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 + Δ) for some Δ > 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 + Δ) 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 (iOLSδ)

Consider an iid sample of observations \( \{Y_i, X_i\}_{i=1}^n \), generated by the “true” model

\[
Y_i = \exp(X_i' \beta) U_i, \tag{1}
\]

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

\[
\log (Y_i + \delta \exp(X_i' \beta)) = X_i' \beta + \upsilon_i. \tag{2}
\]

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 \( E[\upsilon_i|X_i] = (1 + \delta) \), 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. \( \delta \) 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 \( E(U_i|X_i) = 1 \).
   - iOLSδ nests both: as \( \delta \to 0 \), \( E(\upsilon_i|X_i) \to E(\upsilon_i|X_i) \) and as \( \delta \to \infty \), \( E(\upsilon_i|X_i) \to E(U_i|X_i) \).

Endogeneity (i2LSδ): 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 \( \delta \) 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 (iOLSδ, i2LSδ, Poisson, etc.)
2. Estimate model of a non-zero dependent variable (with logit or non-parametric model)
4. Reject if regression coefficient \( \lambda \) far from 1.


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

\[
\begin{array}{ccc}
                  & \text{Logit Model of Non-Zero Dependent Variable} & \text{iOLSδ}_{0.100} & \text{Poisson} \\
\lambda & 0.04 & 0.46 & 1.26 \\
(s.e.) & (0.02) & (0.06) & (0.39) \\
t-Stat. & [-53.45] & [-9.82] & [0.68] \\
\end{array}
\]

\[
\begin{array}{ccc}
                  & \text{kNN Model of Non-Zero Dependent Variable} & \\
\lambda & 0.27 & 0.93 & 1.72 \\
(s.e.) & (0.01) & (0.05) & (0.25) \\
\end{array}
\]

This table displays the \( \lambda \)-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.

7. Conclusion

1. No single method is always correct.
2. Need to compare models through testing.
3. iOLSδ is a good starting point: flexible and computationally simple.

8. References