Session 2: Estimating the Basic Gravity Model

Ben Shepherd

Trade Economist & International Development Consultant
www.Developing-Trade.com

ARTNeT Capacity Building Workshop for Trade Research:
“Behind the Border” Gravity Modeling

Monday, September 21, 2009
Outline

1 Introduction

2 The Gravity Model as a Regression Problem
   - Setting up the Problem
   - Statistical Properties of the OLS Estimator
   - Making OLS “Talk”

3 Summary
We have seen that the basic elements of the gravity model accord with economic sense, and agree with some important stylized facts.

However, our analysis was only based on bivariate analysis, i.e. not controlling for the impact of GDP when assessing the impact of distance, and vice versa.

In this session, we look at how to estimate the basic model more rigorously. We then look at how theory influences estimation.
Outline

1. Introduction

2. The Gravity Model as a Regression Problem
   - Setting up the Problem
   - Statistical Properties of the OLS Estimator
   - Making OLS “Talk”

3. Summary
Recall the log-linear gravity model discussed in the previous session:

\[
\log(X_{ij}) = b_0 + b_1 \log(Y_i) + b_2 \log(Y_j) + b_3 \log(d_{ij}) + e_{ij}
\]

\[b_1, b_2 > 0; b_3 < 0\]

By taking the model to the data, we would like to find out the following:

- How well do distance and GDP explain bilateral trade flows?
- Do the data support the expected coefficient signs?
- How sensitive is bilateral trade to distance and GDP, controlling for the simultaneous influence of the other?
The Gravity Model as a Regression Problem

Summary

Setting up the Problem

Statistical Properties of the OLS Estimator

Making OLS “Talk”

Introduction

The Gravity Model

Setting up the Problem

The Gravity Model

\[ \log (X_{ij}) = b_0 + b_1 \log (Y_i) + b_2 \log (Y_j) + b_3 \log (d_{ij}) + e_{ij} \]

To get the information we want from the data, we need a method for estimating the \( b \) parameters.

One sensible candidate, and the usual place to start, is ordinary least squares (OLS):

- Choose \( b_0 \), \( b_1 \), and \( b_2 \) so as to minimize the sum of squared errors \( \sum_i \sum_j e_{ij}^2 \).
Outline

1. Introduction

2. The Gravity Model as a Regression Problem
   - Setting up the Problem
   - Statistical Properties of the OLS Estimator
   - Making OLS “Talk”

3. Summary
The solution to the OLS minimization problem, $\hat{B}$ is not just a sensible set of slope coefficients, in the sense of satisfying an intuitive criterion.

Under specific assumptions as to the properties of the random error term $E$, the OLS estimator also has some very useful statistical properties.

We will make extensive use of these properties to:

- Argue that our estimates are reasonable and reliable; and
- To conduct formal tests of interesting economic hypotheses.
Specifically, OLS is consistent, unbiased, and efficient as an estimator of $\mathbf{B}$ if the following conditions hold:

- None of the dependent variables are perfectly correlated (multicollinearity).
- $\mathbf{E}$ is an independently distributed normal error with mean zero, and with constant variance (homoskedasticity).
- The underlying model relating the dependent and independent variables is linear.
- $\mathbf{E}$ is uncorrelated with any of the independent variables.

If these conditions hold, we can be confident that our estimates are reliable, and that hypothesis tests are informative. If they do not, we cannot!
If the OLS assumptions hold, we can also use the model for hypothesis testing:

- The estimated coefficients approximately normally distributed, with standard errors that can be easily calculated.
- We can test hypotheses on a particular variable using its t-statistic.
- We can test compound linear hypotheses (more than one variable) using the F-statistic.
We can use $R^2 = 1 - \frac{SSR}{TSS}$ as a simple measure of how well the model “fits” the data:

- A higher score is better. (SSR is the sum of squared residuals, and TSS is total sum of squares.)
- Interpret it as the percentage of observed variation in the dependent variable that is accounted for by the model.

Ramsey’s RESET test is a good workhorse test for model specification (possible non-linearities) and omission of important variables:

- Get fitted values from the gravity model, $\hat{X}$.
- Calculate $\hat{X}^2$, $\hat{X}^3$, and $\hat{X}^4$.
- Include them as additional regressors, and check joint significance using an F-test.
- H0: the model is correctly specified. A large test statistic indicates a problem.
Not all violations of the OLS conditions are equal:

- Heteroskedasticity is usually relatively minor, and easily dealt with: always use a “robust” estimator for the variance-covariance matrix. But we will investigate a situation later where it is much more sinister...
- One type of correlation in the error terms can be fixed by using the “cluster ()” option, and specifying the highest level of data aggregation.
- Perfect multicollinearity rarely arises in practice, although it can have some implications in panel data models with fixed effects.

For the purposes of this workshop, we will be spending the most time dealing with violations of the last two conditions: they have serious consequences (bias and inconsistency), and are much harder to fix.
Introduction

The Gravity Model as a Regression Problem

Setting up the Problem
Statistical Properties of the OLS Estimator
Making OLS “Talk”

Summary

Outline

1. Introduction

2. The Gravity Model as a Regression Problem
   - Setting up the Problem
   - Statistical Properties of the OLS Estimator
   - Making OLS “Talk”

3. Summary
We have seen where OLS comes from, and have gotten a first taste of when it can be most useful.
Let’s use STATA to estimate a very simple gravity model by OLS, and then focus on interpreting the results (making it “talk”).
We are interested in:
- Giving economic interpretations to parameter estimates;
- Testing simple (one parameter) hypotheses;
- Testing compound (multiple parameter) hypotheses; and
- Assessing how well the model fits the data.
A Simple OLS Gravity Model

- Data on history and geography (international distance, colonial links, natural barriers, language, etc.). Source: CEPII.
- Data on exporter and importer GDP. Source: World Bank (WDI).
A Simple OLS Gravity Model

Stata Output: `reg ln_trade ln_gdp_imp [etc.], robust cluster(dist)`

| Variable    | Coef.   | Std. Err  | t      | P>|t| |
|-------------|---------|-----------|--------|------|
| ln_imports  | 1.030655| 0.0083676 | 123.17 | 0    |
| ln_gdp_imp  | 1.232748| 0.008379  | 147.12 | 0    |
| ln_gdp_exp  | -1.277118| 0.0244743 | -52.18 | 0    |
| colony      | 0.9585085| 0.1147203 | 8.36   | 0    |
| comcol      | 1.055211| 0.0829653 | 12.72  | 0    |
| comlang_off | 0.9042383| 0.0593294 | 15.24  | 0    |
| contig      | 0.8106201| 0.1359654 | 5.96   | 0    |
| _cons       | -36.79228| 0.4201254 | -87.57 | 0    |
A Simple OLS Gravity Model

Market Size Effects

|                  | Coef.    | Std. Err | t        | P>|t| |
|------------------|----------|----------|----------|-----|
| ln_exports       | 1.030655 | 0.0083676| 123.17   | 0   |
| ln_gdp_exp       | 1.232748 | 0.008379 | 147.12   | 0   |

- A 1% increase in importer size is associated with about a 1% increase in bilateral trade.
- A 1% increase in exporter size is associated with a bit over a 1% increase in bilateral trade.
- Both effects are statistically significant at the 1% level, i.e. it is unlikely that either effect is “really” zero given the data we have.
A Simple OLS Gravity Model

Interpretation

<table>
<thead>
<tr>
<th>Geography Effects</th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln_imports</td>
<td>Coef.</td>
</tr>
<tr>
<td>ln_dist</td>
<td>-1.277118</td>
</tr>
<tr>
<td>contig</td>
<td>0.8106201</td>
</tr>
</tbody>
</table>

| ln_imports        | Std. Err       | t     | P>|t| |
|-------------------|---------------|-------|------|
| ln_dist           | 0.0244743     | -52.18| 0    |
| contig            | 0.1359654     | 5.96  | 0    |

- A 1% increase in the distance between markets is associated with just over a 1% decrease in bilateral trade.
- Sharing a common border is associated with about a 120% increase in bilateral trade ($e^{0.8} - 1 \approx 1.2$).
- Both effects are statistically significant at the 1% level, i.e. it is very unlikely that either effect is “really” zero given the data we have.
### History Effects

| In_imports     | Coef.       | Robust Std. Err | t    | P>|t| |
|----------------|-------------|-----------------|------|-----|
| colony         | 0.9585085   | 0.1147203       | 8.36 | 0   |
| comcol         | 1.055211    | 0.0829653       | 12.72| 0   |
| comlang_off    | 0.9042383   | 0.0593294       | 15.24| 0   |

- Having once been a colony is associated with an increase in bilateral trade with the colonizer of about 160%.
- A common colonizer is associated with an increase in bilateral trade of about 190%.
- A common official language is associated with an increase in bilateral trade of about 150%.
In drawing conclusions about the statistical significance of particular parameters, we have been conducting simple hypothesis tests.

We can also test some compound hypotheses:

- Both GDP coefficients are zero: $F=13736.88$, $P<0.01$
- The two GDP coefficients are both equal to one: $F=407.01$, $P<0.01$
- All geographical and historical variables have zero coefficients: $F=866.77$, $P<0.01$
Stata reports an $R^2$ of 0.6627, i.e. the model accounts for about $\frac{2}{3}$ of the observed variance in the log of bilateral trade.

However, the Ramsey RESET test strongly rejects $H_0$:

- $F = 206.59$
- $Pr. < 0.01$

This suggests that there might be more to bilateral trade than meets the eye.

What other variables might be influencing bilateral trade, but are left out of our simple model?

Are there any non-linearities we need to worry about?
Summary

- The basic gravity model is a sensible, intuitive place to start.
- When estimating, always:
  - Use a robust variance-covariance estimator
  - Adjust the standard errors for clustering at the highest level of aggregation in the data. Usually, clustering by country pair is sufficient in gravity models.
- The trick is in getting OLS to “talk”, i.e. tell us interesting things about the determinants of bilateral trade.