Linders, Gert-Jan M.; de Groot, Henri L.F.

Working Paper

Estimation of the Gravity Equation in the Presence of Zero Flows

Tinbergen Institute Discussion Paper, No. 06-072/3

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Linders, Gert-Jan M.; de Groot, Henri L.F. (2006) : Estimation of the Gravity Equation in the Presence of Zero Flows, Tinbergen Institute Discussion Paper, No. 06-072/3, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/86589

TI 2006-072/3
Tinbergen Institute Discussion Paper

Estimation of the Gravity Equation in the Presence of Zero Flow

Gert-Jan M. Linders
*Henri L.F. de Groot**

Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam.

* *Tinbergen Institute.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Estimation of the gravity equation in the presence of zero flows

Gert-Jan M. Linders[a, *] and Henri L.F. de Groot[b]

[a] Department of Spatial Economics, Vrije Universiteit Amsterdam

De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

[b] Department of Spatial Economics, Vrije Universiteit Amsterdam and Tinbergen Institute

De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Abstract

The gravity model is the workhorse model to describe and explain variation in bilateral trade patterns. Consistent with both Heckscher-Ohlin models and models of imperfect competition and trade, this versatile model has proven to be very successful, explaining a large part of the variance in trade flows. However, the log-linear model cannot straightforwardly account for the occurrence of zero-valued trade flows between pairs of countries. This paper investigates the various approaches suggested to deal with zero flows. Apart from the option to omit the zero flows from the sample, various extensions of Tobit estimation, truncated regression, probit regression and substitutions for zero flows have been suggested. We argue that the choice of method should be based on both economic and econometric considerations. The sample selection model appears to fit both considerations best. Moreover, we show that the choice of method may matter greatly for the results. In the end, the results surprisingly suggest that the simplest solution, to omit zero flows from the sample, often leads to acceptable results, although the sample selection model is preferred theoretically and econometrically.

JEL codes: F14, F15

Keywords: bilateral trade, gravity model, zero flows, sample selection model

[*] Corresponding author: G.J.M. Linders, tel.: +31-20-5986198; e-mail: glinders@feweb.vu.nl. The authors thank Jaap de Vries, Andrés Rodríguez-Pose, Maria Abreu, Johannes Voget, Thomas de Graaff, Aart de Vos, and participants at the NAKE seminar 2005 and FADO seminar for valuable comments and suggestions.

1 Introduction

The gravity model has become the workhorse model to analyze patterns of bilateral trade (Eichengreen and Irwin, 1998). Originally inspired by Newton's gravity equation in physics, the gravity model has become common knowledge in regional science for describing and analyzing spatial flows, and was pioneered in the analysis of international trade by Tinbergen (1962), Pöyhönen (1963) and Linneman (1966). The model works well empirically, yielding sensible parameter estimates and explaining a large part of the variation in bilateral trade (Rose, 2005). However, it has long been disputed for a lack of theoretical foundation. More recently, the gravity model has made a comeback in the international trade literature. Developments in the modelling of bilateral trade that provided the model with a more satisfying theoretical underpinning in trade theory have been crucial in this revival (see, e.g., Feenstra, 2004, and Anderson and Van Wincoop, 2004, for an overview).

In conjunction with the expanding theoretical literature on the gravity model, a number of recent contributions have addressed issues concerning the correct specification and interpretation of the gravity equation in empirical estimation. These deal with, for example, the specification of panel gravity equations, the estimation of cross-section gravity equations, and the correct interpretation of the distance effect on patterns of bilateral trade (e.g., Buch et al., 2004, Egger, 2000, Egger and Pfaffermayr, 2005, and Matyas, 1998). All in all, these developments have improved our understanding of the gravity equation as a tool to model and analyze bilateral trade patterns. However, a number of questions with regard to bilateral trade and the gravity equation remain to be investigated (see Anderson and Van Wincoop, 2004). One of these is the question how to deal with zero-valued bilateral trade flows. The standard gravity model cannot easily deal with zero flows. This has resulted in a widespread practice in the literature to ignore zero flows in the analysis of bilateral trade. However, zero-valued observations contain important information for understanding the patterns of bilateral trade, and should not be discarded *a priori*.

This paper deals with the question how to amend the gravity model in order to be able to deal with zero flows. Section 2 describes the gravity equation that we estimate to analyze

bilateral trade, and the data set used in the analysis. Section 3 discusses the theoretical and econometric problems for the gravity model generated by the occurrence of zero flows, and presents an overview of the solutions commonly proposed and applied in the literature. We argue that these solutions are at odds with both a sound theoretical treatment of zero flows in the gravity model and with proper econometric modeling of zero flows in bilateral trade. In Section 4, we propose an alternative method to deal with zero-valued trade flows. The sample selection model, which has been widely used in other fields of applied economics, is rather novel to the literature on bilateral trade. Because the sample selection model offers a theoretically sound and econometrically elegant solution to include zero flows in the gravity model of bilateral trade, it deserves more attention in applied work. Section 5 presents empirical results of estimating a sample selection model of bilateral trade. Moreover, we compare the results to various alternative approaches suggested to address zero flows in bilateral trade, thus providing an explicit check of the sensitivity of the empirical outcomes to the approach chosen. This allows us to asses whether the general consensus in the literature that zero flows do not have much impact on the estimation results (see, e.g., Baldwin, 1994 and Frankel, 1997) is corroborated. Finally, Section 6 discusses our main findings, and provides some conclusions.

2 The gravity model

The gravity model relates bilateral trade flows to the GDP levels of the countries and their geographic distance. GDP reflects the market size in both countries, as a measure of 'economic mass'. The market size of the importing country reflects the potential demand for bilateral imports, while GDP in the exporting country represents the potential supply and diversity of goods from that country; geographic distance reflects resistance to bilateral trade. Usually, the gravity equation is expressed in logarithmic form. We will follow the literature in extending the basic gravity equation with several variables that proxy different aspects of economic distance. These comprise, among others, dummies for common language and colonial history, which capture cultural familiarity, a dummy for membership in a common

trade bloc that reflects economic integration, and a religion dummy that indicates similarity in cultural values and norms. The benchmark version of the gravity equation estimated below looks as follows:

$$\begin{aligned}\ln(T_{ij}) &= \beta_0 + \beta_1 \ln(Y_i) + \beta_2 \ln(Y_j) + \beta_3 \ln(D_{ij}) + \beta_4 Adj_{ij} \\ &+ \beta_5 RIA_{ij} + \beta_6 Lan_{ij} + \beta_7 Col_{ij} + \beta_8 Rel_{ij} + \varepsilon_{ij}\end{aligned}, \quad (1)$$

where ε_{ij} is a stochastic disturbance term that is assumed to be well-behaved. The dependent variable T_{ij} is merchandise exports (in '000 US$) from country *i* to *j*, for 1999. The independent variables are: GDP (*Y*), the distance between *i* and *j* (D_{ij}) and dummies reflecting whether *i* and *j*: share a land border (*Adj*), are both member in a regional integration agreement (*RIA*), have the same primary language (*Lan*) or were part of a common colonial empire (*Col*), and whether they share the same main religion (*Rel*). The data set comprises 127 countries. For further details on the variables and countries in our data set, see Appendix B.

3 Dealing with zero flows

The gravity model predicts that countries have positive trade in both directions, even if this predicted trade may be small. Moreover, the conventional log-linear formulation of the gravity model cannot include zero-valued bilateral trade flows, because the logarithm of zero is undefined. However, in our data set of bilateral trade, some of the trade flows are recorded as zero or missing.[1] At the aggregate level, zero flows mostly occur for trade between small or distant countries, which are expected to trade little (Frankel, 1997). However, disregarding zero flows can bias the empirical results, if they do not occur randomly. Specifically, if geographic distance, low levels of national income, and a lack of cultural or historical links reduce trade, omitting zero flows from the analysis tends to result in an underestimation of the

[1] Most of these flows are recorded as missing in the source database (UN COMTRADE); some have explicitly been recorded as zero. We assume that all missing observations in principle indicate that bilateral exports are considered to be absent by the reporting country. Countries that do not report any trade statistics in the database have been omitted from our sample.

effects of these variables on trade (see Rauch, 1999, pp. 18–19). Omitting zero-flow observations implies that we loose information on the causes of (very) low trade.

Several approaches have been applied or suggested in the literature to address the problem of zero flows (see, e.g., Frankel, 1997, pp. 145–146; Bikker, 1982, pp. 371–372). The most common solution in the literature confines the sample to non-zero observations to avoid the estimation problems related to zero flows. Alternatively, (part of the) zero values may be substituted by a small constant, so that the double-log model can be estimated without throwing these country pairs out of the sample. Examples in the literature that followed this approach are Linnemann (1966), Van Bergeijk and Oldersma (1990), Wang and Winters (1991) and Raballand (2003). Substituting small values prevents omission of observations from the sample, but is essentially ad hoc. The inserted value is arbitrary and does not necessarily reflect the underlying expected value. Thus, inserting arbitrary values close to zero does not provide any formal guarantee that the resulting estimates of the gravity equation are consistent. Both approaches are hence generally unsatisfactory.

Dealing properly with zero flows requires that the information provided by these flows is taken into account, without using ad-hoc methods. The censored regression model (Tobit model) is often employed to analyse data sets in which a substantial fraction of the observations cluster at zero. Several studies have used the standard Tobit model to estimate the gravity equation with zero flows (e.g., Rose, 2004; Soloaga and Winters, 2001; Anderson and Marcouiller, 2002). The Tobit model describes a situation in which part of the observations on the dependent variable is censored (unobservable) and represented instead by mapping them to a specific value, generally zero. The model applies to situations in which outcomes cannot be observed over some range, either because actual outcomes cannot reflect desired outcomes (e.g., actual outcomes cannot be negative), or because of measurement inaccuracy (e.g., rounding). Thus, whether the Tobit model can be applied to study zero flows in the conventional gravity framework depends on two questions. Firstly, ‘Can desired trade be negative?’ and secondly, ‘Is rounding of trade flows an important concern?’.

The gravity model as conventionally specified under the assumption of a log-normally distributed disturbance term would only predict zero trade if the GDP of one or both countries equaled zero. This is a hypothetical situation, of course, which will not occur in practice.[2] If we specified the gravity model with an additive, normally distributed disturbance term, instead of a log-normal error structure, the gravity model could in principle generate negative trade, by means of the random error. This negative trade would then be censored at zero, and actual zero trade might reflect desired negative trade. Note, however, that the underlying expected trade determined by the gravity model can never be negative. This non-stochastic part of the gravity model can be consistently derived from economic optimization (see, e.g., Deardorff, 1998, and Feenstra, 2004). The disturbance term allows for optimization outcomes that differ randomly from the expected outcome, but it is unclear which optimizing framework would justify negative desired trade, even if caused by randomly distributed factors not explicitly identified in the model.[3] We thus answer the first question negatively: desired trade cannot be negative. Rounding to zero of trade flows below some positive value is a second possible reason for censoring of trade flows. In this case, the Tobit model with a positive threshold value would be appropriate. However, censoring of trade flows from below in general does not seem to occur in our data set. Trade flows are reported in the COMTRADE database up to an accuracy of US$ 1 (although this differs somewhat across countries). Therefore, the second question regarding the suitability of censored regression can be answered negatively as well. As a consequence, the Tobit model is not the appropriate model to explain why some trade flows are missing.

Given that the conventional gravity model does not predict zero-valued bilateral trade nor desired negative trade, and in the absence of rounding below some positive value, zero flows have to be interpreted otherwise. In this context, zero flows result from binary decision

[2] One could imagine this to describe the tautological situation of trade with an uninhibited island, which would be zero almost by definition.

[3] In fact, this suggests that an additive disturbance term might better be regarded as truncated from below. Zero flows then always represent desired zero flows, and the model is consistent with economic optimization. However, this solution does not accord with the Tobit model anymore.

making rather than censoring (Sigelman and Zeng, 1999). The appropriate way to proceed, then, is “to model the decisions that produce the zero observations rather than use the Tobit model mechanically” (Maddala, 1992, cf. Sigelman and Zeng, 1999, p. 170). This can be done by modelling the decision whether or not to trade as a Probit model. The outcome of that decision determines whether or not we observe actual trade flows in the sample. The size of potential trade is determined by the gravity model. This structure has been framed in the sample selection model (see, e.g., Greene, 2000, section 20.4; Verbeek, 2000, section 7.4), to which we will now turn for a solution to the problems associated with zero flows in a gravity model context.

4 The sample selection model

The model, also known as the Heckman selection model (Heckman, 1979), is often used in microeconometric research, especially in labour economics. Its use can be traced back, for example, to Gronau (1974). A rather small number of gravity model studies of bilateral trade have used the selection model to deal with zero flows. For example, Bikker (1982) and Bikker and De Vos (1992) make extensive use of a selection model, similar to the one used here. Rose (2000) estimates a variant of the model in a robustness section of the paper, without explicating the model. Hillberry (2002) motivates and estimates a more restricted variant, in which an independent selection and, as he prefers to call it, truncated regression equation are estimated (cf. Cragg, 1971). The sample selection model of bilateral trade is specified as follows:

Selection mechanism:

$$\tilde{\pi}_{ij=}\gamma_0 + \gamma_1 \ln(Y_i) + \gamma_2 \ln(Y_j) + \gamma_3 \ln(y_i) + \gamma_4 \ln(y_j) + \gamma_5 \ln(D_{ij}) + \gamma_6 Adj_{ij}$$
$$+\gamma_7 RIA_{ij} + \gamma_8 Lan_{ij} + \gamma_9 Col_{ij} + \gamma_{10} Rel_{ij} + \gamma_{11} IQ_i + \gamma_{12} IQ_j + \gamma_{13} ID_{ij} + \mu_{ij}$$
$$s_{ij} = 1 \quad \text{if } \tilde{\pi}_{ij} > 0$$
$$s_{ij} = 0 \quad \text{if } \tilde{\pi}_{ij} \leq 0$$

Regression model:

$$\ln(\tilde{T}_{ij}) = \beta_0 + \beta_1 \ln(Y_i) + \beta_2 \ln(Y_j) + \beta_3 \ln(y_i) + \beta_4 \ln(y_j) + \beta_5 \ln(D_{ij}) + \beta_6 Adj_{ij} \qquad (2)$$
$$+\beta_7 RIA_{ij} + \beta_8 Lan_{ij} + \beta_9 Col_{ij} + \beta_{10} Rel_{ij} + \beta_{11} IQ_i + \beta_{12} IQ_j + \beta_{13} ID_{ij} + \varepsilon_{ij}$$

$$\ln(T_{ij}) = \ln(\tilde{T}_{ij}) \text{ if } s_{ij} = 1$$
$$\ln(T_{ij}) = \text{not observed} \quad \text{if } s_{ij} = 0$$

$$(\mu_{ij}, \varepsilon_{ij}) \sim \text{bivariate normal}[0,0,1,\sigma_\varepsilon^2, \rho_{\varepsilon\mu}].$$

The model in equation (2) can be estimated using Maximum Likelihood (ML) estimation (for further details, see Appendix A). The selection equation determines whether or not we observe bilateral trade between two countries in the sample. The regression model determines the potential size of bilateral trade. In general, the selection equation should at least contain all variables that are reflected in the regression equation (Verbeek, 2000). We assume that the selection process reflects decisions made at the microeconomic level on the basis of comparing costs and benefits of bilateral transactions (see Bikker and De Vos, 1992). Anderson and Van Wincoop (2004) point at the importance of fixed costs associated with international trade to explain zero flows in trade, such as border costs (Hillberry, 2002), search costs and other specific investments to enter foreign markets (Romer, 1994). At the macroeconomic level, we assume an underlying latent variable, say profitability, which depends on the same variables as the gravity equation. This can be motivated by the fact that profitability will generally increase if the potential size of trade gets larger. However, this does not imply that profitability only reflects the potential size of the flow. For example, some variables may be more important in determining the profitability of flows rather than the

potential size of these flows. Moreover, the disturbance term of the selection equation will capture all (microeconomic) factors that influence profitability of bilateral transactions. Therefore, we expect that the coefficients in the selection and regression equation will not perfectly match and that the correlation between the disturbance terms will be positive, but not necessarily one.[4]

The basic idea behind the sample selection model is as follows. If a variable such as geographic distance becomes so small that firms decide to stop exporting to a country, because it is no longer profitable, we do not observe potential bilateral trade. Therefore, OLS regression for the observed data on bilateral trade could underestimate the effect of distance, if the correlation between the disturbance terms of both equations in the selection model is positive (cf. Verbeek, 2000, p. 207). Those trade flows that we do observe for small distances will have a positive value for the disturbance term in the selection equation, μ_{ij}, in order for the selection decision to be positive. Because of the positive correlation, $\rho_{\varepsilon\mu}$, the expected disturbance term in the regression model, ε_{ij}, will be positive as well. As a result, observed trade will be expected to be higher than potential trade, which is unconditional on being observed or not. The observed sample will be biased upward at low levels of geographic distance, and OLS estimates of the regression coefficients, for the observed sample of positive trade, will be biased toward zero if $\rho_{\varepsilon\mu} > 0$. The two-staged sample selection model takes this into account, by controlling for what is technically known as sample selection bias. Thus, the

[4] As noted by Bikker and De Vos (1992), for $\gamma_k = \beta_k / \sigma_\varepsilon, k \in \{1,..K\}$, $\gamma_0 = (\beta_0 - c)/\sigma_\varepsilon$ (where c is the censoring limit in the Tobit model for logged trade), and $\rho_{\varepsilon\mu} = 1$, the sample selection model transforms into the Tobit model (see also Verbeek, 2000, and Greene, 2000 for similar observations for the standard Tobit model). The only difference between the sample selection model and the conventional Tobit model, in this case, is that the selection equation has a variance normalized to one and includes a linear transformation with the censoring threshold, because the selection limit is set at zero. Because, in the Tobit model, the latent selection variable and the potential size of the action are perfectly correlated, we can map the latent variable to the observed variable and do not need to normalize the selection equation. Note that, if the estimated sample selection model would (approximately) lead to the relations regarding parameters and cross-equation correlation as put forward here, we would observe trade *as if* it were censored at a positive value. Strictly speaking, this is not a case of censoring, because the observed sample is not limited by non-observability (e.g., due to rounding) of trade below this value.

sample selection model allows us to tackle the problem, noted earlier in the paper, that disregarding zero flows may lead to an underestimation of the regression coefficients of, e.g., distance and GDP.

5 Empirical results

The previous sections have argued that, on theoretical grounds, the sample selection model is preferred to other approaches often used in the literature to deal with zero flows, such as censored regression (Tobit), truncated regression, and substitution of arbitrary small values. This section estimates the gravity equation using these different approaches for zero flows, to assess the sensitivity of the results for using different methods.

The regression results presented in Table 1 compare the various solutions for dealing with zero flows. The first specification represents simple OLS regression on a sample excluding the zero flow observations. All variables have the expected sign, and are highly significant statistically. These findings are in line with the existing literature. Trade increases with GDP and decreases with physical distance.Common language, common border, and trade agreement, as proxies for proximity, positively affect trade.

Specification (2) represents the sample selection model set forward in the previous section. Column (2a) presents the regression equation, and column (2b) the corresponding selection equation. The results are surprisingly similar to the straight OLS results. There is only marginal indication that OLS is biased downwards due to sample selection bias. The correlation between both stages in the selection model ($\rho_{\varepsilon\mu}$) is positive, as expected, but small (although significantly different from zero at $p<0.05$). The impact of some independent variables in the selection stage is quite comparable to the regression stage, after correcting for the re-scaling involved in the selection stage (see footnote 5). This implies that the effect of these variables on the expected potential size of bilateral trade corresponds to their effect on expected profitability. However, this does not hold for several regressors, notably adjacency, language, religion and common trade bloc membership. These findings suggest that the extent

of sample selection bias is relatively small, and that, apart from its theoretical unsuitability, the Tobit model is not supported as a reduced form either.

Specification (3) shows the results of a Tobit estimation that imposes artificial censoring on our trade data. A possible advantage of artificially censoring positive but small trade flows is that these flows are relatively prone to measurement errors, and may be too influential in the regression analysis (Frankel, 1997; Rose, 2000). We have substituted 1 (=$1000) for the zeros, and subsequently put the censoring limit to ln(1)=0, censoring all flows below $1000 including the zero observations. The imposed censoring limit is arbitrary, because of the absence of actual rounding of trade flows. Therefore, even though we treat the zero flows as if they were censored, there is no direct causal relation between the zero flows and the imposed censoring limit. The parameter estimates generally tend to overestimate the results from the sample selection model. This reflects that maximizing the Tobit likelihood function implies that the expected value for all zero flows is forced as closely as possible to (or below) $1000. Clearly, this value is arbitrary and not representative for all zero flows.

Specification (4) uses truncated regression. All actual flows (including the zero flows) below $1000 are truncated from the sample. This approach disregards all truncated flows, and captures that the flows observed just above the truncation limit will on average have positive disturbance terms. As a result, it should correct for a downward bias in OLS estimation. The outcomes from truncated regression (4) are more in line with the Heckman results than the corresponding Tobit model in specification (3), because they are not burdened with the zero flows that are ill-fit to the imposed censoring or truncation limit. However, truncated regression does not appear to correct sufficiently for the selection bias that results from the arbitrarily imposed truncation at $1000. The estimates are lower in absolute terms than the benchmark estimates in specifications (1) and (2).

The final specification (5) in Table 1 performs OLS after substituting an arbitrary, small value for all zero flows. As argued before, OLS in a sample that excludes zero flows yields inconsistent estimates that are biased towards zero. Therefore, it is not straightforward which value (or values) should be substituted for zero flows to best correct for sample selection bias.

To correct for the downward bias in OLS estimators, we have chosen to substitute a single, small value for zero flows. We arbitrarily opt for the smallest integer value recorded in the COMTRADE database, viz. $1. The results in Table 1 illustrate, however, that the approach leads to an overcorrection of the assumed bias. Most parameter estimates are unrealistically high in absolute terms, and overestimate the benchmark results from the sample selection model. Of course, the results from this approach are not robust to the value chosen to substitute for zeros.

Table 1: Estimation Results

	(1) OLS	(2a) Heckman: regression	(2b) Heckman: selection	(3) Censored at $1000	(4) Truncated at $1000	(5) OLS: $1 for zeros
Log GDP exporter	1.23***	1.24***	0.49***	1.48***	1.17***	1.76***
	(133.93)	(139.90)	(41.81)	(140.56)	(137.26)	(129.73)
Log GDP importer	1.01***	1.02***	0.40***	1.21***	0.97***	1.45***
	(109.45)	(114.37)	(37.87)	(116.03)	(113.58)	(106.82)
Log Distance	–1.12***	–1.14***	–0.46***	–1.39***	–1.09***	–1.68***
	(50.08)	(50.95)	(17.08)	(49.71)	(52.39)	(47.98)
Border Dummy	0.93***	0.92***	–0.36	0.69***	0.85***	0.51**
	(7.25)	(7.13)	(1.36)	(4.33)	(6.85)	(2.26)
Language Dummy	0.38***	0.39***	0.51***	0.57***	0.32***	0.76***
	(4.15)	(4.24)	(4.83)	(5.23)	(3.60)	(5.34)
Colonial Dummy	0.81***	0.83***	0.41***	1.15***	0.77***	1.53***
	(10.30)	(10.53)	(4.73)	(12.63)	(10.28)	(12.14)
Religion Dummy	0.13***	0.13***	0.14***	0.28***	0.14***	0.42***
	(2.64)	(2.79)	(3.12)	(4.87)	(3.31)	(5.60)
Trade area Dummy	0.57***	0.56***	0.76***	0.41***	0.61***	0.18*
	(7.94)	(7.77)	(5.13)	(4.22)	(9.20)	(1.66)
Constant	–36.91***	–37.41***	–15.73***	–46.43***	–34.84***	–56.88***
	(96.35)	(100.49)	(36.89)	(107.83)	(98.05)	(100.83)
Observations	13682	16002		16002	13249	16002
'censored'		2320		2753	2753	
Adjusted R^2	0.68					0.64
log likelihood	–30282.40	–34313.15		–34253.03	–27572.54	–44071.15
F-statistic	3950.22			19470.05		3530.48
Wald-statistic		37094.18				33407.61
$\rho_{\varepsilon\mu}$		0.08				
σ_{ε}		2.21				
Inverse Mills ratio (λ)†		0.18				

Notes: Robust t-statistics in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%.
Dependent variable: log bilateral export (1999). †: Inverse Mills ratio computed at the mean value of the regressor variables.

Table 2 includes some additional estimations, as a robustness check. Specifications (1) and (2) again apply Tobit and truncated regression. The lower limit has been put equal to the average value of zero flows following from the benchmark OLS estimation for the non-zero sample. The results show that these methods are not robust for the chosen censoring limit. The Tobit results are now more in line with the benchmark outcomes from the sample selection model, because the censoring limit imposed is a more realistic representation of the zero-flow observations. However, these approaches remain empirically unsatisfactory as well as theoretically unfounded for the situation at hand. Arbitrary censoring and truncation is an ad-hoc, crude method that does not guarantee any quantitative accurateness in terms of results, compared to the preferred and flexible sample selection model. Because of the absence of actual censoring from below, the estimation results will depend on the (arbitrarily chosen) lower limit. Only if the chosen censoring value is sufficiently high to capture potential trade for all zero flows, these approaches would yield consistent estimates. However, this does not help us to understand how zero flows arise, and it would imply that a large number of positive observations are censored as well. Hence, the information contained in these observations would be largely lost.

Specifications (3) and (4) provide robustness checks using country-specific fixed effects in the regression equation. Fixed effects correct for the potential misspecification bias in the estimates of the traditional gravity equation, which does not include country-specific price levels (see Anderson and Van Wincoop, 2004; Feenstra, 2004). Although the results indeed differ quantitatively from the conventional gravity outcomes, the OLS and sample selection models remain highly comparable. The correlation term between regression and selection equation does not differ statistically from zero once country-specific effects have been controlled for. This suggests that the Probit selection model and the linear regression model are independent, which implies that performing fixed-effects OLS on the non-zero sample does not bias the results.

Table 2: Robustness

	(1) Tobit at mean exp. value†	(2) Truncated at mean exp. value†	(3) OLS FE	(4a) Heckman FE: regression ‡	(4b) Heckman: selection
Log GDP exporter	1.32***	1.08***			0.49***
	(147.84)	(131.61)			(67.32)
Log GDP importer	1.09***	0.92***			0.40***
	(123.12)	(112.82)			(67.30)
Log Distance	–1.23***	–1.00***	–1.31***	–1.31***	–0.46***
	(53.23)	(52.13)	(41.68)	(42.31)	(32.79)
Border Dummy	0.75***	0.85***	0.87***	0.87***	–0.32***
	(5.80)	(7.63)	(6.70)	(6.75)	(3.36)
Language Dummy	0.47***	0.35***	0.49***	0.49***	0.51***
	(5.16)	(4.22)	(5.21)	(5.28)	(10.64)
Colonial Dummy	0.93***	0.71***	0.72***	0.72***	0.41***
	(12.37)	(10.20)	(8.73)	(8.84)	(11.74)
Religion Dummy	0.22***	0.10**	0.35***	0.35***	0.14***
	(4.64)	(2.48)	(6.99)	(7.07)	(6.17)
Trade area Dummy	0.55***	0.69***	0.24***	0.24***	0.75***
	(6.84)	(11.43)	(3.11)	(3.12)	(13.22)
Constant	–40.56***	–31.92***	10.98***	10.98***	–15.58***
	(111.45)	(93.15)	(27.59)	(27.86)	(58.71)
Observations	16002	12039	13682	16002	
'censored'	3963	3963		2320	
log likelihood	–29120.83	–22801.03	–28752.54	–32788.54	
F-statistic	20998.82		173.79		
Wald-statistic		30423.08		48028.97	
$\rho_{\varepsilon\mu}$				0.01	
σ_{ε}				1.98	
Inverse Mills ratio (λ)				0.03	
Adjusted R-squared			0.74		

Notes: Absolute value of t-statistic in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%. Dependent variable: log bilateral export (1999).

†: Mean expected value for zero flows ($18916) is based on the OLS results for the non-zero sample. ‡: The selection equation (4b) had to be estimated without fixed effects, including GDPs instead. The fixed effects in the regression equation (4a) capture all country-specific effects, including market size as conventionally reflected by GDP. Therefore, the regression-stage estimation does not suffer from omitted variables bias vis-à-vis the selection equation.

6 Conclusions

Zero flows may bias the estimation results for the gravity equation of bilateral trade. This paper has argued that a careful choice of the method to deal with zero flows is needed. The solutions often applied, substituting small values for zero flows or using Tobit or truncated regression, are not suited to the gravity model. First, zeros do not reflect unobservable trade values. In the gravity model with lognormal disturbance term, desired trade cannot be negative, which excludes censoring at zero as an explanation for observed zeros. Second,

rounding of trade flows as a cause of censoring does not appear to be an important explanation for zero flows either. Instead, zero flows are the result of economic decision-making based on the potential profitability of engaging in bilateral trade at all. Apart from the decision to trade or not, the size of expected potential trade is determined by the conventional gravity model. In case of actual zero trade, potential trade is unobserved. This combination of simultaneous and partly interdependent economic decisions regarding bilateral trade should be explicitly modelled at the macroeconomic level. The sample selection model forms a well-established approach to model bilateral trade in the presence of zero flows. It allows for correlation between both decisions, as the profitability of trade depends on the size of potential flows, but does not require that profitability perfectly reflects potential trade. Other microeconomic factors that do not affect the size of trade can be important for profits.

We have estimated a sample selection model as well as alternative approaches to deal with zero flows. The empirical findings show the sensitivity of the results with respect to the method chosen to deal with zero flows. Because the regression outcomes differ, it is important to make a well-motivated decision on how to deal with zero flows. The paper shows that censored or truncated regression, and replacement of zero flows with arbitrary numbers are not preferable. These approaches may yield misleading results, as they rely on ad-hoc assumptions, and artificial censoring. The sample selection model, on the other hand, allows zero flows and the size of potential trade to be explained jointly. This method correctly takes into account the information provided by zero-valued observations. Moreover, it encompasses censored regression as well as independent Probit and (truncated) regression as special cases. Starting from an explicit theoretical framework on the causes of zero flows, sample selection allows for all kinds of data structures to emerge in practice, and provides information on the decision processes underlying zero flows as well.

Apart from the extra information provided by the selection model, the regression results sugges that OLS on a non-zero sample may not lead to much bias in practice. The results have shown only limited residual correlation between the decision whether to trade at all and the decision how much to trade. Hence, OLS does not suffer greatly from selection bias. As a

result, we draw the conclusion that omitting zero flows from the regression sample leads to satisfactory results in our case, and is preferred to the use of a Tobit model or ad-hoc substitutions for zero flows. One has to keep in mind, however, that the OLS estimates only consider the non-zero sample. In this context, Greene (2000) notes that the extent of bias in OLS estimates depends on the distribution of the regressors in this sub-sample. So, it is not possible to determine beforehand whether the bias of OLS is likely to be serious. Therefore, even though the OLS results prove to be fairly close to the results in the sample selection model, it is preferable to use the sample selection model.

References

Anderson, J.E. and D. Marcouiller (2002): Insecurity and the Pattern of Trade: An Empirical Investigation, *Review of Economics and Statistics*, 84, pp. 342–352.

Anderson, J.E. and E. Van Wincoop (2004): Trade Costs, *Journal of Economic Literature*, 42, pp. 691–751.

Baldwin, R.E. (1994): *Towards an Integrated Europe*, London: CEPR.

Bikker, J.A. (1982): *Vraag-Aanbodmodellen voor Stelsels van Geografisch Gespreide Markten. Toegepast op de Internationale Handel en op Ziekenhuisopnamen in Noord-Nederland*, Amsterdam: VU Boekhandel/Uitgeverij.

Bikker, J.A. and A.F. De Vos (1992): An International Trade Flow Model With Zero Observations: An Extension of the Tobit Model, *Cahiers Economiques de Bruxelles*, pp. 379–404.

Buch, C.M., J. Kleinert and F. Toubal (2004): The Distance Puzzle: On the Interpretation of the Distance Coefficient in Gravity Equations, *Economics Letters*, 83, pp. 293–98.

Cragg, J.G. (1971): Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods, *Econometrica*, 39, pp. 829–844.

Deardorff, A.V. (1998): Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?, in: J. Frankel (ed.), *The Regionalization of the World Economy*, pp. 7–28, Chicago: University of Chicago Press.

Egger, P. (2000): A Note on the Proper Econometric Specification of the Gravity Equation, *Economics Letters*, 66, pp. 25–31.

Egger, P. and M. Pfaffermayr (2005): The Proper Panel Econometric Specification of the Gravity Equation: A Three-Way Model with Bilateral Interaction Effects, *Empirical Economics*, 28, pp. 571–580.

Eichengreen, B. and D.A. Irwin (1998): The Role of History in Bilateral Trade Flows, in: J.A. Frankel (ed.), *The Regionalization of the World Economy*, pp. 33–57, Chicago and London: The University of Chicago Press.

Feenstra, R.C. (2004): *Advanced International Trade: Theory and Evidence*, Princeton and Oxford: Princeton University Press.

Frankel, J.A. (1997): *Regional Trading Blocs in the World Economic System*, Washington D.C.: Institute for International Economics.

Greene, W.H. (2000): *Econometric Analysis*, London: Prentice-Hall International.

Gronau, R. (1974): Wage Comparisons: A Selectivity Bias, *Journal of Political Economy*, 82, pp. 1119–1143.

Heckman, J.J. (1979): Sample Selection Bias as a Specification Error, *Econometrica*, 47, pp. 153–161.

Hillberry, R.H. (2002): Aggregation Bias, Compositional Change, and the Border Effect, *Canadian Journal of Economics*, 35, pp. 517–530.

Linnemann, H. (1966): *An Econometric Study of International Trade Flows*, Amsterdam: North-Holland.

Maddala, G.S. (1992): *Introduction to Econometrics*, New York: Macmillan.

Matyas, L. (1998): The Gravity Model: Some Econometric Considerations, *The World Economy*, 21, pp. 397–401.

Pöyhönen, P. (1963): A Tentative Model for the Volume of Trade between Countries, *Weltwirtschaftliches Archiv*, 90, pp. 93–99.

Raballand, G. (2003): Determinants of the Negative Impact of Being Landlocked on Trade: An Empirical Investigation through the Central Asian Case, *Comparative Economic Studies*, 45, pp. 520–536.

Rauch, J.E. (1999): Networks versus Markets in International Trade, *Journal of International Economics*, 48, pp. 7–35.

Romer, P. (1994): New Goods, Old Theory, and the Welfare Costs of Trade Restrictions, *Journal of Development Economics*, 43, pp. 5–38.

Rose, A.K. (2000): One Money, One Market: The Effect of Common Currencies on Trade, *Economic Policy*, 15, pp. 8–45.

Rose, A.K. (2004): Do We Really Know That the WTO Increases Trade?, *American Economic Review*, 94, pp. 98–114.

Sigelman, L. and L. Zeng (1999): Analyzing Censored and Sample-Selected Data with Tobit and Heckit Models, *Political Analysis*, 8, pp. 167–182.

Soloaga, I. and L.A. Winters (2001): Regionalism in the Nineties: What Effect on Trade?, *North American Journal of Economics and Finance*, 12, pp. 1–29.

Tinbergen, J. (1962): *Shaping the World Economy*, New York: The Twentieth Century Fund.

Van Bergeijk, P.A.G. and H. Oldersma (1990): Detente, Market-Oriented Reform and German Unification: Potential Consequences for the World Trade System, *Kyklos*, 43, pp. 599–609.

Verbeek, M. (2000): *A Guide to Modern Econometrics*, Chichester, UK: Wiley.

Wang, Z.K. and L.A. Winters (1991): The Trading Potential of Eastern Europe, *CEPR Discussion Paper*, no. 610, London.

Appendix A. Estimation of the sample selection model

In this appendix, we present the likelihood function of the sample selection model estimated in Section 5. We will illustrate sample selection bias when the correlation between the selection and regression model is positive.

A.1. Maximum Likelihood estimation

In general terms, the sample selection model of bilateral trade can be defined as follows:

$$
\begin{aligned}
&\ln(T_{ij}) = \ln(\tilde{T}_{ij}); s_{ij} = 1 \qquad \text{if } \tilde{\pi}_{ij} > 0 \\
&\ln(T_{ij}) = \text{not observed}; s_{ij} = 0 \quad \text{if } \tilde{\pi}_{ij} \leq 0 \\
&\text{where:} \\
&\ln(\tilde{T}_{ij}) \;\; = x'_{1i}\beta_1 + x'_{2j}\beta_2 + x'_{3ij}\beta_3 + \varepsilon_{ij} \\
&\tilde{\pi}_{ij} \qquad = x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3 + \mu_{ij} \\
&x_1, x_2 \text{ and } x_3 \text{ are vectors of exporter- and importer specific and bilateral regressors} \\
&\beta_k \text{ and } \gamma_k, k \in \{1,2,3\} \text{ are vectors of regression and selection parameters, and:} \\
&(\varepsilon, \mu) \sim \text{bivariate normal}\left(0, 0, \sigma_\varepsilon, \sigma_\mu, \rho_{\varepsilon\mu}\right).
\end{aligned}
\tag{1}
$$

The parameters in equation (1) can be estimated using Maximum Likelihood. We follow Verbeek (2000, section 7.4.2) to derive the likelihood functions for an individual observation. Although both decisions in the model are most naturally thought of as occurring simultaneously, it is instructive to view the two parts separately when constructing the likelihood function. The selection equation essentially describes a binary choice problem. Therefore, the contribution to the likelihood is the probability of observing $s_{ij} = 1$ ($\tilde{\pi}_{ij} > 0$), if trade is non-zero, and $s_{ij} = 0$ ($\tilde{\pi}_{ij} \leq 0$), if trade is zero. The contribution for non-zero trade furthermore consists of the conditional probability density of observed trade given that trade is actually taking place, $f(\ln(T_{ij}) \mid s_{ij} = 1)$. This results in the following log-likelihood function:

$$
\ln L(\beta, \gamma, \sigma_\varepsilon, \rho_{\varepsilon\mu}) = \sum_{T_{ij}=0} \ln P\{s_{ij} = 0\} + \sum_{T_{ij}>0} \left[\ln f\left(\ln(T_{ij}) \mid s_{ij} = 1\right) + \ln P\{s_{ij} = 1\} \right]. \tag{2}
$$

The conditional distribution of $\ln(T_{ij})$, given that $s_{ij}=1$, is rather complicated. However, a reformulation simplifies matters substantially (Verbeek, 2000; Bikker and De Vos, 1992). We can use a general rule for joint distributions:

$$f\left(\ln(T_{ij}) \mid s_{ij}=1\right) P\{s_{ij}=1\} = P\{s_{ij}=1 \mid \ln(T_{ij})\}\, f\left(\ln(T_{ij})\right). \tag{3}$$

The probability density of log trade follows a normal distribution, whereas the probability in the first term on the right-hand side is from a conditional normal density function. Using the underlying latent selection variable, this conditional normal density function has the following mean and variance.

$$\begin{aligned} E\{\tilde{\pi}_{ij} \mid \ln(T_{ij})\} &= x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3 + E\{\mu_{ij} \mid \varepsilon_{ij}\} \\ &= x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3 + \frac{\sigma_{\varepsilon\mu}}{\sigma_\varepsilon^2}\left(\ln(T_{ij}) - x'_{1i}\beta_1 - x'_{2j}\beta_2 - x'_{3ij}\beta_3\right) \\ V\{\tilde{\pi}_{ij} \mid \ln(T_{ij})\} &= 1 - \frac{\sigma_{\varepsilon\mu}^2}{\sigma_\varepsilon^2} = 1 - \rho_{\varepsilon\mu}^2 \end{aligned} \tag{4}$$

Thus:

$$\tilde{\pi}_{ij} \mid \ln(T_{ij}) = x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3 + \frac{\sigma_{\varepsilon\mu}}{\sigma_\varepsilon^2}\left(\ln(T_{ij}) - x'_{1i}\beta_1 - x'_{2j}\beta_2 - x'_{3ij}\beta_3\right) + \eta_{ij}$$

$$\eta_{ij} \sim \text{independent } N\left(0, (1-\rho_{\varepsilon\mu}^2)\right).$$

With the modification in equation (3) and the conditional distribution in equation (4), the log likelihood can be written as follows.

$$\ln L(\beta, \gamma, \sigma_\varepsilon, \rho_{\varepsilon\mu}) = \sum_{T_{ij}=0} \ln P\{s_{ij}=0\} + \sum_{T_{ij}>0} \left[\ln f\left(\ln(T_{ij})\right) + \ln P\{s_{ij}=1 \mid \ln(T_{ij})\}\right]. \tag{5}$$

The relevant probabilities and probability density for an individual observation, with either observed trade or zero trade, directly result from equations (1) and (4):

$$
\begin{aligned}
P\{s_{ij}=0\} &= P\{\tilde{\pi}_{ij}\le 0\} = P\{\mu_{ij}\le -x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3\} \\
&= 1-\Phi\left(x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3\right) \\
P\{s_{ij}=1\,|\,\ln(T_{ij})\} &= P\{\tilde{\pi}_{ij}>0\,|\,\ln(T_{ij})\} = P\{\eta_{ij} > -x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3 \\
&\quad -\frac{\sigma_{\varepsilon\mu}}{\sigma_\varepsilon^2}\left(\ln(T_{ij}) - x'_{1i}\beta_1 - x'_{2j}\beta_2 - x'_{3ij}\beta_3\right)\} = \\
&= \Phi\left(\frac{x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3 + \left(\sigma_{\varepsilon\mu}/\sigma_\varepsilon^2\right)\left(\ln(T_{ij}) - x'_{1i}\beta_1 - x'_{2j}\beta_2 - x'_{3ij}\beta_3\right)}{\sqrt{1-\rho_{\varepsilon\mu}^2}}\right) \\
f\left(\ln(T_{ij})\right) &= \frac{1}{\sigma_\varepsilon}\phi\left(\frac{\ln(T_{ij}) - x'_{1i}\beta_1 - x'_{2j}\beta_2 - x'_{3ij}\beta_3}{\sigma_\varepsilon}\right),
\end{aligned}
\tag{6}
$$

where $\phi(.)$ and $\Phi(.)$ stand for the standard normal probability density and cumulative distribution function, respectively.

The log likelihood function in equation (5), maximized with respect to the unknown parameters from the sample selection model, leads to consistent and asymptotically efficient estimators for the parameters of the selection and regression equations (Verbeek, 2000, p. 211).

A.2. Sample selection bias

The most important property of the sample selection model is its flexibility with respect to the influence of zero-trade observations. The model includes separate explanatory equations for selection and potential size of the action of primary interest, but allows correlation between both stages. If the residuals in both stages are correlated, the non-random sampling implied by the selection equation leads to sample selection bias in the observed (i.e., positive trade) sample. We can illustrate this by confining ourselves to the model in equation (1), as it applies to the non-zero observations in our sample. In particular, consider the conditional expectation of log trade, given that trade is profitable to begin with (for further details, see Greene, 2000; Verbeek, 2000):

$$
\begin{aligned}
E\{\ln(T_{ij}) \mid \ln(T_{ij}) \text{is observed}\} &= E\{\ln(T_{ij}) \mid \tilde{\pi}_{ij} > 0\} \\
&= E\{\ln(T_{ij}) \mid \mu_{ij} > -x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3\} \\
&= x'_{1i}\beta_1 + x'_{2j}\beta_2 + x'_{3ij}\beta_3 + E\{\varepsilon_{ij} \mid \mu_{ij} > -x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3\} \\
&= x'_{1i}\beta_1 + x'_{2j}\beta_2 + x'_{3ij}\beta_3 + \frac{\sigma_{\varepsilon\mu}}{\sigma_\mu^2} E\{\mu_{ij} \mid \mu_{ij} > -x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3\} \\
&= x'_{1i}\beta_1 + x'_{2j}\beta_2 + x'_{3ij}\beta_3 + \frac{\sigma_{\varepsilon\mu}}{\sigma_\mu} \frac{\phi\left(x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3/\sigma_\mu\right)}{\Phi\left(x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3/\sigma_\mu\right)} \qquad (7) \\
&= x'_{1i}\beta_1 + x'_{2j}\beta_2 + x'_{3ij}\beta_3 + \rho_{\varepsilon\mu}\sigma_\varepsilon\lambda\left(\alpha_{ij}\right) \\
\text{with } \sigma_\mu &\equiv 1; \alpha_{ij} = -x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3 \\
\text{and } \lambda\left(\alpha_{ij}\right) &= \frac{\phi\left(x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3\right)}{\Phi\left(x'_{1i}\gamma_1 + x'_{2j}\gamma_2 + x'_{3ij}\gamma_3\right)}.
\end{aligned}
$$

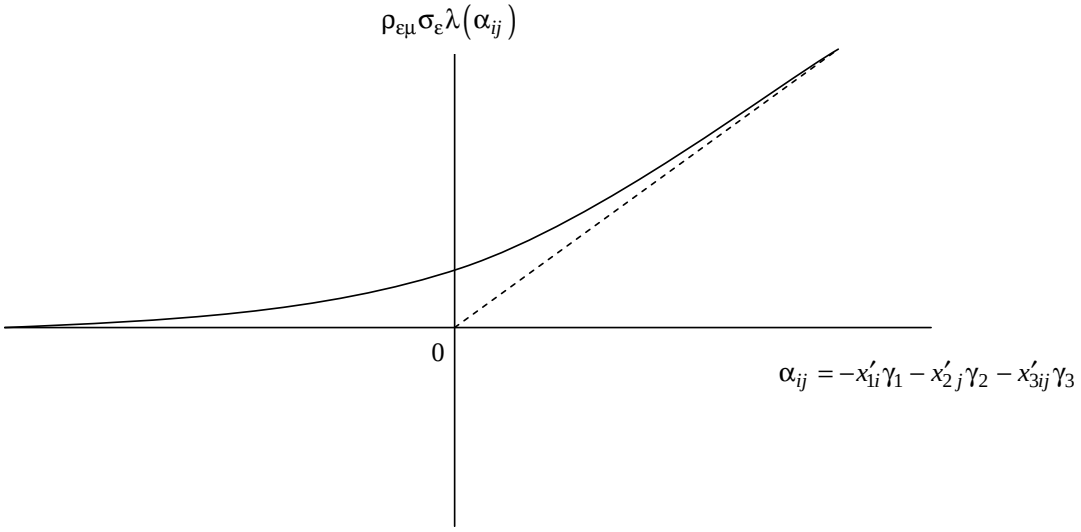
The expectation of the conditional disturbance term in the selection equation (μ_{ij}) exceeds zero, given that it is truncated from below in the observed-trade sample. To judge whether this leads to sample selection bias in the regression equation, we have to consider the expectation of the regression disturbance term (ε_{ij}), conditional on the truncation in the selection equation. From equation (7), the expectation of ε_{ij}, given that μ_{ij} is truncated from below, exceeds zero if $\rho_{\varepsilon\mu}$ is positive. The estimates in the main text of this paper indeed show a positive correlation between ε_{ij} and μ_{ij}. Thus, the conditional expected value of (log) trade, given that trade is observed, exceeds expected potential trade, unconditional on being observed or not. In other words, OLS regression of log trade on the regressor variables, using only non-zero trade observations, produces inconsistent estimates of the regression parameters in $\beta_k, k \in \{1,2,3\}$. This bias is known as sample selection bias. It can be seen most intuitively by summarizing the complete model as it applies to the non-zero sub-sample.

$$
\begin{aligned}
\ln\left(T_{ij}\right) \mid \left(s_{ij} = 1\right) &= E\{\ln\left(T_{ij}\right) \mid \left(s_{ij} = 1\right)\} + \nu_{ij} \\
&= x'_{1i}\beta_1 + x'_{2j}\beta_2 + x'_{3ij}\beta_3 + \beta_\lambda\lambda\left(\alpha_{ij}\right) + \nu_{ij}, \qquad (8) \\
\text{where } \beta_\lambda &= \rho_{\varepsilon\mu}\sigma_\varepsilon.
\end{aligned}
$$

If $\beta_\lambda \neq 0$, an OLS regression omitting λ from the model suffers from omitted variable bias.[5] To determine the direction of bias in OLS results due to sample selection, we have to take a closer look at the relation between selection and regression in the non-zero sample.

As shown by equations (7) and (8), the conditional expectation of log trade is different from the unconditional expectation of potential trade, because of the term $\lambda(\alpha_{ij}) = \lambda(-x'_{1i}\gamma_1 - x'_{2j}\gamma_2 - x'_{3ij}\gamma_3) > 0$. For positive $\rho_{\varepsilon\mu}$, the conditional expected value exceeds unconditional expected potential trade. Figure A.2.1 below illustrates how the size of this difference depends on the expected value of the latent selection variable (profitability).[6]

Figure A.2.1 $\mathrm{E}\left[\ln(\mathrm{T_{ij}}) \mid \ln(\mathrm{T_{ij}}) \text{ is observed}\right] - \mathrm{E}\left[\ln(\tilde{\mathrm{T}}_{\mathrm{ij}})\right]$ *as a function of* $-E\left[\tilde{\pi}_{ij}\right]$.



[5] On the other hand, if we can include λ in the specification, OLS will produce consistent estimates of β_k $(k \in \{1,2,3\})$, although inefficient because v_{ij} is heteroskedastic (see Greene, 2000, section 20.4.1 for more details). Equation (8) is the basis for an alternative method often used in empirical applications to estimate the selection model, without the need to estimate the full model by maximum likelihood. The two-step estimation procedure, due to Heckman (1979) and also known as the 'Heckit' estimator, estimates equation (8) by OLS. First, the selection equation is estimated as a Probit model, to determine $\hat{\lambda}_{ij}$, as estimates of λ_{ij}. These estimated values are subsequently inserted in the second-step OLS regression.

[6] The figure is based on Figure 20.2 in Greene (2000).

The figure shows that conditional expected trade is highest, compared to unconditional expected potential trade, for low values of expected profitability. Given the positive correlation $\rho_{\varepsilon\mu}$, this makes sense. In order to assure profitability, the realization for the disturbance term μ_{ij} should be high. Given the truncation in the selection equation, the expected value of trade will be high as well.

Apart from the relationship between expected profitability and conditional expected trade, it is important to establish the potential consequences of truncation in the selection equation for sample selection bias of OLS. We may conclude from our estimation results in Section 5 that the difference between conditional and unconditional expected trade is highest for low values of unconditional expected trade, because most explanatory variables in our model have the same sign in both the selection and the regression equation. This corresponds to the intuitive argument in the main text. A low expected profitability coincides with low unconditional expected trade. Therefore, trade flows that we observe between countries that are more distant will be relatively more above their unconditional expected value, on average. The regression plane tends to be flattened by the sample selection process. As a result, the OLS regression coefficients for the 'observed' sample of non-zero bilateral trade will underestimate the true effect on unconditional expected potential trade.

Appendix B. Description of the data

This appendix describes the data used in the paper, and their sources. A table that lists all the countries included in the analysis is presented at the end of the Appendix.

B.1. Data sources and variables used in the empirical analysis

The empirical analysis uses both country-specific and bilateral data from various sources. The GDPs of the exporting and importing countries are examples of country-specific variables, while geographic distance, adjacency, and common language and religion, among others, are examples of bilateral characteristics for each pair of countries. Below we have described the data and sources in more detail. The analysis applies to 1999.

Trade

The dependent variable in the gravity model is the log of the value of bilateral merchandise exports, which results in two observations for each country pair, i.e. the export flows from country *i* to *j*, and those from *j* to *i*. We have used the UN COMTRADE database for bilateral trade flows in 1999. We have used reported imports rather than reported exports, because import data provide a better coverage. We have used mirror import flows between *i* and *j*; the direction of these mirror import flows corresponds to that of the export flows from *i* to *j*. Although mirror import data have fewer missing trade observations than export data, some trade flow observations are reported missing in mirror imports whereas corresponding exports are non-zero. We have confronted missing observations in reported mirror imports with corresponding flows in reported exports; when corresponding reported exports were non-zero, these values have been substituted in reported mirror imports. Thus, only trade flows that are missing in both reported mirror imports and reported exports have been treated as zero-entried trade values (or non-availables, in regressions that omit zero flows).

GDP

The source of GDP data is the World Development Indicators (World Bank, 2000 - on CD Rom). GDP levels are in constant US $ at 1995 prices and refer to 1999.

Bilateral characteristics: distance, adjacency, trade area, language, colonial history and religion

The data on geographic distance, common border, common official language, common regional trade agreement, common dominant religion and common colonial history have been collected from diverse sources, which have kindly been made available by several researchers and research institutes on the internet. We have used OECD data for regional integration agreements, Sala-i-Martin's (1997)[7] database for religions and colonial backgrounds, and Jon

[7] See: http://www.columbia.edu/~xs23/data.htm.

Haveman's International Trade Data[8] for distance, contiguity and language. This part of the database is available upon request. Some remarks on these variables are:

- Distance is measured as straight line distance ('as the crow flies') between nation capitals. The data are from the data website of Jon Haveman. In line with previous research, geographic distance is measured as the distance from home to foreign 'as the bird flies', using the principal city of each country as its centre of gravity. This implies that the distance between the two centres of gravity of neighboring countries is likely to overestimate the average distance of trade between them. The relative impact of mismeasurement is much larger in neighboring countries than in countries that are located far away from each other. For a discussion on the use and usefulness of other, more sophisticated measures of geographic distance, we refer to Frankel (1997, chapter 4). In general, more sophisticated geographic distance measures produce similar results, and cannot eliminate the measurement error for contiguous countries either.
- The border dummy takes the value of one if two countries are adjacent. Adjacency requires either a land border or a small body of water as border. Measurement error in the distance variable, as well as the effect of historical relations between adjacent countries are captured by this dummy variable. The contiguity data are from the website of Jon Haveman.
- Whether pairs of countries take part in a common regional integration agreement (RIA) has been determined on the basis of OECD data on major regional integration agreements.[9] A dummy variable indicates whether a pair of countries enters into at least one common RIA.
- To assess whether two countries have the same official language, we use a database collected by Jon Haveman, that distinguishes fourteen languages: Arabic, Burmese,

[8] See: http://www.macalester.edu/research/economics/page/haveman/trade.resources/tradedata.html.
[9] See: http://www.oecd.org/dataoecd/39/37/1923431.pdf

Chinese, Dutch, English, French, German, Greek, Korean, Malay, Persian, Portuguese, Spanish and Swedish. This data has been extended to cover more countries and languages with CIA's World Factbook[10]. In case none of the above applied and no further language data were available, countries were assigned to the categories 'other language' or 'non available'. A language dummy variable reflects whether or not two countries have a common language.

- Cultural and/or historical ties between countries may also consist of a shared colonial past or a common dominant religion. Data for these variables come from Sala-i-Martin (1997).The colonial dummy variable reflects whether country pairs share a colonial history. The data consider the British, French and Spanish empires only. In contrast to the original data source, we also included these colonizers themselves into the respective empires. In this way, the figures identify shared colonial relations for pairs of countries.
- Based on the percentage of the population adhering to one of seven major religions (i.e., Buddhism, Catholicism, Confucianism, Hinduism, Jewish religion, Islam, and Protestantism), country pairs score a value of one on the religion dummy if their dominant religion is the same. For some countries, two religions were equally dominant over the others. In these cases, both religions were considered to be dominant.

[10] See: http://www.cia.gov/cia/publications/factbook/.

B.2. List of countries included in the sample

The database includes 127 countries, listed in the table below.

Country			
Albania	Gabon	Mauritius	Togo
Algeria	Gambia	Mexico	Trinidad & Tobago
Argentina	Georgia	Moldova	Tunisia
Armenia	Germany	Mongolia	Turkey
Australia	Ghana	Morocco	Turkmenistan
Austria	Greece	Nepal	Uganda
Azerbaijan	Guatemala	Netherlands	Ukraine
Bahamas, The	Guinea	New Zealand	United Kingdom
Belarus	Guyana	Nicaragua	United States
Belgium	Honduras	Niger	Uruguay
Belize	Hong Kong, China	Nigeria	Venezuela
Benin	Hungary	Norway	Vietnam
Bhutan	Iceland	Pakistan	Yemen, Rep.
Bolivia	India	Panama	Yugoslavia
Brazil	Indonesia	Paraguay	Zambia
Bulgaria	Iran, Islamic Rep.	Peru	Zimbabwe
Burkina Faso	Ireland	Philippines	
Burundi	Israel	Poland	
Cameroon	Italy	Portugal	
Canada	Jamaica	Romania	
Chile	Japan	Russian Federation	
China	Jordan	Rwanda	
Colombia	Kazakhstan	Saudi Arabia	
Costa Rica	Kenya	Senegal	
Cote d'Ivoire	Korea, South (Rep.)	Singapore	
Croatia	Kuwait	Slovak Republic	
Cyprus	Kyrgyzstan	Slovenia	
Czech Republic	Latvia	South Africa	
Denmark	Lebanon	Spain	
Dominican Rep.	Lithuania	Sri Lanka	
Ecuador	Luxembourg	Sudan	
Egypt, Arab Rep.	Macedonia, FYR	Suriname	
El Salvador	Madagascar	Sweden	
Estonia	Malawi	Switzerland	
Ethiopia	Malaysia	Syrian Arab Republic	
Finland	Mali	Tanzania	
France	Malta	Thailand	