with descriptive statistics, we can generate some interesting graphs to illustrate our tested relationships:

twoway (lfit exports conflict, range(0 3000))

Where twoway is the graph type**, lfit** means we want the predicted values of exports based on our regression model (instead of the observed values) , and we want to graph exports as our dependent variable on the y-axis and conflict as our independent variable on the x-axis. There are a variety of options you can specify, including the range of the x-values as done above. Exploring the graphics dropdown menu is an excellent way to better understand both your data and the relationships within your models.



**Dichotomous Dependent Variables**

 Not all variables are continuous. If we are interested in explaining the onset of a militarized interstate dispute there are only two options, either there is a dispute, or there is not. If a variable is not continuous, but instead dichotomous (0 or 1) it becomes impossible for us to discuss the effect of a unit change in an independent variable on a dependent variable with only two outcomes. Instead, we have to talk about the effect of an independent variable on the odds of something happening. For this, we’ll use **a *logistic regression*** instead of the *ordinary least squares* regression we used above. So, let us seek to explain the onset of a militarized interstate dispute using trade dependence and control for GDP. First, we have to create the variables:

. gen onset=0

. replace onset=1 if mids>0

(2216 real changes made)

We generate an onset variable for our dependent variable of the occurrence of a militarized interstate dispute. We can’t use the mids variable, which is the number of mids a state is engaged in a year, and ranges from 0 to 51. We simply want to know if a state is involved in the onset of any mid in a year or not. First we generate the variable and set it equal to 0. Second, we replace all observations of the onset variable with a 1 if the mids variable is greater than 0. This newly generated dichotomous variable is our dependent variable, or what we are seeking to explain.

Secondly, we have exports and imports as percentages of GDP, but we don’t have an overall measure of trade dependence. So we’ll create another variable which is simply the sum of exports and imports:

. gen trade=exports+imports

Now we are ready to run our model, which we do using the **“logit”** command:

. logit onset trade penncons

Iteration 0: log likelihood = -3243.3716

Iteration 1: log likelihood = -3180.9548

Iteration 2: log likelihood = -3167.8094

Iteration 3: log likelihood = -3166.4093

Iteration 4: log likelihood = -3166.3974

Logistic regression Number of obs = 4699

 LR chi2(2) = 153.95

 Prob > chi2 = 0.0000

Log likelihood = -3166.3974 Pseudo R2 = 0.0237

------------------------------------------------------------------------------

 onset | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 trade | .001448 .0006835 2.12 0.034 .0001083 .0027876

 penncons | .0006935 .0000747 9.28 0.000 .0005471 .0008399

 \_cons | -.4147171 .0673827 -6.15 0.000 -.5467848 -.2826493

------------------------------------------------------------------------------

Immediately notice that the relationship between trade and GDP with MID onset is positive: both variables increase the probability, or odds, of a country being involved in a militarized interstate dispute in a given year. However, our coefficients are in logged odds of the relationship. Stata will convert these coefficients into **odds-ratios** for us for ease of interpretation. So we re-run the model, adding to the end “,or”, telling Stata we’d like the results in odds-ratio form.

. logit onset trade penncons, or

Iteration 0: log likelihood = -3243.3716

Iteration 1: log likelihood = -3180.9548

Iteration 2: log likelihood = -3167.8094

Iteration 3: log likelihood = -3166.4093

Iteration 4: log likelihood = -3166.3974

Logistic regression Number of obs = 4699

 LR chi2(2) = 153.95

 Prob > chi2 = 0.0000

Log likelihood = -3166.3974 Pseudo R2 = 0.0237

------------------------------------------------------------------------------

 onset | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 trade | 1.001449 .0006845 2.12 0.034 1.000108 1.002792

 penncons | 1.000694 .0000748 9.28 0.000 1.000547 1.00084

------------------------------------------------------------------------------

To interpret these ratios we can simply interpret the ratio of the coefficient in reference to “1”. A one percentage point increase in trade dependence has a .14% increase in the probability a state experiences a MID onset, on average. This may seem like a very small relationship, but using our summarize command we see that trade dependence ranges from .31 to 438. Doing a little quick math, we can tell that a very dependent state at the maximum value of trade dependence is 61% more likely to be engaged in a militarized dispute in a given year, which is a highly influential effect!

The indicators of statistical significance can be interpreted similar to the regression equation. Note that if you are assessing the logit output in odds-ratio form, you want your 95% confidence interval to not include 1 instead of 0. Similarly, an odds ratio less than one would be interpreted as the inverse of a negative odds ratio. A .99 odds ratio, for example, would be interpreted as 1 unit increase in independent variable decreases the probability of experiencing a militarized dispute onset by 1%, on average.

**Making Your Own Data**

While the data provided lists a series of important and often used variables in international politics, a student may be interested in creating her/his own data to better match the research question. Thankfully, there’s a quick and easy tool that allows international politics researchers to generate their own datasets from scratch called EUGene. To work with EUGene, download it from <http://eugenesoftware.org/>.

For students using Windows 7 x64, remember to run the program in administrator (see the FAQ on the EUGene website).

Once installed, we can create datasets based upon our research question. Click the “Create Data Set” option in the menu once you start the program. You may create country-year data, similar to the data we used in the above exercises. Alternatively, you may select dyad-year data if you’re interested in relationships. If you are interested in the direction of a relationship (what causes country X to send a certain amount of exports to Y) then you’ll want directed dyadic data. Once you select the type of dataset you wish to create, select a destination for the output file, choose the population of cases, both the time span and selection of countries. Finally, choose the variables you are interested in including in the dataset. A full list of available variables is listed on <http://eugenesoftware.org/>.

As an example, let’s say that we’re interested in how trade affects the possibility of conflict *within a relationship between two states*. This means we’re interested in a dyad-year framework that includes variables for trade and conflict. First, open the EUGene program, select “Create Data Set” in the menu bar followed by “Non-Directed Dyad Year”. We want non-directed relationships because we’re not particularly interested in which side in the relationship is doing what, merely how trade and conflict are related within a relationship generally. Second, we select the output destination (anywhere on your computer that you like, just remember where you’re putting the file). The output will be a tab deliminated text file. We’ll use Excel to view the data later. We’re going to select all years, but you can restrict the population to fewer years if you have a research design that calls for it, such as during or after the Cold War. Then, select the dyads of interest. For this example, we’ll select “All Dyads from All Countries”. Again, if you’re research design calls for something different, perhaps you are only interested in the behaviors of major powers, you can select this option. Finally, click the variables tab at the top of the screen. In this interface, you’ll choose the variables you’re interested in including in your model. You’ll notice immediately that the “CCodes” and “Year” variables are already chosen. The CCode variable is the numeric code you see in the data already provided to you for the course. We’ll also select “Abbreviation” as a shortcut to identify countries as we’re browsing our data later. We now have a myriad of variables that we can choose from by navigating the tabs. For this illustration, select the “Conflict Data” tab. Click “COW MID Data (1816+)” and already selected is “Dispute Onset” which is our variable of interest. Next, we’ll click the “User Data” tab, selecting “flow1” and “flow2” under the tab “COWDyadTrade” . Note that as you choose which variables to include in your data, you can mouse over the variable name to see a brief statement about what the variable is. Finally, click okay at the bottom of the screen and EUGene will generate your dataset. Once it is complete, open excel, click open from the menu at the top of the screen. Change the file type in the bottom right of the open interface to “All Files”. Next, find the “.out” file EUGene created for you, select it, and click “Open”. Excel will then ask if the file is “Delimited”, click “Next >”. Select the Delimiters (the default is “Tab”), then click Finish. The File will now open in Excel. Before proceeding to Stata, browse the data in Excel and be sure the data is appropriate to your research design and nothing looks out of the ordinary. Finally, copy the data in Excel and paste into the data editor in Stata. Close the Stata data editor, and now, you’re ready to proceed analyzing your created dataset.

1. Note that the military expenditures data is in 2009 constant dollars while world bank data is in 2000 dollars. While it is reasonable to use the unchanged data to compare distributions between countries, as well as analyze data, if we want accurate figures for our discussion or descriptive statistics, we would need to first convert them into the same constant dollar units. [↑](#footnote-ref-1)