ECON 442: Quantitative Trade Models **Professor Jack Rossbach**

Instructions on Doing Gravity Regressions in STATA

Important: If you don't know how to use a command, use the **help** command in R. For example, type **help reg** in the command line for STATA and it will provide documentation for using the **reg** command.

Step 1: Download the Data

- I uploaded the data on my website under Gravity Data for Problem 1, the file name is gravity_data.dta
- Available directly from CEPII: http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=8 (the "lighter dataset" already has trade flows merged in)

Alternatively: Trade Data can be downloaded from Comtrade, then merged with other data using iso3 country codes.

Step 2: Open the Data in STATA

- Can double click the data file and it should open STATA
- Can also open STATA then load the file using the **use** command to load it (to load non-STATA datasets there are various **import** commands, e.g. **import excel** or **import delimited**)
- Can also open STATA then go to file → open and locate the data file that way

Step 3: Prepare the Data for the Regression

Preparation Part 1: Keeping Data for Year 2000

- In the problem set, we only want to use data for the year 2000 (in general, it is better to use year fixed effects rather dropping years).
- We do this using the **keep** command. To keep data for only the year 2000 we type the command **keep if year == 2000** Note the two sequential equals signs. As with R, a single equals sign denotes assignment (can't use <- for assignment in STATA), and two equals signs denotes a test for equality.
- We could also drop the data we don't want using the **drop** command. The equivalent of the above bullet would be **drop if year != 2000** where != means "is not equal to."
- This step is already done in the data on my website, but if you download the data from CEPII then you'll have to do it yourself.

- It's not necessary to actually drop the data to exclude it in our regressions. We can use the **if** option when running our regression to run the regression for only a particular year, e.g. **reg y x, if year == 2000**

Preparation Part 2: Transforming the Data

- We are going to estimate our data in logs so we can use Ordinary Least Squares (OLS) regression. To do this, we need to transform the relevant variables to logs, these variables are trade flows, GDPs, and distance. We won't transform indicator, dummy, or categorical variables.
- To transform the data we have two possible commands: **replace** can be used to replace the data for a given variable. The other option, which I prefer as long as there aren't performance concerns, is to generate a new variable using the **gen** command. The benefit of generating a new variable is that you won't accidentally transform the variable a second time or think you already transformed it when you haven't.
- The **gen** command works as follows: **gen [new var] = expression** for us, the expression will be log(gdp) or log(distance) and [new var] is whatever we decide to name the old variable
- **gen lgdp_o = log(gdp_o)** generates a new variable named lgdp_o that is equal to the log of gdp_o. gdp_o is the variable in the dataset containing the value for GDP of the origin country. In STATA, log refers to natural log, not log base 10 as in Excel.
- The variables we need to transform are **gdp_o** (GDP of origin/exporting country), **gdp_d** (GDP of destination/importing country), **distw** (population weighted distance between the countries), and **flow** (trade flows: total exports from origin to destination).
- *Note: You could estimate the gravity equation without taking logs after downloading the package here, but logs and OLS are fine for our purposes even though our estimated coefficients will be slightly biased.*

Step 4: Run the Regression

Regression Part 1: No Fixed Effects

- We will be running OLS regression, which is done using the **reg** command. The format is **reg [dependant var] = [independent vars]** , the dependent var for us is the LHS of the gravity regression equation and the independent variables are everything on the RHS.
- The gravity equation is

$$\log \text{Trade}_{ij} = \text{Constant} + \beta_{\text{exp}} \log \text{GDP}_i + \beta_{\text{imp}} \log \text{GDP}_j + \beta_{\text{dist}} \log D_{ij} + \text{Controls} + \epsilon_{ij}$$

where Trade_{ij} is exports from country i to j, GDP_i and GDP_j are, respectively, the GDPs of countries i (the exporter) and j (the importer), D_{ij} is the distance between the countries. Controls are things such as whether there is a FTA or a common language.

- Suppose the necessary variables from Step 3 have been transformed by taking logs and the names now have a "l" in front of them. We can run the regression with no controls by doing the command: **reg lflow lgdp_o lgdp_d ldistw** after which we will get output similar to the below

```
. reg lflow lgdp_o lgdp_d ldistw
```

Source	SS	df	MS	Number of obs	= 19997
				F(3, 19993)	=11623.45
Model	191339.135	3	63779.7116	Prob > F	= 0.0000
Residual	109704.716	19993	5.48715632	R-squared	= 0.6356
				Adj R-squared	= 0.6355
Total	301043.851	19996	15.0552036	Root MSE	= 2.3425

lflow	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lgdp_o	1.161225	.0075349	154.11	0.000	1.146456	1.175994
lgdp_d	.9215594	.0073702	125.04	0.000	.9071132	.9360057
ldistw	-1.472611	.0202102	-72.86	0.000	-1.512225	-1.432997
_cons	-7.689504	.1983911	-38.76	0.000	-8.078367	-7.300641

The most important columns are the *Coef.* and *Std. Err.* columns. *Coef* is the point estimate of the coefficient for the gravity regression. For example, we see 1.161225 as the value for lgdp_o, this is our point estimate of $\beta_{\exp}$. The *Std. Err.* column reports the standard error of our estimate, which indicates how precisely we were able to estimate it. Standard errors are used to test for statistical significance and to construct confidence intervals.

- Most economists prefer to see the standard error reported itself rather than only reporting p-values (the *P>|t|* column) or confidence intervals. A quick rule of thumb is to double the standard error, if it's less than the estimated coefficient, then it is statistically significant. The reason for reporting the standard error is that there is disagreement over the importance of statistical significance, but there is no disagreement over the importance of understanding the precision of our estimates.
- To add controls we simply add them to our dependent variables. For example, to run the regression with controls for FTAs and a common official language we would run the regression: **reg lflow lgdp_o lgdp_d ldistw rta comlang_off** where *rta* indicates whether a FTA (regional trade agreement) is in place and *comlang_off* indicates whether a there is an official shared language.
- For 1.i, the only control should be **rta**. For **1.ii**, we should include the controls **comlang_off**, **contig,** and **col_hist**; where *contig* indicates whether the countries share a border (are contiguous) and *col_hist* indicates whether the countries were ever in a colonial relationship.

Regression Part 2: Fixed Effects

- Fixed effects means that we have a dummy or indicator for each value that variable takes. The most common fixed effects are for countries (which we will be doing) and years.
- Dummy variables for each value are equal to 1 if the variable we are doing fixed effects for takes the value corresponding to a given dummy and 0 otherwise.
- *Example*: Suppose we have three countries that export, the USA, Canada, and Mexico. Then if we do exporter fixed effects we will have three separate dummy variables that we will include in our regression. Our regression will take the form

$$\log \text{Trade}_{ij} = \text{Gravity Vars} + \beta_{USA}\text{Dummy}_{\text{USA}} + \beta_{\text{Can}}\text{Dummy}_{\text{Can}} + \beta_{\text{Mex}}\text{Dummy}_{\text{Mex}} + \epsilon_{ij}$$

  Where the Dummy Variables take the following values:

Exporter	USA Exporter Dummy	Canada Exporter Dummy	Mexico Exporter Dummy
USA	1	0	0
Canada	0	1	0
Mexico	0	0	1

  The interpretation of β_{USA} therefore is how much higher we expect exports to be if the exporting country is the USA. Note that the coefficients can also be negative.

- To run a regression in fixed effects we have to do two things. First, **encode** the 3 digit ISO country codes so that instead of string variables they are integers. We do this using the **encode** command, the format of which is **encode [var to encode], generate([new var name])**
- The commands **encode iso_o, generate(importer)** and **encode iso_d, generate(exporter)** create new variables name *importer* and *exporter* which are encoded versions of the origin and destination country codes.
- Important: Run the command **set more off** and the command **set matsize 800** before running the command in the next bullet. Without the first command, STATA will pause when the screen fills up with output and you'll have to hit a key several times to make it continue. Without the second command, STATA will not be able to create the dummy variables for the fixed effects regression.
- To run the fixed effects regression, we included the encoded variables in the regression with "i." in front of the variable name. The *i.* indicates that each value the variable takes should be treated as an indicator/dummy variable. The regression with both importer and exporter fixed effects is therefore **reg lflow i.exporter i.importer ldistw rta comlang_off contig col_hist**
- The regression output will include estimates for all the dummy coefficients. For us, this is fine. If you want to suppress there are options including the **xtreg** command, **areg** command, and running our OLS regression with all output suppressed by doing **quietly: reg …** and then using the **estout** package (**ssc install estout**). Use the help commands and google if you want to learn more about those options.

More information on Fixed Effects:

- The reason we have importer and exporter fixed effects is because in the model-derived gravity regression we have terms that depend on the price level/remoteness of a given country. We don't typically observe these things directly (we could use CPIs or PPIs, but there are differences in methodologies that make comparing across countries difficult), but we do know that according to the model the price level should depend only on that country.
- Basically, what the fixed effects does is replace terms that depend on only a single country with a fixed effect. We have trouble in that the fixed effect depends on multiple things, but we can often use theory to disentangle the fixed effects, and if our coefficient of interest appears outside of the fixed effect, e.g. β_{dist} or β_{FTA}, then we don't particularly care about the fixed effects and only need them to make sure those coefficients aren't biased.
- Note: In the regression without fixed effects we put in the GDP for each country. In the regression with fixed effects we will have to remove GDP since we only have a single observation for each country due to using only year 2000 data. That means there is a one-to-one mapping between the fixed effects and GDP values in 2000, so we can only have one of the two. We could technically have both GDP and Exporter fixed effects if we had multiple years, but it is standard practice to only have the fixed effects, since the fixed effects can absorb the GDP term.

Step 5: Counterfactuals

- We can use the gravity equation and estimated coefficients to do simple counterfactuals regarding what trade flows would have been if one of the values took a different value
- Suppose a given independent variable X in the regression changes to X' and all the other independent variables stay the same. Then we have that

$$\log \text{Trade}'_{ij} - \log \text{Trade}_{ij} = \beta_X(X' - X),$$

since all the other terms cancel. We can use to above to get

$$\%\ \text{Change in Trade from } i \text{ to } j \equiv 100 \times \left(\frac{\text{Trade}'_{ij}}{\text{Trade}_{ij}} - 1\right) = 100 \times (\exp[\beta_X(X' - X)] - 1)$$

- If X' is a dummy variable, then $(X' - X)$ will be either 1 or -1, depending on the initial value of the dummy.
- To find the value of the variable for a given country pair, we can use the **list** command along with the **if** option and iso country codes. E.g. for the log distance between USA and China we can use the command: **list ldistw if iso_o == "CHN" & iso_d == "USA"**
- If $X = \log \text{GDP}$ or something similar, then we can cancel the log and exp to get

$$\%\ \text{Change in Trade from } i \text{ to } j \equiv\ 100 \times \left(\frac{\text{Trade}'_{ij}}{\text{Trade}_{ij}} - 1\right) = 100 \times \left(\left(\frac{\text{GDP}'}{\text{GDP}}\right)^{\beta_X} - 1\right)$$

- For small changes in X (the changes in the problem set aren't), the coefficient is approximately equal to the % change in the dependent variable from a 1% increase in the independent variable. This is a very bad approximation for large changes in X.
- If we have both X changing to X' and Y changing to Y' then we'd have

$$\%\ \text{Change in Trade from } i \text{ to } j = 100 \times (\exp[\beta_X(X' - X) + \beta_Y(Y' - Y)] - 1)$$

- There are a lot of dangers in using gravity regressions to do counterfactuals. The dangers are similar to the dangers for macroeconomic policy raised by the Lucas critique. These can be partially mitigated by using economic theory. The current standard is to perform counterfactuals within an economic model that directly incorporates the most important elements of the policy issue you're interested in, and then gravity equations can be used to estimate the exogenous parameters associated with that general equilibrium model.

TL;DR version

1. Load data
2. Take logs
3. Run OLS regression using **reg** command
4. **encode** iso_d and iso_o so they are numeric variables instead of string
5. Run OLS regression using **reg** command, include **i.importer** and **i.exporter** for fixed effects
6. Use the gravity equation and estimated coefficients to do simple counterfactuals