Replicating The Log of Gravity

Mauricio Vargas Sepúlveda (ORCID 0000-0003-1017-7574)
Department of Political Science, University of Toronto
Munk School of Global Affairs and Public Policy, University of Toronto

Corresponding author: m.sepulveda@mail.utoronto.ca

Last updated: 2024-09-05 13:22

Contents

1 Abstract

This document replicates the main results from Santos Silva and Tenreyro (2006) in R. The original results were obtained in TSP back in 2006. The idea here is to be explicit regarding the conceptual approach to regression in R. For most of the replication I used base R without external libraries except when it was absolutely necessary. The findings are consistent with the original article and reveal that the replication effort is minimal, without the need to contact the authors for clarifications or incur into data transformations or filtering not mentioned in the article.

2 Introduction

Santos Silva and Tenreyro (2006) used TSP back in 2006 to estimate different functional forms of the gravity model of trade. The article is a classic in the field of international trade and has been cited over 8,500 times according to Google Scholar. Its contribution, which was a new Poisson-based estimator to correct the bias in the Ordinary Least Squares (OLS) estimator under heteroskedasticity, has been widely used in the academic literature. Besides pure academic interests, the OLS estimator bias has implications for policy analysis, and posterior works with clear insights for International Relations academics and Public Policy practitioners, including Head and Mayer (2014), Yotov et al. (2016), and Felbermayr et al. (2020).

Outside of statistical concerns about biases, the field of International Relations has developed interesting critiques about leaving theory behind. Mearsheimer and Walt (2013) argues that the field has been dominated by quantitative studies that present simplistic hypothesis after incentives to present academic findings as more interesting for policymakers. From this critique, Santos Silva and Tenreyro (2006) is relevant not only for its methodological contribution and presenting an estimator that can largely differ from OLS, but also for presenting an estimator that is consistent with micro-founded approaches to the gravity model of trade, including Eaton and Kortum (2001).

The goal in this document is to be explicit regarding the conceptual approach to regression in

R. For most of the replication we used base R (R Core Team 2021) and relying on external libraries (e.g., R packages) when it was absolutely necessary based on the criteria that implementing methods such as the Tobit estimator from scratch would be time consuming and would deviate from the replication effort. The results are consistent with the gravity package (Woelwer et al. 2020) which provides convenient wrappers for gravity estimation, but the idea was to be explicit about the steps involved in the replication process and using wrappers would have made the steps less clear.

3 Original codes and data

We organized the original codes and data from the authors' site, The Log of Gravity, on GitHub and therefore ease replication in the event that the links change in the future.

We obtained the original data and code with the following R code:

```
url <- "https://personal.lse.ac.uk/tenreyro/regressors.zip"
zip <- gsub(".*/", "", url)
if (!file.exists(zip)) try(download.file(url, zip))

dout <- "regressors"
if (!dir.exists(dout)) unzip(zip, exdir = dout)
```

The original dataset obtained in the previous step is available in Stata format, a closed source format that we can read in R thanks to the haven package (Wickham and Miller 2021) without loss of information or other common problems when reading proprietary formats.

We proceeded to import R packages to ease our work and read the data with the following R code:

```
library(haven) # read Stata datasets
library(censReg) # Tobit estimation
library(stargazer) # table of results

log_of_gravity <- read_dta(paste0(dout, "/Log of Gravity.dta"))
```

4 Models replication

4.1 Poisson Pseudo Maximum Likelihood

Table 3 in Santos Silva and Tenreyro (2006) summarises a large portion of the article. Starting from the Poisson Pseudo Maximum Likelihood (PPML) estimation, we can replicate the results with the Generalized Linear Model (GLM) and a quasi-Poisson link in R. The original article estimates the model with and without zero flows.

We could replicate the PPML model results with and without zero flows with the following R code:

```
ppml_formula <- trade ~ lypex + lypim + lyex + lyim + ldist + border +
  comlang + colony + landl_ex + landl_im + lremot_ex + lremot_im +
  comfrt_wto + open_wto

fit_ppml_1 <- glm(
  ppml_formula,
  data = log_of_gravity,
  subset = trade > 0,
  family = quasipoisson()
)

fit_ppml_2 <- glm(
  ppml_formula,
  data = log_of_gravity,
  family = quasipoisson()
)
```

The replication effort for the PPML model was minimal, it was sufficient to look at the summary table in the article and subset the data to drop zero flows. This step did not require to guess data cleaning steps or transformations not mentioned in the article.

4.2 Ordinary Least Squares

The only consideration for the OLS model was to drop zero flows for some of the models with log in the dependent variable even when Table 3 is not explicit about this, otherwise we break the fitting algorithm (e.g. because $\log(0) \to -\infty$).

For estimations of the form $\log(\text{trade}) = \beta_0 + \beta_1 \text{lypex} + \cdots + \varepsilon$, we needed to drop zero flows to replicate the result. On the other hand, for estimations of the type $\log(1 + \text{trade}) = \beta_0 + \beta_1 \text{lypex} + \cdots + \varepsilon$, we did not need to drop zero flows.

```
fit_ols_1 <- lm(
  update.formula(ppml_formula, log(.) ~ .),
  data = log_of_gravity,
  subset = trade > 0
)

fit_ols_2 <- lm(
  update.formula(ppml_formula, log(1 + .) ~ .),
  data = log_of_gravity
)
```

4.3 Tobit

The Tobit estimation required the use of the censReg package (Henningsen 2020). Unlike PPML and OLS models, this required us to extract the right hand side of the model formula to define a vector of zeroes with a length equal to the independent variables plus two as starting point for the Maximum Likelihood estimation (e.g., counting the depending variable and intercept besides the estimating slopes).

In order to obtain the a parameter that matches the results in the article we proceeded with an iteration loop until achieving convergence with respect to one of the estimated slopes. The initial value of $a = 200$ was arbitrary and set after trying reasonable guesses that converge to the slopes in the original article after 9 iterations for a final value of $a = 159$.

```
a <- 200
lypex_ref <- 1.058
tol <- 0.001
lypex_estimate <- 2 * lypex_ref
iter <- 0

while (abs(lypex_estimate - lypex_ref) > tol) {
  log_of_gravity$log_trade_cens <- log(a + log_of_gravity$trade)
  log_trade_cens_min <- min(log_of_gravity$log_trade_cens, na.rm = TRUE)
```

```
  fit_tobit <- censReg(
    formula = update.formula(ppml_formula, log_trade_cens ~ .),
    left = log_trade_cens_min,
    right = Inf,
    data = log_of_gravity,
    start = rep(0, 2 + length(attr(terms(ppml_formula), "term.labels"))),
    method = "BHHH"
  )

  lypex_estimate <- coef(fit_tobit)[2]
  if (abs(lypex_estimate - lypex_ref) > 2 * tol) {
    a <- a - 5
  } else {
    a <- a - 1
  }
  iter <- iter + 1
}
```

4.4 Non-Linear Least Squares

The starting values were retrieved from the PPML model results with zero flows and then passed to the GLM function with a Gaussian log-link in R.

```
fit_ppml_eta <- fit_ppml_2$linear.predictors
fit_ppml_mu <- fit_ppml_2$fitted.values
fit_ppml_start <- fit_ppml_2$coefficients

fit_nls <- glm(
  ppml_formula,
  data = log_of_gravity,
  family = gaussian(link = "log"),
  etastart = fit_ppml_eta,
  mustart = fit_ppml_mu,
  start = fit_ppml_start,
  control = list(maxit = 200, trace = FALSE)
)
```

5 Replication results

Santos Silva and Tenreyro (2006) is very close to full replication according to the criteria defined in Peng (2011). The replication results are consistent with the original article and reveal equivalent figures for all the models estimated in the original article.

The results are presented in the following table:

```
stargazer(
  fit_ols_1, fit_ols_2, fit_tobit, fit_nls, fit_ppml_1, fit_ppml_2,
  header = FALSE, font.size = "footnotesize", model.names = F,
  omit.table.layout = "d", omit.stat = c(
    "f", "ser", "ll", "aic", "bic", "rsq", "adj.rsq"
  ),
  title = "Replication results for OLS (1-2), Tobit (3), NLS (4) and
  PPML (5-6)."
)
```

Table 1: Replication results for OLS (1-2), Tobit (3), NLS (4) and PPML (5-6).

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
lypex	0.938***	1.128***	1.059***	0.738***	0.721***	0.732***
	(0.012)	(0.011)	(0.011)	(0.004)	(0.008)	(0.006)
lypim	0.798***	0.866***	0.848***	0.862***	0.732***	0.741***
	(0.011)	(0.011)	(0.010)	(0.005)	(0.008)	(0.006)
lyex	0.207***	0.277***	0.228***	0.396***	0.154***	0.157***
	(0.017)	(0.017)	(0.014)	(0.010)	(0.013)	(0.010)
lyim	0.106***	0.217***	0.178***	−0.033***	0.133***	0.135***
	(0.017)	(0.017)	(0.014)	(0.007)	(0.013)	(0.010)
ldist	−1.166***	−1.151***	−1.160***	−0.924***	−0.776***	−0.784***
	(0.034)	(0.037)	(0.029)	(0.008)	(0.018)	(0.013)
border	0.314**	−0.241	−0.225**	−0.081***	0.202***	0.193***
	(0.143)	(0.164)	(0.109)	(0.010)	(0.034)	(0.026)
comlang	0.678***	0.742***	0.759***	0.689***	0.751***	0.746***
	(0.064)	(0.064)	(0.052)	(0.016)	(0.037)	(0.028)
colony	0.397***	0.392***	0.416***	0.036**	0.020	0.025
	(0.068)	(0.068)	(0.056)	(0.018)	(0.043)	(0.032)
landl_ex	−0.062	0.106*	−0.038	−1.367***	−0.872***	−0.863***
	(0.065)	(0.060)	(0.060)	(0.031)	(0.057)	(0.043)
landl_im	−0.665***	−0.278***	−0.478***	−0.471***	−0.703***	−0.696***
	(0.063)	(0.060)	(0.059)	(0.022)	(0.054)	(0.040)
lremot_ex	0.467***	0.526***	0.563***	1.188***	0.647***	0.660***
	(0.078)	(0.089)	(0.077)	(0.018)	(0.048)	(0.036)
lremot_im	−0.205**	−0.109	−0.032	1.010***	0.549***	0.562***
	(0.081)	(0.089)	(0.074)	(0.018)	(0.048)	(0.036)
comfrt_wto	0.491***	1.289***	0.728***	0.443***	0.179***	0.181***
	(0.105)	(0.143)	(0.113)	(0.014)	(0.036)	(0.027)
open_wto	−0.170***	0.739***	0.310***	0.928***	−0.139***	−0.107***
	(0.049)	(0.048)	(0.040)	(0.024)	(0.039)	(0.029)
logSigma			0.677***			
			(0.007)			
Constant	−28.492***	−39.909***	−36.626***	−45.098***	−31.530***	−32.326***
	(1.088)	(1.221)	(1.059)	(0.239)	(0.596)	(0.444)
Observations	9,613	18,360	18,360	18,360	9,613	18,360

Note: *p<0.1; **p<0.05; ***p<0.01

6 Conclusion

The replication effort was minimal, without the need to contact the authors for clarifications or incur into data transformations or filtering not mentioned in the article. The results are consistent with the original article and reveal proper and transparent scholarship. Ideally, all disciplines that rely on quantitative research and the use of statistical methods to test hypotheses should follow similar or higher standards of transparency, but that is not always the case as initiatives such as Simonsohn, Nelson, and Simmons (2024) reveal.

7 Acknowledgements

Thanks to Joao Santos Silva for pointing that out, in a previous version of this document I mentioned that the original results were obtained with Stata.

References

Eaton, Jonathan, and Samuel Kortum. 2001. "Technology, Trade, and Growth: A Unified Framework." *European Economic Review* 45 (4): 742–55. https://doi.org/10.1016/S0014-2921(01)00129-5.

Felbermayr, Gabriel, Aleksandra Kirilakha, Constantinos Syropoulos, Erdal Yalcin, and Yoto V. Yotov. 2020. "The Global Sanctions Data Base." *European Economic Review* 129 (October): 103561. https://doi.org/10.1016/j.euroecorev.2020.103561.

Head, Keith, and Thierry Mayer. 2014. "Chapter 3 - Gravity Equations: Workhorse,Toolkit, and Cookbook." In *Handbook of International Economics*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 4:131–95. Handbook of International Economics. Elsevier. https://doi.org/10.1016/B978-0-444-54314-1.00003-3.

Henningsen, Arne. 2020. *censReg: Censored Regression (Tobit) Models*. https://CRAN.R-project.org/package=censReg.

Mearsheimer, John J., and Stephen M. Walt. 2013. "Leaving Theory Behind: Why Simplistic Hypothesis Testing Is Bad for International Relations." *European Journal of In-*

ternational Relations 19 (3): 427–57. https://doi.org/10.1177/1354066113494320.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–27.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Santos Silva, Joao, and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics* 88 (4): 641–58.

Simonsohn, Uri, Leif Nelson, and Joe Simmons. 2024. "Data Colada." *Data Colada*. https://datacolada.org/.

Wickham, Hadley, and Evan Miller. 2021. *haven: Import and Export SPSS, Stata and SAS Files*. https://CRAN.R-project.org/package=haven.

Woelwer, Anna-Lena, Jan Pablo Burgard, Joshua Kunst, and Mauricio Vargas. 2020. *Gravity: Estimation Methods for Gravity Models*. http://pacha.dev/gravity.

Yotov, Yoto V., Roberta Piermartini, JosÃ©-Antonio Monteiro, and Mario Larch. 2016. *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*. WTO iLibrary. https://doi.org/10.30875/abc0167e-en.