Variable Scaling and Hypothesis Testing in the Gravity Model*

Anton Yang† Russell Hillberry‡

Yale University Purdue University

September 1, 2023

Abstract

Statistical inference around hypothesis tests in Poisson Pseudo Maximum Likelihood (PPML) models is sensitive to data scaling choices. We show this analytically and demonstrate it using a simple application of the gravity model of trade. The scale of the data on the independent variable affects the scale of both the Likelihood statistic and the Likelihood Ratio test statistic. Lagrange Multiplier tests are similarly sensitive to data scaling choices. When considering Wald tests, we find some nuance. Data scaling affects the Wald statistic when it depends upon the asymptotic variance-covariance matrix, but not when the variance-covariance matrix depends upon residuals from robust estimation. Testing of joint hypotheses in PPML gravity models should therefore rely on Wald tests constructed from robust standard errors.

Keywords: Gravity model; Poisson Pseudo Maximum Likelihood; Variable scaling; Likelihood Ratio test; Wald test

JEL classification: C12; C13; C81; F14

*Declarations of interest: None.

†A307, 28 Hillhouse Avenue, New Haven, CT 06511. Email: anton.yang@yale.edu.

‡*Corresponding author*: 692 Krannert Building, 403 W. State Street, West Lafayette, IN 47907-2053. Email: rhillber@purdue.edu.

1 Introduction

The enormous literature on the gravity model of trade has adopted the Poisson Pseudo Maximum Likelihood (PPML) function as its preferred econometric objective.[1] The primary focus of this literature has been estimation of model parameters, parameters that map cleanly onto prominent theories of bilateral trade that feature constant elasticity of substitution (CES) demands for national factor bundles.[2] The canonical CES theories assume homotheticity, *ad valorem* trade costs, constant trade elasticities, and, typically, that all trade is in final goods. Given the recent emphasis on moving from gravity model estimation to welfare analysis, it is important that the canonical theories be subject to further testing against richer theories of bilateral trade.[3]

A nascent literature has shown that the canonical CES theories are overly restrictive. One approach to testing theoretical restrictions uses the PPML objective function in a two-step estimation procedure: first estimate a PPML gravity specification with fixed effects, and then use the fixed effects to construct variables for subsequent hypothesis testing via t-tests in a second PPML specification.[4] While this approach is persuasive, it requires adjustments for simultaneity, and only allows testing for the significance of variables individually rather than collectively. It would be preferable to be able to estimate simultaneously, and to jointly test restrictions on sets of variables relevant to a theoretical restriction.

A promising alternative approach to testing theoretical restrictions employs a Mathe-

[1] Santos Silva and Tenreyro (2006a) introduce the PPML framework to the gravity-model-of-trade literature. In August of 2023, the article had been cited nearly 8,000 times in Google Scholar.

[2] The most prominent of these theories are Anderson (1979), Eaton and Kortum (2002), and Melitz (2003) (as interpreted by Chaney (2008)). Arkolakis, Costinot and Rodríguez-Clare (2012) show that all of these theories map to a CES factor demand model. Fally (2015) and Anderson and Yotov (2012) demonstrate the ease of mapping between the PPML estimation model (with origin and destination fixed effects) and the standard CES gravity theory.

[3] A related reason for additional empirical scrutiny is that the canonical theories were devised *ex post* to fit a robust empirical relationship that was known *ex ante*. The empirical success of the model is not therefore validation of the theory, though it is sometimes interpreted that way.

[4] Caron, Fally and Markusen (2014) transform first stage estimates from a PPML specification to estimate income elasticities of demand at the sector level. Most income elasticities are different than one. Chen and Novy (2022) use a PPML framework to first estimate fixed effects, and then construct interactions of trade frictions with fitted trade shares to demonstrate the presence of heterogeneity in trade cost elasticities, as their translog demand function predicts.

matical Program with Equilibrium Constraints (MPEC) to conduct structural estimation of gravity models of trade.[5] In this context an MPEC maximizes an econometric objective function (e.g. a likelihood function) subject to a set of constraints that fully define the proposed theory of bilateral trade. Because the MPEC separates the econometric objective from the particular economic theory under investigation, it allows (in principle) direct testing of parameter restrictions in richer theories that nest the CES gravity model. One question that has arisen in this context is whether the properties of the PPML objective function allow it to be used for evaluating model restrictions jointly.[6]

In this paper we show that the results of hypothesis tests in the PPML specification are, in many cases, sensitive to data scaling choices. This is unfortunate because empirical estimation of the gravity model of trade involves choices of data scale.[7] We show that data scaling - in particular scaling of the data on the independent variable (e.g. bilateral trade) - affects hypothesis testing of PPML models in non-trivial ways. The values of both the likelihood function and the associated LR test statistic depend on data scale. Inferences drawn from Lagrange Multiplier tests and the standard Wald test are also sensitive to data scaling.

We find only one approach that is suitable for tests of joint hypotheses. We show that Wald test statistics calculated from an information matrix defined by robust (Huber-White) standard errors are independent of the scale of the y-variable employed in PPML estimation. The test is also robust to scaling of the x-variables. We conclude that Wald tests that rely

[5] Su and Judd (2012) describe the benefits of MPEC for structural estimation. Balistreri and Hillberry (2007) first use an MPEC to estimate a structural gravity model. Balistreri, Hillberry and Rutherford (2011) estimate the structural parameters of the Melitz (2003) model for the manufacturing sector. Tan (2013) uses a similar approach to estimate the parameters of gravity model with a flexible translog demand system, and conducts counterfactual analysis with the richer parameter set. These papers focus on parameter estimation rather than testing of restrictions.

[6] Yang (2021) develops and estimates a generalized Armington model of trade that allows for flexibility in trade and income elasticities, and shows that his general model nests several more restrictive models, including the canonical CES model as well as richer theories of gravity. The econometric objective in his MPEC is PPML.

[7] For example, in Santos Silva and Tenreyro (2006a) the dependent variable (the value of bilateral trade) is expressed in units of one thousand U.S. dollars. Another prominent paper, Anderson and van Wincoop (2003), scales both trade flows and Gross Domestic Products so they are expressed in millions of dollars. The log distance variable in the gravity model also involves a choice of scale (miles, kilometers, etc.).

on test statistics calculated with robust standard errors are thus appropriate for the setting we imagine.

In order to illustrate our key insights, we conduct an application in which we test the hypothesis that colonial links between two trading partners affect trade flows in a PPML specification that also includes a variable indicating that the two countries share a common language. Consistent with our mathematical derivations, both Likelihood statistics and Likelihood Ratio tests of excluding the colonial links variable are sensitive to data scaling choices. As our theory shows, a Wald test that depends on asymptotic standard errors is sensitive to scaling, while a Wald test that relies on robust standard errors is not. Our application tests a restriction on single variable, in order to allow transparent comparisons with t-tests. We show analytically that our results extend to tests of joint restrictions of multiple parameters, and to structural models estimated with an MPEC.

The remainder of the paper is organized as follows. Section 2 briefly reviews the PPML estimator in the context of the gravity model of trade. In Section 3 we demonstrate the main issue by showing - analytically and empirically - that scaling affects statistical inference in LR tests. In Section 4 we show that the effects of scaling on Wald tests depend on the form of the error term used to construct the information matrix, and validate these results with our empirical example. Section 5 shows that Lagrange Multiplier tests are also sensitive to data scaling. Section 6 considers other topics, including the scaling of the model's independent variables and the extension of the main results to an MPEC setting. Section 7 concludes.

2 The PPML Estimator

Consider the estimation specification suggested by Santos Silva and Tenreyro (2006a):

$$y_i = \exp(x_i'\beta) + \epsilon_i, \tag{1}$$

where y_i, $i = 1, .., n$, can be observed from data, x_i is a vector of exogenous variables, β a vector of associated parameters, and ϵ_i an error term with $E[\epsilon_i|x_i] = 0$. Following Gourieroux, Monfort and Trognon (1984*a*,*b*), choose β to maximize the log-likelihood function:

$$L(\beta) = K - \sum_i^n \exp(x_i'\beta) + \sum_i^n y_i(x_i'\beta), \tag{2}$$

where K is a constant term.[8]

Our analytic exercises will use this general framework, but to make our example concrete consider a simple PPML gravity model of trade:

$$y_{ij} = \exp[\beta_0 + \beta_1 \log(x_i) + \beta_2 \log(x_j) + \beta_3 z_{ij}] + \epsilon_{ij}, \tag{3}$$

where y_{ij} are observed bilateral trade flows between locations i and j, the gross domestic product of exporting and importing locations are denoted by x_i and x_j, respectively, and z_{ij} is a variable that applies bilaterally to locations i and j. The log transformation on the right-hand side of (3) is standard in the gravity literature.

[8] In the Poisson likelihood function used to estimate count models, K takes an explicit form $K = \sum_i^n (y_i!)$. In the PPML model, the y-variable does not appear in the likelihood function, a central reason for the issues we confront here.

3 Scale and Testing using Likelihood Ratios

Consider two models: a 'small' parsimonious model, and a 'large' model with more parameters. We wish to consider a hypothesis test of the form:

Hypothesis $\mathbf{H_0}$: *the 'small' model is more consistent with the data.*

Hypothesis $\mathbf{H_a}$: *the 'large' model is more consistent with the data.*

The test statistic for an LR test of $\mathbf{H_0}$ appears as:

$$\Lambda = -2[L^{H_0}(\beta) - L^{H_a}(\beta')], \tag{4}$$

where β and β' are the parameter vectors associated with the 'small' and 'large' models, respectively. Asymptotically, Λ is distributed χ^2 with k degrees of freedom, where k is the reduction in the number of parameters when moving from the 'large' to the 'small' model. Given a critical value c for this distribution, the null hypothesis will be rejected if $\Lambda > c$.[9]

3.1 Implications of Scaling for Likelihood

Let the log-likelihood function for the 'small' model take the form:

$$L^{H_0}(\beta) = -\sum_i^n \exp[\beta_0 + \beta_1 \log(x_i)] + \sum_i^n y_i[\beta_0 + \beta_1 \log(x_i)], \tag{5}$$

where β_0 is a constant term, β_1 is the coefficient on $\log x_i$.[10]

In order to show that the scaling of y_i affects Λ, we introduce an arbitrary scalar $S \neq 1$ that we apply to y_i. The associated log-likelihood function is:

$$\widetilde{L^{H_0}}(\tilde{\beta}) = -\sum_i^n \exp[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i)] + \sum_i^n S y_i[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i)] \tag{6}$$

where $\widetilde{L^{H_0}}(\tilde{\beta})$ is the log-likelihood involved with the scaling factor S, and $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are

[9] See Gouriéroux, Holly and Monfort (1982).

[10] We follow the standard PPML-gravity literature and suppress the parameter K. Equation (4) differences K from Λ anyway.

the model's parameters given the scaling. Taking the first-order conditions with respect to β terms in Equations (5) and (6) and solving the associated system of equations reveals relationships between the parameters before and after the scaling: $\beta_0 = \tilde{\beta}_0 - \log(S)$ and $\beta_1 = \tilde{\beta}_1$. Intuitively, scaling affects the size of the constant term, but other coefficients in the model are unaffected. However, the log-likelihoods before and after the scaling of y_i are not equivalent:

$$\begin{aligned}
\widetilde{L^{H_0}}(\tilde{\beta}) =& -\sum_i^n \exp[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i)] + S\sum_i^n y_i[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i)] \\
=& -S\sum_i^n \exp[\beta_0 + \beta_1 \log(x_i)] + S\sum_i^n y_i[\beta_0 + \log S + \beta_1 \log(x_i)] \\
\neq& -\sum_i^n \exp[\beta_0 + \beta_1 \log(x_i)] + \sum_i^n y_i[\beta_0 + \beta_1 \log(x_i)] \\
=& L^{H_0}(\beta).
\end{aligned} \tag{7}$$

The relationship only holds with equality when $S = 1$.

3.2 Implications of Scaling for Likelihood Ratio Tests

Now shift to the large model by introducing another variable, z_i. The associated log-likelihood function is:

$$L^{H_a}(\beta') = -\sum_i^n \exp[\beta'_0 + \beta'_1 \log(x_i) + \beta'_2 \log z_i] + \sum_i^n y_i[\beta'_0 + \beta'_1 \log(x_i) + \beta'_2 \log z_i], \tag{8}$$

with $'$ indicating parameters of the larger model. The log-likelihood of the large model with scaled data appears as:

$$\widetilde{L^{H_a}}(\tilde{\beta}') = -\sum_i^n \exp[\tilde{\beta}'_0 + \tilde{\beta}'_1 \log(x_i) + \tilde{\beta}'_2 \log(z_i)] + S\sum_i^n y_i[\tilde{\beta}'_0 + \tilde{\beta}'_1 \log(x_i) + \tilde{\beta}'_2 \log(z_i)]. \tag{9}$$

It can be shown that the parameters in the scaled and unscaled large models have the relationships: $\beta_0' = \tilde{\beta}_0' - \log(S)$, $\beta_1' = \tilde{\beta}_1'$ and $\beta_2' = \tilde{\beta}_2'$.

The difference between the log-likelihoods using the scaled data is given by:

$$\begin{aligned}\widetilde{L^{H_0}}(\tilde{\beta}) - \widetilde{L^{H_a}}(\tilde{\beta}') = &- \sum_i^n \exp[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i)] + S \sum_i^n y_i[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i)] \\ &+ \sum_i^n \exp[\tilde{\beta}_0' + \tilde{\beta}_1' \log(x_i) + \tilde{\beta}_2' \log(z_i)] - S \sum_i^n y_i[\tilde{\beta}_0' + \tilde{\beta}_1' \log(x_i) + \tilde{\beta}_2' \log(z_i)].\end{aligned} \tag{10}$$

Dividing both sides by S in equation (10) and substituting the derived parametric relationships into the test statistics produces:

$$\begin{aligned}\frac{\widetilde{\Lambda}(\tilde{\beta})}{S} = \frac{-2\left[\widetilde{L^{H_0}}(\tilde{\beta}) - \widetilde{L^{H_a}}(\tilde{\beta}')\right]}{S} = &-2\Bigg\{ -\frac{1}{S}\sum_i^n \exp[\beta_0 + \log(S) + \beta_1 \log(x_i)] \\ &+ \sum_i^n y_i[\beta_0 + \log(S) + \beta_1 \log(x_i)] \\ &+ \frac{1}{S}\sum_i^n \exp[\beta_0' + \log(S) + \beta_1' \log(x_i) + \beta_2' \log(z_i)] \\ &- \sum_i^n y_i[\beta_0' + \log S + \beta_1' \log(x_i) + \beta_2' \log(z_i)]\Bigg\}.\end{aligned} \tag{11}$$

The right-hand side of equation (11) is equal to the test statistic with unscaled data: $\Lambda(\beta) = -2\Big[L^{H_0}(\beta) - L^{H_a}(\beta')\Big]$. The lesson is that $\Lambda(\beta)$ varies inversely with the scale applied to the y variable. The critical value of the χ^2 distribution is unchanged, but the test statistic changes with S, polluting statistical inference.

3.3 Empirical Example

To illustrate the relevance of this issue, we offer a simple example using the data and empirical specification proposed by Santos Silva and Tenreyro (2006a). Those authors estimate a gravity model with 14 independent variables and an unreported constant term. Their specification contains two related variables: an indicator that the origin and destination countries share a colonial tie, and another indicator that the two countries share a common language. Since common historical forces often drove these two outcomes, it can be difficult for an applied researcher to know whether or not both variables should be included in a gravity regression. In Santos Silva and Tenreyro (2006a) the coefficient on the common language dummy is statistically significant, while the coefficient on the colonial tie variable is not. Our empirical example is a test of the hypothesis that a model without the colonial tie variable is equivalent to a model that includes it, thereby justifying estimation of the smaller model (without the colonial tie dummy).

Table 1: PPML estimates with different scalings of trade flows

	S=1000		S=1*		S=0.001		S=1e-6	
Model	w/ Colony	w/o Colony	w/ Colony	w/o Colony	w/ Colony	w/o Colony	w/ Colony	w/o Colony
Shipment in	USD	USD	1,000 USD	1,000 USD	mil. USD	mil. USD	bil.USD	bil.USD
Comlang	0.75**	0.76**	0.75**	0.76**	0.75**	0.76**	0.75**	0.76**
	(0.13)	(0.08)	(0.13)	(0.08)	(0.13)	(0.08)	(0.13)	(0.08)
Colony	0.03	-	0.03	-	0.03	-	0.03	-
	(0.15)	-	(0.15)	-	(0.15)	-	(0.15)	-
$L^a(\beta')$	-8.70197e11		-870246443		-888289.69		-2630.97	
$L^0(\beta)$	-8.7025e11		-870297478.5		-888340.72		-2631.02	
Λ	1.02e8**		102071**		102.07**		0.10207	
P > χ^2	0.00		0.00		0.00		0.75	

Results from replications of the full specification in Santos Silva and Tenreryo (2006a) with different scalings of the trade variable; *indicates the scale used Santos Silva and Tenreyro; **indicates statistical significance at the 5% level; standard errors in parentheses.

We download the data from Santos Silva and Tenreyro (2006b). We consider different scalings of the bilateral trade data, premultiplying it with different values of a scalar S. We estimate the model in Stata using the *ppmlhdfe* command.[11] Since our focus is on

[11] Correia, Guimarães and Zylkin (2020) develop this command to estimate the PPML model in the presence of high dimensional fixed effects. While our model does not contain high dimensional fixed effects, the *ppmlhdfe* command is still suitable. A key advantage of the package for our purposes is that it calculates and reports a likelihood statistic.

the two dummy variables and the likelihood statistics, we report only statistics related to these outcomes in Table 1. For each scaling of the data, we report results for PPML models estimated with and without the colonial tie variable. We arrange the results in increasing size of S, which corresponds to different choices of units for the value of bilateral trade. Columns 1-2 use one dollar units. Columns 3-4 measure trade in thousands of dollars. Columns 5-6 use million dollar units; columns 7-8 use one billion dollar units.

Column 3 is a replication of Santos Silva and Tenreyro (2006a), with almost exactly identical results. The coefficient on the common colony variable (0.03) is economically small and statistically insignificant, suggesting that perhaps it can be excluded from the model. Column 4 reports the common language coefficient in the model that excludes the colonial tie variable. Log-likelihoods for the two models are reported below the coefficient estimates, as is the test statistic Λ. The value of Λ for Santos Silva and Tenreyro's scaling of the data is 102,071, providing a clear rejection of the model without the colonial tie. This is surprising, as the coefficient on the *Colony* variable is not statistically different from zero in Column 3.

Columns 1-2 report estimates of the two models when the data are scaled in single dollar units. The choice of smaller units scales Λ upward by a factor of 1,000, incorrectly increasing the level of confidence in rejecting the null hypothesis. In columns 5-6, trade flows are scaled in millions of dollars; Λ is therefore scaled downward by 1,000 relative to the base case. The null hypothesis continues to be soundly rejected, but by a smaller margin than with the initial scaling. In columns 7-8, we scale the data into units of \$1 billion. In this case the computed value of Λ fails to reject the hypothesis that the two models are equivalent.

Note that the parameter estimates in all specifications are indifferent to the scale of the data. As in the mathematics above, scaling the data affects the constant term, but is otherwise unimportant for parameter estimation. Only the values of $L(\beta)$, Λ and the associated statistical inference are affected by scaling.

4 Wald Test

We next turn our attention to the Wald test, where we observe that the effects of scaling depend upon the errors that are used to calculate the Fisher information matrix. We show analytically that test statistics calculated with model-based asymptotic standard errors depend on data scaling, while test statistics calculated with robust standard errors do not.[12] In order to simplify the notation, in both cases our analytical exercise uses the 'small' model and a restriction on β_1. Results for restrictions on a specification with more parameters can be logically extended, as can be seen in subsection 4.3.[13]

4.1 Non-Robust Asymptotic Standard Errors

To calculate the Wald test statistic with asymptotic standard errors we refer the reader to the unscaled 'small' model in equation (5). The associated Hessian matrix $\boldsymbol{H}(\beta)$ is given by:

$$\boldsymbol{H}(\beta) = \frac{\partial^2 L^{H_0}(\beta)}{\partial\beta \cdot \partial\beta^T} = -\sum_{i=1}^{n} \begin{bmatrix} \exp(\beta_0 + \beta_1 \log(x_i)) & \log(x_i)\exp(\beta_0 + \beta_1 \log(x_i)) \\ \log(x_i)\exp(\beta_0 + \beta_1 \log(x_i)) & (\log(x_i))^2 \exp(\beta_0 + \beta_1 \log(x_i)) \end{bmatrix}. \tag{12}$$

The Fisher information matrix $\boldsymbol{I^F}(\beta)$ is the negative of $\boldsymbol{H}(\beta)$. The asymptotic variance-covariance matrix is defined as the inverse of $\boldsymbol{I^F}(\beta)$. Wald test statistics depend on the variance-covariance matrix. Our central result in this section is that scaling the y_i data produces an asymptotic variance-covariance matrix that is a proportional scaling of the asymptotic variance-covariance matrix associated with the unscaled model. We demonstrate this relationship in Theorem 1.

[12] The "model-based" qualifier on the asymptotic standard errors is formal language acknowledging that the errors depend upon the model specification as well as upon the estimated parameters. In the interest of brevity, we omit the "model-based" adjective hereafter.

[13] In Appendix A.2, we offer a proof for the 'large' model and further extend this case to an n x n matrix.

Theorem 1 *Given a twice-differentiable Poisson Pseudo-Maximum Likelihood function in equation (5) with dependent variable y_i for $i = 1, \ldots, n$ and independent variable $\boldsymbol{x}_i \in \mathbb{R}^n$ in the model matrix $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$, and residuals $\boldsymbol{\epsilon} = (y_1 - \exp(x_1'\beta), \ldots, y_n - \exp(x_n'\beta))^\top \in \mathbb{R}^n$. Assume that the Hessian matrix of the function with respect to model parameters $\boldsymbol{\mathcal{B}} = \{\beta_0; \beta_1, ..., \beta_n\}^T$ in the space $\Gamma \subseteq \mathbb{R}^n$ is invertible (nonsingular). If the dependent variable y_i for $i = 1, \ldots, n$ is scaled by a factor $S \neq 1$, then the relationship between the asymptotic variance-covariance matrix $\boldsymbol{V}^{asym}$ of the unscaled model and the asymptotic variance-covariance matrix $\widetilde{\boldsymbol{V}}^{asym}$ of the scaled model must be $\boldsymbol{V}^{asym} = S\widetilde{\boldsymbol{V}}^{asym}$.*

Proof. The asymptotic variance-covariance matrix $\boldsymbol{V}^{\text{asym}}$ stated in Theorem 1 is

$$\boldsymbol{V}^{\text{asym}} = \left[\boldsymbol{I}^F(\beta)\right]^{-1} = \left[-\boldsymbol{H}(\beta)\right]^{-1} = \left\{-\left[\frac{\partial^2 L(\beta)}{\partial\beta \cdot \partial\beta^T}\right]\right\}^{-1}. \tag{13}$$

$\boldsymbol{V}^{\text{asym}}$ denotes the asymptotic variance-covariance matrix for the unscaled model, while $\widetilde{\boldsymbol{V}}^{\text{asym}}$ represents the asymptotic variance-covariance matrix for the scaled model in equation (6).

It is straightforward to show that:

$$\begin{aligned}
\boldsymbol{V}^{\text{asym}} &= \frac{1}{\det(\boldsymbol{I}^F)} \begin{bmatrix} \frac{\partial^2 L}{\partial\beta_1\beta_1^T} & -\frac{\partial^2 L}{\partial\beta_0\partial\beta_1} \\ -\frac{\partial^2 L}{\partial\beta_1\partial\beta_0} & \frac{\partial^2 L}{\partial\beta_0\beta_0^T} \end{bmatrix} \\
&= \frac{1}{\det(\boldsymbol{I}^F)} \begin{bmatrix} \sum_{i=1}^n (\log(x_i))^2 \exp(\beta_0 + \beta_1 \log(x_i)) & -\sum_{i=1}^n \log(x_i) \exp(\beta_0 + \beta_1 \log(x_i)) \\ -\sum_{i=1}^n \log(x_i) \exp(\beta_0 + \beta_1 \log(x_i)) & \sum_{i=1}^n \exp(\beta_0 + \beta_1 \log(x_i)) \end{bmatrix} \\
&= \frac{\begin{bmatrix} \sum_{i=1}^n (\log(x_i))^2 \exp(\beta_0 + \beta_1 \log(x_i)) & -\sum_{i=1}^n \log(x_i) \exp(\beta_0 + \beta_1 \log(x_i)) \\ -\sum_{i=1}^n \log(x_i) \exp(\beta_0 + \beta_1 \log(x_i)) & \sum_{i=1}^n \exp(\beta_0 + \beta_1 \log(x_i)) \end{bmatrix}}{\frac{\partial^2 L}{\partial\beta_0\beta_0^T}\frac{\partial^2 L}{\partial\beta_1\beta_1^T} - \left(\frac{\partial^2 L}{\partial\beta_0\partial\beta_1}\right)^2}.
\end{aligned} \tag{14}$$

Following results in Section 3.1, we have $\beta_0 = \tilde{\beta}_0 - \log(S)$ and $\beta_1 = \tilde{\beta}_1$. This allows us

to show that $\det(\widetilde{\boldsymbol{I}^F}) = S^2\det(\boldsymbol{I}^F)$:

$$
\begin{aligned}
\det(\widetilde{\boldsymbol{I}^F}) &= \frac{\partial^2\widetilde{L}}{\partial\widetilde{\beta}_0^2}\frac{\partial^2\widetilde{L}}{\partial\widetilde{\beta}_1^2} - \left(\frac{\partial^2\widetilde{L}}{\partial\widetilde{\beta}_0\partial\widetilde{\beta}_1}\right)^2 \\
&= \sum_{i=1}^n \exp(\widetilde{\beta}_0 + \widetilde{\beta}_1\log(x_i)) \cdot \sum_{i=1}^n (\log(x_i))^2 \exp(\widetilde{\beta}_0 + \widetilde{\beta}_1\log(x_i)) \\
&- \sum_{i=1}^n \log(x_i)\exp(\widetilde{\beta}_0 + \widetilde{\beta}_1\log(x_i)) \cdot \sum_{i=1}^n (\log(x_i))^2 \exp(\widetilde{\beta}_0 + \widetilde{\beta}_1\log(x_i)) \\
&= \sum_{i=1}^n \exp(\beta_0 + \log(S) + \beta_1\log(x_i)) \cdot \sum_{i=1}^n (\log(x_i))^2 \exp(\beta_0 + \log(S) + \beta_1\log(x_i)) \\
&- \sum_{i=1}^n \log(x_i)\exp(\beta_0 + \log(S) + \beta_1\log(x_i)) \cdot \sum_{i=1}^n (\log(x_i))^2 \exp(\beta_0 + \log(S) + \beta_1\log(x_i)) \\
&= S^2\left[\frac{\partial^2 L}{\partial\beta_0\beta_0^T}\frac{\partial^2 L}{\partial\beta_1\beta_1^T} - \left(\frac{\partial^2 L}{\partial\beta_0\partial\beta_1}\right)^2\right]. \\
&= S^2\det(\boldsymbol{I}^F)
\end{aligned}
\tag{15}
$$

We apply this result to the definition of $\widetilde{\boldsymbol{V}}^{\text{asym}}$:

$$
\begin{aligned}
\widetilde{\boldsymbol{V}}^{\text{asym}} &= \frac{1}{\det(\widetilde{\boldsymbol{I}^F})}\begin{bmatrix} \sum_{i=1}^n (\log(x_i))^2\exp(\beta_0 + \log(S) + \beta_1\log(x_i)) & -\sum_{i=1}^n \log(x_i)\exp(\beta_0 + \log(S) + \beta_1\log(x_i)) \\ -\sum_{i=1}^n \log(x_i)\exp(\beta_0 + \log(S) + \beta_1\log(x_i)) & \sum_{i=1}^n \exp(\beta_0 + \log(S) + \beta_1\log(x_i)) \end{bmatrix} \\
&= \frac{S\begin{bmatrix} \sum_{i=1}^n (\log(x_i))^2\exp(\beta_0 + \beta_1\log(x_i)) & -\sum_{i=1}^n \log(x_i)\exp(\beta_0 + \beta_1\log(x_i)) \\ -\sum_{i=1}^n \log(x_i)\exp(\beta_0 + \beta_1\log(x_i)) & \sum_{i=1}^n \exp(\beta_0 + \beta_1\log(x_i)) \end{bmatrix}}{S^2\left[\frac{\partial^2 L}{\partial\beta_0\beta_0^T}\frac{\partial^2 L}{\partial\beta_1\beta_1^T} - \left(\frac{\partial^2 L}{\partial\beta_0\partial\beta_1}\right)^2\right]} \\
&= \frac{\begin{bmatrix} \sum_{i=1}^n (\log(x_i))^2\exp(\beta_0 + \beta_1\log(x_i)) & -\sum_{i=1}^n \log(x_i)\exp(\beta_0 + \beta_1\log(x_i)) \\ -\sum_{i=1}^n \log(x_i)\exp(\beta_0 + \beta_1\log(x_i)) & \sum_{i=1}^n \exp(\beta_0 + \beta_1\log(x_i)) \end{bmatrix}}{S\left[\frac{\partial^2 L}{\partial\beta_0\beta_0^T}\frac{\partial^2 L}{\partial\beta_1\beta_1^T} - \left(\frac{\partial^2 L}{\partial\beta_0\partial\beta_1}\right)^2\right]}.
\end{aligned}
\tag{16}
$$

Equations (14) and (16) immediately yield that $\boldsymbol{V}^{\text{asym}} = S\widetilde{\boldsymbol{V}}^{\text{asym}}$. ■

Since the variance is scaled by S, in the case of test of a restriction on β_1, the Wald test statistic calculated from the scaled data ($\widetilde{W}$) is inversely proportional to the Wald test statistic calculated from the unscaled data:

$$\widetilde{W} = \frac{\left(\widetilde{\hat{\beta}}_1 - \beta_1^{\text{restricted}}\right)^2}{\widetilde{V}^{\text{asym}}} = \frac{\left(\hat{\beta}_1 - \beta_1^{\text{restricted}}\right)^2}{SV^{\text{asym}}} = \frac{1}{S}W. \tag{17}$$

This result extends to tests of multiple restrictions on the vector $\boldsymbol{\beta}$ (Greene, 2012):

$$W^{\text{multiple restrictions}} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{restricted}})^T \widehat{\boldsymbol{COV}}^{\text{asym}^{-1}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{restricted}}), \tag{18}$$

where

$$\widehat{\boldsymbol{COV}}^{\text{asym}} = \begin{bmatrix} \frac{\partial^2 L^{H_0}}{\partial \beta_1^2} & \frac{\partial^2 L^{H_0}}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 L^{H_0}}{\partial \beta_1 \partial \beta_n} \\ \frac{\partial^2 L^{H_0}}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 L^{H_0}}{\partial \beta_2^2} & \cdots & \frac{\partial^2 L^{H_0}}{\partial \beta_2 \partial \beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L^{H_0}}{\partial \beta_n \partial \beta_1} & \frac{\partial^2 L^{H_0}}{\partial \beta_n \partial \beta_2} & \cdots & \frac{\partial^2 L^{H_0}}{\partial \beta_n^2} \end{bmatrix}. \tag{19}$$

By Corollary 1.1 (in appendix A.1), since each element of the Hessian matrix $\boldsymbol{H}(\beta)$ is scaled by S when the dependent variable y_i is scaled by S: $\widetilde{\widehat{COV}}^{\text{asym}} = S \cdot \widehat{COV}^{\text{asym}}$, we have:

$$\widetilde{W}^{\text{multiple restrictions}} = \frac{1}{S} W^{\text{multiple restrictions}}. \tag{20}$$

Since the hypothesis tests involving $\widetilde{W}$ and W compare both statistics to the same critical value, the result of the test depends on the scale of S.

4.2 Wald Statistics Calculated from Robust Standard Errors

An alternative approach to calculating the variance-covariance matrix is to use the residuals from an estimation that returns Huber-White robust standard errors.[14] Consider a Huber-White bias-corrected variance estimator $\boldsymbol{V}^{\text{H-W}}$ defined as follows:

$$\boldsymbol{V}^{\text{H-W}} = \left[\boldsymbol{I}^F(\beta)\right]^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma} \boldsymbol{X} \left[\boldsymbol{I}^F(\beta)\right]^{-1} = \left[\boldsymbol{I}^F(\beta)\right]^{-1} \boldsymbol{X}^T \operatorname{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2) \boldsymbol{X} \left[\boldsymbol{I}^F(\beta)\right]^{-1}, \quad (21)$$

where $\boldsymbol{\Sigma} = \boldsymbol{X}^T \operatorname{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon} n^2) \boldsymbol{X}$ is the *correction matrix* based on raw residuals (Maas and Hox, 2004). The raw residuals vector, $\boldsymbol{\epsilon} = (y_1 - \exp(x_1'\beta), \ldots, y_n - \exp(x_n'\beta))^\top \in \mathbb{R}^n$, consists of the differences between observed values y_i and their corresponding predicted values $\exp(x_i'\beta)$. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ for $i = 1, ..., n$ represent the explanatory variable vectors, and let $\mathbf{1}_n$ be an $n \times 1$ vector of ones. The model matrix $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$ is defined as the concatenation of $\mathbf{1}_n$ and the matrix formed by taking each $\boldsymbol{x}_i$ as a row. The diagonal matrix $\text{diag}_{ij} \in \mathbb{R}^{n \times n}$ has the squared residuals on its main diagonal and zeros on the off-diagonal elements. The $(p+1) \times (p+1)$ matrix $\boldsymbol{\Sigma}$ accounts for the influence of the raw residuals on the standard errors of the estimated parameters in the space $\Upsilon \subseteq \mathbb{R}^p$.[15] The *degrees-of-freedom* adjustment is unaffected by the variable scale:

$$\boldsymbol{V}^{\text{H-W, adjusted}} = \frac{n}{n-1} \boldsymbol{V}^{\text{H-W}}, \quad (22)$$

where n is the number of observations in the model. Note that since the y-variables are scaled by $S \neq 1$, and $\log(S)$ is absorbed into the constant (thereby incorporating S into the exponential function), we have $\widetilde{\boldsymbol{\epsilon}} = (\widetilde{y}_1 - \exp(x_1'\widetilde{\beta}), \ldots, \widetilde{y}_n - \exp(x_n'\widetilde{\beta}))^\top = (S \cdot y_1 - S \cdot \exp(x_1'\beta), \ldots, S \cdot y_n - S \cdot \exp(x_n'\beta))^\top = S \cdot \boldsymbol{\epsilon}$.

[14] See Huber (1967) and White (1982).

[15] For the model specified in equation (2), which follows the standard PPML-gravity framework, it is easy to show that the parameter vector $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$ remains invariant under both x-variable and y-variable scaling across the entire parameter space $\Upsilon \subseteq \mathbb{R}^p$.

Therefore, given Equations (14), (16), (21) and (22), we have:

$$
\begin{aligned}
\boldsymbol{V}^{\text{H-W, adjusted}} &= \frac{n}{n-1}\left\{\boldsymbol{V}^{\text{asym}}\boldsymbol{X}^T \operatorname{diag}(\hat{\epsilon}_1^2,\ldots,\hat{\epsilon}_n^2)\boldsymbol{X}\boldsymbol{V}^{\text{asym}}\right\} \\
&= \frac{n}{n-1}\left\{\frac{1}{S}\left[\boldsymbol{I}(\beta)\right]^{-1}\boldsymbol{X}^T \operatorname{diag}(S^2\hat{\epsilon}_1^2,\ldots,\hat{S^2\epsilon}_n^2)\boldsymbol{X}\frac{1}{S}\left[\boldsymbol{I}(\beta)\right]^{-1}\right\} \\
&= \frac{n}{n-1}\left\{\left[\widetilde{\boldsymbol{I}}(\widetilde{\beta})\right]^{-1}\boldsymbol{X}^T \operatorname{diag}(S^2\hat{\epsilon}_1^2,\ldots,S^2\hat{\epsilon}_n^2)\boldsymbol{X}\left[\widetilde{\boldsymbol{I}}(\widetilde{\beta})\right]^{-1}\right\} \\
&= \frac{n}{n-1}\left\{\widetilde{\boldsymbol{V}}^{\text{asym}}\boldsymbol{X}^T \operatorname{diag}(\widetilde{\hat{\epsilon}}_1^2,\ldots,\widetilde{\hat{\epsilon}}_n^2)\boldsymbol{X}\widetilde{\boldsymbol{V}}^{\text{asym}}\right\} \\
&= \widetilde{\boldsymbol{V}}^{\text{H-W, adjusted}}.
\end{aligned}
\tag{23}
$$

That is, the robust variance estimator does not change with $S \neq 1$ on dependent variables, and neither do the Wald statistics, as can be inferred directly using equation (17).[16]

4.3 Empirical Application of the Wald Tests

To demonstrate empirically the results for Wald statistics we consider the same exercises as in Table 1, testing a single restriction that sets the coefficient on the *Colony* variable to zero. We estimate the models in STATA, using commands that employ each of our approaches to calculating the variance-covariance matrix. For each command we report results for two different scalings of bilateral trade.

In Table 2, columns 1 and 2 report results from a Wald test calculated after estimating with the *Poisson* command in STATA. The Wald test following this command uses V^{asym} to calculate the variance-covariance matrix. As in our derivations, the χ^2 statistic depends on the scale. Since the critical value is fixed, the p-values also depend on scale. In column 1, bilateral trade is scaled in actual U.S. dollars, the Wald test rejects the hypothesis that the colony variable can be excluded from the regression. In column 2, trade is scaled in billions of dollars, and the same test does not allow rejection of the null hypothesis.

Columns 3 and 4 report results from the same specifications estimated with the *ppmlhdfe*

[16]Note that the k x 1 vector x_t (where $t = \{1, ..., n\}$) of fixed variables are strictly exogenous, while $[x_1, x_2, ..., x_n]^T$ remains unaffected by the scaling choices.

command that we applied in Section III. This command uses $V^{H-W,adjusted}$ to calculate the test statistic. As in our derivations, the test statistic is independent of scaling. When robust standard errors are used, the model produces the same test statistic in both cases, and fails to reject the null hypothesis.

Table 2: Wald Tests using Asymptotic and Robust Standard Errors

Test of "Colony"	**Asymptotic SE**		**Robust SE**	
	$S = 1$	$S = 10^6$	$S = 1$	$S = 10^6$
Shipment in	USD	bil. USD	USD	bil. USD
χ^2	1.0e+05	0.10	0.03	0.03
$P > \chi^2$	0.00	0.7493	0.8674	0.8674

Results from replications of the full specification in Santos Silva and Tenreyro (2006a) with different scalings of the trade variable and different STATA commands. The hypothesis test considered here excludes the *Colony* variable from the benchmark model. Results in the *Asymptotic SE* column were calculated by postestimation testing after the "Poisson" STATA command. *Robust SE* results were calculated by testing after estimation with the "ppmlhdfe" STATA command.

5 Lagrange Multiplier (LM) Test

We now consider the implications of scaling for the LM test. As with the Wald statistic, the LM statistic relies on the asymptotic variance-covariance matrix defined in (13).

Using the 'small' model (equation (5)), the Hessian matrix is as follows:

$$\mathbf{H} = \frac{\partial^2 L^{H_0}(\beta)}{\partial \beta^2} = -\sum_{i=1}^{n} \begin{bmatrix} \exp(\beta_0 + \beta_1 \log(x_i)) & \log(x_i)\exp(\beta_0 + \beta_1 \log(x_i)) \\ \log(x_i)\exp(\beta_0 + \beta_1 \log(x_i)) & (\log(x_i))^2 \exp(\beta_0 + \beta_1 \log(x_i)) \end{bmatrix}. \tag{24}$$

Substituting $\boldsymbol{H}$ into the LM test statistics:

$$\mathrm{LM}_\beta = \nabla^T(\beta)[-\mathbf{H}]^{-1}\nabla(\beta). \tag{25}$$

where

$$\nabla(\beta) = \begin{bmatrix} -\sum_{i=1}^{n} \exp[\beta_0 + \beta_1 \log(x_i)] + \sum_{i=1}^{n} y_i \\ -\sum_{i=1}^{n} \exp[\beta_0 + \beta_1 \log(x_i)] \cdot \log(x_i) + \sum_{i=1}^{n} y_i \cdot \log(x_i) \end{bmatrix}$$

is the gradient of the log-likelihood function with respect to β_0 and β_1, respectively. Building upon the result we obtained from Equations (14) and (16), if we scale the dependent variable by a factor of $S \neq 1$:

$$\boldsymbol{V}^{\text{asym}} = S\widetilde{\boldsymbol{V}}^{\text{asym}}, \tag{26}$$

and given the parametric relationship before and after the scaling (derived from the first order conditions): $\beta_0 = \tilde{\beta}_0 - \log(S)$ and $\beta_1 = \tilde{\beta}_1$, we have:

$$\begin{aligned} \text{LM}_\beta &= \nabla^T(\beta)[-\boldsymbol{H}]^{-1}\nabla(\beta) \\ &= \nabla^T(\beta)\boldsymbol{V}^{\text{asym}}\nabla(\beta) \\ &= \frac{1}{S} \cdot S \cdot \nabla^T(\beta)\frac{1}{S}\boldsymbol{V}^{\text{asym}} S \cdot \nabla(\beta) \\ &= \frac{1}{S} \cdot S \cdot \nabla^T(\beta)\widetilde{\boldsymbol{V}}^{\text{asym}} S \cdot \nabla(\beta) \\ &= \frac{1}{S} \cdot \widetilde{\nabla^T}(\widetilde{\beta})\widetilde{\boldsymbol{V}}^{\text{asym}}\widetilde{\nabla}(\widetilde{\beta}) \\ &= \frac{1}{S} \cdot \widetilde{\text{LM}_{\tilde{\beta}}}. \end{aligned} \tag{27}$$

As with the other cases, hypothesis tests using the LM test are sensitive to scaling because the test statistic is sensitive to scaling while the critical value is not.

6 Other Topics

6.1 Other Scalings of the Data

We have shown that the scale of the data on the y-variable matters for the scale of the likelihood function and the test statistic of an LR test. Does scaling of x-variables matter in the same way? We investigate this question using the same analytical methods.

We take the same log-likelihood function, equation (5). Rather than scaling y_i by the arbitrary scalar S, we now apply S to x_i. The new log-likelihood appears as:

$$\widetilde{L^{H_0}}(\tilde{\beta}) = -\sum_{i}^{n} \exp[\tilde{\beta}_0 + \tilde{\beta}_1 \log(Sx_i)] + \sum_{i}^{n} y_i[\tilde{\beta}_0 + \tilde{\beta}_1 \log(Sx_i)], \tag{28}$$

where variables and parameters are defined as above. Once again, we take the first order conditions with respect to the β terms, and solve. The parameter before and after the scaling are related as follows: $\beta_0 = \tilde{\beta}_0 + \tilde{\beta}_1 \log(S)$ and $\beta_1 = \tilde{\beta}_1$. As above, scaling affects the constant term but not the other coefficients.

Analytical methods of the same kind that we use to study scaling of the y-variable show that three of the four tests we study are insensitive to scaling of the x-variable. Only the Lagrange Multiplier test is sensitive to scaling of the x-variable. The Wald test using a Huber-White variance-covariance matrix is the only test that is robust to scaling of data on both y- and x-variables.

6.2 Extension to Structural Estimation via MPECs

A key reason to explore the viability of joint hypothesis tests in the PPML framework is that recent developments in the structural estimation literature allow clean tests of the CES gravity model against richer theories of trade. In these problems, theories of bilateral trade are expressed through the constraints on the econometric objective. Viable tests of joint parameter restrictions would allow formal evaluation of rich theories of bilateral trade against models that are nested within those theories. For example, Yang (2021) develops a parsimonious Armington model of trade using the Constant Difference of Elasticities (CDE) framework proposed by Hanoch (1975). This CDE model nests the non-homothetic CES preferences that have recently become popular in the structural transformation literature (e.g. Comin, Lashkari and Mestieri (2021)), as well as the canonical CES gravity model estimated by Anderson and van Wincoop (2003).

Formal evaluation of restrictions imposed by the CDE model we leave to Yang (2021). In this paper we only wish to understand if the lessons from the reduced form specification can be transferred to an MPEC setting. We consider a test of a single parameter restriction in the context of Balistreri and Hillberry (2007)'s MPEC estimation of the canonical CES gravity model. This is lengthy, so we include it in an appendix, Appendix B. In brief, we find that our results do transfer to this particular MPEC setting. Statistical inference in the LR test is polluted by sensitivity to data scaling, while the Wald test based on Huber-White standard errors is insensitive to scaling. Our analytical work evaluates only a single parameter restriction, but we suspect that this result is general. Generalizing this insight to multiple restrictions in other MPECs is beyond the scope of this particular paper.[17]

7 Conclusion

The CES gravity model is widely accepted as canon, even though the model's assumptions are restrictive and - in light of the theory's prominence - under-tested. The few prominent papers that test the theory apply t-tests to individual parameters, parameters associated with variables that are constructed after the first stage of a two-step estimation process involving PPML. A framework that allows joint testing of multiple parameter restrictions would be preferable. Estimation of highly general models of gravity is possible via MPECs, but formal testing requires an econometric procedure that supports such tests.

In this paper we investigate the conditions under which data scaling choices affect hypothesis testing of parameter restrictions imposed on estimates from a PPML gravity model. Tests of individual parameter restrictions can be evaluated with t-tests, so we focus on tests that are capable of evaluating joint restrictions. We use both analytical exercises and empirical examples to demonstrate our results.

[17] In most circumstances a numerical proof is likely to be sufficient. If the results of a hypothesis test can be shown to be robust to alternative scalings of the data (as in Table 2), the problems we have identified here would seem to be resolved. We suspect that LR tests will be polluted by scaling in other settings, while the particular Wald test we identify as robust to scaling will prove robust to scaling in other related settings.

Our study considers four kinds of hypothesis tests: an LR test, Wald tests with two different methods for calculating the variance-covariance matrix, and an LM test. Of these, we find only one test that is robust to scaling the data on the y-variable, the Wald test that uses Huber-White corrected standard errors to construct the variance-covariance matrix. This test is also robust to scaling of x-variables. In an appendix, we also show that the test can be used to evaluate hypotheses in the context of an MPEC.

Our results indicate that a Wald test based on upon a variance-covariance matrix constructed from robustly estimated standard errors is appropriate for testing joint restrictions on parameters estimated in a PPML specification. This finding is of particular relevance to empirical studies of bilateral trade, because the PPML specification is commonly employed there and because data scaling is common. The finding is likely to be especially useful for a nascent literature that tests theoretical constraints that canonical models of bilateral trade impose on more general theories. The result may also be useful in other applications.

References

Anderson, James E. 1979. "A Theoretical Foundation for the Gravity Equation." *The American Economic Review*, 69(1): 106–116.

Anderson, James E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review*, 93(1): 170–192.

Anderson, James E, and Yoto V Yotov. 2012. "Gold Standard Gravity." National Bureau of Economic Research Working Paper 17835.

Arkolakis, Costas, Arnaud Costinot, and Andrés Rodríguez-Clare. 2012. "New Trade Models, Same Old Gains?" *American Economic Review*, 102(1): 94–130.

Balistreri, Edward J., and Russell H. Hillberry. 2007. "Structural Estimation and the Border Puzzle." *Journal of International Economics*, 72(2): 451–463.

Balistreri, Edward J., Russell H. Hillberry, and Thomas F. Rutherford. 2011. "Structural Estimation and Solution of International Trade Models with Heterogeneous Firms." *Journal of International Economics*, 83(2): 95–108.

Caron, Justin, Thibault Fally, and James R Markusen. 2014. "International Trade Puzzles: A Solution Linking Production and Preferences." *The Quarterly Journal of Economics*, 129(3): 1501–1552.

Cayley, Arthur. 1858. "A Memoir on the Theory of Matrices." *Philosophical Transactions of the Royal Society of London*, 148: 17–37.

Chaney, Thomas. 2008. "Distorted Gravity: The Intensive and Extensive Margins of International Trade." *American Economic Review*, 98(4): 1707–21.

Chen, Natalie, and Dennis Novy. 2022. "Gravity and Heterogeneous Trade Cost Elasticities." *The Economic Journal*, 132(644): 1349–1377.

Comin, Diego, Danial Lashkari, and Martí Mestieri. 2021. "Structural Change with Long-run Income and Price Effects." *Econometrica*, 89(1): 311–374.

Correia, Sergio, Paulo Guimarães, and Tom Zylkin. 2020. "Fast Poisson Estimation with High-dimensional Fixed Effects." *The Stata Journal*, 20(1): 95–115.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica*, 70(5): 1741–1779.

Fally, Thibault. 2015. "Structural Gravity and Fixed Effects." *Journal of International Economics*, 97: 76–85.

Goldstein, Harvey. 2011. *Multilevel Statistical Models.* Vol. 922, John Wiley & Sons.

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984*a*. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica*, 52(3): 701–720.

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984*b*. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica*, 52(3): 681–700.

Gouriéroux, Christian, Alberto Holly, and Alain Monfort. 1982. "Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters." *Econometrica*, 50(1): 63–80.

Greene, William H. 2012. *Econometric Analysis.* . 7th ed., Prentice Hall.

Hanoch, Giora. 1975. "Production and Demand Models with Direct or Indirect Implicit Additivity." *Econometrica: Journal of the Econometric Society*, 43(3): 395–419.

Huber, Peter J. 1967. "Under Nonstandard Conditions." Berkeley, CA, USA:University of California Press.

Maas, Cora JM, and Joop J Hox. 2004. "The Influence of Violations of Assumptions on Multilevel Parameter Estimates and their Standard Errors." *Computational Statistics & Data Analysis*, 46(3): 427–440.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity." *Econometrica*, 71(6): 1695–1725.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006a. "The Log of Gravity." *The Review of Economics and Statistics*, 88(4): 641–658.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006b. "Data for 'The Log of Gravity'." London School of Economics. http://personal.lse.ac.uk/tenreyro/regressors.zip, accessed June 8, 2022.

Su, Che-Lin, and Kenneth L. Judd. 2012. "Constrained Optimization Approaches to Estimation of Structural Models." *Econometrica*, 80(5): 2213–2230.

Tan, Shawn W. 2013. "Structural Estimation of a Flexible Translog Gravity Model." *University of Melbourne, Department of Economics Working Paper, 1164.*

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica: Journal of the Econometric Society*, 1–25.

Yang, Anton C. 2021. "Structural Estimation of a Gravity Model of Trade with the Constant-Difference-of-Elasticities Preferences." *mimeo*.

Appendix: Variable Scaling and Hypothesis Testing in the Gravity Model

A Extension to Multivariate Settings

In this appendix we show that the results in the paper extend to a multivariate setting with multiple restrictions. These results are structured as Corollaries to Theorem 1.

A.1 Corollary 1.1 and Proof

Corollary 1.1 *Let $L(\beta)$ be the twice-differentiable log-likelihood function:*

$$\begin{aligned} L(\beta) = &- \sum_{i}^{n} \exp[\beta_0 + \beta_1 \log(x_i) + \beta_2 \log(z_i) + ... + \beta_m \log(w_i)] \\ &+ \sum_{i}^{n} y_i[\beta_0 + \beta_1 \log(x_i) + \beta_2 \log(z_i) + ... + \beta_m \log(w_i)], \end{aligned} \tag{A.1}$$

where $\beta_0, \beta_1, ..., \beta_m$ are parameters, and $x_i, z_i, ..., w_i$ are observed from the data. Consider a scalar $S \neq 1$ on y_i, which yields the following log-likelihood function:

$$\begin{aligned} \widetilde{L}(\tilde{\beta}) = &- \sum_{i}^{n} \exp[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i) + \tilde{\beta}_2 \log(z_i) + ... + \tilde{\beta_m} \log(w_i)] \\ &+ \sum_{i}^{n} (Sy_i)[\tilde{\beta}_0 + \tilde{\beta}_1 \log(x_i) + \tilde{\beta}_2 \log(z_i) + ... + \tilde{\beta_m} \log(w_i)], \end{aligned} \tag{A.2}$$

where $\widetilde{\mathcal{B}} \equiv \{\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta_m}\}$ are the parameters given the scale, and the relationship between the original parameters $\mathcal{B}$ and the scaled parameters $\widetilde{\mathcal{B}}$ is as follows:

$$\begin{cases} \beta_0 = \widetilde{\beta}_0 - \log(S) \\ \beta_1 = \widetilde{\beta}_1 \\ \beta_2 = \widetilde{\beta}_2 \\ \vdots \\ \beta_m = \widetilde{\beta}_m, \end{cases} \tag{A.3}$$

then the Hessian matrix of log-likelihood function $\widetilde{L}(\widetilde{\beta})$*, denoted by* $\boldsymbol{H}(\widetilde{\beta})$*, is related to the original Hessian matrix* $\boldsymbol{H}(\beta)$ *by:*

$$\widetilde{\boldsymbol{H}}(\tilde{\beta}) = S \cdot \boldsymbol{H}(\beta). \tag{A.4}$$

Corollary 1.1 asserts that each element of the Hessian matrix $\boldsymbol{H}(\beta)$ (in the PPML function) is scaled by S when the dependent variable y_i is scaled by S.

Proof. Using the chain rule, it is easy to show that the Hessian matrix of the log-likelihood function, for this more general case, is:

$$\boldsymbol{H}(\beta) = \frac{\partial^2 L^{H_0}(\beta)}{\partial \beta \cdot \partial \beta^T} = -\sum_{i=1}^{n} \exp(\beta_0 + \beta_1 \log(x_i) + \beta_2 \log(z_i) + ... + \beta_m \log(w_i)) \cdot \boldsymbol{M}_i, \tag{A.5}$$

where $\boldsymbol{M}_i$ (corresponding to each observation i in the data) is a matrix of the second-order partial derivatives with respect to model parameters $\mathcal{B} \equiv \{\beta_0, \beta_1, ..., \beta_m\}$ as follows:

$$\boldsymbol{M}_i = \begin{bmatrix} 1 & \log(x_i) & \log(z_i) & \cdots & \log(w_i) \\ \log(x_i) & (\log(x_i))^2 & (\log(x_i)\log(z_i)) & \cdots & (\log(x_i)\log(w_i)) \\ \log(z_i) & (\log(z_i)\log(x_i)) & (\log(z_i))^2 & \cdots & (\log(z_i)\log(w_i)) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \log(w_i) & (\log(w_i)\log(x_i)) & (\log(w_i)\log(z_i)) & \cdots & (\log(w_i))^2 \end{bmatrix}. \tag{A.6}$$

Note that the scaling factor $S \neq 1$ on y_i does not affect $\boldsymbol{M}_i$, but only the first term of the right-hand side in equation (A.5), where the relationships between the original parameters $\{\beta_0, \beta_1, \ldots, \beta_m\}$ and the scaled parameters $\{\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_m\}$ are given by $\beta_0 = \tilde{\beta}_0 - \log(S)$ and $\beta_j = \tilde{\beta}_j$ for $j = 1, 2, \ldots, m$. Hence, we have

$$\begin{aligned} \widetilde{\boldsymbol{H}}(\widetilde{\beta}) = \frac{\widetilde{\partial^2 L^{H_0}}(\beta)}{\partial\widetilde{\beta} \cdot \partial\widetilde{\beta}^T} &= -\sum_{i=1}^{n} \exp(\widetilde{\beta}_0 + \widetilde{\beta}_1 \log(x_i) + \widetilde{\beta}_2 \log(z_i) + ... + \widetilde{\beta_m} \log(w_i)) \cdot \boldsymbol{M}_i \\ &= -\sum_{i=1}^{n} \exp(\beta_0 + \log(S) + \beta_1 \log(x_i) + \beta_2 \log(z_i) + ... + \beta_m \log(w_i)) \cdot \boldsymbol{M}_i \\ &= -S \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 \log(x_i) + \beta_2 \log(z_i) + ... + \beta_m \log(w_i)) \cdot \boldsymbol{M}_i \\ &= S \cdot \boldsymbol{H}(\beta). \end{aligned} \tag{A.7}$$

■

A.2 Corollary 1.2 and Proof

Corollary 1.2 *Given the conditions of Corollary 1.1, and assuming the Fisher Information Matrix $\boldsymbol{I}^F$, defined over a commutative ring with identity R, is invertible (nonsingular), we can generalize the result in equation (23) to accommodate multiple independent variables $\boldsymbol{x}_i \in \mathbb{R}^n$, where $i \geq 2$. This applies to a parameter set $\boldsymbol{\mathcal{B}} = \beta_0, \beta_1, ..., {\beta_n}^T$ within the parameter space $\Gamma \subseteq \mathbb{R}^n$.*

Corollary 1.2 asserts that the results are applicable for joint restrictions on multiple variables.

Proof. We start with the 'large' model introduced in equation (8):

$$L^{H_a}(\beta) = -\sum_{i}^{n} \exp[\beta_0 + \beta_1 \log(x_i) + \beta_2 \log z_i] + \sum_{i}^{n} y_i[\beta_0 + \beta_1 \log(x_i) + \beta_2 \log z_i], \tag{A.8}$$

where $z_i \subseteq \boldsymbol{x}_i \in \mathbb{R}^n$ is the additional independent variable, β_2 is its corresponding model parameter.[18]

Since equation (2) is twice-differentiable, the Hessian matrix is given by:

$$\boldsymbol{H}(\beta) = \begin{bmatrix} \frac{\partial^2 L^{H_a}}{\partial \beta_0 \beta_0^T} & \frac{\partial^2 L^{H_a}}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 L^{H_a}}{\partial \beta_0 \partial \beta_2} \\ \frac{\partial^2 L^{H_a}}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 L^{H_a}}{\partial \beta_1 \beta_1^T} & \frac{\partial^2 L^{H_a}}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 L^{H_a}}{\partial \beta_2 \partial \beta_0} & \frac{\partial^2 L^{H_a}}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 L^{H_a}}{\partial \beta_2 \beta_2^T} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \tag{A.9}$$

where each entry is a second order partial derivative of the log-likelihood function with respect to the parameters β_0, β_1, and β_2. This is a symmetric matrix because the order of differentiation does not matter.

Given invertibility, the inverse of the negative Hessian can be simplified as:

$$\boldsymbol{V} = \left(\boldsymbol{I}^F\right)^{-1} = [-\boldsymbol{H}(\beta)]^{-1} = \left\{ - \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right\}^{-1} = -\frac{1}{\det(\boldsymbol{I}^F)} \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}^T, \tag{A.10}$$

where C_{ij} is the cofactor of a_{ij} in the matrix $\boldsymbol{H}(\beta)$, $\det(\boldsymbol{I}^F)$ is the determinant of the Fisher information matrix. In this case, we apply the Rule of Sarrus for the calculation of the determinant of variance-covariance matrix $\boldsymbol{V}$:

[18] We eliminate prime notation used originally in the large model because they are unnecessary here.

$$\begin{aligned}\det(\boldsymbol{I}^F) &= a_{11}\cdot C_{11} - a_{12}\cdot C_{12} + a_{13}\cdot C_{13} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}).\end{aligned} \tag{A.11}$$

From Corollary 1.1, it can be shown that each element a_{ij} is scaled by $S \neq 1$, hence:

$$\begin{aligned}\det(\widetilde{\boldsymbol{I}^F}) &= \widetilde{a}_{11}(\widetilde{a}_{22}\widetilde{a}_{33} - \widetilde{a}_{23}\widetilde{a}_{32}) - \widetilde{a}_{12}(\widetilde{a}_{21}\widetilde{a}_{33} - \widetilde{a}_{23}\widetilde{a}_{31}) + \widetilde{a}_{13}(\widetilde{a}_{21}\widetilde{a}_{32} - \widetilde{a}_{22}\widetilde{a}_{31}) \\ &= S^3\cdot[a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})] \\ &= S^3\cdot\det(\boldsymbol{I}^F).\end{aligned} \tag{A.12}$$

By the Cayley–Hamilton theorem shown in Cayley (1858), we have

$$A^n - \mathrm{tr}_1 A^{n-1} + \ldots + (-1)^{n-1}\mathrm{tr}_{n-1}A + (-1)^n\det(A)I_{\mathrm{nxn}} = \mathbf{0}_{\mathrm{nxn}}, \tag{A.13}$$

where $A = \boldsymbol{I}^F = \{a_{ij}\}_{\mathrm{nxn}}$ is a n x n square matrix, satisfying its characteristic polynomial; the trace at the k^{th} level, denoted tr_k, is the sum of all principal minors of A of order k. Since A is invertible, we substitute with $n = 3$, and then multiply through by the inverse of A in equation (A.13):

$$A^2 - \mathrm{tr}(A)A + \mathrm{tr}(A^2)\boldsymbol{I}_{\mathrm{nxn}} - \det(A)A^{-1} = \mathbf{0}_{\mathrm{nxn}}. \tag{A.14}$$

Then the non-robust asymptotic variance-covariance matrix $\boldsymbol{V}$ can be written as follows:

$$\boldsymbol{V}^{\mathrm{asym}} = \left(\boldsymbol{I}^F\right)^{-1} = \frac{1}{\det(\boldsymbol{I}^F)}\left\{\boldsymbol{I}^{F^2} - \mathrm{tr}(\boldsymbol{I}^F)\boldsymbol{I}^F + \mathrm{tr}(\boldsymbol{I}^{F^2})\boldsymbol{I}_{\mathrm{nxn}}\right\}, \tag{A.15}$$

where $\mathrm{tr}(\boldsymbol{I}^F)$ is the trace of the matrix $\boldsymbol{I}^F$.

Since each element of the variance-covariance matrix is scaled by a factor $S \neq 1$, the effect on both the squares of the matrices $\boldsymbol{I}^F$ and the sum of the elements on the main diagonal

will be a scaling by S^2 ($\text{tr}(\widetilde{\boldsymbol{I^F}})$ is a linear combination of S and unscaled diagonal elements). Given equation (A.12), we have $\boldsymbol{V}^{\text{asym}} = S\widetilde{\boldsymbol{V}}^{\text{asym}}$. As in equation (23), the scaling factor $1/S^2$, which arises from the product $1/S = \widetilde{\boldsymbol{V}}^{\text{asym}}/\boldsymbol{V}^{\text{asym}}$, is canceled out by the S^2 term originating from the raw residuals $\text{diag}(S^2\hat{\epsilon}_1^2, \ldots, S^2\hat{\epsilon}_n^2)$. Thus, $\boldsymbol{V}^{\text{H-W, adjusted}} = \widetilde{\boldsymbol{V}}^{\text{H-W, adjusted}}$.

A.2.1 Extension to an n x n Fisher Information Matrix

Using the results in Corollary 1.1, we define $\widetilde{\boldsymbol{I}^F}$ ($S \neq 1$) to be the $n \times n$ matrix obtained by multiplying every element in the Fisher information matrix by S:

$$\widetilde{\boldsymbol{I}^F} = \begin{bmatrix} Sv_{11} & Sv_{12} & \ldots & Sv_{1n} \\ Sv_{21} & Sv_{22} & \ldots & Sv_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Sv_{n1} & Sv_{n2} & \ldots & Sv_{nn} \end{bmatrix}. \tag{A.16}$$

Applying Laplace's expansion along the first row:

$$\det(\boldsymbol{I}^F) = \sum_{j=1}^{n} (-1)^{1+j} v_{1j} \boldsymbol{I}_{1j}^F, \tag{A.17}$$

where $\boldsymbol{I}_{1j}^F$ is the determinant of the $(n-1) \times (n-1)$ submatrix that results from deleting the first row and jth column from $\boldsymbol{I}^F$. The determinant of the scaled submatrix is given by:

$$
\begin{aligned}
\det(\widetilde{\boldsymbol{I}^F}) &= \sum_{j=1}^{n}(-1)^{1+j}(\widetilde{v}_{1j})\det(\widetilde{\boldsymbol{I}^F}_{1j}) \\
&= \sum_{j=1}^{n}(-1)^{1+j}(Sv_{1j})\det(\widetilde{\boldsymbol{I}^F}_{1j}) \\
&= \sum_{j=1}^{n}(-1)^{1+j}(Sv_{1j})S^{n-1}\det(\boldsymbol{I}^F_{1j}) \\
&= S^n\sum_{j=1}^{n}(-1)^{1+j}v_{1j}\det(\boldsymbol{I}^F_{1j}) \\
&= S^n\det(\boldsymbol{I}^F).
\end{aligned}
\tag{A.18}
$$

Here we used the fact that $\widetilde{\boldsymbol{I}^F}_{1j}$ is the $(n-1)\times(n-1)$ matrix $\boldsymbol{I}^F_{1j}$ with all of its entries multiplied by S (by applying Corollary 1.1), so the determinant of $\widetilde{\boldsymbol{I}^F}_{1j}$ is $S^{n-1}\det(\boldsymbol{I}^F_{1j})$.

The scaled variance-covariance can be written as:

$$
\widetilde{\boldsymbol{V}} = \begin{bmatrix} Sv_{11} & Sv_{12} & \dots & Sv_{1n} \\ Sv_{21} & Sv_{22} & \dots & Sv_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Sv_{n1} & Sv_{n2} & \dots & Sv_{nn} \end{bmatrix}^{-1} = -\frac{1}{\det(\widetilde{\boldsymbol{I}^F})}\begin{bmatrix} \widetilde{C}_{11} & \widetilde{C}_{12} & \dots & \widetilde{C}_{1n} \\ \widetilde{C}_{21} & \widetilde{C}_{22} & \dots & \widetilde{C}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{C}_{n1} & \widetilde{C}_{n2} & \dots & \widetilde{C}_{nn} \end{bmatrix}^{T} = -\frac{\mathrm{adj}(\widetilde{\boldsymbol{I}})}{\det(\widetilde{\boldsymbol{I}^F})}.
\tag{A.19}
$$

Since each determinant of the $(n-1)\times(n-1)$ submatrix of $\boldsymbol{I}^F$ ($\det(\boldsymbol{I}^F_{ij})$) is scaled by S^{n-1}, thus each cofactor C_{ij} is scaled by S^{n-1}: $\widetilde{C}_{ij} = (-1)^{i+j}\det(\widetilde{\boldsymbol{I}^F}_{ij}) = S^{n-1}C_{ij}$. It follows that the adjugate is scaled by S^{n-1}: $\mathrm{adj}(\widetilde{\boldsymbol{I}}) = S^{n-1}\cdot\mathrm{adj}(\boldsymbol{I})$. Since the adjugate is scaled by S^{n-1} in the numerator, and the determinant is scaled by S^n; hence, we have $\boldsymbol{V}^{\mathrm{asym}} = S\widetilde{\boldsymbol{V}}^{\mathrm{asym}}$. The remainder of the proof proceeds as in equation (23). ■

B Wald Test under Structural Estimation

Now consider estimating the non-linear problem with constrained optimization using an MPEC. We will use the example demonstrated in Balistreri and Hillberry (2007), but will apply PPML as the objective function instead. The PPML maximization function is given by:

$$\begin{aligned}\mathcal{L}(\beta, U, \rho, b) = &- \sum_{\substack{(i,j)\in\mathbb{S} \\ i\neq j}} \exp\left[\log\left(\beta_i\right) + (1-\sigma)\log\left(U_j\right) + (1-\sigma)\log\left(p_i\right) + (1-\sigma)\log\left(t_{ij}\right) + \sigma\log\left(Y_j\right)\right] \\ &+ \sum_{\substack{(i,j)\in\mathbb{S} \\ i\neq j}} y_{ij}\cdot\left[\log\left(\beta_i\right) + (1-\sigma)\log\left(U_j\right) + (1-\sigma)\log\left(p_i\right) + (1-\sigma)\log\left(t_{ij}\right) + \sigma\log\left(Y_j\right)\right].\end{aligned} \tag{B.1}$$

The associated constraints appear as:

$$\begin{aligned}
(1): &\quad \left[\sum_{i=1}^{n}\beta_i(p_i\tau_{ij})^{1-\sigma}\right]^{\frac{1}{1-\sigma}} \geq P_j \quad \perp \quad U_j \geq 0 \\
(2): &\quad q_i \geq \sum_{j=1}^{n}\left[\beta_i\frac{Y_j}{p_i}\left(\frac{p_i\tau_{ij}}{P_j}\right)^{1-\sigma}\right] \quad \perp \quad p_i \geq 0 \\
(3): &\quad U_jP_j \geq Y_j \quad \perp \quad P_j \geq 0 \\
(4): &\quad Y_i = p_iq_i \\
(5): &\quad \tau_{ij} = d_{ij}^{\rho}\cdot b^{1-\delta_{ij}} \\
(6): &\quad \sum_i \beta_i = \Sigma \quad \text{(normalization)},
\end{aligned}$$

where $\perp$ indicates complementary slackness, d_{ij} is the distance between i and j observed from the data, ρ is the elasticity of trade costs with respect to distance; b is the border coefficient equaling one plus tariff equivalent of border costs. δ_{ij}'s are the dummy variables equaling zeros if shipments cross international border and equaling ones if the shipments are

taken place domestically in locations $\{i, j\} \in \mathcal{S}$.

Note that σ is not easily identified in this framework. As a result, the conventional approach in the trade literature is to fix σ at a specific assumed value. Prices $\boldsymbol{p}$ are also exogenous (either in the data or chosen for the purpose of counterfactual exercises). For simplicity, let us assume that all constraints in the optimization problem are satisfied with equality and are binding at the optimal solution.

Conditions (1), (3) and (5) in the system of constraints give:

$$U_j \left[\sum_{i=1}^{n} \beta_i (b \cdot d_{ij})^{\rho(1-\sigma)} \right]^{\frac{1}{1-\sigma}} = Y_j. \tag{B.2}$$

Conditions (2) and (4) ensure that q_i is exogenous (given the choice of normalized p_i's), thus

$$q_i = \sum_{j=1}^{n} \left[\beta_i \frac{Y_j}{p_i} \left(\frac{p_i \tau_{ij}}{P_j} \right)^{1-\sigma} \right]. \tag{B.3}$$

$$\implies \quad q_i = \sum_{j=1}^{n} \left[\beta_i p_i^{-\sigma} (b \cdot d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right]. \tag{B.4}$$

To make our analysis simple, we fix $p_i = 1 \ \forall i$ and suppress the border costs so $\tau_{ij} = d_{ij}^{\rho}$. Without loss of generality (in terms of parameter estimation), we transform the objective function (B.1) to the following:

$$\begin{aligned} \mathcal{L}(K, U, \rho) = &- \sum_{\substack{(i,j) \in \mathcal{S} \\ i \neq j}} \exp \left[K_i + W_j + (1-\sigma) \log (U_j) + \rho \log (\Delta_{ij}) \right] \\ &+ \sum_{\substack{(i,j) \in \mathcal{S} \\ i \neq j}} y_{ij} \cdot \left[K_i + W_j + (1-\sigma) \log (U_j) + \rho \log (\Delta_{ij}) \right], \end{aligned} \tag{B.5}$$

where $K_i = \log(\beta_i)$ is unobserved by econometrician, and $W_j = \sigma \log(Y_j)$ will be swept into K_i if σ is exogenous, so $\{K_i + W_j\}$ is also unobserved (because K_i is not observed) which is naturally a constant matrix for $\{i, j\} \in \mathcal{S}$ pairs of Poisson regression in this problem. $\Delta_{ij} = d_{ij}^{1-\sigma}$ is observed because d_{ij}'s are physical distances. If U_j's are also observed, then

this is similar to a standard PPML specification without problem constraints. In this case, one can use our mathematics above to test if the distance variable should be included in the model.

Given $\boldsymbol{U}$ is unobserved, we can rearrange equation (B.2) as follows:

$$U_j^{1-\sigma}\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right] = Y_j^{1-\sigma}. \tag{B.6}$$

$$\implies \quad U_j^{1-\sigma} = \frac{Y_j^{1-\sigma}}{[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}]}. \tag{B.7}$$

Taking the log of both sides:

$$\implies \quad (1-\sigma)\log(U_j) = \frac{Y_j^{1-\sigma}}{[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}]}. \tag{B.8}$$

$$\implies \quad (1-\sigma)\log(U_j) = (1-\sigma)\log(Y_j) - \log\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right]. \tag{B.9}$$

Substituting $W_j = \sigma\log(Y_j)$:

$$\implies \quad (1-\sigma)\log(U_j) = \log(Y_j) - W_j - \log\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right]. \tag{B.10}$$

Combining Equations (B.10) and (B.5) cancels out W_j, we get:

$$\begin{aligned}\mathcal{L}(K,U,\rho) = &-\sum_{\substack{(i,j)\in S\\ i\neq j}}\exp\left\{\Omega_{ij} - \log\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right] + \rho\log\left(\Delta_{ij}\right)\right\}\\ &+\sum_{\substack{(i,j)\in S\\ i\neq j}} y_{ij}\cdot\left\{\Omega_{ij} - \log\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right] + \rho\log\left(\Delta_{ij}\right)\right\}.\end{aligned} \tag{B.11}$$

The term Ω_{ij}, which is defined as $\{K_i + \log(Y_j)\}$ or equivalently $\{\log(\beta_i) + \log(Y_j)\}$, is not observed in our dataset. The variable Y_{ij}, representing each pair of i and j, is observed.

However, the parameter β_i, which is unique for every i, is not observed. Also, d_{ij} is a known physical distance for each pair of i and j. The parameter σ takes its assumed value. The parameter ρ is to be estimated. Again, Δ_{ij}, which pertains to each pair of i and j, is observed in our data. Since we have n countries, and y_{ij} are bilateral flows among them, we have $n(n-1)$ equations in terms of Poisson specification (because for each country, we have $n-1$ trade partners). Meanwhile we have n unknowns from the β_i (one for each country) and 1 unknown.

So far, we have used conditions (1), (3), and (5), but not (2), (4), and (6). If we combine (2) and (4), while using the normalization of p_i's to 1, we get:

$$Y_i = \sum_{j=1}^{n} \left[\beta_i (d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma}\right]. \tag{B.12}$$

Since β_i does not have an index of j, it is essentially a constant with respect to the summation. Factoring β_i out of the summation:

$$\implies \quad \beta_i = \frac{Y_i}{\sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma}\right]}. \tag{B.13}$$

This implies that if we can estimate β's from the Poisson regression, then equation (B.13) merely serves as the required n equations to calculate n unknown U_j's. Alternatively, we may choose to substitute (B.13) into (B.5). In this case, the empirical specification will be similar to Anderson and van Wincoop (2003)'s "inward and outward multilateral resistances" with the following components:

$$\implies \quad (1-\sigma)\log(U_j) = (1-\sigma)\log(Y_j) - \log\left[\frac{\sum_{i=1}^{n} Y_i \cdot d_{ij}^{\rho(1-\sigma)}}{\sum_{j=1}^{n} Y_j^{\sigma} U_j^{1-\sigma} \cdot d_{ij}^{\rho(1-\sigma)}}\right]. \tag{B.14}$$

Note that the principle of Rank-Nullity Theorem suggests that the model is just identified using the constrained optimization approach, since we have effectively $2n+1$ equations: n equations from (B.10), n equations from (B.13) and the one objective function (B.5). We

have $2n+1$ unknowns: n unknowns from β_i's, n unknowns from U_j's and ρ. Note that the normalization condition (6) is typically applied in this non-linear system when (1) there are excess degrees of parameter space, and (2) other parameters will be potentially badly scaled if we only choose one of the β_i's given high non-linearity.

B.1 Calculating the Robust Standard Errors

To handle the model constraints, one common method is to incorporate them into the optimization objective using the method of Lagrange multipliers, which transforms a constrained optimization problem into an unconstrained problem:

$$\begin{aligned} L(\beta, U, \rho, \lambda, \mu) = &- \sum_{\substack{(i,j)\in\mathcal{S} \\ i\neq j}} \exp\left[K_i + W_j + (1-\sigma)\log\left(U_j\right) + \rho\log\left(\Delta_{ij}\right)\right] \\ &+ \sum_{\substack{(i,j)\in\mathcal{S} \\ i\neq j}} y_{ij}\cdot\left[K_i + W_j + (1-\sigma)\log\left(U_j\right) + \rho\log\left(\Delta_{ij}\right)\right] \\ &+ \lambda\cdot\left\{(1-\sigma)\log(U_j) - \log(Y_j) + W_j + \log\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right]\right\} \\ &+ \mu\cdot\left(\beta_i - \frac{Y_i}{\sum_{j=1}^{n}\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{1-\sigma}\right]}\right). \end{aligned} \tag{B.15}$$

We now calculate the Hessian, by differentiating with respect to the model parameters, which are β, U, ρ, and the Lagrange multipliers λ and μ:

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 L}{\partial\beta\partial\beta^T} & \frac{\partial^2 L}{\partial\beta\partial U} & \frac{\partial^2 L}{\partial\beta\partial\rho} & \frac{\partial^2 L}{\partial\beta\partial\lambda} & \frac{\partial^2 L}{\partial\beta\partial\mu} \\ \frac{\partial^2 L}{\partial U\partial\beta} & \frac{\partial^2 L}{\partial U\partial U^T} & \frac{\partial^2 L}{\partial U\partial\rho} & \frac{\partial^2 L}{\partial U\partial\lambda} & \frac{\partial^2 L}{\partial U\partial\mu} \\ \frac{\partial^2 L}{\partial\rho\partial\beta} & \frac{\partial^2 L}{\partial\rho\partial U} & \frac{\partial^2 L}{\partial\rho\partial\rho^T} & \frac{\partial^2 L}{\partial\rho\partial\lambda} & \frac{\partial^2 L}{\partial\rho\partial\mu} \\ \frac{\partial^2 L}{\partial\lambda\partial\beta} & \frac{\partial^2 L}{\partial\lambda\partial U} & \frac{\partial^2 L}{\partial\lambda\partial\rho} & \frac{\partial^2 L}{\partial\lambda^2} & \frac{\partial^2 L}{\partial\lambda\partial\mu} \\ \frac{\partial^2 L}{\partial\mu\partial\beta} & \frac{\partial^2 L}{\partial\mu\partial U} & \frac{\partial^2 L}{\partial\mu\partial\rho} & \frac{\partial^2 L}{\partial\mu\partial\lambda} & \frac{\partial^2 L}{\partial\mu^2} \end{bmatrix}. \tag{B.16}$$

We shall derive the score functions in order to compute the Hessian matrix:

$$\frac{\partial L}{\partial \beta} = -\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} \beta_i^{-1} \exp\left[K_i + W_j + (1-\sigma)\log\left(U_j\right) + \rho\log\left(\Delta_{ij}\right)\right] + \sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} y_{ij}\beta_i^{-1} + \frac{\lambda(d_{ij})^{\rho(1-\sigma)}}{\sum_{i=1}^n \beta_i (d_{ij})^{\rho(1-\sigma)}} + \mu$$

$$= -\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} \exp\left[W_j + (1-\sigma)\log\left(U_j\right) + \rho\log\left(\Delta_{ij}\right)\right] + \sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} y_{ij}\beta_i^{-1} + \frac{\lambda(d_{ij})^{\rho(1-\sigma)}}{\sum_{i=1}^n \beta_i (d_{ij})^{\rho(1-\sigma)}} + \mu$$

$$\frac{\partial L}{\partial U} = -\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} \frac{1-\sigma}{U_j} \exp\left[K_i + W_j + (1-\sigma)\log\left(U_j\right) + \rho\log\left(\Delta_{ij}\right)\right]$$

$$+ \sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} y_{ij}\cdot\frac{1-\sigma}{U_j} + \lambda\cdot\frac{1-\sigma}{U_j} + \frac{\mu\cdot(1-\sigma)\cdot Y_i\cdot\sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{-\sigma}\right]}{\left\{\sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{1-\sigma}\right]\right\}^2},$$

$$\frac{\partial L}{\partial \rho} = -\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} \exp\left[K_i + W_j + (1-\sigma)\log\left(U_j\right) + \rho\log\left(\Delta_{ij}\right)\right]\cdot\log(\Delta_{ij})$$

$$+ \sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} \log\left(\Delta_{ij}\right)\cdot y_{ij} - \lambda(1-\sigma)\sum_{i=1}^n \frac{\beta_i(d_{ij})^{\rho(1-\sigma)}\log(d_{ij})}{\sum_{i=1}^n \beta_i(d_{ij})^{\rho(1-\sigma)}}$$

$$+ \mu\frac{Y_i(1-\sigma)\sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}\log(d_{ij})Y_j^{\sigma}U_j^{1-\sigma}\right]}{\left[\sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{1-\sigma}\right]\right]^2},$$

$$\frac{\partial L}{\partial \lambda} = (1-\sigma)\log(U_j) - \log(Y_j) + W_j + \log\left[\sum_{i=1}^n \beta_i(d_{ij})^{\rho(1-\sigma)}\right],$$

$$\frac{\partial L}{\partial \mu} = \beta_i - \frac{Y_i}{\sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{1-\sigma}\right]}.$$

Proof.

Note that in order to derive the first-order derivative with respect to U for the function $-\frac{\mu Y_i}{\sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{1-\sigma}\right]}$, we first rewrite the denominator of the function to make the derivative clearer. Let us set $g(U_j) = \sum_{j=1}^n\left[(d_{ij})^{\rho(1-\sigma)}Y_j^{\sigma}U_j^{1-\sigma}\right]$. Then, the function becomes $f(U_j) = -\frac{\mu Y_i}{g(U_j)}$. The derivative of f with respect to U_j (denoted as $f'(U_j)$ or $\frac{df}{dU_j}$) can be obtained using the chain rule and the formula for the derivative of $1/g(U_j)$:

$$
\begin{aligned}
f'(U_j) &= -\mu Y_i \cdot (-1) \cdot g'(U_j) \cdot [g(U_j)]^{-2} \\
&= \frac{\mu Y_i}{[g(U_j)]^2} \cdot \frac{d}{dU_j} \left[\sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right] \right] \\
&= \frac{\mu Y_i}{\left[\sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right] \right]^2} \cdot \sum_{j=1}^{n} \left[(1-\sigma)(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{-\sigma} \right].
\end{aligned}
$$

For the function: $\lambda \cdot \left\{ (1-\sigma)\log(U_j) - \log(Y_j) + W_j + \log\left[\sum_{i=1}^{n} \beta_i (d_{ij})^{\rho(1-\sigma)}\right] \right\}$, the part that depends on ρ is $\log\left[\sum_{i=1}^{n} \beta_i (d_{ij})^{\rho(1-\sigma)}\right]$. So, we need to take the derivative of this term with respect to ρ. Applying the chain rule, we get the derivative as follows:

$$
\begin{aligned}
&\frac{d}{d\rho} \left(\lambda \cdot \left\{ (1-\sigma)\log(U_j) - \log(Y_j) + W_j + \log\left[\sum_{i=1}^{n} \beta_i (d_{ij})^{\rho(1-\sigma)} \right] \right\} \right) \\
&\quad = \lambda(1-\sigma) \sum_{i=1}^{n} \frac{\beta_i (d_{ij})^{\rho(1-\sigma)} \log(d_{ij})}{\sum_{i=1}^{n} \beta_i (d_{ij})^{\rho(1-\sigma)}}.
\end{aligned}
\tag{B.17}
$$

Similarly, for the function: $\mu \cdot \left(\beta - \frac{Y_i}{\sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right]} \right)$, the part that depends on ρ is in the denominator of the fraction, specifically, $\left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right]$. Applying the chain rule of derivatives, we get:

$$
\frac{d}{d\rho} \left[\mu \cdot \left(\beta - \frac{Y_i}{\sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right]} \right) \right] = \mu \frac{Y_i (1-\sigma) \sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} \log(d_{ij}) Y_j^{\sigma} U_j^{1-\sigma} \right]}{\left[\sum_{j=1}^{n} \left[(d_{ij})^{\rho(1-\sigma)} Y_j^{\sigma} U_j^{1-\sigma} \right] \right]^2}.
\tag{B.18}
$$

■

From the first-derivatives, it is straightforward to see if we multiply y_{ij} by a factor of $S \neq 1$, then $\widetilde{K}_i = K_i$ or $\widetilde{\beta}_i = \beta_i$, $\widetilde{U}_j = U_j$, and $\widetilde{\rho} = \rho$, but the Lagrange multipliers are scaled in the structural estimation: $\widetilde{\lambda} = S \cdot \lambda$, $\widetilde{\mu} = S \cdot \mu$.[19] We can see that W_j (a destination fixed

[19]Similar to the standard case equation (5), the parametric relationship is established through division of S on both sides of the equation (since the right-hand side in the first-order condition is zero) to take out the S factor on y_{ij}.

effect) is driving the raw residual (which is the "constant" term in our canonical example). In the case with scale S, since β's are unaffected, the scaling effect on the residual term is absorbed by $\widetilde{W}_j = W_j + \log(S)$. Note that we can also generalize $K_i = \log(\beta_i)$ and take derivatives with respect to K_i. However, the parametric relationship holds the same.

By inspecting the matrix $\boldsymbol{H}$ and the first-order conditions, it is easy to observe entries that are zeros and ones, so the Hessian matrix is updated as follows:

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 L}{\partial\beta\partial\beta^T} & \frac{\partial^2 L}{\partial\beta\partial U} & \frac{\partial^2 L}{\partial\beta\partial\rho} & \frac{\partial^2 L}{\partial\beta\partial\lambda} & 1 \\ \frac{\partial^2 L}{\partial U\partial\beta} & \frac{\partial^2 L}{\partial U\partial U^T} & \frac{\partial^2 L}{\partial U\partial\rho} & \frac{\partial^2 L}{\partial U\partial\lambda} & \frac{\partial^2 L}{\partial U\partial\mu} \\ \frac{\partial^2 L}{\partial\rho\partial\beta} & \frac{\partial^2 L}{\partial\rho\partial U} & \frac{\partial^2 L}{\partial\rho\partial\rho^T} & \frac{\partial^2 L}{\partial\rho\partial\lambda} & \frac{\partial^2 L}{\partial\rho\partial\mu} \\ \frac{\partial^2 L}{\partial\lambda\partial\beta} & \frac{\partial^2 L}{\partial\lambda\partial U} & \frac{\partial^2 L}{\partial\lambda\partial\rho} & 0 & \frac{\partial^2 L}{\partial\lambda\partial\mu} \\ 1 & \frac{\partial^2 L}{\partial\mu\partial U} & \frac{\partial^2 L}{\partial\mu\partial\rho} & \frac{\partial^2 L}{\partial\mu\partial\lambda} & 0 \end{bmatrix}. \tag{B.19}$$

The variance of the score is the diagonal elements of $\boldsymbol{H}$, which is isomorphic to the case introduced in (12) except that there are $2n + 1$ model parameters. To obtain each entry in the Hessian matrix, one can calculate the derivative of each term with respect to the model parameters. For $\partial^2 L/\partial\beta\partial\beta^T$, we have the following:

$$
\begin{aligned}
&\frac{\partial}{\partial \beta}\left\{-\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} \beta_i^{-1}\exp\left[\log(\beta_i)+W_j+(1-\sigma)\log\left(U_j\right)+\rho\log\left(\Delta_{ij}\right)\right]\right\}\\
&=-\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}}\left\{\exp\left[\log(\beta_i)+W_j+(1-\sigma)\log\left(U_j\right)+\rho\log\left(\Delta_{ij}\right)\right]\left(-\beta_i^{-2}+\beta_i^{-1}\beta_i^{-1}\right)\right\}=0,\\
&\frac{\partial}{\partial \beta}\left(\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}} y_{ij}\beta_i^{-1}\right)=\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}}\left(-y_{ij}\beta_i^{-2}\right),\\
&\frac{\partial}{\partial \beta}\left[\frac{\lambda(d_{ij})^{\rho(1-\sigma)}}{\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}}\right],\\
&=\frac{\lambda}{\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right]^2}\left[0-(d_{ij})^{\rho(1-\sigma)}(d_{ij})^{\rho(1-\sigma)}\right]\\
&=\frac{-\lambda\cdot\left[(d_{ij})^{\rho(1-\sigma)}\right]^2}{\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right]^2},\\
&\frac{\partial \mu}{\partial \beta}=0.
\end{aligned}
$$

Finally, we get:

$$
\frac{\partial^2 L}{\partial\beta\partial\beta^T}=\sum_{\substack{(i,j)\in\mathcal{S}\\ i\neq j}}\left(y_{ij}\beta_i^{-2}\right)-\frac{\lambda\cdot\left[(d_{ij})^{\rho(1-\sigma)}\right]^2}{\left[\sum_{i=1}^{n}\beta_i(d_{ij})^{\rho(1-\sigma)}\right]^2}. \tag{B.20}
$$

Since $\widetilde{\lambda}=S\cdot\lambda$, $\widetilde{\mu}=S\cdot\mu$, and $\widetilde{y}_{ij}=y_{ij}$, it is easy to see that the corresponding entry in $\boldsymbol{H}$ is scaled by S, and this result immediately holds for the diagonal entries $\partial^2/\partial U\partial U^T$ and $\partial^2/\partial\rho\partial\rho^T$ without proof. Recall that the correction matrix is based on the observed raw residuals from the standard Poisson regression (Maas and Hox, 2004)[20]:

$$
\boldsymbol{\Sigma}^{\text{Poisson}}=\boldsymbol{X}^T\,\text{diag}(\hat{\epsilon}_{12}^2,\ldots,\hat{\epsilon}_{ij}^2)\boldsymbol{X}, \tag{B.21}
$$

[20]See Goldstein (2011) for the Huber-White correction matrix for the multilevel structure, which we can apply to this gravity case.

where $\hat{\epsilon}_{ij} = y_{ij} - \exp\left[\widehat{C}_{ij} + (1-\sigma)\log\left(\widehat{U}_j\right) + \rho\log\left(\Delta_{ij}\right)\right]$, where $\widehat{C}_{ij} = \widehat{K}_i + W_i$, $\widetilde{U}_j$ is nothing but j's vector of model coefficients (since $1-\sigma$ is given in the same fashion that data is provided). Because the scaling effect is absorbed by $W_i = \widetilde{W}_j - \log(S)$, therefore:

$$\begin{aligned}
\widetilde{\hat{\epsilon}}_i &= \widetilde{y}_{ij} - \exp\left[\widetilde{\widehat{C}}_{ij} + (1-\sigma)\log\left(\widehat{U}_j\right) + \rho\log\left(\Delta_{ij}\right)\right] \\
&= S\cdot y_{ij}^2 - S\cdot\exp\left[\widehat{C}_{ij} + (1-\sigma)\log\left(\widehat{U}_j\right) + \rho\log\left(\Delta_{ij}\right)\right] \\
&= S\cdot\hat{\epsilon}_{ij}.
\end{aligned} \tag{B.22}$$

Invariance of the Robust $\boldsymbol{V}$ Matrix in MPEC

Since the corresponding entries in $\boldsymbol{H}$ are scaled by $S \neq 1$ and $[\boldsymbol{I}(\beta)]^{-1} = [-\boldsymbol{H}(\beta)]^{-1}$, while the correction matrix is scaled by S (thus $\text{diag}(\hat{\epsilon}_{12}^2, \ldots, \hat{\epsilon}_{ij}^2)$ by S^2), we have:

$$\begin{aligned}
\boldsymbol{V}^{\text{Structural}} &= \left\{\boldsymbol{V}^{\text{asym}}\boldsymbol{X}^T\,\text{diag}(\hat{\epsilon}_{12}^2, \ldots, \hat{\epsilon}_{ij}^2)\boldsymbol{X}\boldsymbol{V}^{\text{asym}}\right\} \\
&= \left\{\frac{1}{S}\left[\boldsymbol{I}(\beta)\right]^{-1}\boldsymbol{X}^T\,\text{diag}(S^2\hat{\epsilon}_{12}^2, \ldots, S^{\hat{2}}\hat{\epsilon}_{ij}^{\,2})\boldsymbol{X}\frac{1}{S}\left[\boldsymbol{I}(\beta)\right]^{-1}\right\} \\
&= \left\{\left[\widetilde{\boldsymbol{I}(\widetilde{\beta})}\right]^{-1}\boldsymbol{X}^T\,\text{diag}(S^2\hat{\epsilon}_{12}^2, \ldots, S^2\hat{\epsilon}_{ij}^2)\boldsymbol{X}\left[\widetilde{\boldsymbol{I}(\widetilde{\beta})}\right]^{-1}\right\} \\
&= \left\{\widetilde{\boldsymbol{V}}^{\text{asym}}\boldsymbol{X}^T\,\text{diag}(\widetilde{\hat{\epsilon}}_{12}^2, \ldots, \widetilde{\hat{\epsilon}}_{ij}^2)\boldsymbol{X}\widetilde{\boldsymbol{V}}^{\text{asym}}\right\} \\
&= \widetilde{\boldsymbol{V}}^{\text{Structural, adjusted}}.
\end{aligned} \tag{B.23}$$

which passes through the MPEC procedure as in equations (17) and (20). The Wald statistic built from robust standard errors is invariant to scaling.

C Constructing the Variance-Covariance Matrix in MPEC

Finally, we illustrate the detailed procedure used to construct the variance matrix in the MPEC procedure.

Conceptually, the procedure for obtaining robust standard errors in structural models is similar to that used in standard Poisson regressions, where the raw residuals are obtained after using the Stata command `ppmlhdfe` or `poisson depvar indvars, vce(robust)` to

estimate the model first, except that these commands will not give us the cardinal values for $\boldsymbol{U}$ that is required in order to estimate ρ. The MPEC procedure in Balistreri and Hillberry (2007) allows us to estimate ρ and more complex models which reduced-form regressions are not feasible. However, if one wishes to understand whether including ρ fits the data better, we show that the LR tests produced from the structural procedure is biased when one scale the independent variables (or both independent and dependent variables) of the model.

Furthermore, our PPML procedure replicates Balistreri and Hillberry (2007)'s results and obtains that $\rho = 0.36$ when we fix $\sigma = 5$. The raw residuals in equation (B.22) are immediately computable when $\boldsymbol{U}$, ρ and $\boldsymbol{\beta}$ are estimated from the structural model.

The transformed observable distance matrix, $\boldsymbol{\Delta}$, is defined as:

$$\boldsymbol{\Delta} = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \dots & \Delta_{1n} \\ \Delta_{21} & \Delta_{22} & \dots & \Delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{n1} & \Delta_{n2} & \dots & \Delta_{nn} \end{bmatrix}, \tag{C.1}$$

The matrix $C_{ij} = \log(\beta_i) + W_j$ is defined as:

$$C = \begin{bmatrix} \log(\beta_1) + W_1 & \log(\beta_2) + W_1 & \dots & \log(\beta_n) + W_1 \\ \log(\beta_1) + W_2 & \log(\beta_2) + W_2 & \dots & \log(\beta_n) + W_2 \\ \vdots & \vdots & \ddots & \vdots \\ \log(\beta_1) + W_n & \log(\beta_2) + W_n & \dots & \log(\beta_n) + W_n \end{bmatrix}, \tag{C.2}$$

which serves as the entries for "constant" terms and can be obtained after β's are estimated.

The vector U is then given by:

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}, \tag{C.3}$$

which is obtained after the estimation.

The model matrix X is defined as:

$$X = \begin{bmatrix} C & \Delta & U \end{bmatrix} = \begin{bmatrix} \log(\beta_1) + W_1 & \dots & \log(\beta_n) + W_n & \Delta_{11} & \dots & \Delta_{1n} & U_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \log(\beta_1) + W_n & \dots & \log(\beta_n) + W_n & \Delta_{n1} & \dots & \Delta_{nn} & U_n \end{bmatrix}. \tag{C.4}$$

The transpose of the model matrix X, denoted by X^T, is:

$$X^T = \begin{bmatrix} C^T \\ \Delta^T \\ U^T \end{bmatrix} = \begin{bmatrix} \log(\beta_1) + W_1 & \dots & \log(\beta_1) + W_n \\ \vdots & \ddots & \vdots \\ \log(\beta_n) + W_1 & \dots & \log(\beta_n) + W_n \\ \Delta_{11} & \dots & \Delta_{n1} \\ \vdots & \ddots & \vdots \\ \Delta_{1n} & \dots & \Delta_{nn} \\ U_1 & \dots & U_n \end{bmatrix}. \tag{C.5}$$

Finally, the inverse of the negative of matrix $\boldsymbol{H}$, denoted as $(-\boldsymbol{H})^{-1}$, is given by:

$$(-\boldsymbol{H})^{-1} = \begin{bmatrix} -\frac{\partial^2 L}{\partial\beta\partial\beta^T} & -\frac{\partial^2 L}{\partial\beta\partial U} & -\frac{\partial^2 L}{\partial\beta\partial\rho} & -\frac{\partial^2 L}{\partial\beta\partial\lambda} & -\frac{\partial^2 L}{\partial\beta\partial\mu} \\ -\frac{\partial^2 L}{\partial U\partial\beta} & -\frac{\partial^2 L}{\partial U\partial U^T} & -\frac{\partial^2 L}{\partial U\partial\rho} & -\frac{\partial^2 L}{\partial U\partial\lambda} & -\frac{\partial^2 L}{\partial U\partial\mu} \\ -\frac{\partial^2 L}{\partial\rho\partial\beta} & -\frac{\partial^2 L}{\partial\rho\partial U} & -\frac{\partial^2 L}{\partial\rho\partial\rho^T} & -\frac{\partial^2 L}{\partial\rho\partial\lambda} & -\frac{\partial^2 L}{\partial\rho\partial\mu} \\ -\frac{\partial^2 L}{\partial\lambda\partial\beta} & -\frac{\partial^2 L}{\partial\lambda\partial U} & -\frac{\partial^2 L}{\partial\lambda\partial\rho} & -\frac{\partial^2 L}{\partial\lambda^2} & -\frac{\partial^2 L}{\partial\lambda\partial\mu} \\ -\frac{\partial^2 L}{\partial\mu\partial\beta} & -\frac{\partial^2 L}{\partial\mu\partial U} & -\frac{\partial^2 L}{\partial\mu\partial\rho} & -\frac{\partial^2 L}{\partial\mu\partial\lambda} & -\frac{\partial^2 L}{\partial\mu^2} \end{bmatrix}^{-1}. \tag{C.6}$$

This completes the structural procedure with scaling choices in the MPEC setting.