THE LOG OF GRAVITY

J. M. C. Santos Silva and Silvana Tenreyro*

Abstract—Although economists have long been aware of Jensen's inequality, many econometric applications have neglected an important implication of it: under heteroskedasticity, the parameters of log-linearized models estimated by OLS lead to biased estimates of the true elasticities. We explain why this problem arises and propose an appropriate estimator. Our criticism of conventional practices and the proposed solution extend to a broad range of applications where log-linearized equations are estimated. We develop the argument using one particular illustration, the gravity equation for trade. We find significant differences between estimates obtained with the proposed estimator and those obtained with the traditional method.

I. Introduction

ECONOMISTS have long been aware that Jensen's inequality implies that $E(\ln y) \neq \ln E(y)$, that is, the expected value of the logarithm of a random variable is different from the logarithm of its expected value. This basic fact, however, has been neglected in many econometric applications. Indeed, one important implication of Jensen's inequality is that the standard practice of interpreting the parameters of log-linearized models estimated by ordinary least squares (OLS) as elasticities can be highly misleading in the presence of heteroskedasticity.

Although many authors have addressed the problem of obtaining consistent estimates of the conditional mean of the dependent variable when the model is estimated in the log linear form (see, for example, Goldberger, 1968; Manning & Mullahy, 2001), we were unable to find any reference in the literature to the potential bias of the elasticities estimated using the log linear model.

In this paper we use the gravity equation for trade as a particular illustration of how the bias arises and propose an appropriate estimator. We argue that the gravity equation, and, more generally, constant-elasticity models, should be estimated in their multiplicative form and propose a simple pseudo-maximum-likelihood (PML) estimation technique. Besides being consistent in the presence of heteroskedasticity, this method also provides a natural way to deal with zero values of the dependent variable.

Using Monte Carlo simulations, we compare the performance of our estimator with that of OLS (in the log linear specification). The results are striking. In the presence of heteroskedasticity, estimates obtained using log-linearized models are severely biased, distorting the interpretation of the model. These biases might be critical for the comparative assessment of competing economic theories, as well as for the evaluation of the effects of different policies. In contrast, our method is robust to the different patterns of heteroskedasticity considered in the simulations.

We next use the proposed method to provide new estimates of the gravity equation in cross-sectional data. Using standard tests, we show that heteroskedasticity is indeed a severe problem, both in the traditional gravity equation introduced by Tinbergen (1962), and in a gravity equation that takes into account multilateral resistance terms or fixed effects, as suggested by Anderson and van Wincoop (2003). We then compare the estimates obtained with the proposed PML estimator with those generated by OLS in the log linear specification, using both the traditional and the fixed-effects gravity equations.

Our estimation method paints a very different picture of the determinants of international trade. In the traditional gravity equation, the coefficients on GDP are not, as generally estimated, close to 1. Instead, they are significantly smaller, which might help reconcile the gravity equation with the observation that the trade-to-GDP ratio decreases with increasing total GDP (or, in other words, that smaller countries tend to be more open to international trade). In addition, OLS greatly exaggerates the roles of colonial ties and geographical proximity.

Using the Anderson–van Wincoop (2003) gravity equation, we find that OLS yields significantly larger effects for geographical distance. The estimated elasticity obtained from the log-linearized equation is almost twice as large as that predicted by PML. OLS also predicts a large role for common colonial ties, implying that sharing a common colonial history practically doubles bilateral trade. In contrast, the proposed PML estimator leads to a statistically and economically insignificant effect.

The general message is that, even controlling for fixed effects, the presence of heteroskedasticity can generate strikingly different estimates when the gravity equation is log-linearized, rather than estimated in levels. In other words, Jensen's inequality is quantitatively and qualitatively important in the estimation of gravity equations. This suggests that inferences drawn on log-linearized regressions can produce misleading conclusions.

Despite the focus on the gravity equation, our criticism of the conventional practice and the solution we propose extend to a broad range of economic applications where the equations under study are log-linearized, or, more generally, transformed by a nonlinear function. A short list of examples includes the estimation of Mincerian equations for wages, production functions, and Euler equations, which are typically estimated in logarithms.

The remainder of the paper is organized as follows. Section II studies the econometric problems raised by the estimation of gravity equations. Section III considers constant-elasticity models in general; it introduces the PML estimator and specification tests to check the adequacy of the proposed estimator. Section IV presents the Monte Carlo simulations. Section V provides new estimates of both the traditional and the Anderson–van Wincoop gravity equation. The results are compared with those generated by OLS, nonlinear least squares, and tobit estimations. Section VI contains concluding remarks.

II.   The Econometrics of the Gravity Equation

A.   The Traditional Gravity Equation

The pioneering work of Jan Tinbergen (1962) initiated a vast theoretical and empirical literature on the gravity equation for trade. Theories based on different foundations for trade, including endowment and technological differences, increasing returns to scale, and Armington demands, all predict a gravity relationship for trade flows analogous to Newton's law of universal gravitation.[1] In its simplest form, the gravity equation for trade states that the trade flow from country i to country j, denoted by T_{ij}, is proportional to the product of the two countries' GDPs, denoted by Y_i and Y_j, and inversely proportional to their distance, D_{ij}, broadly construed to include all factors that might create trade resistance. More generally,

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}, \tag{1}$$

where α_0, α_1, α_2, and α_3 are unknown parameters.

The analogy between trade and the physical force of gravity, however, clashes with the observation that there is no set of parameters for which equation (1) will hold exactly for an arbitrary set of observations. To account for deviations from the theory, stochastic versions of the equation are used in empirical studies. Typically, the stochastic version of the gravity equation has the form

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij}, \tag{2}$$

where η_{ij} is an error factor with $E(\eta_{ij}|Y_i, Y_j, D_{ij}) = 1$, assumed to be statistically independent of the regressors, leading to

$$E(T_{ij}|Y_i,Y_j,D_{ij}) = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}.$$

There is a long tradition in the trade literature of log-linearizing equation (2) and estimating the parameters of interest by least squares, using the equation

$$\ln T_{ij} = \ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j \\ + \alpha_3 \ln D_{ij} + \ln \eta_{ij}. \tag{3}$$

The validity of this procedure depends critically on the assumption that η_{ij}, and therefore $\ln \eta_{ij}$, are statistically independent of the regressors. To see why this is so, notice that the expected value of the logarithm of a random variable depends both on its mean and on the higher-order moments of the distribution. Hence, for example, if the variance of the error factor η_{ij} in equation (2) depends on Y_i, Y_j, or D_{ij}, the expected value of $\ln \eta_{ij}$ will also depend on the regressors, violating the condition for consistency of OLS.[2]

In the cases studied in section V we find overwhelming evidence that the error terms in the usual log linear specification of the gravity equation are heteroskedastic, which violates the assumption that $\ln \eta_{ij}$ is statistically independent of the regressors and suggests that this estimation method leads to inconsistent estimates of the elasticities of interest.

A related problem with the analogy between Newtonian gravity and trade is that gravitational force can be very small, but never zero, whereas trade between several pairs of countries is literally zero. In many cases, these zeros occur simply because some pairs of countries did not trade in a given period. For example, it would not be surprising to find that Tajikistan and Togo did not trade in a certain year.[3] These zero observations pose no problem at all for the estimation of gravity equations in their multiplicative form. In contrast, the existence of observations for which the dependent variable is zero creates an additional problem for

---

[1] See, for example, Anderson (1979), Helpman and Krugman (1985), Bergstrand (1985), Davis (1995), Deardoff (1998), and Anderson and van Wincoop (2003). A feature common to these models is that they all assume complete specialization: each good is produced in only one country. However, Haveman and Hummels (2001), Feenstra, Markusen, and Rose (2000), and Eaton and Kortum (2001) derive the gravity equation without relying on complete specialization. Examples of empirical studies framed on the gravity equation include the evaluation of trade protection (for example, Harrigan, 1993), regional trade agreements (for example, Frankel, Stein, & Wei, 1998; Frankel, 1997), exchange rate variability (for example, Frankel & Wei, 1993; Eichengreen & Irwin, 1995), and currency unions (for example, Rose, 2000; Frankel & Rose, 2002; and Tenreyro & Barro, 2002). See also the various studies on *border effects* influencing the patterns of intranational and international trade, including McCallum (1995), and Anderson and van Wincoop (2003), among others.

[2] As an illustration, consider the case in which η_{ij} follows a log normal distribution, with $E(\eta_{ij}|Y_i, Y_j, D_{ij}) = 1$ and variance $\sigma_{ij}^2 = f(Y_i, Y_j, D_{ij})$. The error term in the log-linearized representation will then follow a normal distribution, with $E[\ln \eta_{ij}|Y_i, Y_j, D_{ij}] = -\frac{1}{2}\ln(1 + \sigma_{ij}^2)$, which is also a function of the covariates.

[3] The absence of trade between small and distant countries might be explained, among other factors, by large variable costs (for example, bricks are too costly to transport) or large fixed costs (for example, information on foreign markets). At the aggregate level, these costs can be best proxied by the various measures of distance and size entering the gravity equation. The existence of zero trade between many pairs of countries is directly addressed by Hallak (2006) and Helpman, Melitz, and Rubinstein (2004). These authors propose a promising avenue of research using a two-part estimation procedure, with a fixed-cost equation determining the cutoff point above which a country exports, and a standard gravity equation. Their results, however, rely heavily on both normality and homoskedasticity assumptions, the latter being the particular concern of this paper. A natural topic for further research is to develop and implement an estimator of the two-part model that, like the PML estimator proposed here, is robust to distributional assumptions.

the use of the log linear form of the gravity equation. Several methods have been developed to deal with this problem [see Frankel (1997) for a description of the various procedures]. The approach followed by the large majority of empirical studies is simply to drop the pairs with zero trade from the data set and estimate the log linear form by OLS. Rather than throwing away the observations with $T_{ij} = 0$, some authors estimate the model using $T_{ij} + 1$ as the dependent variable or use a tobit estimator. However, these procedures will generally lead to inconsistent estimators of the parameters of interest. The severity of these inconsistencies will depend on the particular characteristics of the sample and model used, but there is no reason to believe that they will be negligible.

Zeros may also be the result of rounding errors.[4] If trade is measured in thousands of dollars, it is possible that for pairs of countries for which bilateral trade did not reach a minimum value, say \$500, the value of trade is registered as 0. If these rounded-down observations were partially compensated by rounded-up ones, the overall effect of these errors would be relatively minor. However, the rounding down is more likely to occur for small or distant countries, and therefore the probability of rounding down will depend on the value of the covariates, leading to the inconsistency of the estimators. Finally, the zeros can just be missing observations that are wrongly recorded as 0. This problem is more likely to occur when small countries are considered, and again the measurement error will depend on the covariates, leading to inconsistency.

B. The Anderson–van Wincoop Gravity Equation

Anderson and van Wincoop (2003) argue that the traditional gravity equation is not correctly specified, as it does not take into account multilateral resistance terms. One of the solutions for this problem that is suggested by those authors is to augment the traditional gravity equation with exporter and importer fixed effects, leading to

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} e^{\theta_i d_i + \theta_j d_j}, \tag{4}$$

where α_0, α_1, α_2, α_3, θ_i, and θ_j are the parameters to be estimated and d_i and d_j are dummies identifying the exporter and importer.[5]

Their model also yields the prediction that $\alpha_1 = \alpha_2 = 1$, which leads to the unit-income-elasticity model

$$T_{ij} = \alpha_0 Y_i Y_j D_{ij}^{\alpha_3} e^{\theta_i d_i + \theta_j d_j},$$

whose stochastic version has the form

$$E(T_{ij}|Y_i, Y_j, D_{ij}, d_i, d_j) = \alpha_0 Y_i Y_j D_{ij}^{\alpha_3} e^{\theta_i d_i + \theta_j d_j}. \tag{5}$$

As before, log-linearization of equation (5) raises the problem of how to treat zero-value observations. Moreover, given that equation (5) is a multiplicative model, it is also subject to the biases caused by log-linearization in the presence of heteroskedasticity. Naturally, the presence of the individual effects may reduce the severity of this problem, but whether or not that happens is an empirical issue.

In our empirical analysis we provide estimates for both the traditional and the Anderson–van Wincoop gravity equations, using alternative estimation methods. We show that, in practice, heteroskedasticity is quantitatively and qualitatively important in the gravity equation, even when controlling for fixed effects. Hence, we recommend estimating the augmented gravity equation in levels, using the proposed PML estimator, which also adequately deals with the zero-value observations.

III. Constant-Elasticity Models

Despite their immense popularity, empirical studies involving gravity equations still have important econometric flaws. These flaws are not exclusive to this literature, but extend to many areas where constant-elasticity models are used. This section examines how the deterministic multiplicative models suggested by economic theory can be used in empirical studies.

In their nonstochastic form, the relationship between the multiplicative constant-elasticity model and its log linear additive formulation is trivial. The problem, of course, is that economic relations do not hold with the accuracy of physical laws. All that can be expected is that they hold on average. Indeed, here we interpret economic models like the gravity equation as yielding the expected value of the variable of interest, $y \geq 0$, for a given value of the explanatory variables, x (see Goldberger, 1991, p. 5). That is, if economic theory suggests that y and x are linked by a constant-elasticity model of the form $y_i = \exp(x_i\beta)$, the function $\exp(x_i\beta)$ is interpreted as the conditional expectation of y_i given x, denoted $E[y_i|x]$.[6] For example, using the notation in the previous section, the multiplicative gravity relationship can be written as the exponential function $\exp[\ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j + \alpha_3 \ln D_{ij}]$, which is interpreted as the conditional expectation $E(T_{ij}|Y_i, Y_j, D_{ij})$.

Because the relation $y_i = \exp(x_i\beta)$ holds on average but not for each i, an error term is associated with each observation, which is defined as $\varepsilon_i = y_i - E[y_i|x]$.[7] Therefore, the stochastic model can be formulated as

---

[4] Trade data can suffer from many other forms of errors, as described in Feenstra, Lipsey, and Bowen (1997).

[5] Note that, throughout the paper, T_{ij} denotes exports from i to j.

[6] Notice that if $\exp(x_i\beta)$ is interpreted as describing the conditional median of y_i (or some other conditional quantile) rather than the conditional expectation, estimates of the elasticities of interest can be obtained estimating the log linear model using the appropriate quantile regression estimator (Koenker & Bassett, 1978). However, interpreting $\exp(x_i\beta)$ as a conditional median is problematic when y_i has a large mass of zero observations, as in trade data. Indeed, in this case the conditional median of y_i will be a discontinuous function of the regressors, which is generally not compatible with standard economic theory.

[7] Whether the error enters additively or multiplicatively is irrelevant for our purposes, as explained below.

$$y_i = \exp(x_i\beta) + \varepsilon_i, \tag{6}$$

with $y_i \geqq 0$ and $E[\varepsilon_i|x] = 0$.

As we mentioned before, the standard practice of log-linearizing equation (6) and estimating β by OLS is inappropriate for a number of reasons. First of all, y_i can be 0, in which case log-linearization is infeasible. Second, even if all observations of y_i are strictly positive, the expected value of the log-linearized error will in general depend on the covariates, and hence OLS will be inconsistent. To see the point more clearly, notice that equation (6) can be expressed as

$$y_i = \exp(x_i\beta) \, \eta_i,$$

with $\eta_i = 1 + \varepsilon_i/\exp(x_i\beta)$ and $E[\eta_i|x] = 1$. Assuming for the moment that y_i is positive, the model can be made linear in the parameters by taking logarithms of both sides of the equation, leading to

$$\ln y_i = x_i\beta + \ln \eta_i. \tag{7}$$

To obtain a consistent estimator of the slope parameters in equation (6) estimating equation (7) by OLS, it is necessary that $E[\ln \eta_i|x]$ does not depend on x_i.[8] Because $\eta_i = 1 + \varepsilon_i/\exp(x_i\beta)$, this condition is met only if ε_i can be written as $\varepsilon_i = \exp(x_i\beta) \, v_i$, where v_i is a random variable statistically independent of x_i. In this case, $\eta_i = 1 + v_i$ and therefore is statistically independent of x_i, implying that $E[\ln \eta_i|x]$ is constant. Thus, only under very specific conditions on the error term is the log linear representation of the constant-elasticity model useful as a device to estimate the parameters of interest.

When η_i is statistically independent of x_i, the conditional variance of y_i (and ε_i) is proportional to $\exp(2x_i\beta)$. Although economic theory generally does not provide any information on the variance of ε_i, we can infer some of its properties from the characteristics of the data. Because y_i is nonnegative, when $E[y_i|x]$ approaches 0, the probability of y_i being positive must also approach 0. This implies that $V[y_i|x]$, the conditional variance of y_i, tends to vanish as $E[y_i|x]$ passes to 0.[9] On the other hand, when the expected value of y is far away from its lower bound, it is possible to observe large deviations from the conditional mean in either direction, leading to greater dispersion. Thus, in practice, ε_i will generally be heteroskedastic and its variance will depend on $\exp(x_i\beta)$, but there is no reason to assume that $V[y_i|x]$ is proportional to $\exp(2x_i\beta)$. Therefore, in general, regressing $\ln y_i$ on x_i by OLS will lead to inconsistent estimates of β.

It may be surprising that the pattern of heteroskedasticity and, indeed, the form of all higher-order moments of the conditional distribution of the error term can affect the consistency of an estimator, rather than just its efficiency. The reason is that the nonlinear transformation of the dependent variable in equation (7) changes the properties of the error term in a nontrivial way because the conditional expectation of $\ln \eta_i$ depends on the shape of the conditional distribution of η_i. Hence, unless very strong restrictions are imposed on the form of this distribution, it is not possible to recover information about the conditional expectation of y_i from the conditional mean of $\ln y_i$, simply because $\ln \eta_i$ is correlated with the regressors. Nevertheless, estimating equation (7) by OLS will produce consistent estimates of the parameters of $E[\ln y_i|x]$ as long as $E[\ln (y_i)|x]$ is a linear function of the regressors.[10] The problem is that these parameters may not permit identification of the parameters of $E[y_i|x]$.

In short, even assuming that all observations on y_i are positive, it is not advisable to estimate β from the log linear model. Instead, the nonlinear model has to be estimated.

A. Estimation

Although most empirical studies use the log linear form of the constant-elasticity model, some authors [see Frankel and Wei (1993) for an example in the international trade literature] have estimated multiplicative models using non-linear least squares (NLS), which is an asymptotically valid estimator for equation (6). However, the NLS estimator can be very inefficient in this context, as it ignores the heteroskedasticity that, as discussed before, is characteristic of this type of data.

The NLS estimator of β is defined by

$$\hat{\beta} = \arg \min_b \sum_{i=1}^{n} [y_i - \exp(x_i b)]^2,$$

which implies the following set of first-order conditions:

$$\sum_{i=1}^{n} [y_i - \exp(x_i\hat{\beta})] \exp(x_i\hat{\beta}) \, x_i = 0. \tag{8}$$

These equations give more weight to observations where $\exp(x_i\beta)$ is large, because that is where the curvature of the conditional expectation is more pronounced. However, these are generally also the observations with larger variance, which implies that NLS gives more weight to noisier observations. Thus, this estimator may be very inefficient, depending heavily on a small number of observations.

If the form of $V[y_i|x]$ were known, this problem could be eliminated using a weighted NLS estimator. However, in

---

[8] Consistent estimation of the intercept would also require $E[\ln \eta_i \, |x] = 0$.

[9] In the case of trade data, when $E[y_i|x]$ is close to its lower bound (that is, for pairs of small and distant countries), it is unlikely that large values of trade are observed, for they cannot be offset by equally large deviations in the opposite direction, simply because trade cannot be negative. Therefore, for these observations, dispersion around the mean tends to be small.

[10] When $E[\ln y_i \, | \, x]$ is not a linear function of the regressors, estimating equation (7) by OLS will produce consistent estimates of the parameters of the best linear approximation to $E[\ln y_i \, | \, x]$ (see Goldberger, 1991, p. 53).

practice, all we know about $V[y_i|x]$ is that, in general, it goes to 0 as $E[y_i|x]$ passes to 0. Therefore, an optimal weighted NLS estimator cannot be used without further information on the distribution of the errors. In principle, this problem can be tackled by estimating the multiplicative model using a consistent estimator, and then obtaining the appropriate weights estimating the skedastic function nonparametrically, as suggested by Delgado (1992) and Delgado and Kniesner (1997). However, this nonparametric generalized least squares estimator is rather cumbersome to implement, especially if the model has a large number of regressors. Moreover, the choice of the first-round estimator is an open question, as the NLS estimator may be a poor starting point due to its considerable inefficiency. Therefore, the nonparametric generalized least squares estimator is not appropriate to use as a workhorse for routine estimation of multiplicative models.[11] Indeed, what is needed is an estimator that is consistent and reasonably efficient under a wide range of heteroskedasticity patterns and is also simple to implement.

A possible way of obtaining an estimator that is more efficient than the standard NLS without the need to use nonparametric regression is to follow McCullagh and Nelder (1989) and estimate the parameters of interest using a PML estimator based on some assumption on the functional form of $V[y_i|x]$.[12] Among the many possible specifications, the hypothesis that the conditional variance is proportional to the conditional mean is particularly appealing. Indeed, under this assumption $E[y_i|x] = \exp(x_i\beta) \propto V[y_i|x]$, and β can be estimated by solving the following set of first-order conditions:

$$\sum_{i=1}^{n} [y_i - \exp(x_i\tilde{\beta})]x_i = 0. \tag{9}$$

Comparing equations (8) and (9), it is clear that, unlike the NLS estimator, which is a PML estimator obtained assuming that $V[y_i|x]$ is constant, the PML estimator based on equation (9) gives the same weight to all observations, rather than emphasizing those for which $\exp(x_i\beta)$ is large. This is because, under the assumption that $E[y_i|x] \propto V[y_i|x]$, all observations have the same information on the parameters of interest as the additional information on the curvature of the conditional mean coming from observations with large $\exp(x_i\beta)$ is offset by their larger variance. Of course, this estimator may not be optimal, but without further information on the pattern of heteroskedasticity, it seems

natural to give the same weight to all observations.[13] Even if $E[y_i|x]$ is not proportional to $V[y_i|x]$, the PML estimator based on equation (9) is likely to be more efficient than the NLS estimator when the heteroskedasticity increases with the conditional mean.

The estimator defined by equation (9) is numerically equal to the Poisson pseudo-maximum-likelihood (PPML) estimator, which is often used for count data.[14] The form of equation (9) makes clear that all that is needed for this estimator to be consistent is the correct specification of the conditional mean, that is, $E[y_i|x] = \exp(x_i\beta)$. Therefore, the data do not have to be Poisson at all—and, what is more important, y_i does not even have to be an integer—for the estimator based on the Poisson likelihood function to be consistent. This is the well-known PML result first noted by Gourieroux, Monfort, and Trognon (1984).

The implementation of the PPML estimator is straightforward: there are standard econometric programs with commands that permit the estimation of Poisson regression, even when the dependent variables are not integers. Because the assumption $V[y_i|x] \propto E[y_i|x]$ is unlikely to hold, this estimator does not take full account of the heteroskedasticity in the model, and all inference has to be based on an Eicker-White (Eicker, 1963; White, 1980) robust covariance matrix estimator. In particular, within Stata (StataCorp., 2003), the PPML estimation can be executed using the following command:

poisson export$_{i,j}$ ln(*dist$_{ij}$*)

    ln Y_i ln Y_j (*other variables*)$_{ij}$, *robust*

where *export* (or *import*) is measured in levels.

Of course, if it were known that $V[y_i|x]$ is a function of higher powers of $E[y_i|x]$, a more efficient estimator could be obtained by downweighting even more the observations with large conditional mean. An example of such an estimator is the gamma PML estimator studied by Manning and Mullahy (2001), which, like the log-linearized model, assumes that $V[y_i|x]$ is proportional to $E[y_i|x]^2$. The first-order conditions for the gamma PML estimator are given by

$$\sum_{i=1}^{n} [y_i - \exp(x_i\check{\beta})]\exp(-x_i\check{\beta})\, x_i = 0.$$

In the case of trade data, however, this estimator may have an important drawback. Trade data for larger countries (as gauged by GDP per capita) tend to be of higher quality (see Frankel & Wei, 1993; Frankel, 1997); hence, models assuming that $V[y_i|x]$ is a function of higher powers of $E[y_i|x]$ might give excessive weight to the observations that

---

[11] A nonparametric generalized least squares estimator can also be used to estimate linear models in the presence of heteroskedasticity of unknown form (Robinson, 1987). However, despite having been proposed more than 15 years ago, this estimator has never been adopted as a standard tool by researchers doing empirical work, who generally prefer the simplicity of the inefficient OLS, with an appropriate covariance matrix.

[12] See also Manning and Mullahy (2001). A related estimator is proposed by Papke and Wooldridge (1996) for the estimation of models for fractional data.

[13] The same strategy is implicitly used by Papke and Wooldridge (1996) in their pseudo-maximum-likelihood estimator for fractional data models.

[14] See Cameron and Trivedi (1998) and Winkelmann (2003) for more details on the Poisson regression and on more general models for count data.

are more prone to measurement errors.[15] Therefore, the Poisson regression emerges as a reasonable compromise, giving less weight to the observations with larger variance than the standard NLS estimator, without giving too much weight to observations more prone to contamination by measurement error and less informative about the curvature of $E[y_i|x]$.[16]

B. Testing

In this subsection we consider tests for the particular pattern of heteroskedasticity assumed by PML estimators, focusing on the PPML estimator. Although PML estimators are consistent even when the variance function is misspecified, the researcher can use these tests to check if a different PML estimator would be more appropriate and to decide whether or not the use of a nonparametric estimator of the variance is warranted.

Manning and Mullahy (2001) suggested that if

$$V[y_i|x] = \lambda_0 E[y_i|x]^{\lambda_1}, \tag{10}$$

the choice of the appropriate PML estimator can be based on a Park-type regression (Park, 1966). Their approach is based on the idea that if equation (10) holds and an initial consistent estimate of $E[y_i|x]$ is available, then λ_1 can be consistently estimated using an appropriate auxiliary regression. Specifically, following Park (1966), Manning and Mullahy (2001) suggest that λ_1 can be estimated using the auxiliary model

$$\ln (y_i - \breve{y}_i)^2 = \ln \lambda_0 + \lambda_1 \ln \breve{y}_i + v_i, \tag{11}$$

where $\breve{y}_i$ denotes the estimated value of $E[y_i|x]$. Unfortunately, as the discussion in the previous sections should have made clear, this approach based on the log-linearization of equation (10) is valid only under very restrictive conditions on the conditional distribution of y_i. However, it is easy to see that this procedure is valid when the constant-elasticity model can be consistently estimated in the log linear form. Therefore, using equation (11) a test for $H_0 : \lambda_1 = 2$ based on a nonrobust covariance estimator provides a check on the adequacy of the estimator based on the log linear model.

A more robust alternative, which is mentioned by Manning and Mullahy (2001) in a footnote, is to estimate λ_1 from

$$(y_i - \breve{y}_i)^2 = \lambda_0(\breve{y}_i)^{\lambda_1} + \xi_i, \tag{12}$$

using an appropriate PML estimator. The approach based on equation (12) is asymptotically valid, and inference about λ_1 can be based on the usual Eicker-White robust covariance matrix estimator. For example, the hypothesis that $V[y_i|x]$ is proportional to $E[y_i|x]$ is accepted if the appropriate confidence interval for λ_1 contains 1. However, if the purpose is to test the adequacy of a particular value of λ_1, a slightly simpler method based on the Gauss-Newton regression (see Davidson & MacKinnon, 1993) is available.

Specifically, to check the adequacy of the PPML for which $\lambda_1 = 1$ and $\breve{y}_i = \exp(x_i\tilde{\beta})$, equation (12) can be expanded in a Taylor series around $\lambda_1 = 1$, leading to

$$(y_i - \breve{y}_i)^2 = \lambda_0\breve{y}_i + \lambda_0(\lambda_1 - 1)(\ln y_i)\breve{y}_i + \xi_i.$$

Now, the hypothesis that $V[y_i|x] \propto E[y_i|x]$ can be tested against equation (10) simply by checking the significance of the parameter $\lambda_0(\lambda_1 - 1)$. Because the error term ξ_i is unlikely to be homoskedastic, the estimation of the Gauss-Newton regression should be performed using weighted least squares. Assuming that in equation (12) the variance is also proportional to the mean, the appropriate weights are given by $\exp(-x_i\tilde{\beta})$, and therefore the test can be performed by estimating

$$\begin{aligned}(y_i - \breve{y}_i)^2/\sqrt{\breve{y}_i} = &\ \lambda_0\sqrt{\breve{y}_i} \\ &+ \lambda_0(\lambda_1 - 1)(\ln \breve{y}_i)\sqrt{\breve{y}_i} + \xi_i^*\end{aligned} \tag{13}$$

by OLS and testing the statistical significance of $\lambda_0(\lambda_1 - 1)$ using a Eicker-White robust covariance matrix estimator.[17]

In the next section, a small simulation is used to study the Gauss-Newton regression test for the hypothesis that $V[y_i|x] \propto E[y_i|x]$, as well as the Park-type test for the hypothesis that the constant-elasticity model can be consistently estimated in the log linear form.

IV.   A Simulation Study

This section reports the results of a small simulation study designed to assess the performance of different methods to estimate constant-elasticity models in the presence of heteroskedasticity and rounding errors. As a by-product, we also obtain some evidence on the finite-sample performance of the specification tests presented above. These experiments are centered around the following multiplicative model:

---

[15] Frankel and Wei (1993) and Frankel (1997) suggest that larger countries should be given more weight in the estimation of gravity equations. This would be appropriate if the errors in the model were just the result of measurement errors in the dependent variable. However, if it is accepted that the gravity equation does not hold exactly, measurement errors account for only part of the dispersion of trade data around the gravity equation.

[16] It is worth noting that the PPML estimator can be easily adapted to deal with endogenous regressors (Windmeijer & Santos Silva, 1997) and panel data (Wooldridge, 1999). These extensions, however, are not pursued here.

[17] Notice that to test $V[y_i|x] \propto E[y_i|x]$ against alternatives of the form $V[y_i|x] = \lambda_0 \exp [x_i (\beta + \lambda)]$, the appropriate auxiliary regression would be

$$(y_i - \breve{y}_i)^2/\sqrt{\breve{y}_i} = \lambda_0\sqrt{\breve{y}_i} + \lambda_0\lambda x_i\sqrt{\breve{y}_i} + \xi_i^*,$$

and the test could be performed by checking the joint significance of the elements of $\lambda_0\lambda$. If the model includes a constant, one of the regressors in the auxiliary regression is redundant and should be dropped.

$$E[y_i|x] = \mu(x_i\beta) = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}),$$
$$i = 1, \dots, 1000. \qquad (14)$$

Because, in practice, regression models often include a mixture of continuous and dummy variables, we replicate this feature in our experiments: x_{1i} is drawn from a standard normal, and x_2 is a binary dummy variable that equals 1 with a probability of 0.4.[18] The two covariates are independent, and a new set of observations of all variables is generated in each replication using $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$. Data on y are generated as

$$y_i = \mu(x_i\beta)\eta_i, \qquad (15)$$

where η_i is a log normal random variable with mean 1 and variance σ_i^2. As noted before, the slope parameters in equation (14) can be estimated using the log linear form of the model only when σ_i^2 is constant, that is, when $V[y_i|x]$ is proportional to $\mu(x_i\beta)^2$.

In these experiments we analyzed PML estimators of the multiplicative model and different estimators of the log-linearized model. The consistent PML estimators studied were: NLS, gamma pseudo-maximum-likelihood (GPML), and PPML. Besides these estimators, we also considered the standard OLS estimator of the log linear model (here called simply OLS); the OLS estimator for the model where the dependent variable is $y_i + 1$ [OLS($y + 1$)]; a truncated OLS estimator to be discussed below; and the threshold tobit of Eaton and Tamura (1994) (ET-tobit).[19]

To assess the performance of the estimators under different patterns of heteroskedasticity, we considered the four following specifications of σ_i^2:

Case 1: $\sigma_i^2 = \mu(x_i\beta)^{-2}$; $V[y_i|x] = 1$.
Case 2: $\sigma_i^2 = \mu(x_i\beta)^{-1}$; $V[y_i|x] = \mu(x_i\beta)$.
Case 3: $\sigma_i^2 = 1$; $V[y_i|x] = \mu(x_i\beta)^2$.
Case 4: $\sigma_i^2 = \mu(x_i\beta)^{-1} + \exp(x_{2i})$; $V[y_i|x] = \mu(x_i\beta) + \exp(x_{2i})\,\mu(x_i\beta)^2$.

In case 1 the variance of ε_i is constant, implying that the NLS estimator is optimal. Although, as argued before, this case is unrealistic for models of bilateral trade, it is included in the simulations for completeness. In case 2, the conditional variance of y_i equals its conditional mean, as in the Poisson distribution. The pseudo-maximum-likelihood estimator based on the Poisson distribution is optimal in this situation. Case 3 is the special case in which OLS estimation of the log linear model is consistent for the slope parameters of equation (14). Moreover, in this case the log linear model

not only corrects the heteroskedasticity in the data, but, because η_i is log normal, it is also the maximum likelihood estimator. The GPML is the optimal PML estimator in this case, but it should be outperformed by the true maximum likelihood estimator. Finally, case 4 is the only one in which the conditional variance does not depend exclusively on the mean. The variance is a quadratic function of the mean, as in case 3, but it is not proportional to the square of the mean.

We carried out two sets of experiments. The first set was aimed at studying the performance of the estimators of the multiplicative and the log linear models under different patterns of heteroskedasticity. In order to study the effect of the truncation on the performance of the OLS, and given that this data-generating mechanism does not produce observations with $y_i = 0$, the log linear model was also estimated using only the observations for which $y_i > 0.5$ [OLS($y > 0.5$)]. This reduces the sample size by approximately 25% to 35%, depending on the pattern of heteroskedasticity. The estimation of the threshold tobit was also performed using this dependent variable. Notice that, although the dependent variable has to cross a threshold to be observable, the truncation mechanism used here is not equal to the one assumed by Eaton and Tamura (1994). Therefore, in all these experiments the ET-tobit will be slightly misspecified and the results presented here should be viewed as a check of its robustness to this problem.

The second set of experiments studied the estimators' performance in the presence of rounding errors in the dependent variable. For that purpose, a new random variable was generated by rounding to the nearest integer the values of y_i obtained in the first set of simulations. This procedure mimics the rounding errors in official statistics and generates a large number of zeros, a typical feature of trade data. Because the model considered here generates a large proportion of observations close to zero, rounding down is much more frequent than rounding up. As the probability of rounding up or down depends on the covariates, this procedure will necessarily bias the estimates, as discussed before. The purpose of the study is to gauge the magnitude of these biases. Naturally, the log linear model cannot be estimated in these conditions, because the dependent variable equals 0 for some observations. Following what is the usual practice in these circumstances, the truncated OLS estimation of the log-linear model was performed dropping the observations for which the dependent variable equals 0. Notice that the observations discarded with this procedure are exactly the same that are discarded by OLS($y > 0.5$) in the first set of experiments. Therefore, this estimator is also denoted OLS($y > 0.5$).

The results of the two sets of experiments are summarized in table 1, which displays the biases and standard errors of the different estimators of β obtained with 10,000 replicas of the simulation procedure described above. Only results for β_1 and β_2 are presented, as these are generally the parameters of interest.

---

[18] For example, in gravity equations, continuous variables (which are all strictly positive) include income and geographical distance. In equation (14), x_1 can be interpreted as (the logarithm of) one of these variables. Examples of binary variables include dummies for free-trade agreements, common language, colonial ties, contiguity, and access to water.

[19] We also studied the performance of other variants of the tobit model, finding very poor results.

TABLE 1.—SIMULATION RESULTS UNDER DIFFERENT FORMS OF HETEROSKEDASTICITY

	Results Without Rounding Error				Results with Rounding Error			
	β_1		β_2		β_1		β_2	
Estimator:	Bias	S.E.	Bias	S.E.	Bias	S.E.	Bias	S.E.
Case 1: $V[y_i|x] = 1$								
PPML	−0.00004	0.016	0.00009	0.027	0.01886	0.017	0.02032	0.029
NLS	−0.00006	0.008	−0.00003	0.017	0.00195	0.008	0.00274	0.018
GPML	0.01276	0.068	0.00754	0.082	0.10946	0.096	0.09338	0.108
OLS	0.39008	0.039	0.35568	0.054	—	—	—	—
ET-tobit	−0.47855	0.030	−0.47786	0.032	−0.49981	0.030	−0.49968	0.032
OLS($y > 0.5$)	−0.16402	0.027	−0.15487	0.038	−0.22121	0.026	−0.21339	0.036
OLS($y + 1$)	−0.40237	0.014	−0.37683	0.022	−0.37752	0.015	−0.34997	0.024
Case 2: $V[y_i|x] = \mu\ (x_i\beta)$								
PPML	−0.00011	0.019	0.00009	0.039	0.02190	0.020	0.02334	0.041
NLS	0.00046	0.033	0.00066	0.057	0.00262	0.033	0.00360	0.057
GPML	0.00376	0.043	0.00211	0.062	0.13243	0.073	0.11331	0.087
OLS	0.21076	0.030	0.19960	0.049	—	—	—	—
ET-tobit	−0.42394	0.028	−0.42316	0.033	−0.45518	0.028	−0.45513	0.033
OLS($y > 0.5$)	−0.17868	0.026	−0.17220	0.043	−0.24405	0.026	−0.23889	0.040
OLS($y + 1$)	−0.42371	0.015	−0.39931	0.025	−0.39401	0.016	−0.36806	0.028
Case 3: $V[y_i|x] = \mu\ (x_i\beta)^2$								
PPML	−0.00526	0.091	−0.00228	0.130	0.02332	0.091	0.02812	0.133
NLS	0.23539	3.066	0.07323	1.521	0.23959	3.082	0.07852	1.521
GPML	−0.00047	0.041	−0.00029	0.083	0.17134	0.068	0.14442	0.104
OLS	0.00015	0.032	−0.00003	0.064	—	—	—	—
ET-tobit	−0.31908	0.044	−0.32161	0.058	−0.36480	0.043	−0.36789	0.056
OLS($y > 0.5$)	−0.34480	0.039	−0.34614	0.064	−0.41006	0.037	−0.41200	0.060
OLS($y + 1$)	−0.51804	0.021	−0.50000	0.038	−0.48564	0.022	−0.46597	0.040
Case 4: $V[y_i|x] = \mu\ (x_i\beta) + \exp(x_{2i})\ \mu\ (x_i\beta)^2$								
PPML	−0.00696	0.103	−0.00647	0.144	0.02027	0.104	0.01856	0.146
NLS	0.35139	7.516	0.08801	1.827	0.35672	7.521	0.09239	1.829
GPML	0.00322	0.057	−0.00137	0.083	0.12831	0.085	0.10245	0.129
OLS	0.13270	0.039	−0.12542	0.075	—	—	—	—
ET-tobit	−0.29908	0.049	−0.42731	0.063	−0.34351	0.047	−0.46225	0.060
OLS($y > 0.5$)	−0.39217	0.042	−0.41391	0.070	−0.45188	0.040	−0.46173	0.066
OLS($y + 1$)	−0.51440	0.021	−0.58087	0.041	−0.48627	0.022	−0.56039	0.044

As expected, OLS only performs well in case 3. In all other cases this estimator is clearly inadequate because, despite its low dispersion, it is often badly biased. Moreover, the sign and magnitude of the bias vary considerably. Therefore, even when the dependent variable is strictly positive, estimation of constant-elasticity models using the log-linearized model cannot generally be recommended. As for the modifications of the log-linearized model designed to deal with the zeros of the dependent variable—ET-tobit, OLS($y + 1$), and OLS($y > 0.5$)—their performance is also very disappointing. These results clearly emphasize the need to use adequate methods to deal with the zeros in the data and raise serious doubts about the validity of the results obtained using the traditional estimators based on the log linear model. Overall, except under very special circumstances, estimation based on the log-linear model cannot be recommended.

One remarkable result of this set of experiments is the extremely poor performance of the NLS estimator. Indeed, when the heteroskedasticity is more severe (cases 3 and 4),

this estimator, despite being consistent, leads to very poor results because of its erratic behavior.[20] Therefore, it is clear that the loss of efficiency caused by some of the forms of heteroskedasticity considered in these experiments is strong enough to render this estimator useless in practice.

In the first set of experiments, the results of the gamma PML estimator are very good. Indeed, when no measurement error is present, the biases and standard errors of the GPML estimator are always among the lowest. However, this estimator is very sensitive to the form of measurement error considered in the second set of experiments, consistently leading to sizable biases. These results, like those of the NLS, clearly illustrate the danger of using a PML estimator that gives extra weight to the noisier observations.

As for the performance of the Poisson PML estimator, the results are very encouraging. In fact, when no rounding error is present, its performance is reasonably good in all

[20] Manning and Mullahy (2001) report similar results.

cases. Moreover, although some loss of efficiency is noticeable as one moves away from case 2, in which it is an optimal estimator, the biases of the PPML are always small.[21] Moreover, the results obtained with rounded data suggest that the Poisson-based PML estimator is relatively robust to this form of measurement error of the dependent variable. Indeed, the bias introduced by the rounding-off errors in the dependent variable is relatively small, and in some cases it even compensates the bias found in the first set of experiments. Therefore, because it is simple to implement and reliable in a wide variety of situations, the Poisson PML estimator has the essential characteristics needed to make it the new workhorse for the estimation of constant-elasticity models.

Obviously, the sign and magnitude of the bias of the estimators studied here depend on the particular specification considered. Therefore, the results of these experiments cannot serve as an indicator of what can be expected in other situations. However, it is clear that, apart from the Poisson PML method, all estimators will often be very misleading.

These experiments were also used to study the finite-sample performance of the Gauss-Newton regression (GNR) test for the adequacy of the Poisson PML based on equation (13) and of the Park test advocated by Manning and Mullahy (2001), which, as explained above, is valid only to check for the adequacy of the estimator based on the log linear model.[22] Given that the Poisson PML estimator is the only estimator with a reasonable behavior under all the cases considered, these tests were performed using residuals and estimates of μ $(x_i\beta)$ from the Poisson regression. Table 2 contains the rejection frequencies of the null hypothesis at the 5% nominal level for both tests in the four cases considered in the two sets of experiments. In this table the rejection frequencies under the null hypothesis are given in bold.

In as much as both tests have adequate behavior under the null and reveal reasonable power against a wide range of alternatives, the results suggest that these tests are important tools to assess the adequacy of the standard OLS estimator of the log linear model and of the proposed Poisson PML estimator.

V. The Gravity Equation

In this section, we use the PPML estimator to quantitatively assess the determinants of bilateral trade flows, uncovering significant differences in the roles of various measures of size and distance from those predicted by the

TABLE 2.—REJECTION FREQUENCIES AT THE 5% LEVEL FOR THE TWO SPECIFICATION TESTS

	Frequency	
Case	GNR Test	Park Test
Without Measurement Error		
1	0.91980	1.00000
2	**0.05430**	1.00000
3	0.58110	**0.06680**
4	0.49100	0.40810
With Measurement Error		
1	0.91740	1.00000
2	0.14980	1.00000
3	0.57170	1.00000
4	0.47580	1.00000

logarithmic tradition. We perform the comparison of the two techniques using both the traditional and the Anderson–van Wincoop (2003) specifications of the gravity equation.

For the sake of completeness, we also compare the PPML estimates with those obtained from alternative ways researchers have used to deal with zero values for trade. In particular, we present the results obtained with the tobit estimator used in Eaton and Tamura (1994), OLS estimator applied to $\ln(1 + T_{ij})$, and a standard nonlinear least squares estimator. The results obtained with these estimators are presented for both the traditional and the Anderson–van Wincoop specifications.

A. The Data

The analysis covers a cross section of 136 countries in 1990. Hence, our data set consists of 18,360 observations of bilateral export flows (136×135 country pairs). The list of countries is reported in table A1 in the appendix. Information on bilateral exports comes from Feenstra et al. (1997). Data on real GDP per capita and population come from the World Bank's (2002) *World Development Indicators.* Data on location and dummies indicating contiguity, common language (official and second languages), colonial ties (direct and indirect links), and access to water are constructed from the CIA's (2002) *World Factbook.* The data on language and colonial links are presented on tables A2 and A3 in the appendix.[23] Bilateral distance is computed using the great circle distance algorithm provided by Andrew Gray (2001). Remoteness—or relative distance—is calculated as the (log of) GDP-weighted average distance to all other countries (see Wei, 1996). Finally, information on preferential trade agreements comes from Frankel (1997), complemented with data from the World Trade Organization. The list of preferential trade agreements (and stronger forms of trade agreements) considered in the analysis is displayed in table A4 in the appendix. Table A5 in the appendix provides

---

[21] These results are in line with those reported by Manning and Mullahy (2001).

[22] To illustrate the pitfalls of the procedure suggested by Manning and Mullahy (2001), we note that the means of the estimates of λ_1 obtained using equation (11) in cases 1, 2, and 3 (without measurement error) were 0.58955, 1.29821, and 1.98705, whereas the true values of λ_1 in these cases are, respectively, 0, 1, and 2.

[23] Alternative estimates based on Boisso and Ferrantino's (1997) index of language similarity are available, at request, from the authors.

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

a description of the variables and displays the summary statistics.

B. Results

The Traditional Gravity Equation: Table 3 presents the estimation outcomes resulting from the various techniques for the traditional gravity equation. The first column reports OLS estimates using the logarithm of exports as the dependent variable; as noted before, this regression leaves out pairs of countries with zero bilateral trade (only 9,613 country pairs, or 52% of the sample, exhibit positive export flows).

The second column reports the OLS estimates using $\ln(1 + T_{ij})$ as dependent variable, as a way of dealing with zeros. The third column presents tobit estimates based on Eaton and Tamura (1994). The fourth column shows the results of standard NLS. The fifth column reports Poisson estimates using only the subsample of positive-trade pairs. Finally, the sixth column shows the Poisson results for the whole sample (including zero-trade pairs).

The first point to notice is that PPML-estimated coefficients are remarkably similar using the whole sample and using the positive-trade subsample.[24] However, most coef-

ficients differ—oftentimes significantly—from those obtained using OLS. This suggests that in this case, heteroskedasticity (rather than truncation) is responsible for the differences between PPML results and those of OLS using only the observations with positive exports. Further evidence on the importance of the heteroskedasticity is provided by the two-degrees-of-freedom special case of White's test for heteroskedasticity (see Wooldridge, 2002, p. 127), which leads to a test statistic of 476.6 and to a p-value of 0. That is, the null hypothesis of homoskedastic errors is unequivocally rejected.

Poisson estimates reveal that the coefficients on importer's and exporter's GDPs in the traditional equation are not, as generally believed, close to 1. The estimated GDP elasticities are just above 0.7 (s.e. = 0.03). OLS generates significantly larger estimates, especially on exporter's GDP (0.94, s.e. = 0.01). Although all these results are conditional on the particular specification used,[25] it is worth pointing out that unit income elasticities in the simple gravity framework are at odds with the observation that the trade-to-GDP ratio decreases with increasing total GDP, or, in other words, that smaller countries tend to be more open to international trade.[26]

---

[24] The reason why truncation has little effect in this case is that observations with zero trade correspond to pairs for which the estimated value of trade is close to zero. Therefore, the corresponding residuals are also close to zero, and their elimination from the sample has little effect.

[25] This result holds when one looks at the subsample of OECD countries. It is also robust to the exclusion of GDP per capita from the regressions.

[26] Note also that PPML predicts almost equal coefficients for the GDPs of exporters and importers.

TABLE 4.—RESULTS OF THE TESTS FOR TYPE OF HETEROSKEDASTICITY
(p-VALUES)

Test (Null Hypothesis)	Exports > 0	Full Sample
GNR ($V[y_i|x] \propto \mu (x_i\beta)$)	0.144	0.115
Park (OLS is valid)	0.000	0.000

The role of geographical distance as trade deterrent is significantly larger under OLS; the estimated elasticity is -1.17 (s.e. = 0.03), whereas the Poisson estimate is -0.78 (s.e. = 0.06). This lower estimate suggests a smaller role for transport costs in the determination of trade patterns. Furthermore, Poisson estimates indicate that, after controlling for bilateral distance, sharing a border does not influence trade flows, whereas OLS, instead, generates a substantial effect: It predicts that trade between two contiguous countries is 37% larger than trade between countries that do not share a border.[27]

We control for remoteness to allow for the hypothesis that larger distances to all other countries might increase bilateral trade between two countries.[28] Poisson regressions support this hypothesis, whereas OLS estimates suggest that only exporter's remoteness increases bilateral flows between two given countries. Access to water appears to be important for trade flows, according to Poisson regressions; the negative coefficients on the land-locked dummies can be interpreted as an indication that ocean transportation is significantly cheaper. In contrast, OLS results suggest that whether or not the exporter is landlocked does not influence trade flows, whereas a landlocked importer experiences lower trade. (These asymmetries in the effects of remoteness and access to water for importers and exporters are hard to interpret.) We also explore the role of colonial heritage, obtaining, as before, significant discrepancies: Poisson regressions indicate that colonial ties play no role in determining trade flows, once a dummy variable for common language is introduced. OLS regressions, instead, generate a sizeable effect (countries with a common colonial past trade almost 45% more than other pairs). Language is statistically and economically significant under both estimation procedures.

Strikingly, in the traditional gravity equation, preferential-trade agreements play a much smaller—although still substantial—role according to Poisson regressions. OLS estimates suggest that preferential trade agreements raise expected bilateral trade by 63%, whereas Poisson estimates indicate an average enhancement below 20%.

Preferential trade agreements might also cause trade diversion; if this is the case, the coefficient on the trade-

agreement dummy will not reflect the net effect of trade agreements. To account for the possibility of diversion, we include an additional dummy, *openness,* similar to that used by Frankel (1997). This dummy takes the value 1 whenever one (or both) of the countries in the pair is part of a preferential trade agreement, and thus it captures the extent of trade between members and nonmembers of a preferential trade agreement. The sum of the coefficients on the trade agreement and the openness dummies gives the net creation effect of trade agreements. OLS suggests that trade destruction comes from trade agreements. Still, the net creation effect is around 40%. In contrast, Poisson regressions provide no significant evidence of trade diversion, although the point estimates are of the same order of magnitude under both methods.

Hence, even when allowing for trade diversion effects, on average, the Poisson method estimates a smaller effect of preferential trade agreements on trade, approximately half of that indicated by OLS. The contrast in estimates suggests that the biases generated by standard regressions can be substantial, leading to misleading inferences and, perhaps, erroneous policy decisions.

We now turn briefly to the results of the other estimation methods. OLS on $\ln(1 + T_{ij})$ and tobit give very close estimates for most coefficients. Like OLS, they yield large estimates for the elasticity of bilateral trade with respect to distance. Unlike OLS, however, they produce insignificant coefficients for the contiguity dummy. They both generate extremely large and statistically significant coefficients for the trade-agreement dummy. The first method predicts that trade between two countries that have signed a trade agreement is on average 266% larger than that between countries without an agreement. The second predicts that trade between countries in such agreements is on average 100% larger. NLS tends to generate somewhat different estimates. The elasticity of trade with respect to the exporter's GDP is significantly smaller than with OLS, but the corresponding elasticity with respect to importer's GDP is significantly larger than with OLS. The estimated distance elasticity is smaller than with OLS and bigger than with Poisson. Like the other methods, NLS predicts a significant and large effect for free-trade agreements.

It is noteworthy that all methods, except the PPML, lead to puzzling asymmetries in the elasticities with respect to importer and exporter characteristics (especially remoteness and access to water).

To check the adequacy of the estimated models, we performed a heteroskedasticity-robust RESET test (Ramsey, 1969). This is essentially a test for the correct specification of the conditional expectation, which is performed by checking the significance of an additional regressor constructed as $(x'b)^2$, where b denotes the vector of estimated parameters. The corresponding p-values are reported at the bottom of table 3. In the OLS regression, the test rejects the hypothesis that the coefficient on the test variable is 0. This

---

[27] The formula to compute this effect is $(e^{b_i} - 1) \times 100\%$, where b_i is the estimated coefficient.

[28] To illustrate the role of remoteness, consider two pairs of countries, (i, j) and (k, l), and assume that the distance between the countries in each pair is the same ($D_{ij} = D_{kl}$), but i and j are closer to other countries. In this case, the most remote countries, k and $l,$ will tend to trade more between each other because they do not have alternative trading partners. See Deardoff (1998).

TABLE 5.—THE ANDERSON–VAN WINCOOP 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 distance	−1.347**	−1.332**	−1.272**	−0.582**	−0.770**	−0.750**
	(0.031)	(0.036)	(0.029)	(0.088)	(0.042)	(0.041)
Contiguity dummy	0.174	−0.399*	−0.253	0.458**	0.352**	0.370**
	(0.130)	(0.189)	(0.135)	(0.121)	(0.090)	(0.091)
Common-language dummy	0.406**	0.550**	0.485**	0.926**	0.418**	0.383**
	(0.068)	(0.066)	(0.057)	(0.116)	(0.094)	(0.093)
Colonial-tie dummy	0.666**	0.693**	0.650**	−0.736**	0.038	0.079
	(0.070)	(0.067)	(0.059)	(0.178)	(0.134)	(0.134)
Free-trade agreement dummy	0.310**	0.174	0.137**	1.017**	0.374**	0.376**
	(0.098)	(0.138)	(0.098)	(0.170)	(0.076)	(0.077)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9613	18360	18360	18360	9613	18360
RESET test p-values	0.000	0.000	0.000	0.000	0.564	0.112

means that the model estimated using the logarithmic specification is inappropriate. A similar result is found for the OLS estimated using $\ln(1 + T_{ij})$ as the dependent variable and for the NLS. In contrast, the models estimated using the Poisson regressions pass the RESET test, that is, the RESET test provides no evidence of misspecification of the gravity equations estimated using the PPML. With this particular specification, the model estimated using tobit also passes the test for the traditional gravity equation.

Finally, we also check whether the particular pattern of heteroskedasticity assumed by the models is appropriate. As explained in section III B, the adequacy of the log linear model was checked using the Park-type test, whereas the hypothesis $V[y_i|x] \propto \mu(x_i\beta)$ was tested by evaluating the significance of the coefficient of $(\ln \breve{y}_i)\sqrt{\breve{y}_i}$ in the Gauss-Newton regression indicated in equation (13). The p-values of the tests are reported in table 4. Again, the log linear specification is unequivocally rejected. On the other hand, these results indicate that the estimated coefficient on $(\ln \breve{y}_i)$ $\sqrt{\breve{y}_i}$ is insignificantly different from 0 at the usual 5% level. This implies that the Poisson PML assumption $V[y_i|x] = \lambda_0 E[y_i|x]$ cannot be rejected at this significance level.

The Anderson–van Wincoop Gravity Equation: Table 5 presents the estimated coefficients for the Anderson–van Wincoop (2003) gravity equation, which controls more properly for multilateral resistance terms by introducing exporter- and importer-specific effects. As before, the columns show, respectively, the estimated coefficients obtained using OLS on the log of exports, OLS on $\ln(1 + T_{ij})$, tobit, NLS, PPML on the positive-trade sample, and PPML. Note that, because this exercise uses cross-sectional data, we can only identify bilateral variables.[29]

As with the standard specification of the gravity equation, we find that, using the Anderson–van Wincoop (2003) specification, estimates obtained with the Poisson method vary little when only the positive-trade subsample is used. Moreover, we find again strong evidence that the errors of the log linear model estimated using the sample with positive trade are heteroskedastic. With this specification, the two-degree-of-freedom special case of White's test for heteroskedasticity leads to a test statistic of 469.2 and a p-value of 0.

Because we are now conditioning on a much larger set of controls, it is not surprising to find that most coefficients are sensitive to the introduction of fixed effects. For example, in the Poisson method, although the distance elasticity remains about the same and the coefficient on common colonial ties is still insignificant, the effect on common language is now smaller and the coefficient on free-trade agreements is larger. The results of the other estimation methods are generally much more sensitive to the inclusion of the fixed effects.

Comparing the results of PPML and OLS for the positive-trade subsample, the following observations are in order. The distance elasticity is substantially larger under OLS (−1.35 versus −0.75). Sharing a border has a positive effect on trade under Poisson, but no significant effect under OLS. Sharing a common language has similar effects under the two techniques. Common colonial ties have strong effects under OLS (with an average enhancement effect of 100%), whereas Poisson predicts no significant effect. Finally, the

---

[29] Anderson and van Wincoop (2003) impose unit income elasticities by using as the dependent variable the log of exports divided by the product of the countries' GDPs. Because we are working with cross-sectional data and the model specification includes importer and exporter fixed effects, income elasticities cannot be identified, and there is no need to impose restrictions on them. Still, the estimation of the PML models could be performed using as the dependent variable the ratio of exports to the product of the GDPs. This would downweight the observations with large

values of the product of the GDPs, implicitly assuming that the variance of the error term is proportional to the square of this product. This is contrary to what is advocated by Frankel and Wei (1993) and Frankel (1997), who suggest that larger countries should be given more weight in the estimation of gravity equations because they generally have better data. In any case, whether this should be done or not is an empirical question, and the right course of action depends on the pattern of heteroskedasticity. With our data, using the ratio of exports to the product of GDPs as the dependent variable leads to models that are rejected by the specification tests. Therefore, the implied assumptions about the pattern of heteroskedasticity are not supported by our data. Hence, we use exports as the dependent variable of the gravity equation, and not the ratio of exports to the product of GDPs.

TABLE 6.—RESULTS OF THE TESTS FOR TYPE OF HETEROSKEDASTICITY
(*p*-VALUES)

Test (Null Hypothesis)	Exports > 0	Full Sample
GNR ($V[y_i|x] \propto \mu\ (x_i\beta)$)	0.100	0.070
Park (OLS is valid)	0.000	0.000

two techniques produce reasonably similar estimates for the coefficient on the trade-agreement dummy, implying a trade-enhancement effect of the order of 40%.

As before, the other estimation methods lead to some puzzling results. For example, OLS on $\ln(1 + T_{ij})$ now yields a significantly negative effect of contiguity, and under NLS, the coefficient on common colonial ties becomes significantly negative.

To complete the study, we performed the same set of specification tests used before. The *p*-values of the heteroskedasticity-robust RESET test at the bottom of table 5 suggest that with the Anderson–van Wincoop (2003) specification of the gravity equation, only the models estimated by the PPML method are adequate. The *p*-values of the tests to check whether the particular pattern of heteroskedasticity assumed by the models is appropriate are reported in table 6. As in the traditional gravity equation, the log linear specification is unequivocally rejected. On the other hand, these results indicate that the estimated coefficient on $(\ln \check{y}_i) \sqrt{\check{y}_i}$ is insignificantly different from 0 at the usual 5% level. This implies that the Poisson PML assumption $V[y_i|x] = \lambda_0 E[y_i|x]$ cannot be rejected at this significance level.

To sum up, whether or not fixed effects are used in the specification of the model, we find strong evidence that estimation methods based on the log-linearization of the gravity equation suffer from severe misspecification, which hinders the interpretation of the results. NLS is also generally unreliable. In contrast, the models estimated by PPML show no signs of misspecification and, in general, do not produce the puzzling results generated by the other methods.[30]

VI. Conclusions

In this paper, we argue that the standard empirical methods used to estimate gravity equations are inappropriate.

---

[30] It is worth noting that the large differences in estimates among the various methods persist when we look at a smaller subsample of countries that account for most of world trade and, quite likely, have better data. More specifically, we run similar regressions for the subsample of 63 countries included in Frankel's (1997) study. These countries accounted for almost 90% of the world trade reported to the United Nations in 1992. One advantage of this subsample is that the number of zeros is significantly reduced. Heteroskedasticity, however, is still a problem: The null hypothesis of homoskedasticity is rejected in both the traditional and the fixed-effects gravity equations. As with the full sample, PPML generates a smaller role for distance and common language than OLS, and, unlike OLS, PPML predicts no role for colonial ties. In line with the findings documented in Frankel (1997), the OLS estimated coefficient on the free-trade-agreement dummy is negative in both specifications of the gravity equation, whereas PPML predicts a positive and significant effect (slightly bigger than that found for the whole sample). These results are available—on request—from the authors.

The basic problem is that log-linearization (or, indeed, any nonlinear transformation) of the empirical model in the presence of heteroskedasticity leads to inconsistent estimates. This is because the expected value of the logarithm of a random variable depends on higher-order moments of its distribution. Therefore, if the errors are heteroskedastic, the transformed errors will be generally correlated with the covariates. An additional problem of log-linearization is that it is incompatible with the existence of zeros in trade data, which led to several unsatisfactory solutions, including truncation of the sample (that is, elimination of zero-trade pairs) and further nonlinear transformations of the dependent variable.

We argue that the biases are present both in the traditional specification of the gravity equation and in the Anderson–van Wincoop (2003) specification, which includes country-specific fixed effects.

To address the various estimation problems, we propose a simple Poisson pseudo-maximum-likelihood method and assess its performance using Monte Carlo simulations. We find that in the presence of heteroskedasticity the standard methods can severely bias the estimated coefficients, casting doubt on previous empirical findings. Our method, instead, is robust to different patterns of heteroskedasticity and, in addition, provides a natural way to deal with zeros in trade data.

We use our method to reestimate the gravity equation and document significant differences from the results obtained using the log linear method. For example, income elasticities in the traditional gravity equation are systematically smaller than those obtained with log-linearized OLS regressions. In addition, in both the traditional and Anderson–van Wincoop specifications of the gravity equation, OLS estimation exaggerates the role of geographical proximity and colonial ties. RESET tests systematically favor the Poisson PML technique. The results suggest that heteroskedasticity (rather than truncation of the data) is responsible for the main differences.

Log-linearized equations estimated by OLS are of course used in many other areas of empirical economics and econometrics. Our Monte Carlo simulations and the regression outcomes indicate that in the presence of heteroskedasticity this practice can lead to significant biases. These results suggest that, at least when there is evidence of heteroskedasticity, the Poisson pseudo-maximum-likelihood estimator should be used as a substitute for the standard log linear model.

REFERENCES

Anderson, J., "A Theoretical Foundation for the Gravity Equation," *American Economic Review* 69 (1979), 106–116.
Anderson, J., and E. van Wincoop, "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review* 93 (2003), 170–192.
Bergstrand, J., "The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence," this RE-VIEW, 69 (1985), 474–481.

Boisso, D., and M. Ferrantino, "Economic and Cultural Distance in International Trade: Empirical Puzzles," *Journal of Economic Integration* 12 (1997), 456–484.

Cameron, A. C., and P. K. Trivedi, *Regression Analysis of Count Data* (Cambridge: Cambridge University Press, 1998).

Central Intelligence Agency, *World Factbook,* http://www.cia.gov/cia/publications/factbook/ (2002).

Davidson, R., and J. G. MacKinnon, *Estimation and Inference in Econometrics* (Oxford: Oxford University Press, 1993).

Davis, D., "Intra-industry Trade: A Hecksher-Ohlin-Ricardo Approach," *Journal of International Economics* 39 (1995), 201–226.

Deardoff, A., "Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?" in Jeffrey Frankel (Ed.), *The Regionalization of the World Economy* (Chicago: University of Chicago Press, 1998).

Delgado, M., "Semiparametric Generalized Least Squares Estimation in the Multivariate Nonlinear Regression Model," *Econometric Theory* 8 (1992), 203–222.

Delgado, M., and T. J. Kniesner, "Count Data Models with Variance of Unknown Form: An Application to a Hedonic Model of Worker Absenteeism," this REVIEW, 79 (1997), 41–49.

Eaton, J., and S. Kortum, "Technology, Geography and Trade," NBER working paper no. 6253 (2001).

Eaton, J., and A. Tamura, "Bilateralism and Regionalism in Japanese and US Trade and Direct Foreign Investment Patterns," *Journal of the Japanese and International Economics* 8 (1994), 478–510.

Eichengreen, B., and D. Irwin, "Trade Blocs, Currency Blocs, and the Reorientation of World Trade in the 1930's," *Journal of International Economics,* 38 (1995), 1–24.

Eicker, F., "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *The Annals of Mathematical Statistics* 34 (1963), 447–456.

Feenstra, R. C., R. E. Lipsey, and H. P. Bowen, "World Trade Flows, 1970–1992, with Production and Tariff Data," NBER working paper no. 5910 (1997).

Feenstra, R., J. Markusen, and A. Rose, "Using the Gravity Equation to Differentiate among Alternative Theories of Trade," *Canadian Journal of Economics* 34 (2001), 430–447.

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

Frankel, J., and A. Rose, "An Estimate of the Effect of Common Currencies on Trade and Income," *Quarterly Journal of Economics* 117 (2002), 409–466.

Frankel, J., E. Stein, and S. Wei, "Continental Trading Blocs: Are They Natural, or Super-Natural?" in J. Frankel (Ed.), *The Regionalization of the World Economy* (Chicago: University of Chicago Press, 1998).

Frankel, J., and S. Wei, "Trade Blocs and Currency Blocs," NBER working paper no. 4335 (1993).

Goldberger, A. "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica* 36 (1968), 464–472.

Goldberger, A., *A Course in Econometrics* (Cambridge, MA: Harvard University Press, 1991).

Gourieroux, C., A. Monfort, and A. Trognon, "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica* 52 (1984), 701–720.

Gray, A., http://argray.fateback.com/dist/formula.html (2001).

Hallak, J. C., "Product Quality and the Direction of Trade," *Journal of International Economics* 68 (2006), 238–265.

Harrigan, J., "OECD Imports and Trade Barriers in 1983," *Journal of International Economics* 35 (1993), 95–111.

Haveman, J., and D. Hummels, "Alternative Hypotheses and the Volume of Trade: The Gravity Equation and the Extent of Specialization," Purdue University mimeograph, (2001).

Helpman, E., and P. Krugman, *Market Structure and Foreign Trade* (Cambridge, MA: MIT Press, 1985).

Helpman, E., M. Melitz, and Y. Rubinstein, "Trading Patterns and Trading Volumes," Harvard University mimeograph (2004).

Koenker, R., and G. S. Bassett, Jr., "Regression Quantiles," *Econometrica* 46 (1978), 33–50.

Manning, W. G., and J. Mullahy, "Estimating Log Models: To Transform or Not to Transform?" *Journal of Health Economics* 20 (2001), 461–494.

McCallum, J., "National Borders Matter: Canada-US Regional Trade Patterns," *American Economic Review* 85 (1995), 615–623.

McCullagh, P., and J. A. Nelder, *Generalized Linear Models,* 2nd ed. (London: Chapman and Hall, 1989).

Papke, L. E., and J. M. Wooldridge, "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics* 11 (1996), 619–632.

Park, R., "Estimation with Heteroskedastic Error Terms," *Econometrica* 34 (1966), 888.

Ramsey, J. B., "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society B* 31 (1969), 350–371.

Robinson, P. M., "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica* 55 (1987), 875–891.

Rose, A., "One Money One Market: Estimating the Effect of Common Currencies on Trade," *Economic Policy* 15 (2000), 7–46.

StataCorp., *Stata Statistical Software: Release 8* (College Station, TX: StataCorp LP, 2003).

Tenreyro, S., and R. Barro, "Economic Effects of Currency Unions," FRB Boston series working paper no. 02–4 (2002).

Tinbergen, J., *The World Economy. Suggestions for an International Economic Policy* (New York: Twentieth Century Fund, 1962).

Wei, S., "Intra-national versus International Trade: How Stubborn Are Nation States in Globalization?" NBER working paper no. 5331 (1996).

White, H., "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48 (1980), 817–838.

Winkelmann, R., *Econometric Analysis of Count Data,* 4th ed. (Berlin: Springer-Verlag, 2003).

Windmeijer, F., and J. M. C. Santos Silva, "Endogeneity in Count Data Models: An Application to Demand for Health Care," *Journal of Applied Econometrics* 12 (1997), 281–294.

Wooldridge, J. M., "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics* 90 (1999), 77–97.

——— *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002).

World Bank, *World Development Indicators CD-ROM* (The World Bank, 2002).

APPENDIX

TABLE A1.—LIST OF COUNTRIES

Albania	Denmark	Kenya	Romania
Algeria	Djibouti	Kiribati	Russian Federation
Angola	Dominican Rep.	Korea, Rep.	Rwanda
Argentina	Ecuador	Laos P. Dem. Rep.	Saudi Arabia
Australia	Egypt	Lebanon	Senegal
Austria	El Salvador	Madagascar	Seychelles
Bahamas	Eq. Guinea	Malawi	Sierra Leone
Bahrain	Ethiopia	Malaysia	Singapore
Bangladesh	Fiji	Maldives	Solomon Islands
Barbados	Finland	Mali	South Africa
Belgium-Lux.	France	Malta	Spain
Belize	Gabon	Mauritania	Sri Lanka
Benin	Gambia	Mauritius	St. Kitts and Nevis
Bhutan	Germany	Mexico	Sudan
Bolivia	Ghana	Mongolia	Suriname
Brazil	Greece	Morocco	Sweden
Brunei	Guatemala	Mozambique	Switzerland
Bulgaria	Guinea	Nepal	Syrian Arab Rep.
Burkina Faso	Guinea-Bissau	Netherlands	Tanzania
Burundi	Guyana	New Caledonia	Thailand
Cambodia	Haiti	New Zealand	Togo
Cameroon	Honduras	Nicaragua	Trinidad and Tobago
Canada	Hong Kong	Niger	Tunisia
Central African Rep.	Hungary	Nigeria	Turkey
Chad	Iceland	Norway	Uganda
Chile	India	Oman	United Arab Em.
China	Indonesia	Pakistan	United Kingdom
Colombia	Iran	Panama	United States
Comoros	Ireland	Papua New Guinea	Uruguay
Congo Dem. Rep.	Israel	Paraguay	Venezuela
Congo Rep.	Italy	Peru	Vietnam
Costa Rica	Jamaica	Philippines	Yemen
Côte D'Ivoire	Japan	Poland	Zambia
Cyprus	Jordan	Portugal	Zimbabwe

TABLE A2.—COMMON OFFICIAL AND SECOND LANGUAGES

English	French	Spanish	Dutch
Australia	Belgium-Lux.	Argentina	Belgium-Lux.
Bahamas	Benin	Belize	Netherlands
Barbados	Burkina Faso	Bolivia	Suriname
Belize	Burundi	Chile	
Brunei	Cameroon	Colombia	**German**
Cameroon	Canada	Costa Rica	
Canada	Central African Rep.	Dominican Rep.	Austria
Fiji	Chad	Ecuador	Germany
Gambia	Comoros	El Salvador	Switzerland
Ghana	Congo Dem. Rep.	Eq. Guinea	
Guyana	Congo Rep.	Guatemala	**Greek**
Hong Kong	Côte D'Ivoire	Honduras	
India	Djibouti	Mexico	Cyprus
Indonesia	Eq. Guinea	Nicaragua	Greece
Ireland	France	Panama	
Israel	Gabon	Paraguay	**Hungarian**
Jamaica	Guinea	Peru	
Jordan	Haiti	Spain	Hungary
Kenya	Lebanon	Uruguay	Romania
Kiribati	Madagascar	Venezuela	
Malawi	Mali		**Italian**
Malaysia	Mauritania	**Arabic**	
Maldives	Mauritius		Italy
Malta	Morocco	Algeria	Switzerland
Mauritius	New Caledonia	Bahrain	
New Zealand	Niger	Chad	**Lingala**
Nigeria	Rwanda	Comoros	
Oman	Senegal	Djibouti	Congo Dem. Rep.
Pakistan	Seychelles	Egypt	Congo Rep.
Panama	Switzerland	Israel	
Papua New Guinea	Togo	Jordan	**Russian**
Philippines	Tunisia	Lebanon	
Rwanda		Mauritania	Mongolia
Seychelles	**Malay**	Morocco	Russian Federation
Sierra Leone		Oman	
Singapore	Brunei	Saudi Arabia	**Swahili**
South Africa	Indonesia	Sudan	
Sri Lanka	Malaysia	Syria	Kenya
St. Helena	Singapore	Tanzania	Tanzania
St. Kitts and Nevis		Tunisia	
Suriname	**Portuguese**	United Arab Em.	**Chinese**
Tanzania		Yemen	
Trinidad and Tobago	Angola		China
Uganda	Brazil	**Turkish**	Hong Kong
United Kingdom	Guinea-Bissau		Malaysia
United States	Mozambique	Cyprus	Singapore
Zambia	Portugal	Turkey	
Zimbabwe			

TABLE A3.—COLONIAL TIES

United Kingdom		France	Spain
Australia	Mauritius	Algeria	Argentina
Bahamas	New Zealand	Benin	Bolivia
Bahrain	Nigeria	Burkina Faso	Chile
Barbados	Pakistan	Cambodia	Colombia
Belize	Saint Kitts and Nevis	Cameroon	Costa Rica
Cameroon	Seychelles	Central African Rep.	Cuba
Canada	Sierra Leone	Chad	Ecuador
Cyprus	South Africa	Comoros	El Salvador
Egypt	Sri Lanka	Congo	Eq. Guinea
Fiji	Sudan	Djibouti	Guatemala
Gambia	Tanzania	Gabon	Honduras
Ghana	Trinidad and Tobago	Guinea	Mexico
Guyana	Uganda	Haiti	Netherlands
India	United States	Laos	Nicaragua
Ireland	Zambia	Lebanon	Panama
Israel	Zimbabwe	Madagascar	Paraguay
Jamaica		Mali	Peru
Jordan		Mauritania	Venezuela
Kenya		Morocco	**Portugal**
Kuwait		Niger	
Malawi		Senegal	Angola
Malaysia		Syria	Brazil
Maldives		Togo	Guinea-Bissau
Malta		Tunisia	Mozambique
		Vietnam	Oman

TABLE A4.—PREFERENTIAL TRADE AGREEMENTS IN 1990

EEC/EC	CARICOM	CACM
Belgium	Bahamas	Costa Rica
Denmark	Barbados	El Salvador
France	Belize	Guatemala
Germany	Dominican Rep.	Honduras
Greece	Guyana	Nicaragua
Ireland	Haiti	
Italy	Jamaica	**Bilateral Agreements**
Luxembourg	Trinidad and Tobago	
Netherlands	St. Kitts and Nevis	EC-Cyprus
Portugal	Suriname	EC-Malta
Spain		EC-Egypt
United Kingdom	**SPARTECA**	EC-Syria
		EC-Algeria
EFTA	Australia	EC-Norway
	New Zealand	EC-Iceland
Iceland	Fiji	EC-Switzerland
Norway	Kiribati	Canada–United States
Switzerland	Papua New Guinea	Israel–United States
Liechtenstein	Solomon Islands	
CER	**PATCRA**	
Australia	Australia	
New Zealand	Papua New Guinea	

TABLE A5.—SUMMARY STATISTICS

Variable	Full Sample		Exports > 0	
	Mean	Std. Dev.	Mean	Std. Dev.
Trade	172132.2	1828720	328757.7	2517139
Log of trade	—	—	8.43383	3.26819
Log of exporter's GDP	23.24975	2.39727	24.42503	2.29748
Log of importer's GDP	23.24975	2.39727	24.13243	2.43148
Log of exporter's per capita GDP	7.50538	1.63986	8.09600	1.65986
Log of importer's per capita GDP	7.50538	1.63986	7.98602	1.68649
Log of distance	8.78551	0.74168	8.69497	0.77283
Contiguity dummy	0.01961	0.13865	0.02361	0.15185
Common-language dummy	0.20970	0.40710	0.21284	0.40933
Colonial-tie dummy	0.17048	0.37606	0.16894	0.37472
Landlocked exporter dummy	0.15441	0.36135	0.10767	0.30998
Landlocked importer dummy	0.15441	0.36135	0.11401	0.31784
Exporter's remoteness	8.94654	0.26389	8.90383	0.29313
Importer's remoteness	8.94654	0.26389	8.90787	0.28412
Preferential–trade-agreement dummy	0.02505	0.15629	0.04452	0.20626
Openness dummy	0.56373	0.49594	0.65796	0.47442