AGRODEP
African Growth & Development Policy modeling consortium
FACILITATED BY IFPRI

AGRODEP Technical Note TN-05

April 2013

Hands-on gravity estimation with STATA

Version 2

Maria Cipollina and Luca Salvatici

Hands-on gravity estimation with STATA

In this document we give several examples of hands-on estimation to familiarize yourself with the gravity equation methodological choices highlighted in the literature review. This guide provides an illustrative dataset with alternative Stata codes presenting the different possible estimation strategies.

Part 1 describes how estimations are carried out with panel data and are directed to show the relevance of the multilateral resistance term as well as the modeling of the (trade) policy variables. In Part 2, cross-sections estimations show the importance of working with disaggregated data. Finally, Part 3 shows how you can solve the 'zero (trade flows) problem' using either Heckman or Poisson estimators.

As you read this guide, you will use STATA to carry out estimations designed to familiarize you with the software and, more importantly, the gravity model. STATA is a statistical software program and we assume that you have a recent version of STATA (version 11.2 or later). The instructions in this guide are quite detailed. Our aim is to give sufficient detail to enable a new user of this software to follow the examples relying solely on this guide. On the other hand, the guide is strictly related to the literature review that highlights the main theoretical and methodological issues illustrated in the regressions.

Data files

There are two Data files:

- dataset_def.dta: it contains all the essential variables used in the regressions using panel data (Part 1). The dataset covers the period from 1996 to 2006 and includes 154 developed and developing countries.
- us_agr.dta: it contains all the essential variables used in the regressions using cross-section data (Part 2 and Part 3). It refers to year 2004 US agricultural imports from 226 countries. Data are disaggregated at the most detailed level allowed by the international Harmonized System (HS) classification (6 digits) and include 689 products.

The variable names are largely self-explanatory and are described when the labels are created: their generation and construction can thus be directly inspected. The data sources are described in the Appendix.

Do files

There are three Do files:

- regressions aggregated data.do: it runs regressions using panel, aggregated data (Part 1).
- regressions disaggregated data.do: it runs regressions using cross-section, disaggregated as well as aggregated data (Part 2).
- regressions zeroes treatment.do: it runs regressions using non-linear estimators (Heckman or Poisson) dealing with 'zero' trade flows (Part 3).

Part 1: Aggregated data

A. Variable Generation.

Part A brings in the data and generates the variables used in the analysis:

(a) We use the data file dataset_def

use dataset_def.dta

(b) We take the logs of all continuous variables included in the regressions:

g limports=ln(imports)

g lgdp_o=ln(gdp_o)

g lgdp_d=ln(gdp_d)

g ldist= ln(distw)

g ltariff=ln(1+s_average)

(c) We label the variables to be included in the tables.

la var limports "Ln(Imports)"

la var colony "Colonial link"

la var comlang_off "Common language"

la var contig "Border"

la var ldist "Ln(distance)"

la var lgdp_d "Ln(GDP_importer)"

la var lgdp_o "Ln(GDP_exporter)"

la var rta "Regional Trade Agreement"

la var ltariff "Ln(1+Tariff)"

(d) Finally, we generate the different fixed effects.

qui tab imp, g(dimp)

qui tab exp, g(dexp)

qui tab pair, g(dpair)

qui tab year, g(dyear)

B. Regression Specifications

Part B runs panel regressions with aggregated data, and the dummy RTA (i.e., Regional Trade Agreements) as (trade) policy variable. Regressions are based on equation (2) with time, importer, exporter and country-pair fixed effects.

We start by declaring data to be panel.

tsset pair year

In order to show the consequences of ignoring the multilateral resistance term, we firstly estimate equation (2) without fixed effects

eststo: reg limports lgdp_d lgdp_o ldist contig colony comlang_off rta, robust

Then, we introduce the different types of fixed effects:

eststo: reg limports lgdp_d lgdp_o ldist contig colony comlang_off rta dyear, robust*

eststo: reg limports lgdp_d lgdp_o ldist contig colony comlang_off rta dimp dexp*, robust*

eststo: reg limports lgdp_d lgdp_o ldist contig colony comlang_off rta dimp dexp* dyear*,*

robust

eststo: reg limports lgdp_d lgdp_o ldist contig colony comlang_off rta dpair dyear*, robust*

The dummy for pair effects is equal to 1 for all observations of trade occurring between a given pair of countries, for all pairs. Country dummies remove cross-section, but not time-series biases. The latter is a serious shortcoming since omitted factors affecting bilateral trade costs often vary over time. Pair dummies cannot be used in cross-section data since the number of dummies would be equal to the number of observations.

The command "esttab" creates the regression table in a file regressions1.doc

esttab using regressions1.doc, title (aggregate-dummy policy) se ar2 label replace rtf b(2) star (0.10 ** 0.05 *** 0.01) se(2) mti drop (dexp* dimp* dyear* dpair*)*

TABLE: Panel results with different fixed effects

Model 1and 2 report the base regression. Column (1.1) reports results without fixed effects. Column (2.1) reports results where time dummies are added to the regression, to account for

the changing nature of the relationship over time. Column (2.2) and (2.3) show results for time invariant importer and exporter fixed effects and for time varying exporter and importer fixed effects, respectively. Finally, column (2.4) presents a specification where pair effects are also added.

Variables	1. Without FE	2. With FE			
	(1.1)	(2.1)	(2.2)	(2.3)	(2.4)
Ln(GDP_importer)	0.75***	0.74***	0.70***	0.91***	1.05***
	(0.02)	(0.02)	(0.14)	(0.20)	(0.15)
Ln(GDP_exporter)	1.20***	1.22***	0.01	0.08	0.18
	(0.01)	(0.01)	(0.12)	(0.14)	(0.12)
Ln(distance)	-1.50***	-1.55***	-1.47***	-1.47***	-1.59***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.23)
Dummy: Border	0.61***	0.54***	0.54***	0.54***	2.96***
	(0.10)	(0.10)	(0.11)	(0.11)	(0.85)
Dummy: Colonial link	-0.51**	-0.61***	0.09	0.08	6.39***
	(0.23)	(0.24)	(0.25)	(0.25)	(0.62)
Dummy: Common language	1.10***	1.21***	0.78***	0.78***	6.88***
	(0.06)	(0.06)	(0.09)	(0.09)	(0.40)
Dummy: Regional trade agreement	0.54***	0.71***	0.64***	0.62***	0.28
	(0.08)	(0.09)	(0.10)	(0.10)	(0.18)
Constant	6.51***	6.69***	22.37***	20.25***	6.63***
	(0.38)	(0.38)	(2.34)	(2.80)	(0.33)
Time fe	No	Yes	No	Yes	Yes
Exporter fe	No	No	Yes	Yes	No
Importer fe	No	No	Yes	Yes	No
Country-pair fe	No	No	No	No	Yes
Observations	7797	7797	7797	7797	7797
Adjusted R^2	0.605	0.619	0.737	0.739	0.881

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

All coefficients have the expected signs, the only exception is the coefficient of the colonial links that seem to have a negative impact if we do not consider country fixed effects (columns 1.1 and 2.1). Focusing on the most widely used specification (column 2.3), the estimated coefficients should be interpreted as follows: size of importer country has a positive and significant impact with an elasticity of 0.91, so that an increase in GDP of 10% increase trade by 9.1%; an increase in distance of 10% reduce trade by around 15%; the existence of border and language links imply an increase in trade of 72% and 118% ($e^{0.54}$-1=0.72; $e^{0.78}$-1= 1.18), respectively; the estimated coefficient of dummy for RTA of 0.62 implies that regional trade agreements increase trade of 86% ($e^{0.62}$-1=0.86).

Finally, in order to show the relevance of an actual measure we estimate equation (3) using a continuous variable for trade policy

reg limports lgdp_d lgdp_o ldist contig colony comlang_off ltariff dimp dexp* dyear*, robust*

Then we create the results table

esttab using regressions1.doc, title (aggregate-tariff) se ar2 label replac rtf b(2) star (0.10 ** 0.05 *** 0.01) se(2) mti drop (dexp* dimp* dyear*) append*

TABLE: Panel results with continuous policy variable

	(1)
Ln(Gdp_importer)	0.61***
	(0.20)
Ln(Gdp_exporter)	0.12
	(0.14)
Ln(distance)	-1.43***
	(0.04)
Dummy: Border	0.62***
	(0.11)
Dummy: Colonial link	0.05
	(0.24)
Dummy: Common language	0.74***
	(0.09)
Ln(1+tariff)	-0.57***
	(0.05)
Constant	22.87***
	(2.84)
Time fe	Si
Exporter fe	Si
Importer fe	Si
Observations	7797
Adjusted R^2	0.743

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

All coefficients have the expected signs. Trade policies have a negative and significant impact on trade, a tariff factor increase by 10% leads to a 6% reduction of trade.

Part 2: Disaggregated data

A. Variable Generation.

(a) We use the data file us_agr

use us_agr.dta

(b) We take the logs of all continuous variables included in the regressions

g limports=ln(trade)

g lgdp_o=ln(gdp_o)

g lgdp_d=ln(gdp_d)

g ldist= ln(distw)

g ltariff=ln(1+tariff)

(c) We label the variables to be included in the tables.

la var limports "Ln(Imports)"

la var colony "Colonial link"

la var comlang_off "Common language"

la var contig "Border"

la var ldist "Ln(distance)"

la var lgdp_d "Ln(GDP_importer)"

la var lgdp_o "Ln(GDP_exporter)"

la var ltariff "Ln(1+Tariff)"

(d) We generate the exporter and product fixed effects

qui tab exp, g(dexp)

qui tab hs6, g(dhs6)

B. Regression Specifications

(a) As in Part 1, we firstly estimate equation (2) using the OLS estimator without fixed effects

eststo: reg limports lgdp_o ldist contig colony comlang_off ltariff, robust

(b) Then, we introduce the different types of fixed effects:

eststo: reg limports lgdp_o ldist contig colony comlang_off ltariff dhs6, robust*

eststo: reg limports contig colony comlang_off ltariff dexp dhs6*, robust*

(c) Finally, we collapse the dataset in order to obtain aggregated data for a robustness analysis.

collapse (sum) trade (mean) tariff gdp_o distw contig colony comlang_off, by(exp)

g limports=ln(trade)

g lgdp_o=ln(gdp_o)

g ldist= ln(distw)

g ltariff=ln(1+tariff)

(d) and we run again the regression to highlight the relevance of the aggregation issue

eststo: reg limports lgdp_o ldist contig colony comlang_off ltariff, robust

(e) The command "esttab" creates the regression table in a file regressions2.doc

esttab using regressions2.doc, title (dati_us_agr) se ar2 label replac rtf b(2) star (0.10 ** 0.05 *** 0.01) se(2) mti drop (dexp* dhs6*) append*

eststo clear

TABLE: Cross-section results with different fixed effects and different levels of aggregation

Cross-sectional model covering imports in 689 agricultural commodities from 227 countries to US in 2004.

Columns (1) to (3) show results with disaggregated data using different fixed-effects specifications. Column (4) reports result of aggregated data.

	Disaggregation level: Hs6			Aggregated data
	(1)	(2)	(3)	(4)
Ln(Gdp_exporter)	0.34***	0.44***		1.07***
	(0.01)	(0.01)		(0.06)
Ln(distance)	-0.39***	-0.36***		-1.96***
	(0.05)	(0.05)		(0.40)
Dummy: Border	1.79***	1.92***	2.78***	-0.52
	(0.13)	(0.12)	(0.83)	(0.74)
Dummy: Colonial link	0.04	0.03	-0.56	-0.23
	(0.08)	(0.07)	(0.76)	(0.87)
Dummy: Common language	0.29***	0.38***	4.33***	1.09***
	(0.05)	(0.04)	(0.73)	(0.38)
Ln(1+tariff)	0.19	-2.85***	-2.00***	-5.00
	(0.16)	(0.41)	(0.41)	(10.85)
Constant	-4.34***	-11.38***	-7.40***	9.85***
	(0.42)	(0.40)	(2.52)	(3.38)
Product (HS6) fe	No	Si	Si	-
Exporter fe	No	No	Si	No
Observations	20902	20902	21136	176
Adjusted R^2	0.113	0.254	0.303	0.602

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Most empirical analyses use gravity models with aggregated data, but using aggregate trade flows to analyze the effects of trade policies applied at product level seems misleading. As a matter of fact, the estimated coefficient related to trade policy, namely *Ln*(*1+tariff*), is not significant in column (4).

Part 3: Zeroes treatment

A. Variable Generation

(a) We use the data file us_agr

use us_agr.dta

We take the logs of all continuous variables included in the regressions

g limports=ln(trade)

g lgdp_o=ln(gdp_o)

g lgdp_d=ln(gdp_d)

g ldist= ln(distw)

g ltariff=ln(1+tariff)

(c) We label the variables to be included in the tables.

la var limports "Ln(Imports)"

la var colony "Colonial link"

la var comlang_off "Common language"

la var ldist "Ln(distance)"

la var lgdp_d "Ln(GDP_importer)"

la var lgdp_o "Ln(GDP_exporter)"

la var ltariff "Ln(1+Tariff)"

(d) We generate the exporter and product fixed effects

qui tab exp, g(dexp)

qui tab hs6, g(dhs6)

B. Regression Specifications

(a) We firstly run the regression using the Heckman estimator

eststo: heckman limports contig colony comlang_off ltariff , select(contig colony ltariff) mills(lambda)

(b) Then we run the regression using the Poisson Pseudo-Maximum Likelihood estimator

eststo: ppml trade contig colony comlang_off ltariff

(c) The command “esttab” creates the regression table in a file regressions3.doc

esttab using regressions3.doc, title (treatment of zeros) se ar2 label replace rtf b(2) star (0.10 ** 0.05 *** 0.01) se(2) mti*

TABLE: Results with different estimators: Heckman and Poisson

Model (1) reports results obtained using the Heckman two-step procedure. The first column shows the second stage estimates of the trade flow, whereas the second column reports the first-stage Probit selection equation. Model (2) shows results obtained using Poisson Pseudo-Maximum Likelihood estimator.

	Heckman (1)		PPML
	Regression	Selection	(2)
Dummy: Border	2.67***	1.95***	3.20***
	(0.18)	(0.06)	(0.01)
Dummy: Colonial link	0.76***	0.68***	1.30***
	(0.11)	(0.03)	(0.01)
Dummy: Common language	-0.09**		0.12***
	(0.04)		(0.01)
Ln(1+tariff)	0.25	-0.04	-0.58***
	(0.16)	(0.04)	(0.04)
Constant	-3.26***	-0.31***	-0.52***
	(0.18)	(0.01)	(0.01)
lambda	-0.58**		
	(0.18)		
Observations	52340		52340

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The use of disaggregated data raises the "zero trade flows" issue, which introduces obvious problems in the log-linear form of the gravity equation. Several authors consider the Heckman two-step estimator as the best procedure, others argue that gravity type models should be estimated in multiplicative form, and recommend the Poisson Pseudo-Maximum Likelihood (PPML) estimator to deal with the problem of zeros in the trade matrix, in order to achieve unbiased and consistent estimates. The significant coefficient of the Mills ratio confirms that correcting for sample selection bias is justified, however, because of the presence of heteroskedasticity, estimates of the log-linear form of the gravity equation are biased and inconsistent, and this may lead to prefer the Poisson specification of the trade gravity model.

APPENDIX: Data Source

Dataset_def.dta: Dataset is building on extraction from WITS (http://wits.worldbank.org/wits/index.html) and on information provided by the CEPII dataset (http://www.cepii.fr/).

The WITS application gives access to international trade statistics of UN COMTRADE (The United Nations Commodity Trade Statistics database) and tariff database of UNCTAD-TRAINS (Trade Analysis and Information System).

The Cepii dataset includes data on GDP and distances between countries and dummies for contiguity, common language, and former colonial links.

Us_agr.dta: Data on trade and tariffs at the HS6 level of detail are taken from the MAcMapHS6-V2 database (http://www.ifpri.org/publication/picture-tariff-protection-across-world-2004). MAcMapHS6 provides a consistent worldwide assessment of protection, including ad valorem equivalent rates of specific duties and tariff rate quotas (including those introduced at the end of the Uruguay Round), for 2004. Data for the remaining explanatory variables are from the Cepii dataset.