CREST
CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

WORKING PAPER SERIES

Dealing with Logs and Zeros in Regression Models

Christophe Bellego, David Benatia, Louis-Daniel Pape

N°08/April 2022

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

Dealing with Logs and Zeros in Regression Models

Christophe Bellégo* David Benatia† Louis-Daniel Pape‡

March 22, 2022

Abstract

Log-linear models are prevalent in empirical research. Yet, how to handle zeros in the dependent variable remains an unsettled issue. This article clarifies it and addresses the "log of zero" by developing a new family of estimators called iterated Ordinary Least Squares (iOLS). This family nests standard approaches such as log-linear and Poisson regressions, offers several computational advantages, and corresponds to the correct way to perform the popular $\log(Y+1)$ transformation. We extend it to the endogenous regressor setting (i2SLS) and overcome other common issues with Poisson models, such as controlling for many fixed-effects. We also develop specification tests to help researchers select between alternative estimators. Finally, our methods are illustrated through numerical simulations and replications of landmark publications.

Keywords: Contraction mapping, Elasticity, Gravity model, Iterative estimator, Log-linear, Selection bias.

JEL: C26, C52, C55.

*CREST (UMR 9194), ENSAE, Institut Polytechnique de Paris, 5 Avenue Henry Le Chatelier, 91120 Palaiseau, France (e-mail: christophe.bellego@ensae.fr).

†HEC Montréal, Département d'Économie Appliquée, 3000 Chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7, Canada (Corresponding author, e-mail: david.benatia@hec.ca).

‡CREST (UMR 9194), École Polytechnique, Institut Polytechnique de Paris, 5 Avenue Henry Le Chatelier, 91120 Palaiseau, France (e-mail: louis.pape@ensae.fr).

The authors thank Ghazala Azmat, Michael Bognanno, Marine Carrasco, Sergio Correia, Laurent Davezies, Xavier D'Haultfœuille, Philippe Gagnepain, Ben Groom, James McKinnon, Mathieu Marcoux, Ioana Marinescu, Arnaud Maurel, Isabelle Méjean, Joris Pinkse, João Santos Silva, Alain Trognon, Michael Visser, Martin Weidner, Jeffrey Wooldridge, Tom Zylkin, and all participants to seminars and conferences for insightful comments. STATA packages are available at https://github.com/ldpape/iOLS_delta.

1 Introduction

Having to deal with the (natural) logarithm of a zero in the dependent variable is a recurring problem. There is, unfortunately, a lack of consensus regarding the best way to address those zeros in log-linear and log-log models, as evidenced by the many alternative approaches used in recent leading publications. Besides, these existing approaches may not always be suited to the data and may suffer from a range of possible issues.[1]

This paper not only discusses the existing methods and their limitations for addressing the "log of zero", it also develops a new family of estimators and a model selection procedure which overcome these limitations. Our estimators are simple iterative extensions of ordinary least squares (OLS) and two-stage least-squares (2SLS). They are consistent, asymptotically normal, computationally simple, and can accommodate many fixed-effects along with instrumental variables. Our approach is general in the sense that it nests a continuum of models, including the log-linear model and Poisson regression as special cases, and can be interpreted as the correct way to perform the $\log(Y_i + 1)$ transformation. We then develop a specification test aimed at selecting the most suitable approach to address the log of zero. This test consists in verifying the external validity of each model with respect to the observed pattern of zeros in the data. We believe this new family of estimators, combined with the specification tests, provides a natural framework for handling zeros in the dependent variable.

To document the ubiquity of logs and zeros in regression models, and the lack of consensus thereof, we have collected information from three sources: a review of empirical publications, a survey conducted during seminars, and the records from an online research forum. First, we have reviewed all articles published in the American Economic Review (AER) between 2016 and 2020. Figure 1 summarizes our findings. We find that log-linear and log-log models are among the most frequent specifications used in empirical research. Nearly 40% of empirical papers used a log-specification and 36% of these faced the problem of the log of zero. Several solutions are employed in practice. In most publications, the authors chose to keep the zero observations

[1] In this paper, we focus on the log-linear model and address the minor differences of the log-log model as an extension in Appendix B.4.

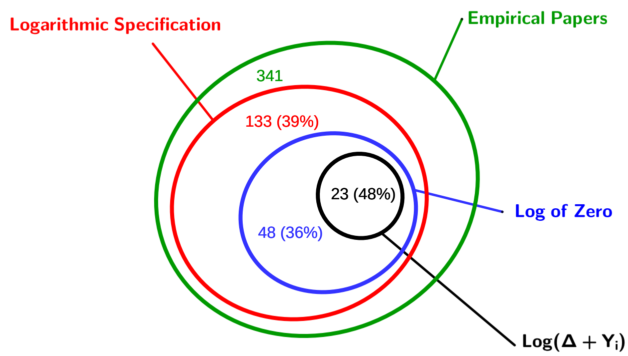


Figure 1: Prevalence of the Log of Zero in the AER (2016-2020)

and opted to either (1) add a positive discretionary value to the dependent variable (48%), (2) use Poisson-type estimators (35%), or (3) apply the inverse hyperbolic sine (IHS) transformation (15%). Discarding non-positive observations occurred in 31% of publications. We also note that in around 20% of cases, the authors compared several methods in order to gauge the robustness of their results.[2] Second, we have also conducted a survey in three online seminars in economic departments asking researchers "What would you do when facing the log of zero?". The distribution of responses mirrors that observed in the AER publications except for the large stated preference for the use of mixture models.[3] Third, the question "Log transformation of values that include 0 (zero) for statistical analyses?" asked in 2014 on ResearchGate, a multidisciplinary research-oriented social network, has been read 120,000 times by August 2020 and has received 38 contributions from researchers in medicine, biology, statistics, engineering and other fields. Again, the prevalence of each suggested solution is comparable to that in the AER.

This evidence suggests three main issues which must be overcome. First, the problem of the log of zero is widespread in empirical research and extends beyond the realm of economics. This paper attempts to solve this problem with a new flexible and computationally-efficient estimator. Second, the prevalence of simple *ad hoc* fixes over theoretically-founded methods suggests the existence of numerical difficulties in some contexts. Our discussions with empirical researchers indeed revealed that many

[2]This excludes cases where the authors decided to use a linear specification by fault of having to use such a fix. See Table D.2 in the Appendix for additional details and information regarding data collection.

[3]Appendix D provides more details about the surveys.

opt for the popular fix approach (i.e. $\log(Y+1)$) for its simplicity and convenience, at the cost of what they assume is a small bias. Our approach, unlike Poisson models, do not suffer from the convergence issues identified in Santos Silva and Tenreyro (2011), or difficulties in handling instrumental variables or many fixed-effects. Third, the choice of one method over another is almost never justified in publications. In contrast, this paper focuses on the identification of model parameters and provides a way to substantiate the choice of a particular model. For many proposed solutions to the log of zero, like Poisson regression, identification rests on assuming some particular moment conditions. We show that these conditions correspond to implicit assumptions about the conditional probability of having a zero in the data. Based on this observation, we develop a statistical test for evaluating which model is consistent with the observed pattern of zeros. Our methodological contribution hence both extends the set of potential solutions while simultaneously limiting the researcher's discretionary power.

To handle these issues, we develop a new class of estimators, called iterated Ordinary Least Squares (iOLS) designed to address the "log of zero". Our approach consists in adding an observation-specific value to the outcome, instead of a constant, which is scaled using a hyper-parameter. This parameter controls the underlying moment condition used for estimation and can be user-selected or data-driven using our model selection approach. The range of admissible moment conditions form a continuum which limits correspond to the restrictions used in the log-linear model and the Poisson model, respectively. The model selection procedure amounts to finding the hyper-parameter value such that the conditional probability of having a zero observation implied by the model is consistent with the data.

We study the theoretical properties of iOLS (and i2SLS), including consistency and asymptotic normality, using the asymptotic theory developed in Dominitz and Sherman (2005). Our estimator corresponds to the fixed point of an (asymptotic) contraction mapping, which is solved for by running OLS (or 2SLS) iteratively.[4] To fix ideas, let us illustrate our estimation method in its simplest form. Assume Y and X are the dependent and independent variables, β is the parameter of interest, and $\delta > 0$ is the hyper-parameter. The simplest iOLS procedure consists of the following

[4]Our model could also be estimated by non-linear least squares at greater computational costs.

steps: 1. Initialize $t = 0$ and obtain an initial estimate $\hat{\beta}_0$, e.g. by regressing $\log(Y+1)$ onto X using OLS; 2. Regress $\log\left(\frac{Y+\delta\exp(X'\hat{\beta}_t)}{1+\delta}\right)$ onto X using OLS to obtain $\hat{\beta}_{t+1}$; 3. Update t to $t+1$; 4. Repeat steps 2 and 3 until $\hat{\beta}_{t+1}$ converges.[5] Remark that this transformed dependent variable corresponds to the log of a weighted average of Y and $\exp(X'\beta)$, where weights depend on the hyper-parameter δ.

This estimator has multiple advantages: (a) it can be estimated by ordinary least squares, hence is computationally fast and easy to implement, with possibly high dimensional fixed effects; (b) it extends naturally to the endogenous setting using iterated 2SLS (i2SLS), also in situations with many fixed effects; (c) it is amenable to different identifying assumptions by varying the hyper-parameter; (d) robust standard errors are readily available; and (e) it does not suffer from highly dispersed response variables or the numerical convergence issues identified for (additive) Poisson estimators.

Our methodological contributions are illustrated through numerical simulations and (partial) replications of Michalopoulos and Papaioannou (2013) and Santos Silva and Tenreyro (2006). The former examine the role of pre-colonial ethnic institutions on economic development by using the popular fix to address the log of zero, whereas the latter has popularized the use of Pseudo-Poisson Maximum Likelihood (PPML) for estimating gravity models in trade. Our approach yields plausible and justifiable estimates in both replications whereas PPML estimates are found to be externally inconsistent with the observed patterns of zeros. .

The remaining of the paper is organized as follows. Section 2 clarifies the log of zero issue and discusses existing practices as well as their limits. Section 3 develops a new family of solutions. Section 4 presents specification tests and a data-driven model selection procedure. Numerical simulations are presented in Section 5. The replication exercise is proposed in Section 6. Section 7 concludes the paper.[6]

[5] Our approach does not use Iterated Reweighted Least-Squares, unlike the logit and Poisson models estimated as generalized linear models, but it can be extended to include a weighting matrix like in Generalized Least Squares.

[6] The online Appendix also contains several useful extensions. Appendix B.1 implements Poisson regression as an iOLS estimator. Appendix B.2 adapts iOLS to the endogenous setting. Appendix B.3 extends it to deal with negative values in Y. Appendix B.4 addresses log-log specifications. Appendix B.5 develops a fast estimator for high-dimensional fixed-effects. Appendix B.6 deals with the log of a ratio. Appendix B.7 shows an alternative iOLS approach equivalent to sample-selection models. Finally, Appendix B.8 details the testing procedures in the endogenous regressors setting.

2 Existing Practices

Log-linear regressions are used in research for many purposes. These include using the log transformation because (1) the parameter estimate is an elasticity or a semi-elasticity;[7] (2) logs can linearize a theoretical model, e.g. a Cobb-Douglas production function; (3) logs can make heteroskedasticity vanish in some settings; (4) the data is sometimes *naturally* related by a log-linear relationship; or even (5) it provides a concave transformation for a highly dispersed outcome.

In this section, we first provide a general setup from which we assume the data to have been generated from. We then review the five main identified solutions to the "log of zero", and discuss their assumptions and limitations. The most popular fix consists in adding a positive constant to all observations. A second solution is to discard the non-positive observations from the sample. A third solution uses transformations of the response variable, such as IHS, akin to the log function. Fourth, the Poisson model handles the presence of zeros well in many settings. It is especially popular in international trade for the estimation of gravity equations (Head and Mayer, 2014). Finally, mixture models (e.g. Tobit or Heckit) also provide a solution by modelling the occurrence of non-positive observations as a sample selection problem.

2.1 The setup

Let us consider an iid sample of observations $\{Y_i, X_i\}_{i=1}^n$, where n denotes the sample size, generated by the "true" model given by

$$Y_i = \exp(X_i'\beta + \varepsilon_i)\xi_i, \tag{1}$$

where β is a fixed parameter of interest in $\mathbb{R}^K$, with $K \geq 1$, ε_i is a random error, and $\xi_i \in \{0, 1\}$ is a Bernoulli random error. Let X denote the $n \times K$ matrix comprised of the K-dimensional column vector X_i with elements X_{ki}, for $1 \leq k \leq K$.

Y_i can either be equal to zero, when $\xi_i = 0$, or take positive values, when $\xi_i = 1$. Taking logs on both sides of (1) is allowed only if Y_i (and thus ξ_i) takes only strictly

[7] In a log-log model such as $\log(y) = \beta \log(x) + \epsilon$, the elasticity of y with respect to x is given by $\frac{\partial \log(y)}{\partial \log(x)} = \frac{\partial y}{\partial x}\frac{x}{y} = \beta$.

positive values. Doing so yields the log-linear model given by

$$\log(Y_i) = X_i'\beta + \varepsilon_i. \tag{2}$$

For parsimony, we will focus on the more compact *multiplicative* representation,

$$Y_i = \exp(X_i'\beta)U_i, \tag{3}$$

where $U_i = \exp(\varepsilon_i)\xi_i$ is the error term, and the related *additive* model

$$Y_i = \exp(X_i'\beta) + \epsilon_i, \tag{4}$$

with $\epsilon_i = \exp(X_i'\beta)(U_i - 1)$ as the error term. We will discuss the difference between the later two forms when presenting Poisson models.

2.2 A wide variety of solutions

The popular fix: $\log(\mathbf{Y}+\mathbf{1})$. The most popular solution is to add a positive constant Δ to all observations Y_i so that $\tilde{Y}_i = Y_i + \Delta > 0$ making the log-transformation feasible for $\tilde{Y}_i$. The choice of Δ is discretionary and may arbitrarily bias the estimates and their standard errors. Moreover, the size of the bias will depend on the data at hand, suggesting that adding the smallest possible constant is not necessarily the least "harmful" choice.[8] We reconcile this approach with theory in Section 3.

To understand the bias, consider the model specified in (1). Adding $\Delta > 0$ and applying the log function yields after rearrangement

$$\log(Y_i + \Delta) = X_i'\beta + \log\left(U_i + \frac{\Delta}{\exp(X_i'\beta)}\right) \tag{5}$$

where the error term $\omega_i = \log\left(U_i + \frac{\Delta}{\exp(X_i'\beta)}\right)$ is correlated with X_i by construction, even when U_i and X_i are statistically independent, and creates an endogeneity bias. Although the choice of Δ matters, $\exp(X_i'\beta)$ can be arbitrarily close to zero hence leading to possibly large biases. Thus, the "popular fix" estimator is in general not

[8] Winkelmann (2008) discusses this approach and Cohn, Liu and Wardlaw (2021) survey its application in finance.

consistent.[9] Furthermore, standard errors will be too small because the zero observations transformed into $\log(\Delta)$ artificially reduce the variance of the transformed dependent variable.

Discarding zeros. The simplest solution is to delete the zero observations and estimate (2) directly with OLS. Formally, discarding zeros introduces a selection bias unless $E[\varepsilon|\xi = 1, X]$ is a constant. Similarly, one could discard zeros and estimate (4) assuming $E[\exp(\varepsilon)|\xi = 1, X]$ to be constant. Doing so assumes away any role played by the zeros and has context-dependent consequences; rendering it inadvisable at least since Young and Young (1975). At the very least, it will change the scope of the study by narrowing down the focus to observations for which $Y_i > 0$.

The economic interpretation of the error term should always be discussed when making such an assumption. For instance, some empirical studies relying on the Mincer equation for the purpose of estimating the returns to schooling use the log wage and discard unemployed individuals. Unemployed agents have unobserved wage rates which can be labelled as zeros. If ε_i captures the unobserved ability of individual i, it will undoubtedly be correlated with her employment outcome $\xi_i = 1$ or $\xi_i = 0$, hence introducing a sample selection bias when discarding the zeros.

Other transformations. An alternative approach relies on log-like transformations applicable to non-positive values. The most popular is the IHS or related transformations (MacKinnon and Magee, 1990; Burbidge, Magee and Robb, 1988; Ravallion, 2017). It consists in transforming Y_i into $\tilde{Y}_i = \log(\theta Y_i + \sqrt{\theta^2 Y_i^2 + 1})/\theta$ and estimating $\tilde{Y}_i = X_i'\beta + \omega_i$ by OLS. If the underlying model writes in log, then this transformation will likely yield biased estimates.[10] Nearly all economic applications set θ to 1 such that $\tilde{Y}$ tends toward $\log(2Y)$ for large values of Y.

This transformation essentially consists in adding a positive *observation-specific* value to the response variable before applying the log function. Its similarity with the log function may lead to treating them interchangeably. However, for small values

[9]This estimator is consistent under $E(\omega_i|X) = constant$ which implies strong assumptions of the joint distribution of U_i and X_i.

[10]Considering model (1), obtaining consistent estimates requires a moment condition like $E(\log(\theta U_i + \frac{\sqrt{\theta Y_i^2+1}}{\exp(X_i'\beta})|X) = 0$, which may be difficult to justify since $\frac{\sqrt{\theta Y_i^2+1}}{\exp(X_i'\beta)}$ is a non-linear function of X.

of Y_i, these transformations can behave differently. Besides, as shown in Bellemare and Wichman (2020), the interpretation of the coefficients is not trivial and the underlying elasticity is potentially biased or undefined.[11] It is hence satisfactory in contexts where applying a concave transformation is the main objective but suffers from similar limitations than the "popular fix" discussed above.

Poisson models. The model presented in (1) is non-linear in variables and parameters. The parameters are identified and non-linear estimators, such as non-linear least squares or Iterated Reweighted Least-Squares (IRLS), yield consistent estimates of β under the strict exogeneity restriction $E(U_i|X_i) = 1$ which implies the unconditional moments[12]

$$E(X_i(Y_i - \exp(X_i'\beta))) = 0. \tag{6}$$

The empirical counterpart yields a solution equivalent to that of the Pseudo log-likelihood of the Poisson model (Gourieroux, Monfort and Trognon, 1984). This approach is computationally efficient because it is a well-defined concave problem. This approach has been popularized by Santos Silva and Tenreyro (2006) for gravity models and is generally referred to as PPML. It is based on the additive representation of the model in (4) assuming $E((U_i-1)\exp(X_i'\beta)|X_i) = 0$, which is equivalent to $E(U_i|X_i) = 1$ in absence of endogenous regressors but leads to a different objective criterion.

Nevertheless, Poisson models have several shortcomings: (1) Existence of a solution is not guaranteed (Santos Silva and Tenreyro, 2010);[13] (2) They can exhibit convergence issues (Santos Silva and Tenreyro, 2011); (3) They can be very imprecise if the log-scale error is heavy-tailed (Manning and Mullahy, 2001); (4) They can be difficult to estimate with many fixed-effects (Correia, Guimarães and Zylkin, 2019); (5) instrumental variables require stronger assumptions and may dramatically in-

[11]The authors show that in $\tilde{Y}_i = X_i'\beta + \epsilon_i$, the elasticity $\hat{\zeta}_{yx} = \hat{\beta}x\frac{\sqrt{y^2+1}}{y}$ is a function of x, y, or is not defined for $y = 0$. β is an elasticity only if $x = 1$ and y is large.

[12]Choosing the "best" unconditional moments, or rather picking the optimal instruments, from a conditional moment restriction is beyond the scope of this paper.

[13]The authors have a dedicated website with helpful resources about PPML (https://personal.lse.ac.uk/tenreyro/lgw.html).

crease computational complexity (Wooldridge, 2015);[14] and (6) they suffer from an incidental parameter problem in the standard errors when including two-way fixed effects (Fernández-Val and Weidner, 2016), and in the point estimates when controlling for three-way fixed-effects. (Weidner and Zylkin, 2021).

Another important consideration about Poisson models is the role played by heteroskedasticity. Santos Silva and Tenreyro (2006) conclude that "under heteroskedasticity, the parameters of the log-linearized models estimated by OLS lead to biased estimates of the true elasticities" and suggest using PPML as a solution. This result is not as general as it may seem. It is only valid under the Poisson restriction $E[U|X] = 1$. To show this, we perform a Taylor expansion of the log-scale error ε_i around 0 to obtain

$$\varepsilon_i = \log(1) + (\exp(\varepsilon_i) - 1) - \frac{(\exp(\varepsilon_i) - 1)^2}{2} + \frac{(\exp(\varepsilon_i) - 1)^3}{3} + ... \tag{7}$$

This expression implies that, if $E[\varepsilon_i|X] = 0$ holds, the identifying assumption of the Poisson model $E[\exp(\varepsilon_i)|X] = 1$ also holds *only if* one makes additional assumptions about the higher-order centered moments of $\exp(\varepsilon_i)$, i.e. its variance, skewness, etc. The reverse conclusion hence applies: under heteroskedasticity, the additional assumptions about the higher-order moments of $\exp(\varepsilon_i)$ are unlikely to hold, therefore the Poisson model leads to biased estimates of the true elasticities. The presence of heteroskedasticity alone hence does not invalidate one approach over the other because it depends on the underlying exogeneity restriction, which is unverifiable.

This is not to say that heteroskedasticity is irrelevant. Indeed, Manning and Mullahy (2001) show that the parameters of the log-linear model, even though well-identified under $E[\varepsilon_i|X] = 0$, implies a conditional mean function $E(Y|X)$ which can be biased under heteroskedasticity, if the function is not appropriately retransformed, e.g. by using the smearing estimator (Duan, 1983) or a gaussian approximation (Manning, 1998). In opposition, Poisson models do not require ex-post retransformation to deliver an elasticity estimate of the mean response, provided $E[\exp(\varepsilon_i)|X] = 1$ holds true.

Finally, there are some reasons to prefer the multiplicative Poisson model over its

[14] Non-linear IV estimators require strict exogeneity between the errors and instruments unlike linear estimators.

additive counterpart. The latter (PPML) provides a computationally efficient solution in many cases but suffers from two significant drawbacks. First, the estimator minimizes the sum of squared *absolute* errors $Y_i - \exp(X_i'\beta)$. Doing so implies that observations with large values of Y are given much more weight than those with smaller values.[15] This is similar to discarding zero observations in many settings. As such, it cannot be considered a satisfactory approach to dealing with zeros in the general setting. This issue is illustrated in our applications.

Second, the additive formulation may be "awkward to motivate directly" (Mullahy, 1997). This is particularly evident when considering an omitted variable problem. In such a case, the additive model assumes this omitted variable to have an additive effect whereas all explanatory variables have multiplicative effects. The moment conditions for identification in the additive case hence differ from the usual setting. Finally, elasticities estimated from the additive model depends on both the exogenous and omitted variables, unlike those obtained in the multiplicative case.[16]

Mixture models and Heckman's correction. Mixture models consist in modeling the selection explicitly, $Y_i = 0$ or $Y_i > 0$, using a latent variable approach under chosen distributional assumptions. This approach is rarely used to address the log of zeros but has been relied upon in the context of gravity equations (Eaton and Tamura, 1994). The Heckman's ("Heckit") correction (Heckman, 1979) applies as follows. In the setting provided by model (1), it assumes that $\xi_i = 1$ if $X_i'\gamma + \nu_i > 0$, and $\xi_i = 0$ otherwise. $X_i'\gamma + \nu_i$ is hence referred to as the "selection equation". The key identifying restriction is that ε_i and ν_i are bivariate normal, so that $E[\varepsilon_i|U_i > 0, X]$ admits the closed-form expression $E[\varepsilon_i|\nu_i > -X_i'\gamma, X] = \lambda\frac{\phi(-X_i'\gamma)}{\Phi(X_i'\gamma)}$, for $\phi(\cdot)$ and $\Phi(\cdot)$ denoting the Gaussian probability density and distribution functions, respectively. λ and γ are estimable parameters. Estimation takes two steps. First, a probit model of $Y_i > 0$ conditional on X_i yields $\hat{\gamma}$. Second, the log-linear regression with an additional term, as specified by

$$\log(Y_i) = X_i'\beta + \lambda\frac{\phi(-X_i'\hat{\gamma})}{\Phi(X_i'\hat{\gamma})} + e_i, \tag{8}$$

[15] They yield different estimates because the additive model uses the unconditional moment restrictions $E(X_i \exp(X_i'\beta)(U_i - 1)) = 0$ instead of $E(X_i(U_i - 1)) = 0$ for the multiplicative model.

[16] These arguments have been made by Mullahy (1997) and Jeffrey Wooldridge (`https://www.statalist.org/forums/forum/general-stata-discussion/general/1536854-multiplicative-versus-additive-iv-poisson`).

is estimated by OLS to obtain β and λ. The relevance of the correction term can be tested using a t-test to check whether $\hat{\lambda}$ is different from zero.[17] Note that, however, this approach is heavily dependent on the distributional assumption in absence of instrumental variables in the selection equation.

3 Iterated Ordinary Least Squares (iOLS)

In this section, we develop a new approach which somehow reconciles the popular fix with econometric theory. This new approach yields a family of estimators requiring only OLS, and more importantly which embeds both the log-linear model and Poisson models as special cases. For clarity, we first show how our estimation procedure arises in the context of the log of zeros. Second, we present the algorithm. Third, we derive its asymptotic properties, and detail how minor modifications can accommodate alternative exogeneity conditions. Finally, we develop a variety of important extensions in the Appendices.

3.1 Fixing the popular fix (iOLS$_\delta$)

We let Δ_i vary across observations such that $Y_i + \Delta_i > 0$. From (5), we have

$$\log(Y_i + \Delta_i) = X_i'\beta + \log\left(U_i + \frac{\Delta_i}{\exp(X_i'\beta)}\right). \tag{9}$$

Letting $\Delta_i = \delta \exp(X_i'\beta)$, for some $\delta > 0$, this equation becomes

$$\log(Y_i + \delta \exp(X_i'\beta)) = X_i'\beta + \upsilon_i. \tag{10}$$

where the new error term $\upsilon_i = \log(\delta + U_i) > \log(\delta)$ is unlikely to be mean-zero. Before proceeding to the more general setting, let us first assume the following strict exogeneity restriction

$$E[\upsilon_i|X] = \log(\delta + 1). \tag{11}$$

[17] We show how this approach relates to ours in Appendix B.7.

It means that the conditional mean of $\log(\delta + U_i)$ is independent of covariates. Under this assumption, (10) can be rewritten into

$$\log\left(\frac{Y_i + \delta\exp(X_i'\beta)}{\delta + 1}\right) = X_i'\beta + \overline{v}_i, \tag{12}$$

where $\overline{v}_i = \log(\delta + U_i) - \log(\delta + 1)$ is a mean-zero error term, and β could be estimated using a non-linear estimator or our iterative least-squares estimator to solve

$$\min_{\beta} \sum_{i=1}^{n} \left(\log(Y_i + \delta\exp(X_i'\beta)) - \log(\delta + 1) - X_i'\beta\right)^2. \tag{13}$$

Remark that this transformation consists in taking the log of a weighted average of Y and its conditional mean, where δ controls the weights.

From log-linear to Poisson. Interestingly, δ acts as a model selection parameter within a continuum of models which includes two important special cases: the log-linear model (as $\delta \to 0$) and the multiplicative Poisson model (as $\delta \to \infty$), along with a range of intermediate models.

On the one hand, the relation with log-linear models in absence of zeros is trivial. Let $\xi_i = 1$ for all $i = 1, ..., n$, the condition in (11) is then equivalent to $E[\varepsilon_i|X] = 0$ for $\delta = 0$ which is the exogeneity restriction used by the log-linear model (2), and so the objective in (13) becomes

$$\min_{\beta} \sum_{i=1}^{n} (\log(Y_i) - X_i'\beta)^2. \tag{14}$$

To show this relation when there are zeros, let us develop the error term as

$$\log(\delta + U_i) = (1 - \xi_i)\log(\delta) + \xi_i \log(\delta + \exp(\varepsilon_i)). \tag{15}$$

Taking a Taylor expansion of the second term around $\exp(\varepsilon_i)$, substituting in the above expression, and rearranging yield

$$\log(\delta + U_i) = (1 - \xi_i)\log(\delta) + \xi_i\varepsilon_i + o(\delta). \tag{16}$$

Imposing condition (11) for δ small enough is approximately equivalent to assuming that

$$Pr(\xi_i = 0|X)\log(\delta) + Pr(\xi_i = 1|X)E[\varepsilon_i|X, \xi_i = 1] = 0 \tag{17}$$

does not depend on X. Therefore, for δ small enough, our approach corresponds to the log-linear model where the log-scale errors for zero observations satisfy the condition $E[\varepsilon_i|X, \xi_i = 0] = \log(\delta)$.[18]

On the other hand, taking a Taylor expansion of $\log(\delta + U_i)$ around $U_i = 1$ gives

$$E[\log(\delta + U)|X] = \log(1 + \delta) + \frac{1}{1+\delta}E[U - 1|X] + o(\delta^{-2}). \tag{18}$$

The terms in $o(\delta^{-2})$ will become negligible faster than $E[U - 1|X]/(1 + \delta)$ as δ increases. Thus, for δ large enough, the above exogeneity condition is approximately equivalent to $E[U|X] = 1$ and (13) goes to

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i \exp(-X_i'\beta) - 1\right)^2, \tag{19}$$

which corresponds to the optimization problem solved by the (multiplicative) Poisson model specified in (3). Section B.1 shows that both multiplicative and additive Poisson models are nested within our general framework and not just as limiting cases.

3.2 Estimation by iOLS

Let us make the following assumptions about the covariates X_i and error term U_i.

Assumption 1 (Covariates) *X has full column rank and $E(X_iX_i') < \infty$.*

Assumption 2 (Errors) *The aggregate error $U_i = \exp(\varepsilon_i)\xi_i$ is independently and identically distributed.*

In the general setting, our approach consists in transforming the response variable as

$$\tilde{Y}_i(\beta, \delta) = \log(Y_i + \delta \exp(X_i'\beta)) - c(\beta, \delta), \tag{20}$$

[18] One can also show that it gives a reasonable approximation of the log-linear model as long as $E[\varepsilon_i|X, \xi_i = 1]$ is proportional to $Pr(\xi_i = 1|X)^{-1}$, as illustrated in Section 5's DGP 2.

where $c(\beta, \delta)$ is a known function of β and δ, possibly different from $\log(\delta + 1)$, so as to obtain a (seemingly) linear model:

$$\tilde{Y}_i(\beta, \delta) = X_i'\beta + \overline{\upsilon}_i, \tag{21}$$

with $E[\overline{\upsilon}_i] = 0$. We refer to this model as iOLS$_\delta$, because it depends on the choice of the parameter δ, which will be further discussed together with exogeneity restrictions. The moment condition $E[X_i\overline{\upsilon}_i] = 0$ yields

$$\beta = E[X_iX_i']^{-1}E\left[X_i\tilde{Y}_i(\beta, \delta)\right], \tag{22}$$

which characterizes β as the solution of a fixed-point problem. Based on this insight, we propose the following iterative least-squares estimator. Thereafter, we drop the dependence of $\tilde{Y}_i$ on δ to alleviate notations.

Algorithm 1 (iOLS estimator) *The iOLS estimator is defined as follows:*

1. *Initialize* $t = 0$ *and let* $\hat{\beta}_0$ *be an initial estimate, e.g. obtained with the "popular fix" estimator* $\hat{\beta}^{PF} = [X'X]^{-1}X'\log(Y + \Delta) \in \mathbb{R}^K$, *for some* $\Delta > 0$;
2. *Transform the dependent variable into* $\tilde{Y}(\hat{\beta}_t)$ *using* (20);
3. *Compute the OLS estimate* $\hat{\beta}_{t+1} = [X'X]^{-1}X'\tilde{Y}(\hat{\beta}_t)$, *and update* t *to* $t + 1$;
4. *Iterate steps 2 and 3 until* $\hat{\beta}_t$ *converges.*

Remark also that $X'X$ needs only be inverted once, making this approach computationally efficient. In addition, the resulting estimate varies continuously with δ which allows using warm starting to solve for a large range of δ's, if desired.

3.3 Identification and asymptotic properties

The simplified version of iOLS presented earlier assumed $E[\upsilon_i|X] = \log(\delta + 1)$. We now turn to the more general setting by considering the assumption below.

Assumption 3 *The error term* υ_i *satisfies the weak exogenous restriction* $E[X_i'(\upsilon_i - c(\delta, \beta))] = 0$ *where the value* $c(\delta, \beta)$ *is unknown. In addition, let* $E[U_i] = 1$.

Like in the Poisson model, the assumption that $E[U_i]$ is known allows identifying the intercept term. It also affects $c(\delta, \beta)$ which is determined by (unconditional) higher-order moments of U_i. To see this, consider the Taylor expansion of $\log(\delta + U_i)$ around $\log(1 + \delta)$ to obtain

$$c(\delta, \beta) = \log(1 + \delta) - \frac{1}{2(1+\delta)^2} E[(U_i - 1)^2] + \frac{1}{3(1+\delta)^3} E[(U_i - 1)^3] + o(\delta), \quad (23)$$

where the second and third terms are respectively the variance and third centered moment of U_i. The first centered moment is zero under Assumption 3. If the infinite weighted sum of higher moments of U_i equals zero then we are back to the simple setting where $c(\delta, \beta) = \log(1 + \delta)$. Hereafter, we omit the dependence of c on δ and β for notational convenience.

Identification. Demeaning the error term is required to identify all parameters. Let us assume the exogeneity condition $E[X_i \overline{\upsilon}_i] = 0$, where $\overline{\upsilon}_i = \upsilon_i - c$ denotes the centered error term of the linearized model. This condition yields the set of $k + 1$ equations

$$E\left[X_i \left(\log(Y_i + \delta \exp(X_i'\beta)) - c\right)\right] = E\left[X_i X_i'\right] \beta, \quad (24)$$

with $k + 2$ unknowns. This system identifies β only if c is known. Fortunately, the model in (1) provides the additional restriction necessary for identification. Let us write $X_i'\beta = \beta^1 + X_i^{r\prime}\beta^r$, where β^1 is the constant term and the other term represents the non-deterministic part. We rewrite (1) into

$$Y_i = \exp(\beta^1 + X_i^{r\prime}\beta^r)U_i = \exp(\beta^1)\exp(X_i^{r\prime}\beta^r)U_i. \quad (25)$$

Rearranging, taking expectations and applying the log function gives the following expression for the intercept given the other parameters

$$\beta^1_\beta = \log(E[Y_i \exp(-X_i^{r\prime}\beta^r)]). \quad (26)$$

Therefore, the parameters are identified and c can be written as[19]

$$c(\delta,\beta) = E[\log(Y_i + \delta \exp(\beta^1_\beta + X_i^{r'}\beta^r)) - \beta^1_\beta - X_i^{r'}\beta^r]. \tag{27}$$

Remark that the estimation of c from the unconditional restriction $E(U) = 1$ bears similarities with the smearing estimator (Duan, 1983). Indeed, our approach integrates this idea directly in the estimation procedure. Doing so allows avoiding the issue that $\exp(X'\hat{\beta})$ is not a consistent estimate of $E(Y|X)$, which arises if β is estimated with OLS in the log-linear model (Manning, 1998). Therefore, the iOLS estimator of β does correspond to the semi-elasticity (or elasticity if X is replaced by $\log(X)$) of Y with respect to X around the mean.

Asymptotic properties. We establish the asymptotic properties of iOLS$_\delta$ in the following theorem.

Theorem 1 (Consistency and Normality of iOLS$_\delta$) *Under Assumptions 1, 2, 3, and suitable regularity conditions, the iOLS estimator using $c(\delta,\beta)$ defined in (27) is consistent and achieves the parametric rate of convergence $n^{-1/2}$ for any $\delta \in (0,+\infty)$. Formally, we have $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ as $n \to \infty$ for any $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0,1)$ is the modulus of the associated contraction mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$. In addition, iOLS$_\delta$ is asymptotically normally distributed such that $\sqrt{n}\left(\hat{\beta}_{t(n)} - \beta\right) \xrightarrow{d} \mathcal{N}(0,\Omega)$, as $n \to \infty$, where the covariance can be consistently estimated using*

$$\hat{\Omega} = \left(\frac{1}{n}X'(I-W)X\right)^{-1} \hat{\Sigma}_0 \left(\frac{1}{n}X'(I-W)X\right)^{-1}, \tag{28}$$

where W is a diagonal weighting matrix with diagonal elements $\frac{\delta}{\delta+U_i} \in [0,1)$, and $\hat{\Sigma}_0$ is a (robust) consistent estimator of the covariance of $X_i(\log(1+U_i) - c)$ across observations. Ω hence corresponds to the asymptotic covariance of the OLS estimator in the last iteration up to reweighting using W.

[19] In our practical implementation, we solve the identification problem by using the consistent estimator defined for any ϕ as $\hat{c}(\phi) = \frac{1}{n}\sum_{i=1}^n \log(Y_i + \delta\exp(\tilde{\phi}^1_\phi + X_i^{r'}\phi^r)) - \frac{1}{n}\sum_{i=1}^n(\tilde{\phi}^1_\phi + X_i^{r'}\phi^r)$, where the constant parameter estimate is replaced by the estimator $\tilde{\phi}^1_\phi = \log(n^{-1}\sum_{i=1}^n Y_i \exp(-X_i^{r'}\phi^r))$.

This asymptotic result guarantees root-n consistent estimates and, for any fixed n, the iterative process converges after a finite number of iterations: $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0, 1)$ is the modulus of the associated contraction mapping. The numerical convergence will hence be slower for larger sample sizes n and modulus κ closer to 1. κ depends on the DGP and is decreasing with δ.

Note that very small values of δ might prevent the algorithm from converging in finite time. This occurs because a very small δ may imply a κ very close to 1, hence a very slow numerical convergence. However, our algorithm checks this condition and recommends a larger δ whenever necessary.[20]

Remark that the initial estimate $\hat{\beta}_0$ can be an important determinant of the number of iterations before convergence. More iterations will be needed as it is further away from the fixed-point. However, this initial estimate only needs to be within the parameter space and does not need to be consistent for the above theorem to hold.

The asymptotic distributions of iOLS$_\delta$ and of OLS in the last iteration (once the estimator has converged) are similar. Although the standard errors of the latter are incorrect for iOLS$_\delta$, a reweighting of the corresponding covariance matrix using simple algebra is sufficient and allows to use any HAC-robust covariance estimator.[21]

3.3.1 Poisson regression as iOLS

We have already shown that iOLS$_\delta$ approaches the multiplicative Poisson model for an arbitrarily large δ. Nevertheless, it is possible to use the iOLS framework to (exactly) estimate both the multiplicative and additive Poisson models. Doing so can be useful in some settings, either to avoid possible convergence issues or for speeding up computations, as will be illustrated in simulations with fixed-effects and endogenous regressors. We refer to those estimators as iOLS$_{MP}$ and iOLS$_{AP}$, respectively, and replace Assumption 3 by

[20] Our algorithm includes a safety check to ensure that κ is sufficiently smaller than 1. In practice, it is estimated as the median across estimates obtained at each iteration by $\hat{\kappa}_{t+1} = |\beta_{t+1} - \beta_t|/|\beta_t - \beta_{t-1}|$.

[21] A simple approximation of the standard errors for iOLS$_\delta$ consists in multiplying those of the last step OLS by a factor $1 + \delta$.

Assumption 4 (Poisson condition) *The error term U_i admits the weak exogeneity restriction $E[X_i'(U_i - 1)] = 0$, which implies $E[U_i] = 1$.*

Under this new assumption, the iOLS procedure yields the multiplicative Poisson estimates if (27) is changed to

$$c_i(\delta, \beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i \exp(-X_i'\beta) - 1), \tag{29}$$

and the PPML estimates if changed to

$$c_i(\delta, \beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i - \exp(X_i'\beta)). \tag{30}$$

These estimators are studied in Appendix B.1, where Theorem 3 shows their consistency and asymptotic normality. These results show that our approach is flexible with respect to the choice of both the identifying restriction and objective criterion without significant consequences in large samples, except for minor modifications to the covariance matrix. Other extensions, including i2SLS and high-dimensional fixed effects, are presented in Appendix B.

4 Specification testing and model selection

Empirical researchers facing the log of zero usually compare several estimators to gauge the sensitivity of their results. Yet, each estimator is only valid under specific unverifiable identifying assumptions. The latter can nonetheless be systematically investigated by evaluating the model's external validity with respect to the observed patterns of zeros.

The tests developed in this section offer an opportunity for an ex-post evaluation of the identifying restrictions used for moment-based estimators. Our tests are specification tests used to evaluate the validity of conditional moment restrictions, like $E(U_i|X_i) = 1$ for Poisson models.[22] They are, as such, similar to the RESET test of Ramsey (1969) for linear regression and its application for Poisson models by

[22] Santos Silva, Tenreyro and Windmeijer (2015) proposed a radically different approach based on non-nested hypothesis tests (Davidson and MacKinnon, 1981) which consist in testing two competing models against each other.

Wooldridge (1997) and Santos Silva and Tenreyro (2006). Our approach is, however, tailored to addressing the log of zero by focusing on the conditional probability of observing $Y = 0$. It also provides a much more powerful test of the conditional restrictions in this context as will be shown in the simulations.

A common limit of these tests, including RESET tests, is their focus on the conditional moment restrictions (e.g. $E(U_i|X_i) = 1$) rather than the unconditional restrictions (e.g. $E((U_i - 1)X_i) = 0$). The former is a *sufficient* condition whereas the latter is a *necessary* condition for consistency. We argue that statistical evidence against a sufficient condition is still valuable information about the associated model. The main issue is, however, that a rejection of the sufficient condition is not evidence against the necessary condition, which remains unverifiable. Bearing these limits in mind, we proceed to present our methods.

4.1 Specification testing

Let us first look at the implicit assumption made by Poisson models about the zeros.[23] A related approach will be applied for other restrictions including iOLS$_\delta$. Noting that a zero can only be observed if $U_i = 0$, the Poisson restriction $E(U_i|X_i) = 1$ can be decomposed into

$$E[U_i|X_i] = E[U_i|X_i, U_i > 0]Pr(U_i > 0|X_i) = E(U_i), \tag{31}$$

since $E[U_i|X_i, U_i = 0] = 0$. There are only two possibilities for this condition to hold true. First, both $E[U_i|X_i, U_i > 0]$ and $Pr(U_i > 0|X_i)$ vary with X_i in such a way that the condition holds. It happens for example if U_i is conditionally Poisson. Second, this condition also holds if, instead, $E[U_i|X_i, U_i > 0]$ and $Pr(U_i > 0|X_i)$ are constant. The former is an exogeneity restriction between X_i and U_i, conditional of the error being positive, which assumes away any selection bias. The latter implies that discarding zeros or not before estimation is irrelevant for identification.

This equation characterizes the implicit relation between zeros and non-zero observations given by

$$E[U_i|X_i, U_i > 0] = \frac{E(U_i),}{Pr(U_i > 0|X_i)}. \tag{32}$$

[23] Appendix B.8 details how these tests can be implemented in the endogenous setting.

It means that the conditional error term for non-zero observations is inversely proportional to the conditional probability of having a non-zero observation.[24] We propose to investigate whether this implication matches what is observed in the data. To do so, we develop a test to assess whether the residuals implied by the chosen model satisfy this relationship where the conditional probability is estimated outside the model.[25] This approach amounts to evaluating the null hypothesis

$$H_0 : E[U_i|X_i, U_i > 0] = \frac{1}{Pr(U_i > 0|X_i)}. \tag{33}$$

Under this null hypothesis, the error term satisfies the following equation

$$U_i = \lambda Pr(U_i > 0|X_i)^{-1} + \nu_i \tag{34}$$

with $\lambda = 1$ and $E[\nu_i|U_i > 0, X_i] = 0$. Therefore, one could evaluate H_0 by testing $\lambda = 1$ after regressing U_i onto $E[U]Pr(U_i > 0|X_i)^{-1}$ for non-zero observations. This test would not be feasible in practice because none of those variables are observable. Instead, we propose the following approach:

1. Estimate of $Pr(U_i > 0|X_i)$ denoted $\hat{P}(X)$.
2. Compute Poisson estimates $\hat{\beta}$, e.g. using PPML.
3. Recover the residuals $\hat{U}_i = Y_i \exp(-X'\hat{\beta})$;
4. Define $\hat{W}_i = \hat{P}(X_i)^{-1}$ and estimate the regression $\hat{U}_i = \lambda\hat{W}_i + \nu_i$, with the $n_1 > 0$ non-zero observations to obtain $\hat{\lambda} = \left(n_1^{-1}\sum_{i=1}^{n_1}\hat{W}_i^2\right)^{-1}\left(\sum_{j=1}^{n_1}\hat{W}_j\hat{U}_j\right)$.
5. Reject or not H_0 using a test statistic, such as the t-stat

$$t = \frac{\hat{\lambda} - 1}{\hat{\sigma}_\lambda}, \tag{35}$$

   where $\hat{\sigma}_\lambda$ denotes an estimate of the standard deviation of $\hat{\lambda}$.

The estimation errors from the first and second steps cannot be neglected. We

[24]Heckman's correction model enforces a comparable conditional moment restriction: $E[\log(U_i)|X_i, U_i > 0] = \frac{\lambda\phi(-X_i'\gamma)}{Pr(U_i>0|X_i)}$. More generally, moment-based methods typically make implicit assumptions about selection, whereas sample-selection models enforce explicit restrictions.

[25]Testing whether the probability of observing a zero depends on any X_i is done by investigating the statistical significance of the X_i's in a conditional probability model like the logit.

study the asymptotic properties of the test statistic (35) in the next theorem using the following assumptions.

Assumption 5 (Consistent first-step estimators) *First-step estimators are consistent under H_0, so that $\hat{P}(X_i) \xrightarrow{p} Pr(U_i > 0|X_i)$ and $\hat{\beta} \xrightarrow{p} \beta$ with $n \to \infty$.*

Assumption 6 (Conditional independence of first-step errors) *Estimation errors from first-step estimates, denoted $\varepsilon_i^{WU} = \hat{W}_i\hat{U}_i - W_iU_i$, and the error term in the second-step, denoted $W_i\nu_i$, are independent conditionally on $U_i > 0$.*

Theorem 2 (Asymptotic properties of the test statistic) *Under H_0, Assumptions 1, 5, 6 and some regularity conditions, we have*

$$\frac{\sqrt{n_1}(\hat{\lambda} - 1)}{\hat{\sigma}_\lambda} \xrightarrow{d} N(0, 1), \tag{36}$$

as $n_1 \to \infty$, and $\hat{\sigma}_\lambda^2$ is a consistent estimator of $\sigma_\lambda^2 = E(W_i^2)^{-2}E(W_i^2\nu_i^2) + E(W_i^2)^{-2}E(\varepsilon_i^{WU^2})$, the asymptotic variance of $\hat{\lambda}$. In addition, the t-stat diverges under any alternative as long as $E[\nu_i|U_i > 0, X_i] \neq 0$.

The practical implementation of this test requires two elements. First, computing the t-stat requires to estimate σ_λ^2 which must account for the additional noise introduced by first-step estimates.[26] We calculate the t-stat after estimating $\hat{\lambda}$'s standard error with the pairs bootstrap. Second, the test necessitates a consistent estimate of the conditional probability function $P(U > 0|\cdot)$. Specifying a parametric model, like the logit, probit or even (ex-post bounded) linear probability model, provides a simple option. However, the misspecification of $P(U > 0|\cdot)$ may distort the test's size and performance. A nonparametric or semiparametric estimate of the conditional probability should hence provide a preferable option. Nevertheless, nonparametric and semiparametric estimates are subject to small-sample bias due to regularization which can lead to similar distortions. In our implementation, we use a k-nearest neighbors (kNN) algorithm (Hastie, Tibshirani and Friedman, 2009). Although consistent, nonparametric estimators exhibit poor small-sample behaviors at

[26] The main difficulty in deriving a closed-form expression for $\hat{\sigma}_\lambda$ is to account for the correlation between $\hat{P}(X_i)$ and $\hat{\beta}^{iOLS_{MP}}$ if separately estimated.

the support's boundaries (Sricharan, Raich and Hero, 2010). We address this issue by truncating (trimming) observations associated with predicted probabilities outside the 5% and 95%. Another issue related to nonparametric estimation is the choice of the regularization parameter, i.e. the number of neighbors in kNN. Selecting a small number of neighbors will lead to non-smooth conditional probability functions, hence a noisy estimate of $P(X)$. It is hence preferable to have a reasonably large number of neighbors. The optimal choice of this parameter for inference is beyond the scope of this paper.

The same logic applies to iOLS$_\delta$. The null hypothesis is changed to

$$H_0 : E[\log(\delta + U_i)|X_i, U_i > 0] = \frac{c(\delta, \beta^{iOLS}) - \log(\delta)(1 - Pr(U_i > 0|X_i))}{Pr(U_i > 0|X_i)}, \tag{37}$$

and the corresponding regression given by $\log(\delta + \hat{U}_i) - \log(\delta) = \lambda W_i + \nu_i$, for strictly positive errors only, where $\hat{U}_i = Y_i \exp(-X'\hat{\beta}^{iOLS_\delta})$ and $W_i = (\hat{c} - \log(\delta))\hat{P}(X_i)^{-1}$ based on $\hat{c}$ obtained from iOLS$_\delta$. The rest of the testing procedure is unchanged.

Remark that a similar procedure could be applied to check for which value of Δ the popular fix estimator in (5) is coherent with the observed patterns of zeros. Unfortunately, choosing Δ using such procedure does not address the endogeneity bias introduced by the $\log(\Delta + Y)$ transformation.

4.2 Data-driven model selection

The wide variety of available methods for addressing the log of zero forces researchers to use a selection procedure. We propose to do so in a data-driven way based on the above results.

First, iOLS$_\delta$ for any $\delta \in (0, \infty)$ is based on the moment condition $E[\log(\delta + U)|X] = c(\delta)$ which depends on the choice of δ. If the "true" DGP belongs to this family of models, then there exists a δ for which H_0 will not be rejected as $n \to \infty$. Similarly, the data may better fit a competing estimator such as PPML or IHS. The tests might however not be sufficiently discriminatory to pinpoint a single model in small samples.

We propose to select the model for which $\hat{\lambda}$ is the closest to 1, among all models (iOLS$_\delta$, PPML, and iOLS$_{MP}$) which are not rejected by the tests. This heuristic is aimed at selecting the model with the least deviation between the implied and observed patterns of zeros, i.e. the "most plausible" model. This selection rule is asymptotically valid because only the true model will be associated with $\lambda = 1$ as $n \to \infty$. In the event that all the considered models are rejected by the tests, we believe that mixture models should be preferred.

5 Simulations

Let us specify the model as $Y_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})U_i$, where $\beta_0 = \beta_1 = \beta_2 = 1$, $U_i = \exp(\varepsilon_i)\xi_i$ with $\xi_i = 0$ or 1, and $Pr(\xi_i = 0|X_i) = P(X_i) = \frac{1}{1+\exp(\gamma_0+\gamma_1 X_{1i}+\gamma_2 X_{2i})}$, with $\gamma_0 = -0.4$, $\gamma_1 = 0.4$ and $\gamma_2 = -0.4$. We consider two DGPs:[27]

- DGP 1: $E[U_i|X_i] = 1$ (Poisson). This DGP is aimed at comparing our approach to Poisson models. We assume that $(X_{1i}, X_{2i})'$ is bivariate normal with mean 1, variance $\sigma^2_{X1} = \sigma^2_{X2} = 1$ and covariance $\sigma_{X1X2} = -0.3$. We further assume that ε_i conditional on $\xi_i = 1$ is Gaussian with mean $-\log(P(X_i)) - 1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(X_i)$. Recall that $E(\exp(\varepsilon_i)|X_i, \xi_i = 1) = 1/P(X)$ is implied by the (conditional) Poisson restriction. Note that the error U_i is heteroskedastic in this DGP.
- DGP 2: $E[\varepsilon_i|X] = 0$ (Log-linear). This DGP is useful to illustrate the flexibility of iOLS$_\delta$. We assume that $(X_{1i}, X_{2i})'$ is like in DGP 1. We further assume that ε is gaussian with mean $\xi_i P(X)^{-1} - (1-\xi_i)(1-P(X))^{-1}$, so that $E[\varepsilon_i|X, \xi_i = 1] = 1/P(X)$. Remark that this condition is the log-scale counterpart of $E(\exp(\varepsilon_i)|X, \xi_i = 1) = 1/P(X)$ implied by Poisson models. The variance parameter is fixed at 0.5. Note that $E(U)$ is different from 1 and the error U_i is also heteroskedastic in this DGP.

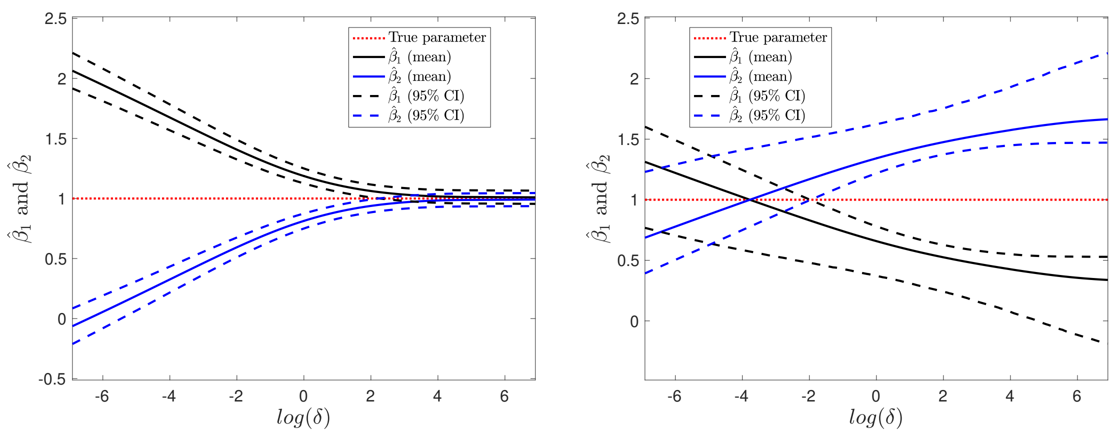
Remark that none of these DGPs assume ideal conditions for iOLS$_\delta$. We simulate 10,000 times each DGP, for two sample sizes ($n = 1,000$ and $n = 10,000$), and report the mean and standard deviations for:[28] (1) iOLS$_\delta$ (best) corresponds to the

[27] Appendix C presents additional simulations, including under the exact $iOLS_\delta$-restriction and with endogenous regressors and fixed-effects.

[28] To save space the results for OLS after discarding zero observations, the IHS transform, and iOLS$_{AP}$ – which is equivalent to PPML – are in Table C.1.

"oracle estimator", i.e. where δ is chosen so that the mean-squared error of $\hat{\beta}_1$ and $\hat{\beta}_2$ is minimized; (2) iOLS$_\delta$ with automatic selection of δ, where the conditional probability is specified as a logit model;[29] (3) iOLS$_{MP}$, the multiplicative Poisson model; (5) PPML, the additive Poisson model estimated using IRLS; and (6) the PF with $\Delta = 1$.[30]

Bias and variance. The mean iOLS$_\delta$ estimates along with 95% confidence intervals are shown as functions of the hyper-parameter δ in Figure 2 for both DGPs. Figure 2a confirms that iOLS$_\delta$ delivers correct estimates under the Poisson DGP provided that δ is sufficiently large (DGP 1). Figure 2b illustrates that the best value for δ is small in the log-linear model (DGP 2).



(a) DGP1 (Poisson) (b) DGP2 (Log-linear)

Figure 2: iOLS estimates as functions of δ (n=10,000)

Table 1 reports the results for DGP 1 (Poisson). We first observe that PPML and iOLS$_{MP}$ are consistent. iOLS$_{MP}$ yields more precise estimates in comparison. Second, iOLS$_\delta$ delivers good estimates with a small bias and a variance comparable to that of iOLS$_{MP}$. The automatic selection of δ introduces some noise in the smaller sample but this effect becomes negligible in the larger sample. Its bias is however slightly larger than for the best iOLS$_\delta$ estimates in both samples. Note that the standard errors of the (best) iOLS$_\delta$ estimate illustrates the root-n consistency result:

[29] Misspecifying the probability function does not affect much the results (Figure C.1).

[30] All simulations were performed in MATLAB 2021b with a 3.6GHz 10-Core Intel Core i9 processor and 32 GB 2667 MHz DDR4 memory.

multiplying the sample size by 10 yields $\sqrt{10} \approx 3$ times smaller standard errors. Finally, the PF exhibits large biases although relatively small errors.

Table 1: Simulations: DGP 1 (Poisson)

		$n = 1000$			$n = 10,000$	
Estim.	β_0	β_1	β_2	β_0	β_1	β_2
$iOLS_\delta$ (best)	0.99	1.02	0.99	1.00	1.01	0.99
	(0.16)	(0.09)	(0.09)	(0.05)	(0.03)	(0.03)
$iOLS_\delta$ (auto)	1.04	1.13	0.87	1.00	1.03	0.97
	(0.26)	(0.20)	(0.20)	(0.05)	(0.03)	(0.03)
$iOLS_{MP}$	0.99	1.01	0.99	1.00	1.00	1.00
	(0.16)	(0.09)	(0.09)	(0.05)	(0.03)	(0.03)
$PPML$	1.01	0.99	0.98	1.00	1.00	1.00
	(0.46)	(0.16)	(0.20)	(0.18)	(0.06)	(0.08)
	(0.11)	(0.07)	(0.07)	(0.03)	(0.02)	(0.02)
PF	0.65	0.61	0.15	0.65	0.61	0.15
	(0.09)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)

Notes: This table shows the mean and standard errors (in parentheses) of parameter estimates across 10,000 simulations based on DGP1.

Table 2 reports the results for DGP 2 (Log-linear). All estimators have large biases except iOLS$_\delta$. Poisson estimators perform poorly in this context. In particular, PPML diverges to extremely large values in many simulations.[31]

Table 2: Simulations: DGP 2 (Log-linear)

		$n = 1000$			$n = 10,000$	
Estim.	β_0	β_1	β_2	β_0	β_1	β_2
$iOLS_\delta$ (best)	4.52	1.15	0.85	5.36	1.01	0.99
	(0.81)	(0.22)	(0.22)	(1.16)	(0.17)	(0.17)
$iOLS_\delta$ (auto)	4.45	1.06	0.94	5.64	1.05	0.95
	(1.34)	(0.17)	(0.17)	(1.68)	(0.11)	(0.10)
$iOLS_{MP}$	3.20	0.44	1.57	3.42	0.30	1.70
	(0.36)	(0.19)	(0.20)	(0.51)	(0.22)	(0.23)
	(4.50)	(1.17)	(1.32)	(10.21)	(1.95)	(2.25)
$PPML$	-Inf	-Inf	Inf	Inf	-Inf	Inf
	(Inf)	(Inf)	(Inf)	(Inf)	(Inf)	(Inf)
PF	1.43	0.67	0.14	1.43	0.67	0.14
	(0.15)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)

Notes: This table shows the mean and standard errors (in parentheses) of parameter estimates across 10,000 simulations based on DGP 2.

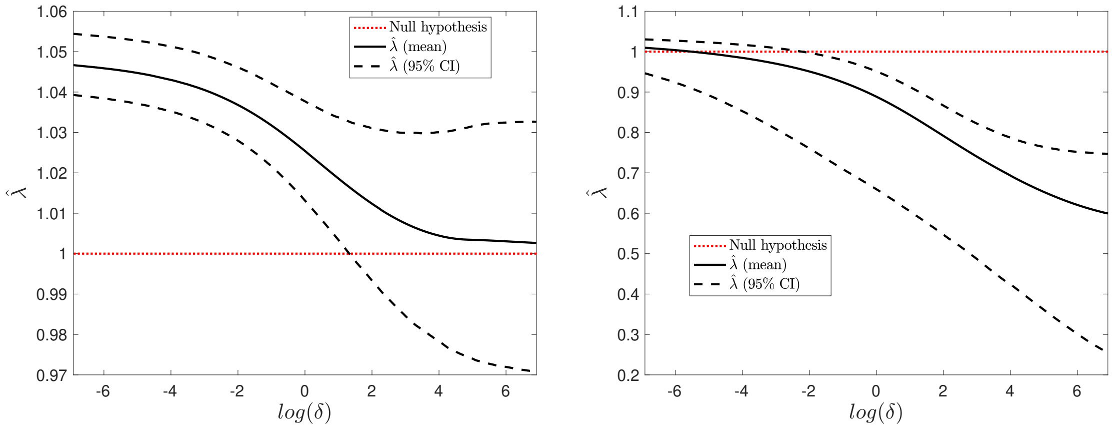
[31] The same result is observed in Table C.1.

Computational gains with endogenous regressors and many fixed-effects. We briefly describe the results of some additional simulations detailed in Appendix C. DGP 1 (Poisson) is modified so that both regressors are now correlated with the error term. In absence of fixed-effects, the i2SLS estimators perform similarly to the non-linear IV estimator (Mullahy, 1997) in terms of bias and variance, and provide the best options. The control function approach of Wooldridge (1997) delivers biased estimates because the endogenous regressors are non-linear functions of the error term in this design. Additive Poisson estimators are not consistent because the unobserved heterogeneity does not enter Y additively.

i2SLS estimators bring important computational gains in settings with fixed-effects. We consider $T = 100$ periods and n/T individuals to model two-way fixed effects. For $n = 1,000$, i.e. 111 regressors, $i2SLS_{MP}$ is computed in about 0.08 second whereas the non-linear IV estimator in Mullahy (1997) takes 45 seconds. For $n = 10,000$, i.e. 201 regressors, $i2SLS_{MP}$ is computed in about 0.8 second whereas the non-linear IV estimator takes above 500 seconds.

Automatic selection and specification tests. These simulations are also useful to study our testing procedures and automatic selection rule. The conditional probabilities to have a zero are logistic in all DGPs. In what follows, we mainly focus on the correct parametric specification to compute the conditional probability of observing zero values (logit). We also report and discuss some results when using a nonparametric approach (kNN). Figure C.1 in Appendix C compares the results when misspecifying the conditional probability function as a Probit or a linear probability model. The differences remain small in all cases.

First, we illustrate the automatic model selection in Figure 3. The plain lines show $\hat{\lambda}$ as a function of $\log(\delta)$, and the dashed lines represent the (pointwise) 95% confidence intervals across simulations. The selection rule is to pick δ such that the parameter λ is as close to 1 as possible. For the Poisson DGP, Figure 3a confirms that $\hat{\lambda}$ tends to be closer to 1 for larger values of δ. For the log-linear DGP, Figure 3b corroborates that $\hat{\lambda}$ tends to be closer to 1 for small values of δ.



(a) DGP1 (Poisson)

(b) DGP3 (Log-linear)

Figure 3: iOLS's $\hat{\lambda}$ as functions of δ (n=10,000)

Second, we investigate the performance of the inference procedure based on the pairs bootstrap. We perform 1000 simulations of each DGP with $n = 10,000$. Table 3 reports the mean estimates of λ, its standards errors across simulations (in parentheses), and the empirical rejection rate of each test for a nominal size of 5% and 300 bootstrap draws. We show the results for various $iOLS_\delta$ illustrating the results in Figures 2 and 3, $iOLS_{MP}$ and $PPML$. All tests, except $RESET$, evaluates the null hypothesis that the underlying model is correct using the approach laid out in Section 4.1. Finally, the last column shows rejection rates for Ramsey's RESET test for PPML including 3 polynomials terms (Santos Silva and Tenreyro, 2006; Ramsey, 1969).[32]

Table 3: Simulations: Specification testing (Logit)

DGP				iOLS$_\delta$			MP	PPML	RESET
	δ	0.01	0.15	1.9	24	86			
	λ	0.98	0.98	0.99	1.00	1.01	1.00	1.01	
1	(se)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.08)	
	Rej%	88.8	72.3	26.1	5.3	15.7	5.7	8.4	7.3
	λ	1.01	1.04	1.10	1.22	1.29	1.43	Inf	
3	(se)	(0.11)	(0.15)	(0.20)	(0.26)	(0.29)	(0.37)	(Inf)	
	Rej	5.6	66.6	94.2	99.7	100.0	100.0	0.2	5.9

Notes: This table shows the relative rejection frequency of each null hypothesis for 1,000 simulations for $n = 10,000$ and DGP 1 to 3.

[32] https://personal.lse.ac.uk/tenreyro/reset.do.

In DGP 1, the tests for $\text{iOLS}_{\delta=24}$ and iOLS_{MP} reject the null hypothesis only 5.3% and 5.7% of the time which is close to the nominal test size. The test for PPML and the RESET test are slightly oversized with 8.4% and 7.3% rejection rates. The other models are rejected with a larger frequency. In DGP 2, iOLS_{δ} is rejected only 5.6% of the time for $\delta = 0.01$, all other models are vastly rejected. PPML estimates behave poorly leading to a non-informative test under the alternative. Finally, findings reveal that the RESET test lacks power against the considered alternative. Similar results are obtained using KNN (Table C.2).

6 Application

6.1 Michalopoulos and Papaioannou (2013)

We now revisit the analysis of Michalopoulos and Papaioannou (2013), a leading empirical study where the log of zero had to be addressed. They examine the relationship between pre-colonial political centralization and contemporary development in African countries. The latter is proxied using light density at night at the regional level and used as the response variable through the "popular fix": $\log(\text{Y}_i + 0.01)$. The effect of political centralization is measured by the coefficient associated with Murdock's 1967 index of jurisdictional hierarchy.[33] The cross-sectional unit is ethnicity-by-country. They control for population density, location, and geography, as well as country fixed effects.[34]

We first search for the optimal value of δ using the test and model selection procedure developed in Section 4. We estimate the conditional probability function[35] using the logit and kNN algorithm respectively.[36] Figure 4a displays the results, for a grid of candidate δ.[37] The choice of δ matters in practice. As shown in Figure 4b, the estimate for jurisdictional hierarchy, denoted by β, ranges from negative to

[33] Ranging between 0 and 4, it provides the number of jurisdictions above the local level for each ethnicity as reported in 1967. A large number indicates a more centralized political organization.

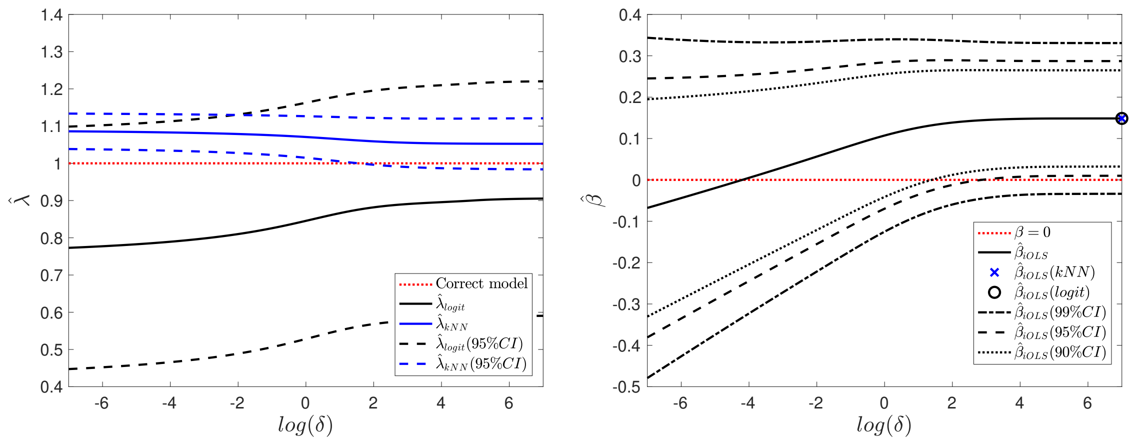
[34] We focus on columns (4) of their Table 3 which include all controls.

[35] We discard the fixed-effects from the probability model to prevent over-fitting and avoid computational difficulties. The results are similar with ethnicity by country fixed effects.

[36] For kNN, we use 100 neighbors and trim observations associated with predicted probabilities below the 5th and above the 95th (empirical) percentiles.

[37] The grid starts at $\exp(-7) \approx 0.001$ with increments of $\exp(0.5)$ up to $\exp(7) \approx 1000$.

positive, as well as non-significant to significant at the 5% level when using standard errors clustered at the ethnicity by country level. The selected δ is large for both probability models and lead to the same estimates of β.



(a) $\hat{\lambda}$ as a function of δ (b) $\hat{\beta}$ as a function of δ

Figure 4: Jurisdictional hierarchy (β) with iOLS$_\delta$ and λ for various candidate δ

Table 4 reports competing estimates of the effect of jurisdictional hierarchy on economic development. Both the author's specification (PF), multiplicative Poisson (iOLS$_{MP}$), and iterated ordinary least squares (iOLS$_\delta$) deliver a positive and statistically significant estimate. In contrast, PPML reports a negative and non-significant point-estimate. The λ's associated with the author's specification and PPML are far from one. Following our model selection procedure, it means that both exhibit poor external validity with respect to the probability of observing a zero. Conversely, λ is close to one for both iOLS$_{\delta \approx 1000}$ and iOLS$_{MP}$, which are statistically the same.

Therefore, the econometric evidence is in favor of a positive and significant effect of political centralization on economic development, as found by the authors.[38] This replication exercise shows that the popular fix, although fundamentally flawed, can at times lead to qualitatively correct results. However, it still delivers artificially lower standard errors by treating zero observations similarly to non-zero observations. We also note that relying solely on PPML can yield misleading results.

[38] In subsequent work, Michalopoulos and Papaioannou (2014) study the link between contemporary political institutions in Africa and economic development. They find an absence of statistical significance using both the popular fix, OLS in level, and PPML (Table 6 in their Appendix).

Table 4: The effect of political centralization

	PF	PPML	iOLS_{MP}	$\text{iOLS}_{\delta=\exp(7)}$
Jurisdictional Hierarchy (β)	0.177***	-0.110	0.149**	0.149**
	(0.0473)	(0.0812)	(0.0634)	(0.0733)
Pop. & Loc. & Geo.	Yes	Yes	Yes	Yes
Country Fixed-Effects	Yes	Yes	Yes	Yes
λ - kNN	-0.163	2.399	1.052	1.052
	(0.029)	(3.768)	(0.034)	(0.035)
p-value	0.000	0.371	0.126	0.134
N	682	682	682	682

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

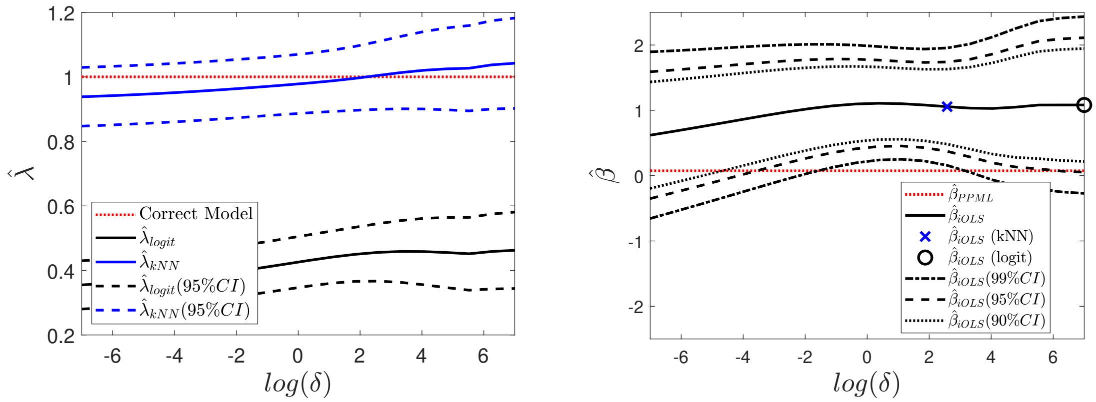
Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors clustered at the country level in parenthesis.

6.2 Santos Silva and Tenreyro (2006)

We now revisit the gravity model of Santos Silva and Tenreyro (2006). Their Table 3 reports models of bilateral trade for data covering 136 countries in 1990. The authors advocate for PPML over a log-linear model arguing that the latter is biased in presence of heteroskedastic errors. For importer (I) and exporter (X) countries, they control for log(Distance), along with dummies for contiguity, shared language, colonial ties, and two variables which, when summed, measure the impact of free trade agreements (FTA) on trade flows.[39] We denote the latter sum as β and focus on its sensitivity to alternative identifying assumptions.

As in the previous illustration, we search for the optimal δ and associated λ, based on both a logit and a kNN estimator. As shown in Figure 5a, The latter provides a series of λ which do not reject the proportionality condition $\lambda = 1$, and leads to select $\delta = 13$ with $\lambda = 1.004$. The measure of trade agreements β evolves from non-significant for small δ's to statistically significant for larger values. In general, those estimates are fairly different from those obtained using PPML.

[39] They also control for log(GDP) and log(GDP per capita), along with a dummies for access to the oceans, remoteness (which measures the access to other trading partners). The estimates for the associated coefficients are available from Table D.5 in the Appendix.



(a) $\hat{\lambda}$ as a function of δ (b) $\hat{\beta}$ as a function of δ

Figure 5: Free Trade Agreement (β) with iOLS$_\delta$ and λ for various candidate δ

Table 5 reports estimates of the effect of Free Trade Agreements for the different models.[40] Whilst the popular fix (with $\Delta = 1$) gives a very large effect of trade agreements on trade flows, PPML's estimate is close to zero and non-significant. For both, λ is far from one, suggesting a lack of external validity of the model with respect to the pattern of zeros in the data. In contrast, iOLS$_{MP}$ and iOLS$_\delta$ deliver economically comparable results with λ close to one. These latter estimates are also close to what the authors obtain using the Tobit approach.

These results illustrate two important effects. On the one hand, the popular fix give too much weight to the zeros by treating them like "normal observations" which artificially inflates the effect of Free Trade Agreements and reduces its standard errors. On the other hand, PPML under-estimates the importance of Free Trade Agreements by putting too much weight on observations associated with large trade values and too little on smaller values, especially the zeros. In contrast, our approach considers a wide range of solutions and allows one to focus on the ones which make sense with respect to the observed zeros.

[40]Table D.5 in Appendix 6 reports the complete results.

Table 5: The effect of Free Trade Agreements

	PF	PPML	iOLS_{MP}	$\text{iOLS}_{\delta=13}$
Free Trade Agreement (β)	2.028***	0.074	1.135***	1.056***
	(0.131)	(0.177)	(0.513)	(0.349)
λ - kNN	0.027	0.449	1.020	1.004
	(0.017)	(0.060)	(0.072)	(0.054)
p-value	0.000	0.000	0.777	0.934
N	18,360	18,360	18,360	18,360

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table displays the coefficient associated with being part of a free trade agreement, standard errors based on 300 pairs bootstrap are in parenthesis.

This application provides a clear illustration of the general result that PPML estimates tend to be driven by observations with large values in Y while neglecting those with small values. In Table 6, we report the estimated effect of trade agreements based on alternative models after discarding observations based on the quantiles of the dependent variable. As expected, the estimates barely change for PPML, even when keeping only the top 10% of trade flows. Comparable results hold for the other explanatory variables (Table D.5). In contrast, all other methods show a decreasing effect when focusing on observations with larger Y. This suggests that Free Trade Agreements have relatively large effects between countries with small trade flows but not for the largest trading nations. In conclusion, both the popular fix and PPML may yield misleading results.

Table 6: Trade Inequality and Estimates of Free Trade Agreements (β)

	PF	PPML	iOLS_{MP}	$\text{iOLS}_{\delta=13}$	Observations
Full Sample	2.028	0.074	1.135	1.048	18360
	(0.131)	(0.177)	(0.513)	(0.349)	
No Zeros	0.325	0.040	0.207	0.141	9613
	(0.109)	(0.179)	(0.216)	(0.151)	
Top 25%	0.155	0.064	-0.036	-0.038	4589
	(0.079)	(0.174)	(0.118)	(0.106)	
Top 10%	-0.045	0.076	-0.248	-0.237	1836
	(0.088)	(0.172)	(0.106)	(0.103)	

Notes: This table displays the coefficient associated with being part of a free trade agreement, based on different subsamples of the data.

7 Conclusion

This paper brought multiple contributions to address a common yet unresolved issue faced in empirical research: the log of zero. Our estimation procedure has several advantages, including: 1) computational simplicity, 2) a natural extension to instrumental variables, 3) robustness to the inclusion of many fixed effects, and 4) their flexibility to exogeneity restrictions. Our testing procedures allows verifying the underlying exogeneity restrictions imposed on the occurrences of zeros. Our replications of leading publications have shown how these tests can guide empirical research.

Hopefully, empirical researchers are now better equipped to address the log of zero and justify their chosen method. The methodology developed in this paper should help find a consensus among practitioners about the best practice to address the log of zero. There are also many possible extensions, like regularized models, which we leave for future research.

References

Bellemare, M. F., and C. J. Wichman. 2020. "Elasticities and the Inverse Hyperbolic Sine Transformation." *Oxford Bulletin of Economics and Statistics*, 82(1): 50–61.

Burbidge, J. B., L. Magee, and A. L. Robb. 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *Journal of the American Statistical Association*, 83(401): 123–127.

Cohn, Jonathan B., Zack Liu, and Malcolm Wardlaw. 2021. "Count Data in Finance." *Available at SSRN.*

Correia, S., P. Guimarães, and T. Zylkin. 2019. "ppmlhdfe: Fast Poisson Estimation with High-Dimensional Fixed Effects." *arXiv e-prints*.

Davidson, Russell, and James G. MacKinnon. 1981. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." *Econometrica: Journal of the Econometric Society*, 781–793.

Dominitz, Jeff, and Robert P. Sherman. 2005. "Some Convergence Theory for Iterative Estimation Procedures with an Application to Semiparametric Estimation." *Econometric Theory*, 21(4): 838–863.

Duan, Naihua. 1983. "Smearing Estimate: a Nonparametric Retransformation Method." *Journal of the American Statistical Association*, 78(383): 605–610.

Eaton, Jonathan, and Akiko Tamura. 1994. "Bilateralism and Regionalism in Japanese and U.S. Trade and Direct Foreign Investment Patterns." *Journal of the Japanese and International Economies*, 8(4): 478 – 510.

Fernández-Val, Iván, and Martin Weidner. 2016. "Individual and Time Effects in Nonlinear Panel Models with Large N, T." *Journal of Econometrics*, 192(1): 291–312.

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica*, 52(3): 701–20.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. "Prototype Methods and Nearest-Neighbors." *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 459–483. New York, NY:Springer New York.

Head, Keith, and Thierry Mayer. 2014. "Gravity Equations: Workhorse,Toolkit, and Cookbook." In *Handbook of International Economics*. Vol. 4 of *Handbook of International Economics*, 131–195. Elsevier.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the Econometric Society*, 153–161.

MacKinnon, James G, and Lonnie Magee. 1990. "Transforming the Dependent Variable in Regression Models." *International Economic Review*, 315–339.

Manning, Willard G. 1998. "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem." *Journal of Health Economics*, 17(3): 283–295.

Manning, Willard G, and John Mullahy. 2001. "Estimating Log Models: to Transform or not to Transform?" *Journal of Health Economics*, 20(4): 461–494.

Michalopoulos, Stelios, and Elias Papaioannou. 2013. "Pre-Colonial Ethnic Institutions and Contemporary African Development." *Econometrica*, 81(1): 113–152.

Michalopoulos, Stelios, and Elias Papaioannou. 2014. "National Institutions and Subnational Development in Africa." *The Quarterly Journal of Economics*, 129(1): 151–213.

Mullahy, John. 1997. "Instrumental-Variable Estimation of Count Data models: Applications to Models of Cigarette Smoking Behavior." *Review of Economics and Statistics*, 79(4): 586–593.

Ramsey, James Bernard. 1969. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2): 350–371.

Ravallion, M. 2017. "A Concave Log-Like Transformation Allowing Non-Positive Values." *Economics Letters*, 161: 130 – 132.

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

Santos Silva, João M.C., and Silvana Tenreyro. 2010. "On the Existence of the Maximum Likelihood Estimates in Poisson Regression." *Economics Letters*, 107(2): 310 – 312.

Santos Silva, João M.C., and Silvana Tenreyro. 2011. "Poisson: Some Convergence Issues." *The Stata Journal*, 11(2): 207 – 212.

Santos Silva, João M.C., Silvana Tenreyro, and Frank Windmeijer. 2015. "Testing Competing Models for Non-negative Data with Many Zeros." *Journal of Econometric Methods*, 4(1): 29 – 46.

Sricharan, Kumar, Raviv Raich, and Alfred O Hero. 2010. "Boundary Compensated k-NN Graphs." *2010 IEEE International Workshop on Machine Learning for Signal Processing*, 277–282.

Weidner, Martin, and Thomas Zylkin. 2021. "Bias and Consistency in Three-Way Gravity Models." *Journal of International Economics*, 132: 103513.

Whitney, K. Newey, and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." *Handbook of Econometrics: Chapter 36*, 2111–2245.

Winkelmann, Rainer. 2008. *Econometric Analysis of Count Data.* Springer Science

& Business Media.

Wooldridge, Jeffrey M. 1997. "Quasi-Likelihood Methods for Count Data." *Handbook of Applied Econometrics*, 2: 352–406.

Wooldridge, Jeffrey M. 2015. "Control Function Methods in Applied Econometrics." *Journal of Human Resources*, 50(2): 420–445.

Young, K. H., and L. Y. Young. 1975. "Estimation of Regressions Involving Logarithmic Transformation of Zero Values in the Dependent Variable." *The American Statistician*, 29(3): 118–120.

Appendix

Proof 1 (Proof of Theorem 1) *Recall that the parameter $\beta \in \mathbb{R}^K$ is characterized by the fixed-point equation*

$$\beta = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\beta)\right], \tag{38}$$

where $\tilde{Y}_i(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c(\beta, \delta)$ is the transformed dependent variable. To simplify exposition, we focus on $\delta = 1$ in our calculations. The mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$ which characterizes the parameter is hence defined $\forall \phi \in \mathbb{R}^K$ as

$$M(\phi) = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\phi)\right]. \tag{39}$$

The sample counterpart of this mapping is given by

$$\hat{M}_n(\phi) = [X'X]^{-1} X'\hat{\tilde{Y}}_i(\phi), \tag{40}$$

where $\hat{\tilde{Y}}_i(\phi) = \log(Y_i + \exp(X_i'\phi)) - \hat{c}(\phi)$, with $\hat{c}(\phi) = \frac{1}{n}\sum_{i=1}^n \log(Y_i + \exp(\hat{\phi}_1(\phi) - \phi_1 + X_i'\phi)) - \log(\frac{1}{n}\sum_{i=1}^n (\hat{\phi}_1(\phi) - \phi_1 + X_i'\phi))$ for $\hat{\phi}_1(\phi) = \log(n^{-1}\sum_{i=1}^n Y_i \exp(-X_i\phi + \phi_1))$

Our proof follows Dominitz and Sherman (2005), hereafter denoted DS, who develop a convergence theory for iterative estimators. Following DS, the convergence of iOLS requires that $M(\cdot)$ and $\hat{M}_n(\cdot)$ be contraction mappings, asymptotically.[41]

[41]The reader is referred to DS for a formal definition of an asymptotic contraction mapping.

In order to show the convergence result $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ as $n \to \infty$ by applying Theorem 1 in DS, we need to show that the following conditions hold:

(i) $\left\{\hat{M}_n(\cdot) : n \geq 1, \omega \in \mathcal{S}\right\}$ *is an asymptotic contraction mapping on* (B_0, E_K), *where* $\mathcal{S}$ *is a sample space,* E_K *is the Euclidean metric on* $\mathcal{R}^K$ *and* B_0 *is the closed ball centered at* β_0 *of radius* $|\hat{\beta}_0 - \beta|$;[42]

(ii) $n^{1/2}|\beta_{t(n)} - \beta| = O_p(1)$ *as* $n \to \infty$;

(iii) $n^{1/2} \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M(\phi)| = O_p(1)$ *as* $n \to \infty$; *and*

(iv) $\sup_{\phi \in B_0} ||\hat{V}_n(\phi) - V(\phi)|| = o_p(1)$ *as* $n \to \infty$.

Regularity conditions. *Our proofs rely on the sufficient regularity conditions listed below, in particular for showing the uniform convergence in conditions (iii) and (iv): (1)* $E[X_i] < \infty$; *(2)* $V[X_i] < \infty$; *(3)* $E[X_i \log(Y_i + \exp(X_i'\phi))] < \infty,\ \forall \phi \in B_0$; *(4)* $V[X_i \log(Y_i + \exp(X_i'\phi))] < \infty,\ \forall \phi \in B_0$; *(5)* $c(\phi) < \infty,\ \forall \phi \in B_0$; *(6)* $V[\hat{c}(\phi)] < \infty,\ \forall \phi \in B_0$; *(7)* $E\left[X_i \frac{\exp(X_i'\phi)}{Y_i+\exp(X_i'\phi)} X_i'\right] < \infty;\ \forall \phi \in B_0$; *(8)* $V\left[X_i \frac{\exp(X_i'\phi)}{Y_i+\exp(X_i'\phi)} X_i'\right] < \infty,\ \forall \phi \in B_0$; *(9)* $\nabla_\phi c(\phi) < \infty,\ \forall \phi \in B_0$; *and (10)* $V[\nabla_\phi \hat{c}(\phi)] < \infty,\ \forall \phi \in B_0$.

Condition (i). *Let us adapt the proof of Lemma 5 in DS as follows. The first step is to consider that X is prewhitened so that $X'X = nI_k$. This assumption is useful to establish the local contraction mapping property without loss of generality. From a multivariate Taylor expansion argument, DS show that condition (i) boils down to showing that the largest eigenvalue of $\nabla_\phi \hat{M}_n(\beta) = \hat{V}_n(\beta)$ is strictly less than unity as $n \to \infty$. Note that we have*

$$\begin{aligned} \hat{V}_n(\phi) &= [X'X]^{-1} X' \nabla_\phi \hat{\hat{Y}}(\phi) \\ &= n^{-1} X' \nabla_\phi \hat{\hat{Y}}(\phi), \end{aligned} \tag{41}$$

where the second equality uses prewhitening and $\nabla_\phi \hat{\hat{Y}}_i(\phi)$ has element (i,k) defined as

$$\left[\nabla_\phi \hat{\hat{Y}}(\phi)\right]_{i,k} = \frac{\exp(X_i'\phi) X_{ki}}{Y_i + \exp(X_i'\phi)} - \frac{\partial \hat{c}(\phi)}{\partial \phi_k}. \tag{42}$$

Let us denote $X_{1i} = 1$, for all i as the constant. By prewhitening, we have $\sum_{j=1}^n X_{1j} = n$ and $\sum_{j=1}^n X_{kj} = 0$ for $k > 1$.

---

[42]Note that DS's condition (i) is about $M(\cdot)$ and not $\hat{M}_n(\cdot)$, with $M(\cdot)$ being non-stochastic in our setting. However those conditions imply each other under conditions (iii) and (iv) by applying their Lemma 3 with trivial modifications.

$$\frac{\partial \hat{c}(\phi)}{\partial \phi_k} = n^{-1}\sum_{i=1}^{n} \frac{\exp(X_i^{r\prime}\phi^r + \hat{\phi}^1)(\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki})}{Y_i + \exp(X_i^{r\prime}\phi^r + \hat{\phi}^1)} - n^{-1}\sum_{i=1}^{n}(\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki}), \tag{43}$$

for $k > 1$ and $\frac{\partial \hat{c}(\phi)}{\partial \phi_1} = 0$. This expression simplifies when evaluated at $\phi = \beta$, as shown by

$$\frac{\partial \hat{c}(\beta)}{\partial \phi_k} = n^{-1}\sum_{i=1}^{n} \frac{X_{ki}}{1+U_i} + O_p(1), \tag{44}$$

for $k > 1$ because $\hat{\phi}^1(\beta) = \log(n^{-1}\sum_{i=1} Y_i \exp(-X_i^{r\prime}\beta^r)) = \beta_1 + \log(n^{-1}\sum_{i=1} U_i)$, where $\log(n^{-1}\sum_{i=1} U_i) = O_p(1)$ by iid assumption and $E[U_i] = 1$, and $n^{-1}\sum_{i=1}^{n} X_{ki} = 0$ by prewhitening. Thus, we have $\frac{\partial \hat{\phi}_1(\beta)}{\partial \phi_k} = 0$.

Therefore, each element (k,l) of $\hat{V}_n(\beta)$ writes

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} \frac{X_{ki}X_{li}}{1+U_i} - n^{-1}\sum_{i=1}^{n} X_{ki} n^{-1}\sum_{j=1}^{n} \frac{X_{lj}}{1+U_j}, \tag{45}$$

for $l > 1$ and

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} \frac{X_{ki}}{1+U_i}, \tag{46}$$

for $l = 1$. Remark that for $k = 1, \forall l > 1$ we have $[V_n(\beta)]_{1,l} = 0$, and for $k = 1, l = 1$, we have $\left[\hat{V}_n(\beta)\right]_{1,1} = n^{-1}\sum_{i=1}^{n} \frac{1}{1+U_i} < 1$. Therefore, the eigenvalue associated with the constant term is strictly below 1, and proving the convergence amounts to showing that the largest eigenvalue of the $(K-1)\times(K-1)$ lower right-hand submatrix of $\hat{V}_n(\beta)$ is strictly less than unity. All elements (k,l) for $k,l > 1$ of this matrix writes

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} \frac{X_{ki}X_{li}}{1+U_i}. \tag{47}$$

because of prewhitening. We can write this in matrix form as

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'WX, \tag{48}$$

where W is a diagonal matrix with elements (i,i) acting as weights given by $\frac{1}{1+U_i} \in$

$(0,1]$. *Note that those weights become* $\frac{\delta}{\delta+U_i} \in [0,1)$ *for* $\delta \neq 1$. *We can thus write* $W = W^{1/2}W^{1/2}$, *and rewrite the submatrix of interest as the quadratic form*

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'W^{1/2}W^{1/2}X. \tag{49}$$

Consequently, this matrix is nonnegative definite and so must have all nonnegative eigenvalues. We can alternatively write the weight matrix $W = I_n - D$, *where* D *is also a diagonal matrix with elements* $\frac{U_i}{1+U_i} \in [0,1)$, *or more generally* $\frac{U_i}{\delta+U_i} \in [0,1)$. *Therefore, we have the alternative expression*

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'(I_n - D)X = I_{K-1} - n^{-1}X'D^{1/2}D^{1/2}X, \tag{50}$$

where the second term is also a quadratic form. It follows that as $n \to \infty$, *the maximum eigenvalue is equal to*

$$\max_{|a|=1} a' \left[\hat{V}_n(\beta)\right]_{k,l>1} a = \max_{|a|=1} 1 - a'X'D^{1/2}D^{1/2}Xa. \tag{51}$$

Assuming the data distribution is non-degenerate, $a'X'D^{1/2}D^{1/2}Xa$ *is positive and bounded away from zero for all unit vectors* $a \in \mathcal{R}^{K-1}$. *Thus, as* $n \to \infty$, *the maximum eigenvalue of* $\hat{V}_n(\beta)$ *is strictly less than unity. This proves the result.*

Condition (ii). *We want to show that* $M(\cdot)$ *is a contraction mapping with fixed-point* β. *Following DS, a sufficient condition to satisfy (ii) for contraction mappings exhibiting linear convergence is* $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, *where* $\kappa \in [0,1)$ *is the modulus of the contraction* $M(\cdot)$, *which can be estimated as the mean or median of* $\hat{\kappa} = |\hat{\beta}_{t+1} - \hat{\beta}_t|/|\hat{\beta}_t - \hat{\beta}_{t-1}|$ *across several iterations. We must hence show that* $M(\cdot)$ *is a contraction mapping converging linearly to* β.

First, let us show that β *is a fixed-point of* M. *We have*

$$M(\beta) = E[X_iX_i']^{-1}E\left[X_i\tilde{Y}_i(\beta)\right], \tag{52}$$

from which substituting $\tilde{Y}_i(\beta)$ *yields*

$$M(\beta) = E[X_iX_i']^{-1}E\left[X_i(X_i'\beta + \overline{v}_i)\right]. \tag{53}$$

Rearranging and making use of Assumption 3 gives

$$M(\beta) = \beta. \tag{54}$$

Let us now show that M is a contraction mapping exhibiting linear convergence. Letting β_t be the parameter after t iterations, we have

$$\beta_{t+1} - \beta = M(\beta_t) - \beta = M(\beta_t) - M(\beta), \tag{55}$$

because $M(\beta) = \beta$ by definition. By the mean value theorem, there is a b_t between β_t and β satisfying $\beta_{t+1} - \beta = M(\beta_t) - M(\beta) = (\beta_t - \beta)V(b_t)$, where $V(\cdot)$ denotes the gradient. Let $e_t = ||\beta_t - \beta||$, where $||\cdot||$ denotes the sup norm, and thus $e_{t+1} = e_t||V(c_t)||_{op}$, with $||\cdot||_{op}$ denoting the operator version of the sup norm. A standard algebra result and the symmetry of the matrix $V(\beta)$, composed of $K \times L$ elements $E[\delta X_{ki}X_{li}/(\delta + U_i)]$, imply that $||V(\beta)||_{op}$ is bounded by the largest eigenvalue of $V(\beta)$. Using similar derivations than for condition (i), or by applying the limit as $n \to \infty$, we have that $||V(\beta)||_{op} < 1$. Therefore, by the continuity of $V(\cdot)$ there is a small neighborhood around β for which

$$||V(\beta)||_{op} < \frac{\kappa + 1}{2} < 1. \tag{56}$$

If β_t lies in this neighborhood, then so does c_t. Therefore, we have $||e_{t+1}|| \leq \frac{\kappa+1}{2}||e_t||$, and $\lim_{t\to\infty} \frac{||e_{t+1}||}{||e_t||} = \lim_{t\to\infty} ||V(c_t)||_{op} = ||V(\beta)||_{op} = \kappa < 1$, which provides the desired result.

Condition (iii). *We now want to show that $\hat{M}_n$ converges uniformly to M, i.e. $n^{1/2}\sup_{\phi\in B_0} |\hat{M}_n(\phi) - M(\phi)| = O_p(1)$ as $n \to \infty$.*

For any $\phi \in B_0$, recall that $\hat{M}_n(\phi) = X'X^{-1}X'\hat{\tilde{Y}}_i(\phi)$. Under the iid assumption and assuming $E[X_iX_i'] < \infty$, applying the weak law of large numbers and Slutsky's theorem yield $n^{-1}X'X^{-1} \xrightarrow{p} E[X_iX_i']^{-1}$ and $\hat{c}(\phi) \xrightarrow{p} c(\phi)$ as $n \to \infty$, and thus $n^{-1}X'\hat{\tilde{Y}}_i(\phi) \xrightarrow{p} E\left[X_i\tilde{Y}_i(\phi)\right]$ as $n \to \infty$. Therefore, $\hat{M}_n(\phi) \xrightarrow{p} M(\phi)$ as $n \to \infty$ and the Lindeberg-Levy's central limit theorem gives $|\hat{M}_n(\phi) - M(\phi)| = O_p(n^{-1/2})$ for any $\phi \in B_0$. To show uniform convergence, let us recall that B_0 is a closed ball in a Euclidean space and so is compact. We obtain the following inequality

$$
\begin{aligned}
|\hat{M}_n(\phi) - M(\phi)| \leq & |n^{-1}\sum_{i=1}^{n} X_i \log(Y_i + \exp(X_i'\phi)) - E[X_i \log(Y_i + \exp(X_i'\phi))]| \\
& + |n^{-1}\sum_{i=1}^{n} X_i \hat{c}(\phi) - E[X_i]c(\phi)| \\
\leq & |n^{-1}\sum_{i=1}^{n} X_i \log(Y_i + \exp(X_i'\phi_l)) - E[X_i \log(Y_i + \exp(X_i'\phi_l))]| \\
& + |n^{-1}\sum_{i=1}^{n} X_i - E[X_i]||\hat{c}(\phi_u)| + |E[X_i]||\hat{c}(\phi_u) - c(\phi_u)|
\end{aligned} \tag{57}
$$

where the first inequality follows from prewhitening and the triangular inequality; the second inequality follows from the compactness of B_0, by which there exist ϕ_u and ϕ_l in the parameter set such that the inequality holds, and by the triangular inequality. All three terms on the right-hand-side (RHS) are finite, and consist in averages of zero-mean iid random variables with finite first and second moments by assumption, and thus have order $O_p(n^{-1/2})$. We deduce the uniform convergence result from the continuity of $\hat{M}_n(\phi)$ and $M(\phi)$ in ϕ by applying Lemma 2.4 in Whitney and McFadden (1994).

Condition (iv). *Let us use the derivations obtained earlier and similar arguments than for condition (iii). We have that $\nabla_\phi \hat{c}(\phi) \xrightarrow{p} \nabla_\phi c(\phi)$ and thus $\hat{V}_n(\phi) \xrightarrow{p} V(\phi)$ as $n \to \infty$. Therefore, the condition $||\hat{V}_n(\phi) - V(\phi)|| = o_p(1)$ holds. Uniform convergence follows from similar derivations to obtain*

$$
\begin{aligned}
||\hat{V}_n(\phi) - V(\phi)|| \leq & ||n^{-1}\sum_{i=1}^{n} X_i \frac{\exp(X_i'\phi_l)}{Y_i + \exp(X_i'\phi_l)} X_i' - E[X_i \frac{\exp(X_i'\phi_l)}{Y_i + \exp(X_i'\phi_l)} X_i']|| \\
& + |n^{-1}\sum_{i=1}^{n} X_i - E[X_i]| \cdot ||\nabla_\phi \hat{c}(\phi_u)|| + |E[X_i]| \cdot ||\nabla_\phi \hat{c}(\phi_u) - \nabla_\phi c(\phi_u)||,
\end{aligned} \tag{58}
$$

where all three terms on the RHS are finite and have finite first and second moments by assumption. All conditions being satisfied, we apply Theorem 1 in DS to obtain the convergence result $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ as $n \to \infty$.

Proof 2 (Proof of Theorem 1: Normality) *We now make use of Theorem 4 in DS to derive the asymptotic distribution of iOLS. All conditions have been verified in the previous results except that* $\sqrt{n}(\hat{M}_n(\beta)-\beta) \xrightarrow{d} Z$ *as* $n \to \infty$*, where* Z *is a limit distribution. Note that we have*

$$\hat{c}(\beta) = n^{-1}\sum_{i=1}^{n}\log(n^{-1}\sum_{j=1}^{n}U_j + U_i) - \log(n^{-1}\sum_{j=1}^{n}U_j) \xrightarrow{p} E[\log(1+U_i)] = c, \quad (59)$$

as $n \to \infty$*, and* $\hat{\tilde{Y}}_i(\beta) = \log(1+U_i) + X_i'\beta - \hat{c}(\beta)$*, so that*

$$\sqrt{n}[X'X]^{-1}X'\hat{\tilde{Y}}_i(\beta) = \sqrt{n}\left(\beta + [X'X]^{-1}X'(\log(1+U) - \hat{c}(\beta))\right). \quad (60)$$

Under the iid assumption and the exogeneity condition $E[X_i\log(1+U_i)] = c$*, the Lindeberg-Levy's central limit theorem yields*

$$\sqrt{n}\left(\hat{M}_n(\beta) - \beta\right) \xrightarrow{d} \mathcal{N}(0,\Sigma), \quad (61)$$

as $n \to \infty$*, where* Σ *is the asymptotic covariance matrix. Remark that it is the asymptotic covariance of the OLS estimator of the regression of* $\hat{\tilde{Y}}(\beta)$ *onto* X*. Heteroskedasticity-robust estimators and alike apply exactly as in the standard OLS setting. However, the iOLS estimator has a slightly different asymptotic distribution. Theorem 4 of DS 2005 gives* $\sqrt{n}\left(\hat{\beta}_{i(n)} - \beta\right) \xrightarrow{d} \mathcal{N}(0,\Omega^{-1})$*, as* $n \to \infty$*, where* $\Omega = (I_k - V(\beta))^{-1}\Sigma(I_K - V(\beta))$ *and the gradient* $\nabla_\phi M(\beta) = V(\beta)$ *is defined as*

$$V(\beta) = E[X_iX_i']^{-1}E[\frac{X_iX_i'}{1+U_i}], \quad (62)$$

of which each element is strictly below 1. Therefore sandwich-type covariance estimators are changed from the classical expression $\hat{\Sigma} = (\frac{1}{n}X'X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'X)^{-1}$ *to*

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \quad (63)$$

where W *is a diagonal weighting matrix with diagonal element* $\frac{1}{1+U_i}$*, and* $\hat{\Sigma}_0$ *is an estimator of the covariance of* $X_i'(\log(1+U_i) - c)$ *across observations. For another* $\delta \neq 1$*, we would have the weights* $\frac{\delta}{\delta+U_i} \in [0,1)$*. In layman's terms, the "meat"* *of HAC-robust estimators is unchanged but the "bread" is modified. As before, the weights become* $\frac{\delta}{\delta+U_i}$ *when* $\delta \neq 1$.

Proof 3 (Proof of Theorem 2: Specification tests) *Let us start by assuming some regularity conditions to guarantee existence:* $E(W_i^2|U_i > 0) < \infty$, $E(W_i^2\nu_i^2|U_i > 0) < \infty$, $E(W_i^2U_i^2|U_i > 0) < \infty$, $E(\varepsilon_i^{WU^2}|U_i > 0) < \infty$. *We decompose* $\hat{\lambda} - 1$ *into*

$$\begin{aligned}\hat{\lambda} - 1 =& \tilde{\lambda} - 1 + \left(n_1^{-1}\sum_{i=1}^{n_1} W_i^2\right)^{-1}\left(\sum_{j=1}^{n_1}\hat{W}_j\hat{U}_j - \sum_{j=1}^{n_1} W_jU_j\right) \\ &+ \left[\left(n_1^{-1}\sum_{i=1}^{n_1}\hat{W}_i^2\right)^{-1} - \left(n_1^{-1}\sum_{i=1}^{n_1} W_i^2\right)^{-1}\right]\left(\sum_{j=1}^{n_1}\hat{W}_j\hat{U}_j\right),\end{aligned} \tag{64}$$

where $\tilde{\lambda} = \left(n_1^{-1}\sum_{i=1}^{n_1} W_i^2\right)^{-1}\left(\sum_{j=1}^{n_1} W_jU_j\right)$ *is the OLS estimator in absence of first-step errors.*

In what follows, we omit the conditioning of expectations are conditioned on $U_i > 0$. *If first-step estimators are consistent, and by* $E(W_i^2) < \infty$, *which implies* $E(W_iU_i|U_i > 0) < \infty$ *under* H_0, *the weak law of large numbers for iid observations implies that* $n_1^{-1}\sum_{i=1}^{n_1}\hat{W}_i^2 \xrightarrow{p} E(W_i^2)$, $n_1^{-1}\sum_{i=1}^{n_1}\hat{W}_i\hat{U}_i \xrightarrow{p} E(W_iU_i)$ *as* $n_1 \to \infty$, $n_1^{-1}\sum_{i=1}^{n_1} W_i^2 \xrightarrow{p} E(W_i^2)$, *and* $n_1^{-1}\sum_{i=1}^{n_1} W_iU_i \xrightarrow{p} E(W_iU_i)$ *as* $n_1 \to \infty$. *Moreover,* H_0 *implies* $E[\nu_i|U_i > 0, X_i] = 0$ *which in turn yields* $\tilde{\lambda} \xrightarrow{p} 1$ *as* $n_1 \to \infty$, *leading to the consistency result:* $\hat{\lambda} \xrightarrow{p} 1$.

Let us now characterize the asymptotic distribution of $\tilde{\lambda}$. *We decompose* $\hat{W}_i\hat{U}_i = W_iU_i + \varepsilon_i^{WU}$, *where the estimation error* ε_i^{WU} *is* $o_p(1)$. *Applying the Lindeberg-Levy's central limit theorem given the iid assumption and finite second-order moments* $E(W_i^2\nu_i^2) < \infty$, $E(\varepsilon_j^{WU^2}) < \infty$ *imply that the first two terms in (64) are asymptotically normal as* $n_1 \to \infty$. *In particular, the firm term satisfies*

$$\sqrt{n}\left(\tilde{\lambda} - 1\right) \xrightarrow{d} N(0, \sigma_{\lambda_0}^2), \tag{65}$$

as $n_1 \to \infty$ *with* $\sigma_{\lambda_0}^2 = E(W_i^2)^{-2}E(W_i^2\nu_i^2)$. *The second term in (64) satisfies*

$$\left(n_1^{-1}\sum_{i=1}^{n_1} W_i^2\right)^{-1}\sqrt{n}\sum_{j=1}^{n_1}\varepsilon_j^{WU} \xrightarrow{d} N(0, \sigma_{\lambda_1}^2), \tag{66}$$

as $n_1 \to \infty$ *with* $\sigma^2_{\lambda_1} = E(W_i^2)^{-2} E(\varepsilon_j^{WU^2})$*. Finally, the third term satisfies*

$$\left[\left(n_1^{-1} \sum_{i=1}^{n_1} \hat{W}_i^2 \right)^{-1} - \left(n_1^{-1} \sum_{i=1}^{n_1} W_i^2 \right)^{-1} \right] \sqrt{n} \sum_{j=1}^{n_1} \hat{W}_j \hat{U}_j \xrightarrow{p} 0, \tag{67}$$

as $n_1 \to \infty$*, because* $\sqrt{n} \sum_{j=1}^{n_1} \hat{W}_j \hat{U}_j$ *is stochastically bounded under* $E(W_i^2 U_i^2 | U_i > 0) < \infty$ *and the term in square bracket converges to zero in probability with* n_1*. Therefore, assuming* $W_i \nu_i$ *and* ε_i^{WU} *to be independent random variables, conditionally on* $U > 0$*, we have*

$$\sqrt{n} \left(\hat{\lambda} - 1 \right) \xrightarrow{d} N(0, \sigma^2_\lambda), \tag{68}$$

where $\sigma^2_\lambda = E(W_i^2)^{-2} E(W_i^2 \nu_i^2) + E(W_i^2)^{-2} E(\varepsilon_j^{WU^2})$*, and the test statistic (35) is asymptotically standard normal provided that* $\hat{\sigma}_\lambda$ *is a consistent estimator of* σ_λ*.*

Online Appendix

A Mathematical Appendix

A.1 Theorem 3: Poisson models estimated with iOLS

Proof 4 (Proof of Theorem 3: iOLS$_{MP}$) *This proof is similar to that of Theorem 1, with small modifications. Regularity conditions must be modified accordingly.*

Let us now consider

$$\begin{aligned}\hat{V}_n(\phi) &= [X'X]^{-1}X'\nabla_\phi \hat{\tilde{Y}}(\phi)\\ &= n^{-1}X'\nabla_\phi \hat{\tilde{Y}}(\phi),\end{aligned} \tag{69}$$

where $\nabla_\phi \hat{\tilde{Y}}_i(\phi)$ *has element* (i,k) *defined as*

$$\left[\nabla_\phi \hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\delta \exp(X_i'\phi)X_{ki}}{Y_i + \delta \exp(X_i'\phi)} + \frac{\partial \hat{U}_i(\phi)}{\partial \phi_k}\left(\frac{1}{1+\delta} - \frac{1}{\hat{U}_i(\phi)+\delta}\right). \tag{70}$$

This expression simplifies, when evaluated at $\phi = \beta$, *to*

$$\left[\nabla_\beta \hat{\tilde{Y}}(\beta)\right]_{i,k} = X_{ki}\left(1 - \frac{U_i}{1+\delta}\right), \tag{71}$$

which yields

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} X_{ki}X_{li}\left(1 - \frac{U_i}{1+\delta}\right). \tag{72}$$

Following the same reasoning as in the previous theorem, a sufficient condition for convergence is that $\frac{U_i}{1+\delta}$ *is between* 0 *and* 1 *for all* i. *Therefore, the choice of* δ *will affect both the speed of convergence and whether the estimator converges at all. An efficient strategy for choosing* δ *is to start at a relatively small value and increment it if convergence fails – which can be checked by estimating* κ *as explained above.*

The proof of asymptotic normality is also unchanged, except that now the diagonal weighting matrix W *in*

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{73}$$

has element $1 - \frac{U_i}{1+\delta}$, *and* $\hat{\Sigma}_0$ *is an estimator of the covariance of* $X_i'U_i$ *across obser-*

vations.

Proof 5 (Proof of Theorem 3: iOLS$_{AP}$)) *This proof follows the same lines, with small modifications to the previous one. The gradient $\nabla_\phi \hat{\tilde{Y}}_i(\phi)$ has now element (i,k) defined as*

$$\left[\nabla_\phi \hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\delta \exp(X_i'\phi) X_{ki}}{Y_i + \delta \exp(X_i'\phi)} - \frac{1}{\hat{U}_i(\phi)+\delta}\frac{\partial \hat{U}_i(\phi)}{\partial \phi_k} + \frac{1}{1+\delta}\frac{\partial (Y_i - \exp(X_i'\phi))}{\partial \phi_k}. \tag{74}$$

This expression simplifies, when evaluated at $\phi = \beta$, to

$$\left[\nabla_\beta \hat{\tilde{Y}}(\beta)\right]_{i,k} = X_{ki}\left(1 - \frac{\exp(X_i'\beta)}{1+\delta}\right), \tag{75}$$

which yields

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} X_{ki}X_{li}\left(1 - \frac{\exp(X_i'\beta)}{1+\delta}\right). \tag{76}$$

Following the same reasoning as in the previous theorem, a sufficient condition for convergence is that $\frac{\exp(X_i'\beta)}{1+\delta}$ is between 0 and 1 for all i. We suggest using the same trial and error approach based on estimating κ.

The proof of asymptotic normality is also unchanged, except that now the diagonal weighting matrix W in

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{77}$$

has element $1 - \frac{\exp(X_i'\beta)}{1+\delta}$, and $\hat{\Sigma}_0$ is an estimator of the covariance of $X_i'\epsilon_i$ across observations.

A.2 Theorem 4: i2SLS

Proof 6 (Proof of Theorem 4 : Instrumental Variables Consistency) *Recall that the parameter $\beta \in \mathbb{R}^K$ is characterized by the fixed-point equation*

$$\beta^{IV} = E[\breve{X}_i\breve{X}_i{}']^{-1}E\left[\breve{X}_i\tilde{Y}_i(\beta)\right], \tag{78}$$

where $\breve{X} = P^Z X$, $P^Z = Z(Z'Z)^{-1}Z'$, $Z \in \mathbb{R}^M$ *with* $M \geq K$, $E(Z_i' X_i)$ *has rank* K, *and* $\tilde{Y}_i(\beta) = \log(Y_i + \exp(X_i'\beta)) - c(\beta)$ *is the transformed dependent variable. The mapping from* $\mathbb{R}^K$ *to* $\mathbb{R}^K$ *which characterizes the parameter is hence defined* $\forall \phi \in \mathbb{R}^K$ *as*

$$M^{IV}(\phi) = E[\breve{X}_i \breve{X}_i{}']^{-1} E\left[\breve{X}_i \tilde{Y}_i(\phi)\right]. \tag{79}$$

The sample counterpart of this mapping is given by

$$\hat{M}_n^{IV}(\phi) = [\breve{X}_i{}' \breve{X}_i]^{-1} \breve{X}_i{}' \hat{\tilde{Y}}_i(\phi), \tag{80}$$

where $\hat{\tilde{Y}}_i(\phi)$ *is defined as before.*

Our proof is very similar to the one used to show Theorem 1. We do not state the modified regularity conditions and only focus on showing condition (i) because the others consist in simple extensions. For condition (i), the first step is to consider that Z *is standardized so that* $\breve{X}$ *is prewhitened:* $\breve{X}'\breve{X} = nI_k$*. As before, showing condition (i) boils down to showing that the largest eigenvalue of* $\nabla_\phi \hat{M}_n^{IV}(\beta) = \hat{V}_n^{IV}(\beta)$ *is strictly less that unity as* $n \to \infty$*. Note that we have*

$$\begin{aligned} \hat{V}_n^{IV}(\phi) &= [\breve{X}'\breve{X}]^{-1} \breve{X}' \nabla_\phi \hat{\tilde{Y}}(\phi) \\ &= n^{-1} \breve{X}' \nabla_\phi \hat{\tilde{Y}}(\phi), \end{aligned} \tag{81}$$

where the second equality uses prewhitening on $\breve{X}$*. Moreover,* $\nabla_\phi \hat{\tilde{Y}}_i(\phi)$ *has element* (i,k) *defined as*

$$\left[\nabla_\phi \hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\exp(X_i'\phi) X_{ki}}{Y_i + \exp(X_i'\phi)} - \frac{\partial \hat{c}(\phi)}{\partial \phi_k}. \tag{82}$$

Let us denote $X_{1i} = 1$ *and* $Z_{1i} = 1$*, for all* i *as the constant. By prewhitening* $\breve{X}$*, we have* $\sum_{j=1}^n \breve{X}_{1j} = n$ *and* $\sum_{j=1}^n \breve{X}_{kj} = 0$ *for* $k > 1$*. The derivative of the nuisance parameter estimate writes*

$$\frac{\partial \hat{c}(\phi)}{\partial \phi_k} = n^{-1} \sum_{i=1}^n \frac{\exp(X_i^{r\prime} \phi^r + \hat{\phi}^1)(\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki})}{Y_i + \exp(X_i^{r\prime} \phi^r + \hat{\phi}^1)} - n^{-1} \sum_{i=1}^n (\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki}), \tag{83}$$

for $k > 1$ *and* $\frac{\partial \hat{c}(\phi)}{\partial \phi_1} = 0$*. As before, this expression simplifies when evaluated at*

$\phi = \beta$, as shown by

$$\begin{aligned}\frac{\partial \hat{c}(\beta)}{\partial \phi_k} =& n^{-1}\sum_{i=1}^{n}\frac{X_{ki}}{1+U_i} - n^{-1}\sum_{i=1}^{n}X_{ki} + O_p(1)\\ =& n^{-1}\sum_{i=1}^{n}\frac{X_{ki}U_i}{1+U_i} + O_p(1),\end{aligned} \tag{84}$$

for $k > 1$ because $\hat{\phi}^1(\beta) = \log(n^{-1}\sum_{i=1} Y_i \exp(-X_i^{r\prime}\beta^r)) = \beta_1 + \log(n^{-1}\sum_{i=1} U_i)$, where $\log(n^{-1}\sum_{i=1} U_i) = O_p(1)$ by iid assumption and $E[U_i] = 1$.

Therefore, each element (k,l) of $\hat{V}_n^{IV}(\beta)$ writes

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n}\frac{\breve{X}_{ki}X_{li}}{1+U_i} - (n^{-1}\sum_{i=1}^{n}\breve{X}_{ki})(n^{-1}\sum_{j=1}^{n}\frac{X_{lj}U_j}{1+U_j}), \tag{85}$$

for $l > 1$ and

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n}\frac{\breve{X}_{ki}}{1+U_i}, \tag{86}$$

for $l = 1$. Remark that for $k = 1, \forall l > 1$ we have $\left[V_n^{IV}(\beta)\right]_{1,l} = n^{-1}\sum_{i=1}^{n}\frac{X_{li}}{1+U_i}$, and for $k = 1, l = 1$, we have $\left[\hat{V}_n^{IV}(\beta)\right]_{1,1} = n^{-1}\sum_{i=1}^{n}\frac{1}{1+U_i} < 1$. Therefore, all elements (k,l) for $k,l \geq 1$ of this matrix writes

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n}\frac{\breve{X}_{ki}X_{li}}{1+U_i}. \tag{87}$$

because of prewhitening. We can write this in matrix form as

$$\left[\hat{V}_n^{IV}(\beta^{IV})\right] = n^{-1}X'P_zWX, \tag{88}$$

where W is a diagonal matrix with elements (i,i) acting as weights given by $\frac{1}{1+U_i} \in (0,1]$. The projection matrix P_z being symmetric and idempotent, its eigenvalues are equal to either 0 or 1. P_z is hence a positive semi-definite matrix. The product P_zW is thus a positive semi-definite matrix because it is the product of two symmetric positive semi-definite matrices.

Nevertheless P_zW is not necessarily symmetric. For any vector $a \in \mathbb{R}^K$, $a'X'P_zWXa$

and $a'X'\frac{1}{2}(P_zW+W'Pz)Xa$ *are the same quadratic forms. We have that* $X'\frac{1}{2}(P_zW+W'Pz)X$ *is positive semi-definite matrix and all its eigenvalues are nonnegative and corresponds to those of* $X'P_zWX$.

We can alternatively write the weight matrix $W = I_n - D$, *where* D *is also a diagonal matrix with elements* $\frac{U_i}{1+U_i} \in [0,1)$. *Therefore, we have the alternative expression*

$$\begin{aligned}\left[\hat{V}_n^{IV}(\beta)\right] &= n^{-1}X'P_z(I_n - D)X \\ &= X'P_zX - n^{-1}X'P_zDX \\ &= I_K - n^{-1}X'P_zDX,\end{aligned} \tag{89}$$

where the second equality comes from P_z *being idempotent, and prewhitening. It follows that as* $n \to \infty$, *the maximum eigenvalue is equal to*

$$\max_{|a|=1} a'\left[\hat{V}_n^{IV}(\beta)\right] a = \max_{|a|=1} 1 - a'X'\frac{1}{2}(P_zD + D'P_z)Xa. \tag{90}$$

Assuming the data distribution is non-degenerate, $a'X'\frac{1}{2}(P_zD + D'P_z)Xa$ *is positive and bounded away from zero for all unit vectors* $a \in \mathcal{R}^K$. *Thus, as* $n \to \infty$, *the maximum eigenvalue of* $\hat{V}_n^{IV}(\beta)$ *is strictly less than unity. This proves the result. The other conditions follow similar derivations as for Theorem 1 which complete the proof.*

Proof 7 (Proof of Theorem 4: Instrumental Variables Normality) *We now derive the asymptotic distribution of i2SLS. We must show that* $\sqrt{n}(\hat{M}_n^{IV}(\beta) - \beta) \xrightarrow{d} Z$ *as* $n \to \infty$, *where* Z *is a limit distribution. As before, we have*

$$\hat{c}(\beta) \xrightarrow{p} E[\log(1 + U_i)] = c, \tag{91}$$

as $n \to \infty$, *and*

$$\hat{\breve{Y}}_i(\beta) = \log(1 + U_i) + X_i'\beta - \hat{c}(\beta), \tag{92}$$

so that

$$\sqrt{n}[\breve{X}'\breve{X}]^{-1}\breve{X}'\hat{\breve{Y}}_i(\beta) = \sqrt{n}\left(\beta + [\breve{X}'\breve{X}]^{-1}\breve{X}'(\log(1 + U) - \hat{c}(\beta))\right). \tag{93}$$

Under the iid assumption and the exogeneity condition $E[\breve{X}_i(\log(1 + U_i) - c)] = 0$, *the Lindeberg-Levy's central limit theorem yields*

$$\sqrt{n}\left(\hat{M}_n^{IV}(\beta)-\beta\right) \xrightarrow{d} \mathcal{N}(0,\Sigma), \tag{94}$$

as $n \to \infty$, *where* Σ *is the asymptotic covariance matrix. Remark that it is the asymptotic covariance of the 2SLS estimator of the regression of* $\hat{\tilde{Y}}(\beta)$ *onto* X *using* Z *as IV. Heteroskedasticity-robust estimators apply as in the standard setting. However, the i2SLS estimator has a slightly different asymptotic distribution, because the true* β *is unknown. Using the same reasoning as for iOLS, we obtain*

$$\sqrt{n}\left(\hat{\beta}_{i(n)}^{IV}-\beta^{IV}\right) \xrightarrow{d} \mathcal{N}(0,[\Omega^{IV}]^{-1}), \tag{95}$$

as $n \to \infty$, *where* $\Omega^{IV} = (I_k - V^{IV}(\beta))^{-1}\Sigma(I_K - V^{IV}(\beta))^{-1}$ *and the gradient* $\nabla_\phi M^{IV}(\beta) = V^{IV}(\beta)$ *is defined as*

$$V(\beta) = E[\breve{X}_i\breve{X}_i']^{-1}E[\frac{\breve{X}_iX_i'}{1+U_i}]. \tag{96}$$

Therefore sandwich-type covariance estimators are given by

$$\tilde{\Sigma} = (\frac{1}{n}X'\frac{1}{2}(P_z(I-W)+(I-W)P_z)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'\frac{1}{2}(P_z(I-W)+(I-W)P_z)X)^{-1}, \tag{97}$$

where W *is a diagonal weighting matrix with diagonal element* $\frac{1}{1+U_i}$, *and* $\hat{\Sigma}_0$ *is an estimator of the covariance of* $P_zX'(\log(1+U_i)-c)$ *across observations. Symmetrizing the weight matrix, as explained in the proof of the preceding theorem, is required to have a symmetric positive definite matrix, hence invertible.*

B Model Extensions

B.1 Poisson models as iOLS

Multiplicative Poisson. First, we consider the multiplicative version of the model. It relies on the identifying assumption $E(U_i|X_i) = 1$, but only requires $E((U_i - 1)X_i) = 0$ for consistency. To enforce this condition, we can add $\frac{1}{1+\delta}(U_i - 1)$ on both sides of (10) and rearrange to obtain

$$\log(Y_i + \delta \exp(X_i'\beta)) - \left(\log(\delta + U_i) - \frac{1}{1+\delta}(U_i - 1)\right) = X_i'\beta + \frac{1}{1+\delta}(U_i - 1). \quad (98)$$

with $U_i = Y_i \exp(-X_i'\beta)$, the second term on the left-hand-side can be rewritten into

$$c_i(\delta, \beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i \exp(-X_i'\beta) - 1), \quad (99)$$

to obtain a new transformed dependent variable

$$\tilde{Y}_i(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c_i(\delta, \beta). \quad (100)$$

and associated model

$$\tilde{Y}_i(\beta) = X_i'\beta + \eta_i, \quad (101)$$

where $\eta_i = \frac{1}{1+\delta}(U_i - 1)$ is a mean-zero error term, and is exogenous to X_i. This estimator will be referred to as iOLS$_{MP}$. The choice of δ will be discussed shortly.

Additive Poisson (PPML). Similarly, one can enforce the additive representation based on model (4), which assumes $E[\epsilon_i|X_i] = 0$ with $\epsilon_i = Y_i - \exp(X_i'\beta)$. This assumption is equivalent to $E[U_i|X] = 1$ but leads to a different least-squares objective function. iOLS can be adapted to this setting by adding and substracting $\frac{1}{1+\delta}(Y_i - \exp(X_i'\beta))$ to (10) and defining

$$c_i(\delta, \beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i - \exp(X_i'\beta)). \quad (102)$$

This estimator, hereafter referred to as iOLS$_{AP}$, is equivalent to PPML but can yield numerically different point estimates because the numerical algorithm differs.

We derive the asymptotic properties of both estimators in the following theorem.

Theorem 3 (Consistency and Normality of iOLS$_{MP}$ and iOLS$_{AP}$) *Under Assumptions 1, 2, 4, and suitable regularity conditions, the iOLS estimators using $c(\delta, \beta)$ in (99) and (102) are consistent, achieve the parametric rate of convergence $n^{-1/2}$, and correspond to the multiplicative and additive Poisson regression* *estimates, respectively. Formally, we have* $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ *as* $n \to \infty$ *for any* $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$*, where* $\kappa \in [0, 1)$ *is the modulus of the associated contraction mapping from* $\mathbb{R}^K$ *to* $\mathbb{R}^K$*. In addition, they are asymptotically normally distributed such that* $\sqrt{n}\left(\hat{\beta}_{t(n)} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Omega)$*, as* $n \to \infty$*, where* Ω *is the same than in Theorem 1 except for the diagonal weighting matrix.*

Unlike in the previous case, the parameter δ does not modify the relevant moment condition. Its only purpose is now to control the convergence of the algorithm. The modulus κ is a function of δ with two important features. First, the algorithm will diverge for too small values of δ, which ultimately depends on the underlying DGP, because it implies κ above 1. Second, a too large δ implies κ very close to 1, hence a slow convergence as discussed earlier. Therefore, the optimal δ is large enough to guarantee convergence but small enough so that convergence is fast. We address these issues by defining a grid of values for δ based on a sufficient condition guaranteeing convergence. We start with a relatively small value and sequentially increment it if the algorithm detects divergence.

B.2 Instrumental variables

The estimation of causal relationships is central to social sciences. Yet, doing so is fraught with difficulties. Simultaneity, an omitted variable, or the presence of measurement errors could result in biased estimates. For example, if a researcher is interested in estimating the causal effect of the number of police officers on crime, one may observe that the police is more often deployed in areas where crime is high and conclude that police causes more crime.

A popular solution consists on finding an instrumental variable which affects the outcome only through the endogenous variable. Using variation in the instrument, one can recover the impact of the main variable of interest on the outcome through an estimation procedure known as *Two Stage Least Squares* (2SLS).

Our iterated solution extends directly to this situation and consists, in turn, in running 2SLS iteratively. Let us define Z as a $n \times L$ matrix with $L \geq K$ instrumental variables so that $E[X'Z] \neq 0$. We assume $E(Z'Z) < \infty$ and denote P_z as the projection matrix $Z(Z'Z)^{-1}Z'$. The following algorithm characterizes the i2SLS estimators.

Algorithm 2 (i2SLS estimator) *Let $\hat{\beta}_0$ be an initial estimate, for instance the 2SLS "popular fix" estimator $\hat{\beta}^{2SPF} = [X'P_zX]^{-1}X'P_z\log(Y+\Delta) \in \mathbb{R}^K$, for some $\Delta > 0$. the i2SLS estimator is obtained as follows.*

1. *Initialize t at 0;*
2. *Transform the dependent variable into $\tilde{Y}(\hat{\beta}_t)$;*
3. *Compute the 2SLS estimate $\hat{\beta}_{t+1}^{2SLS} = (X'P_zX)^{-1}(X'P_z\hat{\tilde{Y}}(\hat{\beta}_t))$, and update t to $t+1$;*
4. *Iterate steps 2 and 3 until $\hat{\beta}_t^{2SLS}$ converges.*

This iterative estimator converges under some conditions on $\tilde{Y}(\cdot)$. The same transformations studied earlier apply without further modifications. We prove the consistency of this estimator using the following assumptions in the next theorem.

Assumption 7 (Covariates) *X and Z have full column rank and $E(X_iX_i') < \infty$ and $E(Z_iZ_i') < \infty$.*

Assumption 8 *The error term υ_i satisfies the weak exogenous restriction $E[Z_i'(\upsilon_i - c(\delta,\beta))] = 0$ where the value $c(\delta,\beta)$ is unknown. In addition, let $E[U_i] = 1$.*

Theorem 4 (Consistency and Asymptotic Normality) *Under Assumptions 2, 7, 8, and suitable regularity conditions, the i2SLS estimator is consistent and achieves the parametric rate of convergence $n^{-1/2}$. Formally, we have*

$$n^{1/2}|\hat{\beta}_{t(n)}^{IV} - \beta| = O_p(1) \tag{103}$$

as $n \to \infty$ for any $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0,1)$ is the modulus of the associated contraction mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$. In addition, the i2SLS estimator is asymptotically normally distributed such that

$$\sqrt{n}\left(\hat{\beta}_{t(n)}^{IV} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Omega^{IV}), \tag{104}$$

as $n \to \infty$, where Ω^{IV}, as given in the proof, corresponds to the asymptotic covariance of the 2SLS estimator in the last iteration up to minor modifications.

This asymptotic result reveals several desirable properties of our procedure. First, the i2SLS estimators can be obtained easily using available software. Second, this iterative procedure makes non-linear instrumental variable estimation computationally tractable even when many control variables are included. This is particularly important because current count models are hard to estimate when using instrumental variables.[43] Finally, researchers often rely on the control function approach in non-linear models. This method requires the error in the second stage to be an additively separable function of the first-stage error and an independent error term. It also rules out settings where the endogenous variable is not continuous (Wooldridge, 2015). In contrast, 2SLS (and thus i2SLS) does not require such assumptions and can leverage the Frisch-Waugh-Lovell theorem to greatly alleviate computations when many fixed-effects are included.

Finally, the specification tests developed for iOLS are easily adapted for situations with endogenous regressors. The main difference is that one must estimate $Pr(Y > 0|Z)$ instead of $Pr(Y > 0|X)$. Further details are provided in Appendix B.8.

B.3 Negative values

Our estimator extends to dependent variables taking negative values. However, one needs to specify a slightly different model. We modify (1) into

$$Y_i = \alpha + \exp(X_i'\beta)U_i, \tag{105}$$

where $\exp(X_i'\beta)U_i \geq 0$ is as before but $\alpha < 0$ shifts leftwise the log function's vertical asymptote at zero towards the minimum value of Y. It is hence fundamentally different from the IHS, which imposes a S-shape transformation around zero. The iOLS transformation becomes

$$\log(Y_i - \alpha + \delta \exp(X_i'\beta)) = X_i'\beta + \log(\delta + U_i) \tag{106}$$

where $Y_i - \alpha \geq 0$. Estimation requires an additional step to find α before proceeding with the iOLS algorithm to address observations bunched at the lower bound (instead

[43] For example, to our knowledge, there are no packages in Stata which allow one to estimate instrumental variable count models, as in Mullahy (1997), with many categorical control variables.

of zero). In this model, Y_i is bounded below by α. The bound can be estimated by taking the first-order statistic $\hat{\alpha} = \min_i Y_i$. Here, the convergence rate of $\hat{\alpha}$ is crucial to determine that of the iOLS estimator. For instance, if Y_i is uniformly distributed, the first-order statistic will converge at rate n^{-1} to the true lower bound and the convergence result of the iOLS estimator will remain unaffected. Reversely, slower convergence rates will prevail if the first-order statistic converges at a rate slower than $n^{-1/2}$.

B.4 Log-log specifications

In many econometric applications, the main parameter of interest is an elasticity of Y_i with respect to some variable X_i. Elasticities are often estimated using a log-log specification. However, it is common to have both dependent and independent variables that are equal to zero for some observations. Taking the log-transform of either of these variables is impossible. We propose to address this issue as follows.

Let us consider the following data generating process

$$Y_i = X_i^{\beta} U_i, \tag{107}$$

with $X_i > 0$ and $U_i \geq 0$. The iOLS$_\delta$ estimator directly applies using the transformation

$$\log(Y_i + \delta X_i^{\beta}) = \beta \log(X_i) + \eta_i, \tag{108}$$

under the exogeneity restriction $E[\log(X_i)\eta_i] = 0$, where $\eta_i = \log(\delta + U_i) - c$ is the mean-zero error term of the linearized model. The only difference with the log-linear setting is that the regressors are also in log form.

A potential issue arises when X_i can take zero values with positive probability. For any independent variable, let us rewrite the above restriction as

$$E[\log(X_i)\eta_i | X_i > 0] Pr(X_i > 0) + \lim_{\epsilon \to 0} E[\log(\epsilon)\eta_i | X_i = 0] Pr(X_i = 0) = 0, \tag{109}$$

which can be rewritten into

$$E[\log(X_i)\eta_i | X_i > 0] Pr(X_i > 0) + \lim_{\epsilon \to 0} \log(\epsilon) E[\mathbb{1}_{(X_i=0)}\eta_i] Pr(X_i = 0) = 0. \tag{110}$$

A sufficient condition for this equality to hold is to have both $E[\log(X_i)\eta_i|X_i > 0] = 0$ and $E[\mathbb{1}_{(X_i=0)}\eta_i] = 0$. The former is the standard exogeneity condition stated for non-negative values of X_i, whereas the latter means that the occurrences of zeros in X_i are exogenous to the errors. In the single covariate setting, one can simply discard observations where $X_i = 0$ and estimate the model based on the condition $E[\log(X_i)\eta_i|X_i > 0] = 0$. In the multivariate case, this approach would lead to discard possibly many observations and greatly dampen statistical power. Instead, one can make use of both restrictions and introduce an extra binary variable in the modelas in

$$\log(Y_i + X_i^{\beta}) = \beta_0 \mathbb{1}_{(X_i=0)} + \beta \tilde{X}_i + \eta_i, \tag{111}$$

where $\tilde{X}_i = \log(X_i)$ for $X_i > 0$ and is equal to 0 otherwise.For ease of exposition, we have supposed the existence of a single explanatory variable but this strategy can be used along with an intercept and other covariates.

B.5 High-dimensional fixed-effects

The inclusion of fixed-effects creates some computational issues in non-linear panel data models. A modified version of the iOLS/i2SLS algorithm can be used to accommodate many fixed effects by making use of the Frisch-Waugh-Lovell theorem as follows. Let us decompose the set of regressors $X = [X_0, X_1]$, where X_0 are binary variables capturing all fixed-effects and X_1 the remaining regressors (including the constant term). Define the projection matrix $P_0 = X_0(X_0'X_0)^{-1}X_0'$ and denote the aggregate fixed-effect term by $\Lambda = X_0'\beta_0$.

Algorithm 3 (iOLS estimator with many fixed effects) *Let $\hat{\beta}_0$, and $\hat{\Lambda}_0$ be initial estimates. The iOLS estimator is defined as the following iterative procedure:*

1. *Initialize t at 0;*
2. *Transform the dependent variable into $\tilde{Y}_{iOLS}(\hat{\beta}_t, \hat{\Lambda}_t)$, where the term $X'\hat{\beta}_t$ is replaced by $X_1'\hat{\beta}_t + \hat{\Lambda}_t$;*
3. *Partial out the transformed dependent variable $\acute{\tilde{Y}}_{iOLS}(\hat{\beta}_t, \hat{\Lambda}_t) = (I_n - P_0)\tilde{Y}_{iOLS}(\hat{\beta}_t, \hat{\Lambda}_t)$ and the remaining regressors variable $\acute{X}_1 = (I_n - P_0)X_1$;*
4. *Compute the OLS estimate $\hat{\beta}_{t+1} = [\acute{X}_1{}'\acute{X}_1]^{-1}\acute{X}_1{}'\acute{\tilde{Y}}(\hat{\beta}_t)$, and update t to $t+1$;*

5. *Recover the fixed-effects into the aggregate term* $\hat{\Lambda}_t = (\tilde{Y}(\hat{\beta}_t) - X_1'\hat{\beta}_{t+1}) - (\acute{\tilde{Y}}(\hat{\beta}_t) - \acute{X}_1{}'\hat{\beta}_{t+1})$
6. *Iterate steps 2 to 5 until* $\hat{\beta}_t$ *converges.*

Note that all matrix inversions in this algorithm can be done only once. The presence of fixed-effects has hence almost no effect on the computation speed of the iterative estimator. Remark further that this approach relates to the Poisson estimator with high-dimensional fixed-effects. Indeed, Correia, Guimarães and Zylkin (2019) transform the PPML estimator into an iteratively reweighted least squares problem, then make use of the Frisch-Waugh-Lovell theorem to speed up computations like above. Their approach bears some similarities with our approach for iOLS$_{AP}$ (additive poisson), except that ours involves naturally less matrix inversions. The i2SLS estimator extends similarly to many fixed-effects.

Incidental parameter problem. In non-linear panel data models, individual fixed-effects are not always consistent when the cross-sectional dimension n increases to infinity while the time dimension T remains fixed. This issue is known as the incidental parameters problem (IPP). It is a well-known issue with maximum likelihood estimators, and some solutions have been recently developed (Fernández-Val and Weidner, 2016; Weidner and Zylkin, 2021). Although we do not study this problem from a theoretical viewpoint, our replications of the simulations in Weidner and Zylkin (2021) reveal that the iOLS estimators exhibit an IPP bias very similar to PPML. Adapting the solutions proposed in the literature should hence be possible. We leave this topic for future research.

B.6 The log of a ratio

Researchers are sometimes willing to estimate equations of the form

$$\log(Y_{i1}/Y_{i2}) = X_i'\beta + \varepsilon_i, \qquad (112)$$

where Y_{i1} and Y_{i2} are two outcomes of interest. It may happen that both outcomes can take zero values, hence not only the log is undefined but also the ratio. The "popular fix" estimator in this case consists in transforming the outcomes and focus on the following model

$$\log((Y_{i1} + \Delta)/(Y_{i2} + \Delta)) = X_i'\beta + \omega_i, \tag{113}$$

for some $\Delta > 0$.Needless to explain why this simple fix is not satisfactory. Instead, let us consider an alternative solution where the two following equations are estimated jointly

$$\begin{aligned} \log(Y_{i1} + \Delta) &= X_i'\beta_1 + \varepsilon_{1i} \\ \log(Y_{i2} + \Delta) &= X_i'\beta_2 + \varepsilon_{2i}, \end{aligned} \tag{114}$$

by rewriting the problem as a seemingly unrelated regression problem. Here, we propose to use the popular fix as a starting point, but other methods like iOLS will apply without difficulty. The seemingly unrelated regression model can be easily implemented as a single iOLS regression. The parameter β of interest corresponds to $\beta_1 - \beta_2$ and inference can be conducted using the delta-method. The advantage of this approach is that one can separately check which model is best to address the log of zero in each equation.

B.7 An alternative iOLS transformation

An alternative iOLS transformation would consist in letting δ vary across observations. For example, let $\delta_i = \delta(1 - \xi_i)$ where ξ_i takes a zero value when $Y_i = 0$ and is equal to 1 otherwise. Therefore, the iOLS transform becomes

$$\log(Y_i + (1 - \xi_i)\delta \exp(X_i'\beta)) = X'\beta + \log((1 - \xi_i)\delta + U_i). \tag{115}$$

Let us recall that $U_i = \exp(\varepsilon_i)\xi_i$, thus the error term is $\log((1 - \xi_i)\delta + \exp(\varepsilon_i)\xi_i)$. We now develop its conditional mean into

$$E(\log((1 - \xi_i)\delta + \exp(\varepsilon_i)\xi_i)|X) = E(\varepsilon_i|\xi_i = 1, X)P(X) + \log(\delta)(1 - P(X)) \tag{116}$$

On the other hand, the exogeneity condition imposed in the log-linear model is about

$$E(\varepsilon_i|X) = E(\varepsilon_i|\xi_i = 1, X)P(X) + E(\varepsilon_i|\xi_i = 0, X)(1 - P(X)). \tag{117}$$

Therefore, imposing the restriction $E(\log((1-\xi_i)\delta + \exp(\varepsilon_i)\xi_i)|X) = 0$ under the assumption that $E(\varepsilon_i|X) = 0$ (log-linear) is equivalent to assuming that

$$E(\varepsilon_i|\xi_i = 0, X) = \log(\delta), \tag{118}$$

where δ can be chosen using the testing procedures presented in the paper.

More generally, $E(\varepsilon_i|X) = 0$ implies that

$$E(\varepsilon_i|\xi_i = 1, X) = -E(\varepsilon_i|\xi_i = 0, X)(1 - P(X))P(X)^{-1}. \tag{119}$$

We can hence evaluate any assumption about $E(\varepsilon_i|\xi_i = 0, X)$ by considering a function $\delta(\cdot) > 0$ and test whether the following condition holds

$$E(\varepsilon_i|\xi_i = 1, X) = -\log(\delta(X))(1 - P(X))P(X)^{-1}. \tag{120}$$

This approach can be helpful although the choice of the candidate functions for $\delta(\cdot)$ to be considered is beyond the scope of this paper. Remark that Heckman's selection model corresponds to specifying

$$\delta(X) = \exp(-\lambda \frac{\phi(-X_i'\gamma)}{1 - \Phi(X_i'\gamma)}). \tag{121}$$

B.8 Testing with endogenous regressors

In this section, we explain how our tests adapt to endogenous regressors.

Testing the Poisson condition. For Poisson models, we have

$$E[U_i|Z_i] = E[U_i|Z_i, U_i > 0]Pr(U_i > 0|Z_i) = E(U_i), \tag{122}$$

since $E[U_i|Z_i, U_i = 0] = 0$. Following the same step as with exogenous regressors, the error term U_i under the null is such that

$$U_i = \lambda E[U]Pr(U_i > 0|Z_i)^{-1} + \nu_i \tag{123}$$

for $U_i > 0$ with $\lambda = 1$ and $E[\nu_i | U_i > 0, Z_i] = 0$. There are hence two differences: 1. one needs to estimate $P(U > 0|Z)$ instead of $P(U > 0|X)$, and 2. an IV estimator, like i2SLS, must be used to obtain $\hat{U}$.

Testing the i2SLS restriction. For iOLS$_\delta$, we have $E[\log(\delta + U_i)|Z_i] = c$. The null hypothesis is now

$$H_0 : E[\log(\delta + U_i)|Z_i, U_i > 0] - \log(\delta) = \frac{c - \log(\delta)}{Pr(U_i > 0|Z_i)}, \tag{124}$$

hence the differences are the same than for Poisson models.

Testing other restrictions. Testing for other restrictions introduces some new steps. Developing the associated exogeneity condition yields

$$E[\omega_i|Z_i, U_i > 0]P(Z_i) + E[\omega_i|Z_i, U_i = 0](1 - P(Z_i)) = 0 \tag{125}$$

which can be rearranged into

$$E[\omega_i|Z_i, U_i > 0] = -E[\omega_i|Z_i, U_i = 0](1 - P(Z_i))P(Z_i)^{-1}. \tag{126}$$

For the popular fix estimator, substituting the expression of ω_i on the RHS gives

$$E[\omega_i|Z_i, U_i > 0] = -(\log(\Delta) - E(X'\beta|Z, U > 0))(1 - P(Z_i))P(Z_i)^{-1}, \tag{127}$$

where the new term $E(X'\beta|Z, U > 0)$ can be obtained from the first-stage estimates of the 2SLS procedure neglecting the zero values. For the IHS estimator, we have the similar form

$$E[\omega_i|Z_i, U_i > 0] = E(X'\beta|Z, U > 0)(1 - P(Z_i))P(Z_i)^{-1}. \tag{128}$$

C Additional simulations results

C.1 iOLS DGP

We also considered a simulation where the "true DGP" is such that $E[\log(\delta + U_i) - c)|X] = 0$, as required by iOLS$_\delta$. This DGP is useful to illustrate iOLS$_\delta$ under ideal conditions. We fix $\delta = 0.25$ and $c = -0.7447$ so that $E(U) \approx 1$. We assume that $(X_{1i}, X_{2i})'$ is like in DGP 1. We further assume that $\log(\delta + \exp(\varepsilon))$ follows a truncated Gaussian distribution with mean $\frac{c-\log(\delta))}{P(X)} + \log(\delta)$, and with $\log(\delta)$ as lower bound to guarantee $\exp(\varepsilon) > 0$. The log-scale error ε is hence approximately Gaussian. The variance parameter is fixed at 0.5. Note that the error U_i is also heteroskedastic in this DGP.

Table C.1 reports the results for DGP 3 (iOLS). All estimators but iOLS$_\delta$ are biased. The automatic selection of δ does not seem to introduce much noise in the smaller sample. Interestingly, the precision of PPML and iOLS$_{AP}$ estimates does not seem to improve with the sample size in this DGP.

Table C.1: Simulations: DGP 3 (iOLS)

	$n = 1000$			$n = 10,000$		
Estim.	β_0	β_1	β_2	β_0	β_1	β_2
$iOLS_\delta$ (best)	0.71	1.02	0.98	0.84	1.00	1.00
	(0.33)	(0.13)	(0.13)	(0.28)	(0.05)	(0.05)
$iOLS_\delta$ (auto)	0.78	1.00	1.00	0.88	1.00	1.00
	(0.62)	(0.13)	(0.12)	(0.45)	(0.04)	(0.04)
$iOLS_{MP}$	0.44	0.67	1.33	0.47	0.63	1.37
	(0.17)	(0.11)	(0.11)	(0.10)	(0.08)	(0.08)
$iOLS_{AP}$	0.29	0.43	1.46	0.08	0.24	1.66
	(1.10)	(0.45)	(0.47)	(1.40)	(0.46)	(0.51)
$PPML$	0.30	0.43	1.46	0.08	0.24	1.66
	(1.12)	(0.46)	(0.48)	(1.38)	(0.46)	(0.51)
OLS	1.09	0.51	1.49	1.09	0.51	1.49
	(0.08)	(0.05)	(0.05)	(0.02)	(0.01)	(0.01)
IHS	0.66	0.53	0.26	0.66	0.53	0.26
	(0.10)	(0.07)	(0.07)	(0.03)	(0.02)	(0.02)
PF	0.49	0.46	0.26	0.49	0.46	0.26
	(0.08)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)

Notes: This table shows the mean and standard errors (in parentheses) of parameter estimates across 10,000 simulations based on DGP3.

C.2 KNN-based specification tests

As explained in Section 4.1, the properties of the tests depend on the specified conditional probability function. To address this concern, one can opt for using a non-parametric estimator of the conditional probability. For comparison, we report the same statistics when using kNN instead of the parametric logit in Table C.2. We select a 100 neighbors for the estimation. Our choice is based on the observation that the larger the number of neighbors, the smoother the resulting probability estimates because they are allowed to take values in a larger set, here 101 possible values. Fewer number of neighbors might possibly lead to better classification predictions but introduces more noise in the estimation of λ. The results show that the non-parametric estimator introduces some distortions in the tests' sizes but exhibit reasonably good power.

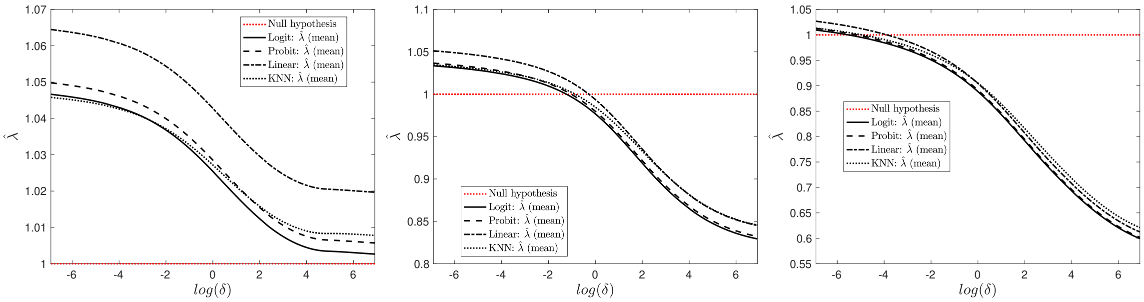
Table C.2: Simulations: Specification testing (kNN)

DGP				$iOLS_\delta$			MP	PPML	RESET
	δ	0.01	0.15	1.9	24	86			
	λ	1.02	1.02	1.01	1.01	1.01	1.00	1.01	
1	(se)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.08)	
	Rej%	92.8	78.1	25.4	6.9	9.7	3.2	6.2	7.3
	λ	1.01	1.00	0.98	0.93	0.91	0.87	Inf	
2	(se)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.06)	(Inf)	
	Rej%	47.8	13.1	44.9	93.3	91.8	97.3	20.1	6.3
	λ	0.99	0.97	0.92	0.82	0.76	0.66	Inf	
3	(se)	(0.04)	(0.06)	(0.09)	(0.10)	(0.11)	(0.15)	(Inf)	
	Rej	11.0	28.2	92.3	98.6	98.0	96.7	0.8	5.9

Notes: This table shows the relative rejection frequency of each null hypothesis for 1,000 simulations for $n = 10,000$ and DGP 1 to 3.

C.3 Misspecification of the conditional probability

For DGPs 1 to 3, we have also used alternative specifications of the conditional probability function to investigate how it affects the automatic selection of δ. Figure C.1 shows that the value of the selected δ is not much affected by the specified function.



(a) DGP1 (Poisson) (b) DGP2 (iOLS) (c) DGP3 (Log-linear)

Figure C.1: iOLS's $\hat{\lambda}$ as functions of δ with different specification of the conditional probability function (n=10,000)

C.4 Endogenous regressors and two-way fixed-effects

We consider two additional DGPs specified as follows:

- DGP 1 with endogenous regressors (IV): $E[U_i|X_i] \neq 1$ but $E[U_i|Z_i] = 1$. Let us assume that $Pr(\xi_i = 0|Z_i) = P(Z_i) = \frac{1}{1+\exp(\gamma_0+\gamma_1 Z_{1i}+\gamma_2 Z_{2i})}$, with the same parameters as in the main text. The instrumental variables Z_{1i} and Z_{2i} are iid normal with mean 1 and variance $\sigma^2_{Z1} = \sigma^2_{Z2} = 1$. We further assume that ε_i is Gaussian with mean $-\log(P(Z_i)) - 1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(Z_i)$. Finally the endogenous regressors are such that $X_{ik} = 0.8Z_{ik} + 0.2\varepsilon_i^2$, for $k = 1, 2$.

- DGP 1 with endogenous regressors and fixed-effects (IV-FE): the DGP is the same as above except for individual and time fixed-effects, denoted α_i and ρ_t, and assumed to be iid uniformly distributed in $[-0.5, 0.5]$. There are n/T individuals with $T = 100$ periods, for a total of $1 + T + n/T$ regressors.

Table C.3 reports the results for DGP IV, where the regressors are endogenous and requires the use of instrumental variables to achieve identification under the assumption that $E[U_i|Z_i] = 1$. The i2SLS estimators perform similarly to the non-linear IV estimator (Mullahy, 1997) $(IV - MP)$ in terms of bias and variance. However, the control function approach of Wooldridge (1997) $(CF - MP)$ delivers biased estimates because the endogenous regressors are non-linear functions of the error term in this design. The estimators denoted $IV - AP$ and $CF - AP$ correspond to the same estimators based on the additive Poisson model. They fail to deliver consistent estimates. $2SPF$ denotes the two-stage popular fix, where $\Delta = 1$ is added to Y before taking the log function in the second stage.

i2SLS estimators bring important computational gains in settings with fixed-effects. We only simulate this model 10 times to evaluate the computational complexity of each estimator. For $n = 1,000$, hence 111 regressors, $i2SLS_{MP}$ is computed in about 0.08 second whereas $IV - MP$ takes 45 seconds. For $n = 10,000$, hence 201 regressors, $i2SLS_{MP}$ is computed in about 0.8 second whereas $IV - MP$ takes 500 seconds.

Table C.3: Simulations: DGP 1 with endogenous regressors (DGP IV)

		$n = 1000$			$n = 10,000$	
Estim.	β_0	β_1	β_2	β_0	β_1	β_2
$i2SLS_\delta$ (best)	1.07	0.99	0.97	1.00	1.01	0.99
	(1.28)	(0.38)	(0.16)	(0.08)	(0.04)	(0.03)
$i2SLS_\delta$ (auto)	1.07	1.15	0.84	1.00	1.03	0.96
	(0.37)	(0.25)	(0.25)	(0.08)	(0.05)	(0.04)
$i2SLS_{MP}$	1.08	0.98	0.98	1.00	1.00	1.00
	(1.25)	(0.38)	(0.16)	(0.09)	(0.05)	(0.03)
$IV - MP$	1.04	0.99	0.98	1.00	1.00	1.00
	(0.36)	(0.17)	(0.10)	(0.09)	(0.05)	(0.03)
$IV - AP$	-0.88	1.37	1.19	0.18	1.24	1.10
	(14.09)	(3.51)	(2.11)	(12.10)	(2.14)	(1.00)
$CF - MP$	0.74	1.08	0.92	0.74	1.07	0.93
	(0.13)	(0.08)	(0.07)	(0.04)	(0.03)	(0.02)
$CF - AP$	-Inf	Inf	Inf	1.32	1.18	0.78
	(Inf)	(Inf)	(Inf)	(2.07)	(0.73)	(0.62)
$2SPF$	0.62	0.69	0.07	0.62	0.69	0.07
	(0.14)	(0.09)	(0.07)	(0.04)	(0.03)	(0.02)

Notes: This table shows the bootstrapped parameter estimates and standard errors calculated on data simulated according to DGP IV, as described in Section 5 and Appendix C. The column "Estim." reports the different estimated models. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard Errors are presented in between parentheses and are calculated using pairs bootstrap based on 10,000 simulations.

D Data Appendix

D.1 American Economic Review (2016-2020)

Table D.1: Solutions to the Log of Zero in the AER (2016-2020)

Log of Zero	$\log(\Delta + Y_i)$	PPML	Drop	IHS
48	23 (48%)	17 (35%)	15 (31%)	7 (15%)

Notes: This table reports the number of articles published in the American Economic Review from 2016 to 2020 where the issue of the log of zero was encountered. "Log of Zero" is the number of publications where at least one regression had to address this issue. "$\log(\Delta + Y_i)$" refers to the common fix of adding some discretionary constant to the dependent variable before taking the logarithmic transformation. "PPML" refers to Pseudo-Poisson Maximum Likelihood or Negative Binomial regression. "Drop" refers to cases where the problematic observations are discarded. "IHS" refers to the Inverse Hyperbolic Sine Transformation of the dependent variable. Some articles used several solutions, as robustness checks, which explains why the sum of solutions is different larger than 48.

Table D.2: American Economic Review Cases per Year

Year	Emp. Pub.	$\log(Y_i)$	$\log(\Delta+Y_i)$	PPML	Drop	IHS
2016	69	27	2	4	7	1
2017	71	28	5	2	4	1
2018	69	32	4	4	2	1
2019	79	27	6	6	2	3
2020	53	19	6	1	0	1

Notes: This table displays the frequency of solutions observed in American Economic Review. The sample extends over the period Jan. 2016 to Oct. 2020. *Emp. Pub.* is the number of empirical papers (includes "data" section). The column $\log(Y_i)$ counts cases where the dependent variable was in logarithmic form or in which a fix (such as $\log(\Delta+Y_i)$, PPML, Drop, or IHS) is used. It excludes cases where the author openly states that a logarithmic specification was preferred but rejected due to the existence of non-positive observations. $\log(\Delta + Y_i)$ is the popular fix. *PPML* refers to Poisson and Negative Binomial regression. *Drop* refers to cases where the author dropped the problematic observations. *IHS* is the Inverse Hyperbolic Transformation.

D.2 ResearchGate

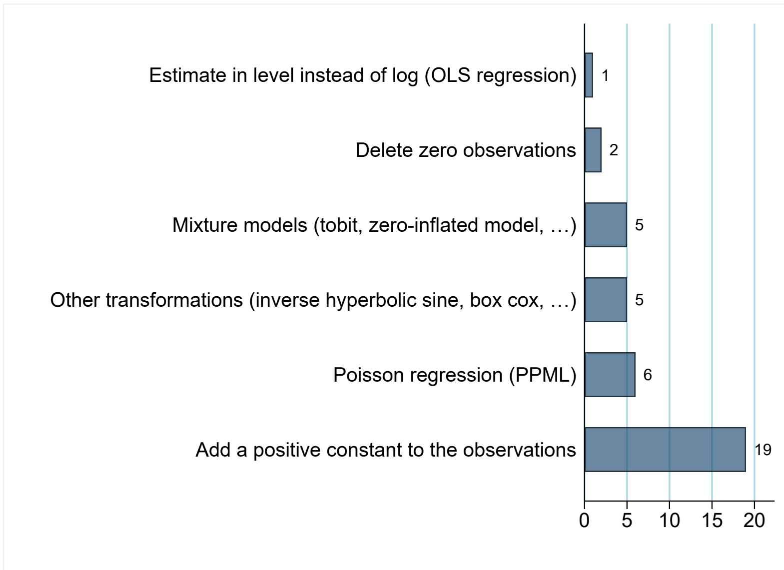


Figure D.1: Proposed solutions by category on ResearchGate (November 2018)

D.3 Wooclap Survey

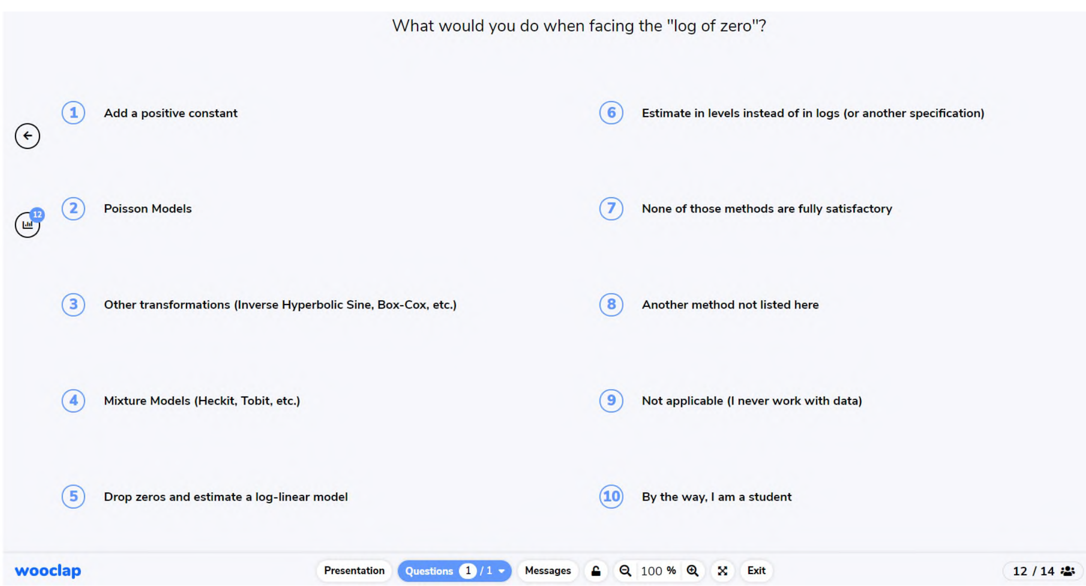


Figure D.2: Wooclap Survey

Description. The survey was implemented during 3 seminars (CREST, HEC Montréal, and University of Montréal) in 2021, before the speaker presented the different approaches. The attendees could provide multiple answers to the questions displayed in Figure D.2 and were invited to indicate if they were a student. Results are presented in Table D.3.

Table D.3: Wooclap Survey Results

	Frequency
Popular fix	42,8 %
Poisson	17,8 %
Other transformation	17,8 %
Mixture	35,7 %
Drop zeros	17,8 %
Levels instead of logs	17,8 %
Another method	3,5 %
None satisfactory	25 %
Not applicable	3,5 %
PhD Student	21,4 %
Nb. Respondents	28

Notes: This table displays relative frequency of answers to the Wooclap Survey. Intrepretation: 42.8% of respondents would use the popular fix (but not necessarily exclusively).

D.4 Michalopoulos and Papaioannou (2013)

Table D.4: Probability Model of Non-Zero Light Density at Night

Logit Probability Model		
Dependent var.	$\mathbb{1}$(Light Density at Night > 0)	
Jurisdictional Hierarchy	0.137	0.196
	(0.137)	(0.167)
Ln(Population Density)	0.690***	1.078***
	(0.112)	(0.215)
Distance to Capital City	-0.429	-0.745
	(0.562)	(1.045)
Distance to Sea Coast	-0.191	1.197
	(0.566)	(1.097)
Distance to Border	-1.513	-0.577
	(3.112)	(2.625)
Ln(1+Water Area)	-0.0918	0.346
	(0.314)	(0.450)
Ln(Land Area)	1.058***	1.274***
	(0.130)	(0.180)
Mean Elevation	-0.715	-0.549
	(0.465)	(0.769)
Soil Suitability	-0.415	-0.0535
	(0.952)	(0.806)
Ecological Suitability	-1.784*	-2.138
	(0.806)	(1.098)
Petroleum Field Dummy	0.933	1.093
	(0.688)	(0.711)
Diamond Mine Dummy	0.394	0.443
	(0.504)	(0.573)
Country Fixed Effect	No	Yes
N	682	621

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table provides the logit estimates and standard errors (s.e), based on the research of Michalopoulos and Papaioannou (2013). Each observation of the data is at the ethnicity by country level. The dependent variable is a dummy equal to one if the recorded light density at night is non-zero. Standard errors are provided in between parenthesis are clustered at the ethnicity by country level.

D.5 Santos Silva and Tenreyro (2006)

Table D.5: Replication of Table (3)

	PF	PPML	PPML (No Zero)	PPML (Top 10%)	iOLS$_{MP}$	iOLS$_{\delta}$
Constant	-39.909	-32.326	-31.530	-28.096	-36.523	-38.707
	(1.225)	(2.135)	(2.234)	(2.739)	(2.997)	(2.431)
Ln(GDP) - (X)	1.128	0.732	0.721	0.650	0.978	1.113
	(0.012)	(0.026)	(0.026)	(0.029)	(0.035)	(0.028)
Ln(GDP) - (I)	0.866	0.741	0.732	0.672	0.859	0.951
	(0.012)	(0.028)	(0.029)	(0.034)	(0.027)	(0.023)
Ln(GDP/Capita) - (X)	0.277	0.157	0.154	0.125	0.297	0.285
	(0.018)	(0.052)	(0.051)	(0.053)	(0.033)	(0.031)
Ln(GDP/Capita) - (I)	0.217	0.135	0.133	0.118	0.254	0.234
	(0.018)	(0.045)	(0.045)	(0.045)	(0.040)	(0.034)
Ln(Distance)	-1.151	-0.784	-0.776	-0.686	-1.477	-1.589
	(0.036)	(0.062)	(0.062)	(0.063)	(0.080)	(0.072)
Contiguity	-0.241	0.193	0.202	0.298	0.204	0.001
	(0.209)	(0.105)	(0.107)	(0.113)	(0.352)	(0.335)
Common Language	0.742	0.746	0.751	0.702	0.692	0.877
	(0.065)	(0.138)	(0.138)	(0.136)	(0.132)	(0.126)
Colonial Ties	0.392	0.025	0.020	0.001	0.390	0.448
	(0.070)	(0.159)	(0.160)	(0.161)	(0.138)	(0.129)
Landlocked - (X)	0.106	-0.863	-0.872	-0.886	-0.589	-0.494
	(0.056)	(0.153)	(0.153)	(0.150)	(0.130)	(0.114)
Landlocked - (I)	-0.278	-0.696	-0.703	-0.673	-0.881	-0.973
	(0.052)	(0.131)	(0.131)	(0.135)	(0.126)	(0.110)
Remoteness - (X)	0.526	0.660	0.647	0.625	1.020	0.888
	(0.089)	(0.142)	(0.143)	(0.154)	(0.178)	(0.169)
Remoteness - (I)	-0.109	0.561	0.549	0.563	0.113	0.036
	(0.090)	(0.125)	(0.126)	(0.133)	(0.218)	(0.162)
Free-trade agreement Dummy	1.289	0.181	0.179	0.245	1.549	1.425
	(0.121)	(0.096)	(0.097)	(0.098)	(0.502)	(0.345)
Openness	0.739	-0.107	-0.139	-0.169	-0.414	-0.369
	(0.050)	(0.128)	(0.129)	(0.136)	(0.102)	(0.095)
N	18,360	18,360	9,613	1,836	18,360	18,360

Notes: This table replicates Table 3 in Santos Silva and Tenreyro (2006). Standard errors based on 300 pairs bootstrap are in parenthesis, and λ for various models of trade. iOLS$_{\delta}$ and iOLS$_{MP}$ are defined in Section 3. PF is the baseline relying on the popular fix ($\Delta = 1$). PPML (No Zero) denotes the estimates from dropping zero flows. PPML (Top 10%) denotes the estimates from keeping the top 25% of trade flows. (X) refers to exporter characteristics and (I) to importer characteristics.

CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

CREST
Center for Research in Economics and Statistics
UMR 9194

5 Avenue Henry Le Chatelier
TSA 96642
91764 Palaiseau Cedex
FRANCE

Phone: +33 (0)1 70 26 67 00
Email: info@crest.science
https://crest.science/

The Center for Research in Economics and Statistics (CREST) is a leading French scientific institution for advanced research on quantitative methods applied to the social sciences.

CREST is a joint interdisciplinary unit of research and faculty members of CNRS, ENSAE Paris, ENSAI and the Economics Department of Ecole Polytechnique. Its activities are located physically in the ENSAE Paris building on the Palaiseau campus of Institut Polytechnique de Paris and secondarily on the Ker-Lann campus of ENSAI Rennes.