

Gravity in Stata: a short workshop

Krisna Gupta

October 28, 2024

This page is dedicated to teach students on running the gravity model in Stata. We uses dataset from CEPII, specifically BACI and the gravity dataset. We rely on `ppmlhdfe` when demonstrating PPML. We try to replicate Silva and Tenreyro (2006), an excellent paper for an introduction to the log gravity model and PPML.

For the latest version [click here](#).

Introduction

The gravity model is probably the most popular model in international trade. Many uses them. It is very intuitive, great predictive power, and most importantly, tweakable (Yotov 2022). But the even most important is that UI students love them. If you're doing trade for your thesis, then you probably going to use the gravity model as your backbone.

This guide is my attempt to help you learn gravity model much easier. The most important part is probably the data and the model itself. What is the minimum things you need in the gravity model, how to arrange the database, run them, and interpret them. You must familiarize yourself with the data and its wrangling (80% of your coding) as well as the main gravity specification to date. I encourage students to pay careful attention to Yotov (2022) as it hosts the recent development in the gravity model, a must read if you're planning to utilize gravity model.

I use Stata here because most economists I know are still using Stata. I myself quite often do R these days. However, the two language aren't very different. You can do the same thing on both, but you may need to google a bit. It's okay to use google a lot. I did as well even right now.

Oh yeah I also informed you guys know Stata already so I won't go into too much basic stuff. For example, you probably already know the function `gen` and `egen`, also `import` and `cd`. Some I will repeat but won't discuss too much. Also, you'd want to learn how to use `outreg2`, a package to save regression results.

By the way, for the R version of gravity, you can consult to [this page](#).

Data

I procure data for this workshop from [CEPII](#). From their website, CEPII is:

The CEPII is the leading French center for research and expertise on the world economy. It contributes to the policy making process through its independent in-depth analyses on international trade, migrations, macroeconomics and finance. The CEPII also produces databases and provides a platform for debate among academics, experts, practitioners, decision makers and other private and public stakeholders. Founded in 1978, the CEPII is part of the network coordinated by France Strategy, within the Prime Minister's services.

I use their BACI dataset (Gaulier and Zignago 2010) and gravity dataset (Conte, Cotterlaz, and Mayer 2022). You can get those from this [link](#). BACI is under "international trade" banner while gravity is under "Gravity" banner. Specifically, I downloaded the 2017-2022 version of BACI and for the gravity dataset I downloaded the R version. You can of course

download whichever version you like but for the purpose of this workshop maybe its best to stick with the same dataset as I.

You can also download from [my drive](#).

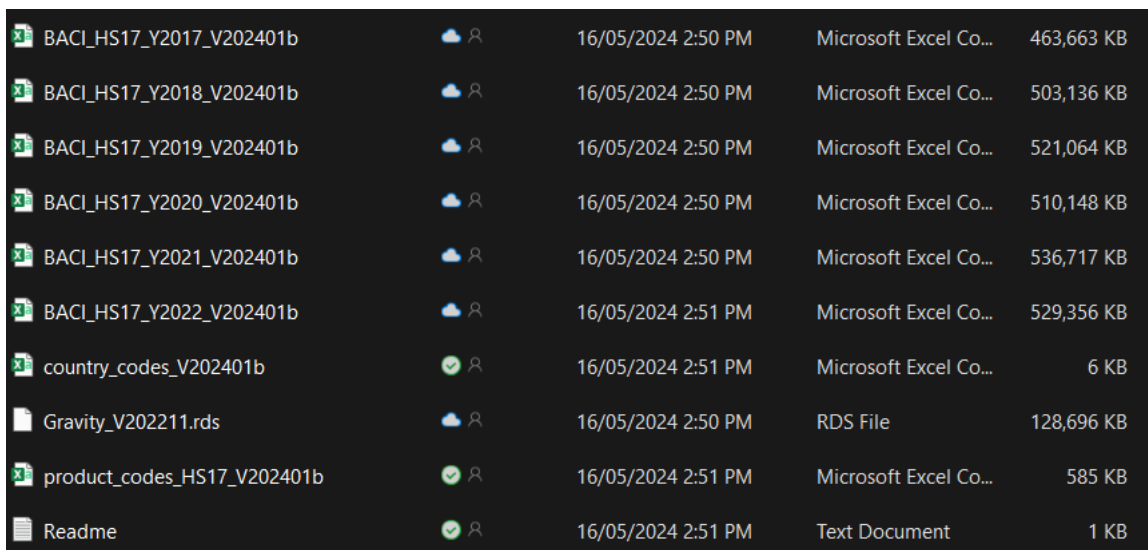
Note that the data here is **extremely large** in size so be mindful. You need hefty internet quota and reasonable speed. Also, you can try opening it with spreadsheet software but unless you have a strong computer, i'd advice against it. Use R instead.

In the CEPII website you can use various other dataset that may be useful for you. At the same time, there are various other source you can utilise for your actual project that's not necessarily from CEPII.

working directory

If you finished downloading data and installing softwares, you then need to set up a working directory. A working directory is basically a folder where you have all the data and your do file. For now what you want is to have a **folder filled with your downloaded data**. Make sure you know the path to this folder. I tend to use easy path for my projects and move it somewhere else when I finished. If you use github or the likes, it'll be even nicer because you can actually wipe out your local repo when you're done.

All in all, you should have a folder with these stuff in it:



BACI_HS17_Y2017_V202401b		16/05/2024 2:50 PM	Microsoft Excel Co...	463,663 KB
BACI_HS17_Y2018_V202401b		16/05/2024 2:50 PM	Microsoft Excel Co...	503,136 KB
BACI_HS17_Y2019_V202401b		16/05/2024 2:50 PM	Microsoft Excel Co...	521,064 KB
BACI_HS17_Y2020_V202401b		16/05/2024 2:50 PM	Microsoft Excel Co...	510,148 KB
BACI_HS17_Y2021_V202401b		16/05/2024 2:50 PM	Microsoft Excel Co...	536,717 KB
BACI_HS17_Y2022_V202401b		16/05/2024 2:51 PM	Microsoft Excel Co...	529,356 KB
country_codes_V202401b		16/05/2024 2:51 PM	Microsoft Excel Co...	6 KB
Gravity_V202211.rds		16/05/2024 2:50 PM	RDS File	128,696 KB
product_codes_HS17_V202401b		16/05/2024 2:51 PM	Microsoft Excel Co...	585 KB
Readme		16/05/2024 2:51 PM	Text Document	1 KB

Notes about the data country_codes, product_codes and Readme are all for reading BACI. Moreover, you wont need Gravity_V202211.rds because it's an R file. Take the dta one.

Simple gravity specification

Theory

The earliest (e.g., naive) gravity model taking directly from Newtonian gravity theory looks something like this:

$$X_{ij} = \tilde{G} \frac{Y_i E_j}{T_{ij}^\theta} \quad (0.1)$$

where X_{it} is the value of trade flow from country i to country j , \tilde{G} is the gravitational constant (aka our usual constant), Y_i is the output in country i , E_j is the value of expenditure in country j and T_{ij} is the total bilateral trade frictions / trade cost between country i and country j .

There are various other types of gravity equations, but let's start with a relatively simple one. One of my favorite simple gravity specification is a budget version of Silva and Tenreyro (2006) which is taken from Anderson and Wincoop (2003) which looks like this:

$$X_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} e^{\theta_i d_i + \theta_j d_j} \quad (0.2)$$

where α_0 is your \tilde{G} , while Y is the output and expenditure which is proxied with GDP. D_{ij} is the distance between the two countries, which can be generalized as a vector of trade cost measures. Typically we use physical distance but also other types of bilateral trade cost. Lastly, the d_i and d_j is country-specific characteristics.

There are various variables used in Silva and Tenreyro (2006). log of exporter's and importer's GDP and GDP per capita. Various "distance" variables is used as well e.g., physical distance and variables like contiguity, common-language dummy, colonial-tie dummy and free trade agreement dummy.

Note that our regression consists only of two indices: exporter i and importer j . We are going to use the gravity data I mentioned earlier, slice the dataset to cover only one year chosen arbitrarily (which is 2019), and run Equation [0.2](#).

Setting data

First of all, let me share you my typical routine at the top:

```
// preamble
clear all
set more off
cd C:\github\statagravity
```

Change the working directory after `cd` with your own working directory. sometimes I set logs but sometimes I am too lazy. Anyway, the next is to read our gravity dataset. First of all, we want to filter countries to match Silva and Tenreyro (2006). First we read our country key, then we use it to filter out the gravity data so we only have those countries.

```
\\ read data
use Countries_V202211,clear
keep if country=="Albania" | country == "Denmark" | country == "Kenya" | country == "Romania"

//// There's certainly a better way to do the above. maybe next time.

/// we need to select destination and origin in two step

gen iso3_o=iso3 // select origin countries

save ctr,replace
clear all
use Gravity_V202211,clear
merge m:1 iso3_o using ctr
keep if _merge==3
drop _merge

save grav,replace

clear all
use ctr,clear

rename iso3_o iso3_d // now select destination country
save ctr,replace
clear all
use grav,clear
merge m:1 iso3_d using ctr
keep if _merge==3
drop _merge

save grav,replace
```

The above is preparation for the gravity dataset we're doing. It is complicated but I put comments to help. This kind of manipulation would be what you will do a lot.

The first thing to do is `sum` your data to see what variables you have and its descriptive statistics. Descriptive statistics helps with visualizing the overall data in your head, and typically a must-do when you write your thesis. Some put it in chapter 3, some in 4. You can use `ds` if you'd like to see variable names.

If you ran either `sum` or `ds`, you will see that the column names are so plenty. Consult to the CEPII website or Conte, Cotterlaz, and Mayer (2022) to learn more. We will only use

some of them, so we will filter these data to make it more concise. Specifically, we will (1) remove some countries, (2) remove non-2019, and (3) remove variables we are not using.

For variables, we will keep iso3_o, iso3_d, distw_harmonic, contig, comcol, comlang_off, gdp_o, gdp_d, gdpco_o, gdpco_d, fta_wto. Note that o means origin / exporter and d means destination / importer. We then add log version of variables.

```
/// Now we keep variable we need

keep iso3num_o iso3num_d year iso3_o iso3_d distw_harmonic contig comcol comlang_off gdp_o

/// add log version

gen ldist=log(distw_harmonic)
gen lgdpo=log(gdp_o)
gen lgdpd=log(gdp_d)
gen lgdpco=log(gdpco_o)
gen lgdpco_d=log(gdpco_d)
gen logtrade=log(1+tradeflow_baci)

save grav,replace
```

you can quickly sum our new grav.

Regression

Let's do 2 types of regression. First we do a regression using a normal ols, and secondly we do ppml.

```
reg1<-feols(data=gravity2,logtrade~lgdpo+lgdpd+lgdpco+lgdpco_d+ldist+contig+
            comcol+comlang_off+fta_wto)
reg2<-feols(data=gravity2,logtrade~lgdpo+lgdpd+lgdpco+lgdpco_d+ldist+contig+
            comcol+comlang_off+fta_wto+iso3_o+iso3_d)
reg3<-fepois(data=gravity2,tradeflow_baci~lgdpo+lgdpd+lgdpco+lgdpco_d+ldist+
            contig+comcol+comlang_off+fta_wto)
reg4<-fepois(data=gravity2,tradeflow_baci~lgdpo+lgdpd+lgdpco+lgdpco_d+ldist+
            contig+comcol+comlang_off+fta_wto+iso3_o+iso3_d)
```

You can call each reg's table with `summary(reg1)`.

```
// regression
/// generate factor variable first
encode iso3_d,generate(iiso3_d)
```

```

encode iso3_o,generate(iiso3_o)
/// we keep 1 year for now, you can experiment later
keep if year==2019
reg logtrade lgdpd lgdpd lgdpco lgdpd ldist contig comcol comlang_off fta_wto
outreg2 using myreg.doc, replace label ctitle(OLS) title(Table XX: my amazing regression)
reg logtrade lgdpd lgdpd lgdpco lgdpd ldist contig comcol comlang_off fta_wto iiso3_d iis
outreg2 using myreg.doc, append label ctitle(FE)
ppmlhdfc logtrade lgdpd lgdpd lgdpco lgdpd ldist contig comcol comlang_off fta_wto
outreg2 using myreg.doc, append label ctitle(PPML)
ppmlhdfc logtrade lgdpd lgdpd lgdpco lgdpd ldist contig comcol comlang_off fta_wto,absorb
outreg2 using myreg.doc, append label ctitle(PPML FE)

```

Results are saved in your working directory as myreg.doc. It should look like this:

You can compare results with Silva and Tenreyro (2006). Note that they don't use fixed effects. Remember, PPML interpretation is percent change for logged independent variable and e^β magnitude (around $\beta \times 100\%$) for level independent variable.

TABLE 3.—THE TRADITIONAL GRAVITY EQUATION

Estimator: Dependent Variable:	OLS $\ln(T_{ij})$	OLS $\ln(1 + T_{ij})$	Tobit $\ln(a + T_{ij})$	NLS T_{ij}	PPML $T_{ij} > 0$	PPML T_{ij}
Log exporter's GDP	0.938** (0.012)	1.128** (0.011)	1.058** (0.012)	0.738** (0.038)	0.721** (0.027)	0.733** (0.027)
Log importer's GDP	0.798** (0.012)	0.866** (0.012)	0.847** (0.011)	0.862** (0.041)	0.732** (0.028)	0.741** (0.027)
Log exporter's GDP per capita	0.207** (0.017)	0.277** (0.018)	0.227** (0.015)	0.396** (0.116)	0.154** (0.053)	0.157** (0.053)
Log importer's GDP per capita	0.106** (0.018)	0.217** (0.018)	0.178** (0.015)	-0.033 (0.062)	0.133** (0.044)	0.135** (0.045)
Log distance	-1.166** (0.034)	-1.151** (0.040)	-1.160** (0.034)	-0.924** (0.072)	-0.776** (0.055)	-0.784** (0.055)
Contiguity dummy	0.314* (0.127)	-0.241 (0.201)	-0.225 (0.152)	-0.081 (0.100)	0.202 (0.105)	0.193 (0.104)
Common-language dummy	0.678** (0.067)	0.742** (0.067)	0.759** (0.060)	0.689** (0.085)	0.752** (0.134)	0.746** (0.135)
Colonial-tie dummy	0.397** (0.070)	0.392** (0.070)	0.416** (0.063)	0.036 (0.125)	0.019 (0.150)	0.024 (0.150)
Landlocked-exporter dummy	-0.062 (0.062)	0.106* (0.054)	-0.038 (0.052)	-1.367** (0.202)	-0.873** (0.157)	-0.864** (0.157)
Landlocked-importer dummy	-0.665** (0.060)	-0.278** (0.055)	-0.479** (0.051)	-0.471** (0.184)	-0.704** (0.141)	-0.697** (0.141)
Exporter's remoteness	0.467** (0.079)	0.526** (0.087)	0.563** (0.068)	1.188** (0.182)	0.647** (0.135)	0.660** (0.134)
Importer's remoteness	-0.205* (0.085)	-0.109 (0.091)	-0.032 (0.073)	1.010** (0.154)	0.549** (0.120)	0.561** (0.118)
Free-trade agreement dummy	0.491** (0.097)	1.289** (0.124)	0.729** (0.103)	0.443** (0.109)	0.179* (0.090)	0.181* (0.088)
Openness	-0.170** (0.053)	0.739** (0.050)	0.310** (0.045)	0.928** (0.191)	-0.139 (0.133)	-0.107 (0.131)
Observations	9613	18360	18360	18360	9613	18360
RESET test p -values	0.000	0.000	0.204	0.000	0.941	0.331

Figure 1: source: Silva and Tenreyro (2006)

PPML HDFE

Here I show a multiyear version of the regression. Instead of keep year at 2019, we use all observation.


```

clear all
use grav,clear
keep if contig!=. //drop multi input

encode iso3_d,generate(iiso3_d)
encode iso3_o,generate(iiso3_o)
egen id=group(iso3_o iso3_d)
xtset id year
xtreg logtrade lgdpd lgdpd lgdpco lgdpco ldlist contig comcol comlang_off fta_wto, r
outreg2 using myreg2.doc, replace label ctitle(OLS) ttitle(Table XX: my amazing regression)
xtreg logtrade lgdpd lgdpd lgdpco lgdpco ldlist contig comcol comlang_off fta_wto,fe r
outreg2 using myreg2.doc, append label ctitle(FE)
ppmlhdfe tradeflow_baci lgdpd lgdpd lgdpco lgdpco ldlist contig comcol comlang_off fta_wto,
outreg2 using myreg2.doc, append label ctitle(PPML)
ppmlhdfe tradeflow_baci lgdpd lgdpd lgdpco lgdpco ldlist contig comcol comlang_off fta_wto,
outreg2 using myreg2.doc, append label ctitle(PPMLHDFE)

```

Product level gravity

Theory

We then proceed to a higher-dimension trade data which you may be interested in. In the field, UI students often interested largely in Indonesian affairs. That is, we are not interested so much in the bilateral flow of all countries, but only on Indonesia. However, we often use more granular dimension than just exporter/importer. Often times we use indices like time, commodities or industries, or even firms (shamelessly inserting my paper here Gupta (2023)).

Now, if you are planning to do these kinds of studies, then you are going to need to tackle higher degree dataset and merging the gravity variables. Most often you can get these variables from [World Development Indicators](#) but CEPII is ok for now (note the main problem of CEPII is its timeliness).

The theory isn't so different compared to our previous gravity model. What we want is an additional indices. We are going to estimate something similar as Equation 0.2 but with more indices. We need to care about multilateral resistance (MR) and we can use dummies since we now have more variations from indices like time and HS code.

According to Yotov (2022), we need at least 3 dummies to run a multi-country, multi-time and multi-goods/sectors¹. We need to have exporter-time dummy, importer-time dummy and country-pair dummy. We need to construct this first. Note that these dummies will likely absorb some of your variables like distance (consistant between pair across time, typically).

¹unless you have domestic trade data which we typically don't. If you do, then there's borders dummy. More on Yotov (2022).

So we will do the HS,time varying version of Equation 0.2:

$$X_{ijpt} = \alpha_0 Y_{it}^{\alpha_1} Y_{jt}^{\alpha_2} D_{ijpt}^{\alpha_3} e^{\theta_1 o_{it} + \theta_2 d_{jt} + \theta_3 p_{ij}} \quad (0.3)$$

Setting data

Maybe later guys. Check out the R version if you wanna.

Closing

OKay now you are ready to run regression yourself. Try to replicate what I do here and you prolly finished 50% of your thesis. You then can work to update this with your own hypothesis, adding more variable and more concentrated.

Running this on R is also excellent. I must confess that R is also speedy (these guys making the package is extremely good), but Stata is a bit more intuitive and compute you with important stats as well such as pseudo-R. Nevertheless, now you should be able to do both!

As you are a student now, I encourage you to explore as much as you can because this is the moment. Once you're a proper adult, you must think more mundane stuff so please value your freedom at this point and explore as much as you can! Go out there make mistakes while you can!

I cannot emphasize enough references in Yotov (2022). Whatever you want to do, a paper prolly covered it already. Learn from them and look for an insight to add. Work with your spv and you'll be fine.

References

- Anderson, James E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." Journal Article. *The American Economic Review* 93 (1): 24.
- Conte, Madallena, Pierre Cotterlaz, and Thierry Mayer. 2022. "The CEPII Gravity Database." Working Papers 2022-05. CEPII. <http://www.cepii.fr/CEPII/fr/publications/wp/abstract.asp?NoDoc=2726>.
- Gaulier, Guillaume, and Soledad Zignago. 2010. "BACI: International Trade Database at the Product-Level. The 1994-2007 Version." Working Papers 2010-23. CEPII. <http://www.cepii.fr/CEPII/fr/publications/wp/abstract.asp?NoDoc=2726>.
- Gupta, Krisna. 2023. "The Heterogeneous Impact of Tariffs and Ntms on Total Factor Productivity for Indonesian Firms." Journal Article. *Bulletin of Indonesian Economic Studies* 59 (2): 269–300. <https://doi.org/10.1080/00074918.2021.2016613>.
- Silva, Santos, and Silvana Tenreyro. 2006. "The Log of Gravity." Journal Article. *The Review of Economics and Statistics* 88 (4): 19.

Yotov, Yoto. 2022. “Gravity at Sixty: The Workhorse Model of Trade.” Journal Article. *CESifo Working Papers* 9584. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4037001.