Dealing with Logs and Zeros in Regression Models

Christophe Bellégo*, David Benatia†, and Louis Pape*

*CREST (UMR 9194), CNRS, École Polytechnique, Institut Polytechnique de Paris, 5 Avenue Henry Le Chatelier, 91120 Palaiseau, France

† HEC Montréal, Département d'Économie Appliquée, 3000 Chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7, Canada

Corresponding author: david.benatia@hec.ca

1. Introduction

- **Log-linear** regressions are **very popular** within and outside of economics.
- Yet, **how to handle zeros in the dependent variable remains obscure**.
- Practitioners often **rely on ad hoc transformations** (e.g, using $\log(Y_i + \Delta)$ for some $\Delta > 0$).

2. Contribution

- We develop a **new model**: **includes both Log-linear and Poisson** Regression as special cases.
- Keeps interpretation of β as **semi-elasticity** and reconciles $\log(Y_i + \Delta)$ with **econometric theory**.
- We **propose a new statistical test** to select the best model given pattern of zeros in the data.

3. Proposed Model: Iterated Ordinary Least Squares (iOLS$_\delta$)

Consider an iid sample of observations $\{Y_i, X_i\}_{i=1}^n$, generated by the "true" model

$$Y_i = \exp(X_i'\beta)U_i, \qquad (1)$$

Like practitioners, we add an **individual-specific** $\Delta_i = \delta \exp(X_i'\beta)$, for some $\delta > 0$, and obtain

$$\log\left(Y_i + \delta \exp(X_i'\beta)\right) = X_i'\beta + \upsilon_i. \qquad (2)$$

with new error term $\upsilon_i = \log(\delta + U_i)$. This **looks like the linear model: is it also estimable by OLS?**

4. Properties

Under the moment condition $E[\upsilon_i|X_i] = \log(1+\delta)$, we show that

1. $\hat{\beta}$ is the **unique fixed-point** of an Asymptotic Contraction Mapping [1]:
   - $\hat{\beta}$ estimable by **running OLS iteratively**, **consistent**, and **asymptotically normal**.
   - **Easy to estimate**: takes **high-dimensional** covariates & **standard errors** from **last-step** OLS.
2. δ provides generality and flexibility: **ends arbitrary choice of moment condition**:
   - Log-linear regression assumes $E(\log(U_i)|X_i) = 0$ when Poisson regression uses $E(U_i|X_i) = 1$.
   - **iOLS$_\delta$ nests both**: as $\delta \to 0$, $E(\upsilon_i|X_i) \to E(\log(U_i)|X_i)$ and as $\delta \to \infty$, $E(\upsilon_i|X_i) \to E(U_i|X_i)$.

Endogeneity (i2LS$_\delta$): with instruments Z_i and $E[\upsilon_i|Z_i] = \log(1+\delta)$, **run 2SLS iteratively**.

5. Statistical test to select correct model and best δ given data

Test exploits the **implicit assumption** placed on the **pattern of zeros by moment** conditions.

- **Example with Poisson:** decomposed into $E(U_i|X_i) = Pr(U_i > 0|X_i) \times E(U_i|X_i, U_i > 0) = 1$.

Simple Testing Procedure:

1. **Estimate candidate** model (iOLS$_\delta$, i2SLS$_\delta$, Poisson, etc.)
2. **Estimate model of a non-zero dependent** variable (with logit or non-parametric model)
3. **Regress** scaled first-step residuals on inverse probability of second-step.
4. **Reject if** regression coefficient λ **far from 1**.

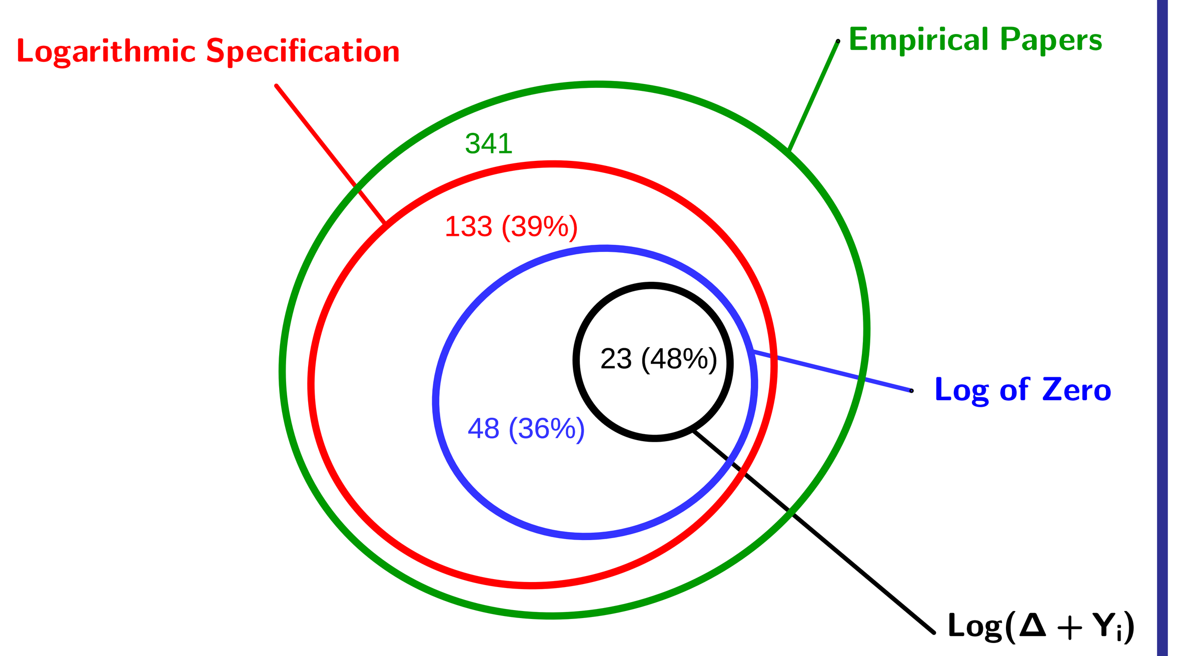
6. Empirical Application: Gravity Equations

Compare iOLS$_\delta$ to Poisson Regression for the purpose of estimating the Gravity Equation of International Trade (136 countries, 1990) of [2]: **Poisson rejected in favor of iOLS$_\delta$** $= 100$.

	$\log(Y_i + \Delta)$	iOLS$_{\delta=100}$	Poisson
Logit Model of Non-Zero Dependent Variable			
$\hat{\lambda}$	0.04	0.46	1.26
(s.e)	(0.02)	(0.06)	(0.39)
t-Stat.	[-53.45]	[-9.82]	[0.68]
kNN Model of Non-Zero Dependent Variable			
$\hat{\lambda}$	0.27	**0.93**	**1.72**
(s.e)	(0.01)	(0.05)	(0.25)
t-Stat.	[-49.85]	**[-1.55]**	[2.86]

This table displays the $\hat{\lambda}$-parameter, standard errors (s.e), and t-statistics (t-Stat.) using 300 pairs bootstrap for the test of three models of international trade, with parametric (logit) and non-parametric (kNN) models of non-zero trade.

A. The Log of Zero in the AER

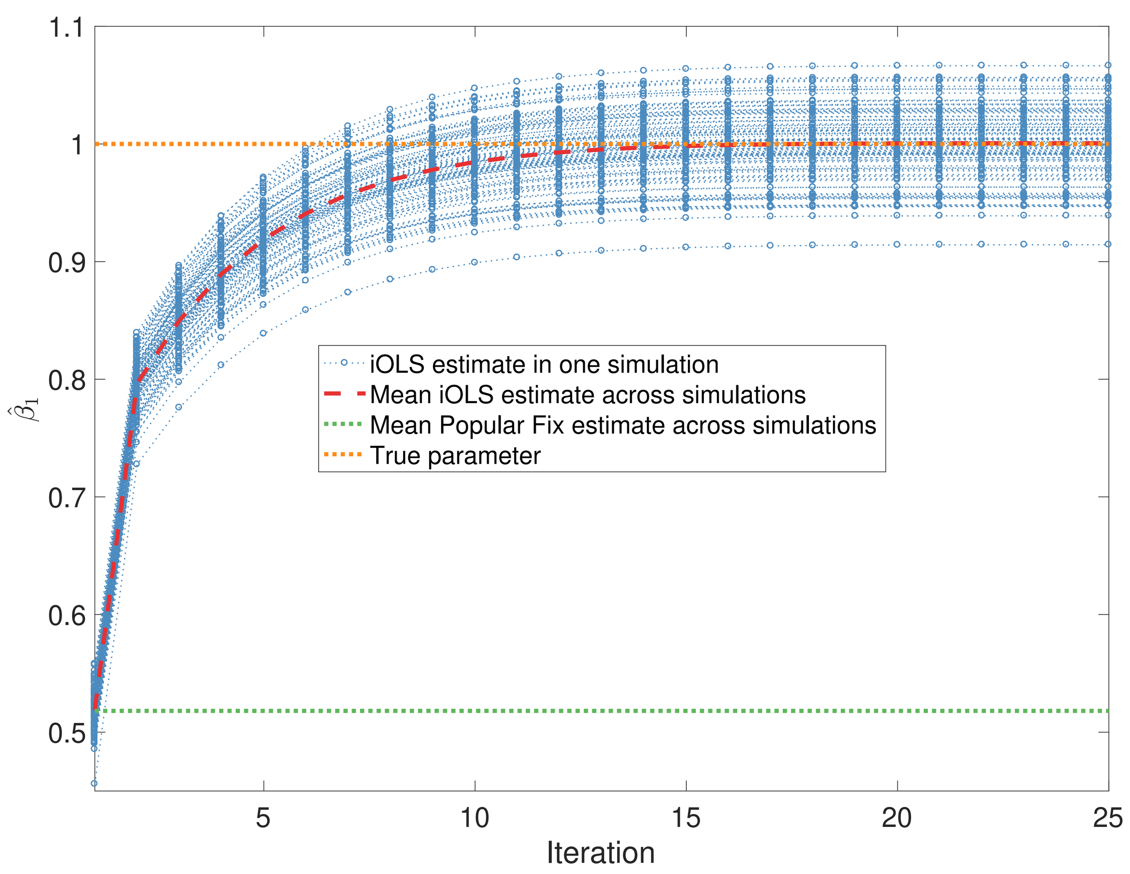


Share of Publications in the AER (2016-2020)

- **The issue is widespread:** 14% of empirical publications in the AER face this issue.
- **48% of the time, authors use** $\log(Y_i + \Delta)$.

B. Computational Procedure

From any starting point, as $n \to \infty$,
OLS run iterativly converges to true value β.



Example of Convergence with iOLS$_{\delta=1}$ (Simulations with n=1,000)

C. Extensions & Software

- iOLS$_U$: estimate Poisson by iterative OLS,
- Deal with zeros in **log-log regression**.
- **Within-Transformation** for estimation with **high-dimensional fixed effects**.
- **Stata programs** for iOLS$_\delta$, i2SLS$_\delta$ and tests.
- Monte Carlo simulations.

7. Conclusion

1. **No single method** is always correct.
2. Need to **compare models** through testing.
3. iOLS$_\delta$ is a **good starting** point: flexible and computationally simple.

8. References

[1] Jeff Dominitz and Robert P. Sherman. Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory*, 21(4):838–863, 2005.

[2] J. M. C. Santos Silva and Silvana Tenreyro. The Log of Gravity. *The Review of Economics and Statistics*, 88(4):641–658, 11 2006.