Dealing with Logs and Zeros in Regression Models

Christophe Bellégo* David Benatia† Louis-Daniel Pape‡

May 28, 2021

Abstract

Log-linear models are prevalent in empirical research. Yet, how to handle zeros in the dependent variable has remained obscure. This article clarifies this issue and develops a new family of estimators, called iterated Ordinary Least Squares (iOLS), which offers multiple advantages to address the "log of zero". We extend it to the endogenous regressors setting (i2SLS) and address common issues like the inclusion of many fixed-effects. In addition, we develop specification tests to help researchers select between alternative estimators. Finally, our methods are illustrated through numerical simulations and replications of recent 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 Marine Carrasco, Decio Coviello, Laurent Davezies, Benoît Dostie, Xavier D'Haultefoeuille, Ben Groom, James McKinnon, Mathieu Marcoux, Isabelle Méjean, Joris Pinkse, Alain Trognon, Michael Visser, and all participants to seminars at CREST, HEC Montréal, and Université de Montréal for insightful discussions and comments. STATA packages are available at https://github.com/ldpape/iOLS and https://github.com/ldpape/i2SLS. Note that this paper has undergone a considerable revision since the previous version was posted on SSRN.

1 Introduction

The log-linear and log-log models are among the most frequent specifications used in empirical research.[1] However, having to deal with the (natural) logarithm of a zero in the response variable is a common issue faced by practitioners. There is, unfortunately, a lack of consensus about the best practice to address those zeros, as evidenced by the many alternative solutions used in recent leading publications. This paper not only clarifies this issue, it also develops a new family of estimators and a model selection procedure. 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. We also develop specification tests aimed at verifying the external validity of the model with respect to the observed patterns of zeros in the data. Those tests prove to be helpful for selecting the most suitable approach to address the log of zero in any given setting.

The log transformation is popular because (1) the parameter estimate is related to an elasticity;[2] (2) logs can linearize a theoretical model, e.g. a Cobb-Douglas production function (Goldberger, 1968) or a gravity equation (Head and Mayer, 2019); (3) logs can make heteroskedasticity vanish in some settings, e.g. when the variance of a variable is proportional to its squared mean (Carroll and Ruppert, 1984); (4) the data is sometimes *naturally* related by a log-linear relationship (Ciani and Fisher, 2018); or even (5) it provides a concave transformation (MacKinnon and Magee, 1990).

However, the variable taken in logs may contain non-positive values. For example, a company can employ no worker, a product can have no sales or two countries zero trade in a given year. In these cases, the log is undefined and a fix is needed. Although this problem is quite common, the solution to be adopted is still unclear to many empirical researchers. We have reviewed all articles published in the American Economic Review (AER) between 2016 and 2020 to support this statement. Figure

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

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

1 summarizes our findings. It shows that nearly 40% of empirical papers used a log-specification and 36% of these faced the problem of the log of zero. It corresponds to an average of 10 publications per year dealing with the log of zero in the AER.

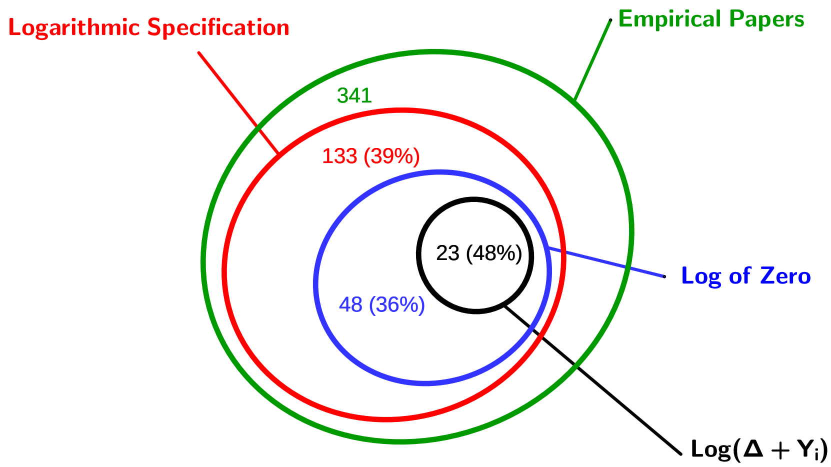

Figure 1: Prevalence of the Log of Zero in the AER (2016-2020)

No *single* solution has achieved consensus. In most publications, the authors chose to keep the zero observations 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.[3]

Moreover, researchers seldom report all their intermediary results leading to the submission of an article. To uncover existing practices, we have conducted a survey in three online seminars in economic departments asking "What would you do when facing the log of zero?".[4] Among the 28 respondents (including 21% of Ph.D students), 42% opt for the popular fix, 35% for mixture models (Tobit, Heckit, etc.), and 18% for abandoning the use of a log-like specification. Putting the latter indi-

[3] This excludes cases where the authors decided to use a linear specification by fault of having to use such a fix. See Table 20 in the Appendix for additional details and information regarding data collection.

[4] See Appendix D.3 for the survey and for the exhaustive set of results.

viduals aside, only 46% would compare multiple approaches, on average 2.7 each. It is interesting to note that the (somewhat large) stated preference for mixture models is not reflected in recent AER publications.

The issue of the log of zero extends well beyond economics. The question "Log transformation of values that include 0 (zero) for statistical analyses?" asked in 2014 on the forum ResearchGate, a multidisciplinary research-oriented social network, has received 38 contributions from researchers in various fields, including, but not limited to, medicine, biology, statistics and engineering. The thread has been read 120,000 times as of August 2020.[5] The prevalence of each solution is comparable to that in the AER. Adding a positive constant is suggested 50% of the time. Poisson, mixture models, and transformations like IHS are each recommended only 12.5% of the time.

There are hence five main solutions to the "log of zero". The most common fix consists in adding a positive constant to all observations (MaCurdy and Pencavel, 1986). This approach will thereafter be referred to as the "popular fix". A second solution is to delete the non-positive observations from the sample (Young and Young, 1975). A third solution uses transformations of the response variable, such as IHS, akin to the log function (MacKinnon and Magee, 1990; Burbidge, Magee and Robb, 1988; Johnson, 1949). A fourth solution consists in adopting mixture models (e.g. Tobit or Heckit) where a sample selection process explains the occurrence of non-positive observations (Heckman, 1979; Eaton and Tamura, 1994; Helpman, Melitz and Rubinstein, 2008). Finally, Poisson models (Gourieroux, Monfort and Trognon, 1984) handle the presence of zeros well in many settings. They are especially popular in international trade where it is the workhorse model for the estimation of gravity equations (Head and Mayer, 2019). To the best of our knowledge, Santos Silva and Tenreyro (2006) were the first to argue in favor of Poisson regression to address the log of zero.

However, to deal with the log of zero and select among these models one must first address the critical question "why do the data contain zeros?".[6] It could be due to either data problems, such as measurement errors of small values, or a "true zero",

[5] See https://www.researchgate.net/post/Log_transformation_of_values_that_include_0_zero_for_statistical_analyses2, and Figure 4 in the Appendix.

[6] This question echoes that of Heckman (1979) about missing data: "why are the data missing?".

for example when a product has exactly zero sale. In any case, one must make distributional assumptions about the zeros either explicitly (e.g. Tobit or Heckit) or implicitly through moment restrictions (e.g. PPML or IHS). We will discuss the assumptions made by existing methods, and propose a model of the latter type. The main advantage of this approach is that it does not require to specify a selection process explaining the occurrence of zeros.

The main focus of our paper is the identification of the model parameters rather than the prediction of an outcome. Identification is key for the estimated parameters to have an economic interpretation. It typically relies on exogeneity restrictions in the form of moment conditions between the errors and regressors, like OLS or Poisson regression. Our discussions with empirical researchers revealed that many opt for the popular fix approach because they do not feel comfortable assuming the exogeneity restriction imposed by Poisson models.[7] Instead, they seem to believe that adding a constant to the outcome before taking the log function yields an error satisfying an exogeneity condition close to that of OLS in a log-linear model. Unfortunately, it does not.

Our approach consists in adding an observation-specific value to the outcome instead of a constant. It makes use of an exogeneity condition, either user-chosen or data-driven, in a range of possible conditions between that of the log-linear model and Poisson. Our estimators are then computed thanks to an iterative procedure. We rely for that on the asymptotic theory developed in Dominitz and Sherman (2005) to prove the consistency and asymptotic normality of our estimators. Iterative estimation methods are frequently encountered in physics and machine learning, where Iterated Reweighted Least-Squares (IRLS) are widely used for robust estimation (Dembinski, Schmelling and Waldi, 2019). Although less popular in economics, iterative estimators are used in some settings. For instance, Blundell and Robin (1999) propose an iterative solution for demand estimation to improve computational effi-

[7] In February 2021, Jeffrey Wooldridge tweeted *"Poisson regression can get one so far with so little trouble, why do so many still resist? [...]"* (https://twitter.com/jmwooldridge/status/1363828456136523779?s=20.) Ten years earlier, the President of StataCorp, William Gould, wrote a blog post arguing that researchers should use Poisson regression rather than OLS with a log outcome: https://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/.

ciency with respect to non-linear methods. Another example is the iterative estimation strategy in Head and Mayer (2014) of the structural gravity model of Anderson and van Wincoop (2003).

We make three principal contributions. First, we clarify the log of zero issue in a didactic way by reviewing existing practices. Second, we develop a new family of solutions, referred to as iterated OLS (iOLS). They consist in adding a data-dependent value to each observation and iterating OLS on the transformed model until convergence. They have multiple advantages: (a) they can be estimated by ordinary least squares, hence are computationally fast and easy to implement[8]; (b) robust standard errors are readily available; (c) they do not suffer from highly dispersed response variables; (d) they extend naturally to the endogenous setting using iterated 2SLS (i2SLS); and (e) they are amenable to different identifying assumptions. Finally, we develop a procedure to select which solution should be preferred in any given setting. This procedure helps choosing the most plausible model(s) given the data at hand. It consists in testing the implicit assumption about the patterns of zeros made by each approach. More formally, it is a test of whether the conditional probability of having a zero implied by the model is consistent with the data.

Our methodological contributions are illustrated through numerical simulations and (partial) replications of three recent publications in top-tier economics journals. First, Santos Silva and Tenreyro (2006) compare various estimators to estimate gravity models of trade and argue in favor of Poisson regression. Second, Michalopoulos and Papaioannou (2013) adds a positive constant to the response variable in order to examine the role of pre-colonial ethnic institutions on economic development. Third, Card and DellaVigna (2020) investigate the preferences of academic journal editors with the IHS transformation. Our tests reveal that no single solution is preferred in all settings. Nevertheless, iOLS tends to be selected more often than other methods in those examples.

The remaining of the paper is organized as follows. Section 2 clarifies the log of zero issue and discusses existing practices found in empirical research. Section 3 develops a new family of solutions. Section 4 presents specification tests and a data-driven

[8] Stata packages for iOLS_δ and i2SLS_δ (with potentially high dimensional fixed effects) are available from www.https://github.com/ldpape.

model selection procedure. Numerical simulations are presented in Section 5. Partial replications of leading publications are proposed in Section 6. Section 7 concludes the paper.

The Appendix section also contains several useful extensions to our methods. First, we adapt it to the endogenous setting in Appendix B.1. Second, we address the case where discarding zeros does not jeopardize identification in Appendix B.2. Third, we show in Appendix B.3 how to deal with negative values in Y. Appendix B.4 discusses log-log specifications with zeros in the independent variables. Appendix B.5 develops a computationally fast "within" iOLS estimator to avoid the incidental parameter problem when many fixed-effects are included. Appendix B.6 shows how to deal with the log of a ratio of two response variables. Appendix B.7 makes use of yet another alternative exogeneity condition close to that of the log-linear model. Finally, Appendix B.8 details the testing procedures in the endogenous regressors setting.

2 Existing Practices

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 an iid mean-zero error term, and $\xi_i \in \{0, 1\}$ is a Bernoulli random error and, without loss of generality, we take $E[\exp(\varepsilon_i)\xi_i] = 1$. Let X denote the $n \times K$ matrix comprised of the K-dimensional column vector X_i with elements X_{ki}, for $1 \leq k \leq K$. Let us assume that $E(X_i X_i') < \infty$, and X has full column rank.

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 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 rely on the more compact *multiplicative* representation,

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

where $U_i = \exp(\varepsilon_i)\xi_i$ has mean one, and refer to the equivalent *additive* model

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

with $\epsilon_i = \exp(X_i'\beta)(U_i - 1)$ treated as a mean-zero error.

2.1 The popular fix: to add a positive constant

The most popular solution is to add a positive constant Δ to all observations Y_i so that $\tilde{Y}_i = Y_i + \Delta > 0$ and the log-transformation becomes feasible. 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.[9]

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 consistent.[10]

Anecdotal evidence reveals that empiricists sometimes believe this bias to be negligi-

[9] Other variants include adding a constant solely to the non-positive values and including an additional dummy variable indicating such a treatment, generating the same kind of troubles. Alternatively, Johnson and Rausser (1971) propose to estimate the constant along with the other parameters. However, their method does not guarantee unbiased estimates.

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

ble for small values of Δ, or for $\Delta = 1$. This belief holds true only under strong and unverifiable restrictions about the underlying DGP. To illustrate this point, we rely on numerical simulations based on the design detailed in Section 5 (DGP 1). The objective is to estimate the parameters $\beta_1 = \beta_2 = 1$. Figure 2 presents the mean estimates using the popular fix, i.e. the OLS estimate of (5), as a function of the value of Δ. For this parameter, the mean squared bias is minimized at $\Delta = 0.7$, but the bias of each parameter varies with the constant and remains substantial. The "best" value for Δ is hence neither arbitrarily small nor equal to 1, contrary to common belief.

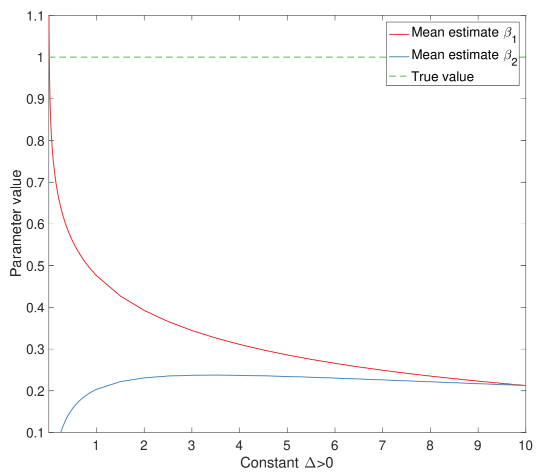

Figure 2: Bias against Δ

2.2 Other methods

Among the models which can address the log of zero, Non-linear methods are popular because they offer a valid approach in many settings. There are also approaches which should usually be avoided.

Poisson models. The model presented in (1) is non-linear in variables and parameters. The parameters are identified and non-linear estimators, such as the generalized method of moments (GMM), yield consistent estimates of β under the strict exogeneity restriction $E(U_i|X_i) = 1$ which implies the unconditional moments[11]

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

which allow the estimation of β by maximizing 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. Santos Silva and Tenreyro (2006) were the first to argue for Pseudo-Poisson Maximum Likelihood (PPML) as a potential solution for the appearance of zeros in Y_i. This approach is based on the additive representation of the model in (4) assuming $E(\epsilon_i|X_i) = 0$, which is equivalent to $E(U_i|X_i) = 1$.

Nevertheless, these Poisson regression has several shortcomings. First, existence of a solution is not guaranteed leading to convergence issues. Second, their precision can be sensitive to the dispersion of Y_i because of the exponential function. Third, they can be difficult to estimate with many fixed-effects. Fourth, instrumental variables require stronger assumptions and may dramatically increase computational complexity.[12] Most of these issues have been discussed and addressed in a series of papers (Santos Silva and Tenreyro, 2010, 2011; Correia, Guimarães and Zylkin, 2019).[13]

Mixture models and Heckman's correction. Censorship models, such as Tobit models (Tobin, 1958), provide another non-linear solution. They 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 not often used to address the log of zeros but has been relied upon in the context of gravity equations. For example, Eaton and Tamura (1994) implement a Tobit approach to model thresholds above which trade starts to be measured.

The Heckman's ("Heckit") correction (Heckman, 1979) is seldom used for the log of

[11]Choosing the "best" unconditional moments, or rather picking the optimal instruments, from a conditional moment restriction is beyond the scope of this paper. The interested reader is referred to Chamberlain (1987).

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

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

zero. 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

$$E[\varepsilon_i|\nu_i > -X_i'\gamma, X] = \lambda \frac{\phi(-X_i'\gamma)}{\Phi(X_i'\gamma)}, \tag{7}$$

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

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

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. When $\hat{\lambda}$ is zero, the mechanism generating the zeros is not correlated to the outcome and OLS regression using the positive values of Y_i will provide a consistent estimate of β. Therefore, this simple two-step approach can be used to investigate whether discarding zeros would threaten identification. Note that, however, this approach is heavily dependent on the distributional assumption in absence of instrumental variables in the selection equation.

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 the following condition holds,

$$E[\varepsilon|\xi = 1, X] = constant. \tag{9}$$

Similarly, one could discard zeros and estimate (4) with PPML assuming

$$E[\exp(\varepsilon)|\xi = 1, X] = constant. \tag{10}$$

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.

Transformations. An alternative approach relies on log-like transformations applicable to non-positive values. The most popular are the "popular fix", presented earlier, and the IHS (MacKinnon and Magee, 1990; Burbidge, Magee and Robb, 1988; Johnson, 1949).[14] It consists in transforming Y_i into $\tilde{Y}_i = \log(\theta Y_i + \sqrt{\theta^2 Y_i^2 + 1})/\theta$ and estimating $\tilde{Y}_i = X'\beta + \omega_i$ by OLS. If the underlying model writes in log, then this transformation will likely yield biased estimates.[15] Nearly all economic applications set θ to 1 such that $\tilde{Y}$ tends toward $\log(2Y)$ for large values of Y. There is also a version with a location parameter as discussed in MacKinnon and Magee (1990). 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 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.[16] It is hence satisfactory in contexts where applying a concave transformation is the main objective, e.g. for prediction models, where identification is not an issue, or when the exogeneity restriction can

[14] An extended concave version of this transformation is provided by Ravallion (2017).

[15] Considering model (1), having 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.

[16] 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.

be justified as discussed later on.

3 Iterated Ordinary Least Squares (iOLS)

In this section, we develop a new approach based on the popular fix. This new approach yields a family of estimators requiring only OLS to implement. For clarity, we first show how our estimation procedure arises in the context of the log of zeros. Second, we derive its asymptotic properties. Third, we detail how minor modifications can accommodate alternative exogeneity conditions.

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

$$\log(Y_i + \Delta_i) = X_i'\beta + \log\left(U_i + \frac{\Delta_i}{\exp(X_i'\beta)}\right). \tag{11}$$

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

$$\log(Y_i + \delta \exp(X_i'\beta)) = X_i'\beta + \upsilon_i. \tag{12}$$

where the new error term $\upsilon_i = \log(\delta + U_i)$ is assumed to satisfy an exogeneity restriction (discussed later). This shows that adding a constant value to Y_i falls short of the varying $\Delta_i = \delta \exp(X_i\beta)$ required to suppress bias.

The DGP specified in (1) assumes $E[U_i] = 1$,[17] implying that the transformed error υ_i is not mean-zero. Instead, we have $E[\log(\delta + U_i)] = c$, where c is an unknown constant depending on higher-order moments of U_i. To see this, consider the Taylor expansion of $\log(\delta + U_i)$ around $\log(1 + \delta)$ to obtain

$$c = \log(1+\delta) - \frac{1}{2(1+\delta)^2}E[(U_i-1)^2] + \frac{1}{3(1+\delta)^3}E[(U_i-1)^3] + ..., \tag{13}$$

where the second and third terms are respectively the variance and third centered

[17] This assumption is only useful to identify the intercept term.

moment of U_i. The first centered moment is assumed to be zero. Thus, this transformation introduces a nuisance parameter in the form of an extra constant term.

3.2 Identification

Demeaning this new error term is required to identify the 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

$$E\left[X_i \left(\log(Y_i + \delta \exp(X_i'\beta)) - c\right)\right] = E\left[X_i X_i'\right]\beta, \tag{14}$$

with $k+2$ unknowns. This system identifies β only if c is known. Fortunately, the multiplicative 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

$$Y_i = \exp(\beta^1 + X_i^{r\prime}\beta^r)U_i = \exp(\beta^1)\exp(X_i^{r\prime}\beta^r)U_i. \tag{15}$$

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

$$\beta^1_\beta = \log(E[Y_i \exp(-X_i^{r\prime}\beta^r)]). \tag{16}$$

Therefore, the parameters are identified and the nuisance c can be written as[18]

$$c(\beta) = E[\log(Y_i + \delta \exp(\beta^1_\beta + X_i^{r\prime}\beta^r)) - \beta^1_\beta - X_i^{r\prime}\beta^r]. \tag{17}$$

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

3.3 Estimation by iOLS

The following transform of the response variable yields a (seemingly) linear model:

$$\tilde{Y}_i^{iOLS_\delta}(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c(\beta) = X_i'\beta + \overline{\upsilon}_i \tag{18}$$

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

$$\beta = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\beta)\right], \tag{19}$$

which characterizes β as the solution of a fixed-point problem. Based on this insight, we propose an iterative least-squares estimator.

Algorithm 1 (iOLS estimator) *The iOLS estimator is defined as the following iterative procedure:*

1. *Initialize t at 0 and let $\hat{\beta}_0$ be an initial estimate, as obtained for example 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 $\tilde{Y}_{iOLS_\delta}(\hat{\beta}_t)$ using (18);*
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.*

We illustrate the algorithm in Figure 3 using the numerical simulations presented in Section 5 (DGP 1). The iterative estimation procedure converges to a solution within 15 to 20 iterations on average. Moreover, only a few iterations are required to suppress most of the bias of the popular fix estimator. Remark also that $X'X$ needs only be inverted once.

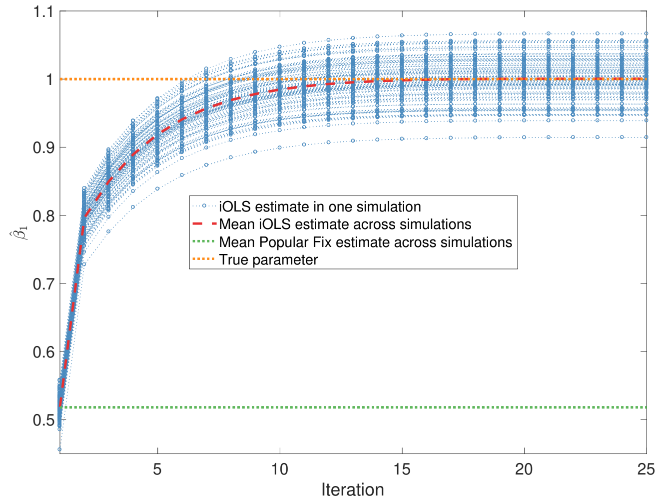

Figure 3: Convergence of iOLS$_{\delta=1}$ (DGP 1, n =1,000)

3.4 Asymptotic Properties

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

Theorem 1 (Consistency and Normality of iOLS$_\delta$) *Under the above assumptions, the iOLS$_\delta$ estimator is consistent and achieves the parametric rate of convergence $n^{-1/2}$. 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 Ω, as given in the proof, corresponds to the asymptotic covariance of the OLS estimator in the last iteration up to minor modifications.*

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 there may exist values of δ such that the algorithm does not converge in finite time. This occurs when δ implies a κ very close to 1, hence a very slow convergence. However, choosing a larger δ will mechanically decrease κ and solve this issue.[19]

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.[20]

3.5 Moment Selection

Our approach relies on an exogeneity condition about the error $\log(\delta + U_i)$. In absence of zeros, the condition is about $\log(\delta + \exp(\varepsilon_i))$ hence about ε_i when $\delta \to 0$, like in the log-linear model (2), whereas the Poisson condition is about $\exp(\varepsilon_i)$. Our understanding of the survey results presented in the introduction is that economists concerned about identification prefer conditions about ε_i rather than $\exp(\varepsilon_i)$, and that is why they often opt for the popular fix.

3.5.1 The role of δ

The condition $E[X_i\overline{v}_i] = 0$ is different from the condition $E[U_i|X_i] = 1$ assumed in Poisson models. Ultimately, which conditional moment restriction is satisfied depends on the context and is unverifiable *ex-ante*. However, as will be detailed later, one can test whether the restriction yields estimates that verify the implicit assumption about the pattern of zeros.

The parameter δ allows selecting a restriction among the *family* of moments $E[X_i \log(\delta + U_i)] = c$. Indeed, we observe at one extreme that when $\delta \to 0$, we have that $\lim_{\delta\to 0} \log(\delta + U)$ is exogenous to X_i, a moment condition similar to the one assumed in the standard log-linear model. At the other extreme, when $\delta \to \infty$, we have a condition equivalent to $E[X_iU_i] = \text{constant}$, which corresponds to the (unconditional) moment condition used in (multiplicative) Poisson regressions. In other words, δ

[19]The algorithm must include a safety check to ensure that κ is sufficiently smaller than 1. In practice, we take the median across estimates obtained at each iteration by $\hat{\kappa} = |\beta_{t+1} - \beta_t|/|\beta_t - \beta_{t-1}|$.

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

allows one to pick any condition (strictly) in-between these two extremes.

To see this, observe the Taylor expansion of $E[X_i \log(\delta + U)]$ around $U_i = 1$

$$E[X_i \log(\delta + U)] = E[X_i] \log(1 + \delta) + \left\{ \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k(1+\delta)^k} E[(U_i - 1)^k X_i] \right\}. \tag{20}$$

Assuming $E[X_i(\log(\delta + U) - c)] = 0$ to be true means the weighted sum of moment conditions between U_i and X_i is constant; where weights depend on the parameter δ. As δ goes to infinity, the weighted sum on the right-hand-side becomes negligible and $\log(\delta + U) \to \log(\delta) + U$, so the limiting moment condition is $E[X_i U_i] =$ constant.

This flexibility is of significant relevance. Indeed, the researcher usually lacks any *a priori* knowledge of the right exogeneity condition (and associated δ). We thus provide a data-driven selection method for δ based on testing the model's validy with respect to the pattern of zeros.

3.5.2 Poisson regression as iOLS

If the *Poisson condition* ($E[U_i|X_i] = 1$) holds in the data, then iOLS$_\delta$ will deliver a reasonable approximation for an arbitrarily large δ. Nevertheless, we now show how to enforce this condition directly in iOLS$_\delta$.

Multiplicative Poisson (iOLS$_U$). 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 (12) and rearrange to obtain

$$\log(Y_i + \delta \exp(X_i'\beta)) - \left(\log(\delta + U_i) - \frac{1}{1+\delta}(U_i - 1) \right) = X_i'\beta + \frac{1}{1+\delta}(U_i - 1). \tag{21}$$

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

$$c_i(\beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i \exp(-X_i'\beta) - 1), \tag{22}$$

to obtain a new transformed dependent variable

$$\tilde{Y}_i^{iOLS_U}(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c_i(\beta). \tag{23}$$

and associated model

$$\tilde{Y}_i^{iOLS_U}(\beta) = X_i'\beta + \eta_i, \tag{24}$$

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

Additive Poisson (iOLS$_\epsilon$). Similarly, one can enforce the additive representation based on model (4), which assumes $E[\epsilon_i|X_i] = 0$, where $\epsilon_i = Y_i - \exp(X_i'\beta)$. 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 (12) and changing $c_i(\beta)$ in (22) into

$$c_i(\beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i - \exp(X_i'\beta)). \tag{25}$$

This estimator, hereafter referred to as iOLS$_\epsilon$, is equivalent to PPML but can yield numerically different estimates. However, it may be less sensitive to the dispersion of the dependent variable since it does not require computing the gradient, as in PPML, or the Hessian, as in the IRLS implementation of Poisson regression (Correia, Guimarães and Zylkin, 2019). We derive the asymptotic properties of both estimators in the following theorem.

Theorem 2 (Consistency and Normality of iOLS$_U$ and iOLS$_\epsilon$) *Under the above assumptions, the two estimators are consistent and achieve the parametric rate of convergence $n^{-1/2}$. 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, as given in the proof, differs for iOLS$_U$ and iOLS$_\epsilon$ but corresponds to the asymptotic covariance* *of the OLS estimator in the last iteration up to minor modifications.*

This result shows 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.

For both estimators, the parameter δ does not modify the relevant moment condition but is key to guarantee 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. Therefore, the optimal δ is large enough to guarantee convergence but small enough so that convergence is fast. We address these issues by starting with a small value which we multiply by 10 if the algorithm diverges, or if our estimate of κ is above 1, and repeat this incrementation until convergence.

Various extensions. In Appendix B, we propose several extensions to the iOLS procedure. First, we adapt it to the endogenous setting in Appendix B.1. Second, we address the case where discarding zeros does not jeopardize identification in Appendix B.2. Third, we show in Appendix B.3 how to deal with negative values in Y. Appendix B.4 discusses log-log specifications with zeros in the independent variables. Appendix B.5 develops a computationally fast "within" iOLS estimator to avoid the incidental parameter problem when many fixed-effects are included. Appendix B.6 shows how to deal with the log of a ratio of two response variables. Finally, Appendix B.7 proposes an alternative solution to use the exogeneity condition of the log-linear model $E(\varepsilon_i|X) = 0$, or an approximation of that condition.

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 identifying assumptions. The latter can be systematically investigated through their implications regarding the patterns of zeros in order to substantiate the choice of an estimation procedure.

The tests developed in this section offer an opportunity for an ex-post evaluation of the identifying restrictions used for moment-based estimators. However, they are not useful to gauge the validity of the explicit distributional assumptions made in mixture models. Our tests are specification tests used to evaluate the validity of conditional moment restrictions, like $E(U_i|X_i) = 1$ for Poisson models.[21] They are, as such, similar to the RESET test of Ramsey (1969) for linear regression and its application for Poisson models by Santos Silva and Tenreyro (2006).[22] Our approach is, however, fundamentally different because it relies on testing the validity of the model with respect to the conditional probability of observing a zero. 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 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 in favor of a sufficient condition is still valuable information that the associated model bears some validity. The main issue is, however, that a rejection of the sufficient condition is not evidence against the necessary condition. In other words, these tests may lead to reject a correct model. Bearing these limits in mind, we proceed to present our methods.

4.1 Specification testing

Testing the Poisson condition. For clarity, we first look at the implications made by Poisson models regarding the pattern of zeros.[23] A related approach will be applied for other restrictions including for iOLS. Noting that a zero can only be

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

[22] Extensions of the RESET test are proposed in Wooldridge (1997).

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

observed if $U_i = 0$, the Poisson restriction $(E(U_i|X_i) = 1)$ can be decomposed into

$$E[U_i|X_i] = E[U_i|X_i, U_i > 0]Pr(U_i > 0|X_i) = E(U_i), \tag{26}$$

since $E[U_i|X_i, U_i = 0] = 0$. There are only two possibilities for this condition to hold true. First, $E[U_i|X_i, U_i > 0]$ 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, or more generally, if it follows a mixture distribution with a mass probability at zero such that the condition holds. 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 means that the occurrence of a zero does not depend on X_i. In this case, discarding zeros or not before estimation is irrelevant for identification.

This equation reveals the implicit relation between zeros and non-zero observations,

$$E[U_i|X_i, U_i > 0] = \frac{E(U_i),}{Pr(U_i > 0|X_i)}, \tag{27}$$

which 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. This amounts to evaluating the null hypothesis[25]

$$H_0 : E[U_i|X_i, U_i > 0] = \frac{E[U]}{Pr(U_i > 0|X_i)}, \tag{28}$$

which implies that $E[U_i|X_i, U_i > 0]$ and $Pr(U_i > 0|X_i)^{-1}$ are proportional.[26]

[24] It is worth noting that 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 the selection process, whereas sample-selection models enforce explicit restrictions.

[25] $E[U]$ is used to address the general framework where $E[U]$ could differ from 1.

[26] We also need to test whether the probability of observing a zero depends on any X_i by assessing $H_{0b} : Pr(U_i > 0|X_i) = p$, for any constant p. The latter is easily checked by estimating a logit

Under the null, one can model the error term U_i as

$$U_i = \lambda E[U] Pr(U_i > 0|X_i)^{-1} + \nu_i \tag{29}$$

for $U_i > 0$ with $\lambda = 1$ and $E[\nu_i|U_i > 0, X_i] = 0$. Therefore, one can evaluate H_0 by testing whether $\lambda = 1$. This test is done in 4 steps: (1) obtain a consistent estimator of $Pr(U_i > 0|X_i)$ denoted $\hat{P}(X)$, which is possible because $U_i > 0$ if and only if $Y_i > 0$; (2) compute Poisson estimates $\hat{\beta}$ for the multiplicative model, for instance with iOLS$_U$; (3) recover the residuals $\hat{U}_i = Y_i \exp(-X'\hat{\beta})$;[27] and (4) estimate the following regression model

$$\hat{U}_i = \lambda W_i + \nu_i, \tag{30}$$

for strictly positive errors only, and where $W_i = \hat{E}[U]\hat{P}(X_i)^{-1}$ and $\hat{E}[U]$ is the unconditional mean of $\hat{U}_i$ across both positive and zero observations.

The following t-stat allows evaluating the model's validity:

$$t = \frac{\hat{\lambda} - 1}{\hat{\sigma}_\lambda}. \tag{31}$$

Under the null, the OLS estimate of λ is consistent since $\hat{U}_i$ and $\hat{P}(X_i)$ are consistent and t will hence converge to zero. In finite samples, however, the standard error estimates will need to be adjusted to account for the additional noise introduced by first-step estimates.[28] We opt for a pairs bootstrap to estimate this test statistic in our practical implementation. This approach yields t_{PPML} t_{iOLS_U} and t_{iOLS_ϵ}.

Testing the iOLS restriction. The same reasoning can be applied to the iOLS$_\delta$ condition $E[\log(\delta + U_i)|X_i] = c$. The null hypothesis is now

or probit model and testing the statistical significance of each X_i's. The null is rejected if any coefficient is found to differ significantly from zero.

[27]For the additive model (PPML), one must use the "additive error" $\hat{\epsilon}_i + 1 = \hat{U}_i / \exp(X_i'\hat{\beta}^{PPML})$, and regress $\hat{\epsilon}_i + 1 = \lambda E[U] Pr(U_i > 0|X_i)^{-1} + \nu_i$.

[28]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 $\hat{\beta}^{iOLS_U}$ which are separately estimated. We do not address this issue.

$$H_0 : E[\log(\delta + U_i)|X_i, U_i > 0] - \log(\delta) = \frac{c - \log(\delta)}{Pr(U_i > 0|X_i)}, \tag{32}$$

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. This approach yields t_{iOLS_δ}.

Testing other restrictions. The same reasoning can be applied to the popular fix or the IHS. Using (5), for the popular fix, the null hypothesis becomes

$$H_0 : E[\omega_i|X_i, U_i > 0] = (X'\beta - \log(\Delta))\frac{1 - Pr(U_i > 0|X_i)}{Pr(U_i > 0|X_i)}, \tag{33}$$

and the corresponding regression model is given by $\hat{\omega}_i = \lambda W_i + \nu_i$, for strictly positive errors only. For the popular fix, we would have $\hat{\omega}_i = \log(Y_i + \Delta) - X'\hat{\beta}^{PF}$ and $W_i = (X'\hat{\beta}^{PF} - \log(\Delta))(1 - \hat{P}(X_i))\hat{P}(X_i)^{-1}$.[29] The t-stat t_{PF} and t_{IHS} are obtained as above.

Testing whether zeros can be dropped. Discarding zeros is not recommended in general but can be valid in some settings. Once zeros are dropped, researchers generally estimate either the log-linear model (2) by OLS, or PPML based on (4).

In the former case, Heckman's model is particularly useful. Statistical significance of the parameter λ associated with the correction term, using the t-stat t_{HECK}, is evidence that dropping zeros introduces a selection bias. In the latter case, one can test for such bias by substituting (29) for $\lambda = 1$ into (3) to obtain

$$Y_i = \exp(X_i'\beta)\frac{E(U)}{Pr(U > 0|X_i)}\eta_i, \tag{34}$$

where $\eta_i - 1 = \nu_i \exp(X_i'\beta)\frac{E(U)}{Pr(U>0|X_i)})^{-1}$. This expression simplifies to $Y_i = \exp(X'\beta - \log(Pr(U > 0|X_i)))\eta_i$. Therefore, testing whether zeros can be discarded from PPML

[29]For the IHS, we would use the null hypothesis $H_0 : E[\omega_i|X_i, U_i > 0] = X'\beta\frac{1-Pr(U_i>0|X_i)}{Pr(U_i>0|X_i)}$. We would have $\hat{\omega}_i = \tilde{Y}_i - X'\hat{\beta}^{IHS}$ and $W_i = X'\hat{\beta}^{IHS}(1 - \hat{P}(X_i))\hat{P}(X_i)^{-1}$.

is possible by estimating the augmented model on the strictly positive observations

$$Y_i = \exp(X'\beta + \theta \log(\hat{P}(X_i))\eta_i, \tag{35}$$

and evaluate $H_0 : \theta = 0$ with the t-statistic t_{PPML0} for the new regressor $\log(\hat{P}(X_i))$. Under the Poisson condition, we should observe $\theta = -1$ but any deviation from $\theta = 0$ may signal that zeros play a non-negligible role, even if $E[U|X] = 1$ does not hold.

Conditional probability estimation. Those tests require 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 estimate of the conditional probability should hence be preferred whenever possible. Although consistent, nonparametric estimate can have poor small-sample behaviors especially at the tails. We use a k-nearest neighbors (kNN) algorithm (Hastie, Tibshirani and Friedman, 2009) and "trim" observations associated with predicted probabilities outside the 10% and 90% quantiles to correct for this issue.[30]

4.2 Model selection

We propose to select the most suitable approach using the previous tests.

First, iOLS$_\delta$ for any $\delta \in (0, \infty)$ is based on condition (A1):

$$E[\log(\delta + U)|X] = constant, \tag{A1}$$

which depends on the choice of δ. The "best" model within this category minimizes t^{iOLS_δ} with respect to δ. This rule will select the model with the least deviation between the implied and observed patterns of zeros. Second, PPML, iOLS$_U$ and

[30]See Cameron and Trivedi (2005) section 9.5.3 for a discussion of this common practice in nonparametric estimation.

iOLS_ϵ are based on condition (A2):

$$E[U|X] = constant, \tag{A2}$$

Third, OLS and PPML without zeros (or iOLS_S in the Appendix B.2) are based on either:

$$E[\log(U)|U > 0, X] = constant, \tag{A3}$$

or $E[U|U > 0, X] = constant$, which states that zeros can be discarded. Fourth, the Heckman's correction model is based on

$$(\varepsilon_i, \nu_i)' \sim \text{bivariate Gaussian}, \tag{A4}$$

which is not readily testable. Fifth, the popular fix or the IHS transformation relies on assumptions of the form

$$E[\omega_i|X] = constant, \tag{A5}$$

where ω_i is a known function of U_i, X_i and β.

We propose a model selection procedure predicated on first using models based on moment conditions rather than explicit distributional assumptions. This implies that we advocate for using more complex estimators only when the simpler ones are rejected. The selection procedure is as follows:

1. Compute t_{iOLS_U}, t_{PPML}, t_{iOLS_δ} for a range of δ, t_{PF} and t_{IHS}, and select the model with the smallest t-statistic in absolute value, denoted $|t_1|$. If $|t_1| < 1.96$, stop and select this model;

2. Else, compute t_{PPML0}, t_{HECK} and take the maximum in absolute value to define $|t_2|$. If $|t_2| < 1.96$, stop and report the estimates of PPML and OLS without zeros;[31]

3. Else, select Heckman's correction model or another mixture model.

[31] By taking the maximum of the two t-stats, we require that both tests suggest that zeros can be dropped before recommending to do so. The rationale is that both tests have power against different alternatives, hence combining them enlarges statistical power.

5 Simulations

Let us specify the dependent variable as

$$Y_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})U_i, \tag{36}$$

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 $\gamma_0 = -0.5$, $\gamma_1 = 0.5$ and $\gamma_2 = -0.5$. We consider six DGPs specified as follows:

- DGP 1 (A1): $E[X_i'(\log(1 + U_i) - c)] = 0$. This DGP is useful to illustrate iOLS$_\delta$ with $\delta = 1$. Let us assume that $\log(1 + \varepsilon_i)$ is uniformly distributed as $U[\frac{c}{2P(X_i)}, \frac{3c}{2P(X_i)}]$ with X_{1i} and X_{1i} also uniformly distributed as $U[-1, 2]$. Choosing $c = 0.41512$ yields the desired condition $E[X_i'(\log(1 + U_i) - c)] = 0$ with $E(U_i) = 1$.

- DGP 2 (A2): $E[U_i|X_i] = 1$. This DGP is aimed at comparing the alternative modelling approaches to PPML. We assume that $(X_{1i}, X_{2i})'$ is bivariate normal with mean zero, variance $\sigma^2_{X1} = \sigma^2_{X2} = 1$ and covariance $\sigma_{X1X2} = -0.3$. We further assume that ε_i is Gaussian with mean $-\log(P(X_i)) - 1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(X_i)$.

- The other DGPs are detailed in the Appendix. DGP 3 (A3) ($E[U_i|U_i > 0, X_i] = 1$) is such that discarding zeros and using PPML yields consistent estimates. DGP 4 (A4) is such that Heckman's model applies. DGP 5 (A5) is designed so that applying the IHS transform yields consistent OLS estimates. Finally, DGP 6 (IV) assumes $E[U_i|X_i] \neq 1$ but $E[U_i|Z_i] = 1$ which corresponds to the Poisson condition with endogenous regressors.

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 the following estimators: iOLS$_{\delta=1}$, iOLS$_{\delta=100}$, iOLS$_U$ (multiplicative Poisson), iOLS$_\epsilon$ (additive Poisson), PPML (additive Poisson), iOLS$_S$ (see Appendix B.5), OLS and PPML without zeros (PPML0), Heckman's corrected model, OLS after performing the inverse hyperbolic sine transform (IHS), and the popular fix with $\Delta = 0.7$ (PF).[32] For DGP 6, where regressors are endogenous, we report the 2SLS analog of the above estimators. The results for DGP 3 to 6 are included in Appendix C.

Table 1: Simulations: DGP 1 (A1: $E[X_i'(\log(1+U_i)-c)] = 0$)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	0.99	1.01	0.99	1.00	1.00	1.00
		(0.09)	(0.10)	(0.10)	(0.03)	(0.03)	(0.03)
	$iOLS_{\delta=100}$	0.92	0.72	1.28	0.92	0.72	1.28
		(0.07)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)
(A2)	$iOLS_U$	0.91	0.71	1.29	0.92	0.70	1.30
		(0.07)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)
	$iOLS_\varepsilon$	1.00	0.61	1.27	1.01	0.60	1.28
		(0.14)	(0.13)	(0.12)	(0.04)	(0.04)	(0.04)
	$PPML$	1.00	0.61	1.27	1.01	0.60	1.27
		(0.14)	(0.13)	(0.12)	(0.04)	(0.04)	(0.04)
(A3)	$iOLS_S$	0.97	0.46	1.54	0.97	0.46	1.54
		(0.06)	(0.04)	(0.04)	(0.02)	(0.01)	(0.01)
	OLS	1.77	0.50	1.50	1.77	0.50	1.50
		(0.04)	(0.03)	(0.04)	(0.01)	(0.01)	(0.01)
	$PPML0$	2.04	0.27	1.58	2.05	0.26	1.58
		(0.08)	(0.09)	(0.07)	(0.02)	(0.03)	(0.02)
(A4)	$Heckman$	−8.18	2.48	−0.48	−7.92	2.47	−0.47
		(3.48)	(0.56)	(0.56)	(1.02)	(0.17)	(0.17)
Others	$IHST$	0.89	0.56	0.18	0.89	0.56	0.18
		(0.05)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)
	PF	0.46	0.52	0.18	0.46	0.52	0.18
		(0.05)	(0.06)	(0.06)	(0.01)	(0.02)	(0.02)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP1. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard errors are presented in parentheses.

Bias and variance. Table 1 reports the results for DGP 1 based on the true identifying condition $E[X_i'(\log(1+U_i)-c)] = 0$. All estimators but iOLS$_{\delta=1}$ are

[32] This is the "best" value for Δ which minimizes the mean square bias, see Section 2.1.

biased, confirming that the identifying conditions of iOLS$_\delta$ indeed differ from those assumed by PPML. This bias is severe for the inverse hyperbolic sine transformation, the popular fix, and the Heckman correction. PPML exhibits a smaller bias than existing alternative estimators and is found to have greater variance than iOLS$_U$. As expected, iOLS$_\epsilon$ corresponds exactly to PPML estimates. These results also illustrate the $\sqrt{n}$-consistency of the estimators as the standard errors are divided by 10 as the sample size increases by 100-fold.

Table 2: Simulations: DGP 2 (A2: $E[U_i|X_i] = 1$)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	1.06	1.26	0.74	1.08	1.25	0.75
		(0.19)	(0.11)	(0.11)	(0.06)	(0.04)	(0.04)
	$iOLS_{\delta=100}$	0.98	1.03	0.97	1.00	1.02	0.98
		(0.17)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)
(A2)	$iOLS_U$	0.98	1.01	0.99	1.00	1.00	1.00
		(0.17)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)
	$iOLS_\varepsilon$	1.02	0.99	0.97	1.00	1.00	1.00
		(0.47)	(0.17)	(0.21)	(0.19)	(0.06)	(0.09)
	$PPML$	1.02	0.99	0.97	1.01	1.00	0.99
		(0.47)	(0.17)	(0.21)	(0.19)	(0.06)	(0.09)
(A3)	$iOLS_S$	1.08	0.74	1.27	1.09	0.73	1.27
		(0.14)	(0.07)	(0.07)	(0.04)	(0.02)	(0.02)
	OLS	1.52	0.73	1.27	1.52	0.73	1.27
		(0.10)	(0.06)	(0.05)	(0.03)	(0.02)	(0.02)
	$PPML0$	2.05	0.69	1.29	2.05	0.69	1.30
		(0.36)	(0.14)	(0.17)	(0.15)	(0.06)	(0.07)
(A4)	$Heckman$	−1.45	1.30	0.70	−1.40	1.30	0.70
		(1.87)	(0.35)	(0.35)	(0.56)	(0.11)	(0.11)
Others	$IHST$	0.82	0.72	0.05	0.82	0.72	0.05
		(0.11)	(0.07)	(0.07)	(0.03)	(0.02)	(0.02)
	PF	0.38	0.68	0.06	0.38	0.68	0.06
		(0.10)	(0.07)	(0.07)	(0.03)	(0.02)	(0.02)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP2. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard errors are presented in parentheses.

Table 2 reports the results for DGP 2 when $E[U_i|X_i] = 1$ is the true identifying condition. We first observe that only PPML, iOLS$_U$ and iOLS$_\epsilon$ are consistent. Second, we nonetheless see that iOLS$_{\delta=100}$ exhibits a small bias but does not exclude the true value from its confidence interval. Third, we note that iOLS$_U$ dominates PPML in terms of precision under this simulation design. Finally, we observe that the estimates of PF and IHS have large biases.

Tests and model selection. The simulations are also useful to study our testing procedures. The conditional probabilities to have a zero are logistic in all DGPs but (A4). 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) or a misspecified model (probit).

First, we report the frequency of selecting each δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ for DGPs 1 and 2, based on the smallest t-stat (section 4.1), When the true restriction is that of iOLS$_{\delta=1}$ (DGP 1), this approach selects $\delta = 1$ correctly 50% of the time for $n = 10,000$ as opposed to 17% of the time when $n = 1000$. In comparison, when the exogeneity condition of PPML is correct (DGP 2), a large δ is selected in most simulations.[33].

Table 3: Simulations: Data-driven selection of δ (iOLS$_\delta$)

		δ						
n	DGP	0.1	0.5	1	5	10	50	100
1000	1	28%	18%	17%	16%	9%	5%	7%
	2	14%	9%	8%	9%	7%	5%	48%
10,000	1	3%	32%	50%	14%	1%	0%	0%
	2	0%	0%	2%	8%	14%	19%	57%

Notes: This table shows the relative frequency with which a given δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ was chosen on the basis of the 10,000 simulations. These simulations vary by sample size n and by DGP. These test assume the probability model to be logistic. Interpretation: when the sample size is n=10,000 and the data was generated using DGP1, $t_{iOLS_{\delta=1}}$ was the smallest 50% of the time, so $\delta = 1$ was selected 50% of the time.

[33] Results are similar for kNN and Probit (Appendix C)

Second, we show the empirical size and power of each test for all DGPs in Table 4 for a nominal size of 5%. In DGP 1, the t-test $t_{\delta=1}$ rejects $\delta = 1$ only 5% of the time which corresponds to the nominal test size. The tests for $iOLS_{\delta=100}$, $iOLS_U$ and $PPML$ have power against this alternative with a 100% rejection rate in large samples. The results for DGP 2 show that the tests for $\text{iOLS}_{\delta=100}$, iOLS_U, PPML and iOLS_ϵ are correctly sized, and that the other tests have satisfactory power.[34] Finally, the test for IHS is correctly sized in DGP 5. It is worth noting that not all tests have power against each of the considered alternative, even in large samples. This is the case for $t_{\delta=1}$ and t_{PF} in DGP 5. Finally, the last column reports the empirical rejection rates of the RESET test for PPML, including 3 polynomials terms (Santos Silva and Tenreyro, 2006; Ramsey, 1969). Findings reveal that the test is slightly oversized and lacks power against all considered alternatives. For comparison, we report the empirical sizes and powers when using kNN instead of logit in Table 5.[35] The tests' sizes are slightly distorted but exhibit satisfactory power.

Table 4: Simulations: Specification testing (Logit)

n	DGP	$t_{\delta=1}$	$t_{\delta=100}$	t_U	t_ε	t_{PPML}	t_{IHST}	$RESET$
1000	1	6%	26%	25%	57%	57%	19%	12%
	2	9%	5%	5%	4%	4%	22%	6%
	3	11%	8%	8%	9%	9%	18%	5%
	4	10%	7%	7%	13%	13%	39%	6%
	5	6%	7%	7%	19%	19%	6%	9%
10,000	1	5%	100%	100%	100%	100%	87%	9%
	2	47%	5%	5%	5%	5%	100%	7%
	3	75%	33%	39%	62%	62%	100%	6%
	4	61%	25%	30%	71%	70%	100%	6%
	5	6%	22%	24%	94%	94%	5%	9%

Notes: This table shows the relative rejection frequency of each null hypothesis for 10,000 simulations. These simulations vary by sample size (as reported by the column "n") and by Data Generating Process (as reported in the column "DGP"). These test assume the probability model to be logistic. RESET refers to the t-statistic associated with the joint significance of three polynomial terms. Interpretation: when the sample size is n=1000 and the data was generated using DGP1, $t_{\delta=1}$ was rejected 6% of the time.

[34] Although not reported here for readability, t_{HECK} and t_{PPML0} are correctly sized with only 5% rejection rates in DGP 3 where zeros can be dropped, and exhibit some power under the alternatives.

[35] Results for Probit are reported in Table 16.

Finally, Table 6 reports selection rates using our procedure. The correct model is chosen more often as the sample size enlarges. However, the estimates' precision largely drives this selection. For example, the Heckman model is only selected 14% of the time when n=10,000 in DGP 4, requiring more observations to approach 100%.

Table 5: Simulations: Specification testing (kNN)

n	DGP	$t_{\delta=1}$	$t_{\delta=100}$	t_U	t_ε	t_{PPML}	t_{IHST}	$RESET$
	1	8%	16%	18%	9%	9%	6%	12%
	2	39%	9%	8%	6%	6%	14%	6%
1000	3	66%	37%	35%	16%	16%	39%	5%
	4	62%	29%	26%	9%	9%	70%	5%
	5	5%	6%	6%	4%	4%	6%	9%
	1	8%	98%	99%	63%	63%	21%	9%
	2	98%	12%	8%	7%	7%	71%	7%
10,000	3	100%	95%	93%	53%	53%	100%	6%
	4	100%	93%	88%	23%	23%	100%	6%
	5	7%	13%	14%	5%	5%	5%	9%

Notes: This table shows the relative rejection frequency of each null hypothesis for 10,000 simulations. These simulations vary by sample size (as reported by the column "n") and by Data Generating Process (as reported in the column "DGP"). These test are based on a non-parametric KNN probability model, trimmed of the top and bottom 10% observations. RESET refers to the t-statistic associated with the joint significance of three polynomial terms. Interpretation: when the sample size is n=1000 and the data was generated using DGP1, $t_{\delta=1}$ was rejected 8% of the time.

Table 6: Simulations: Model selection

n	DGP	$(A1)$	$(A2)$	$(A3)$	$(A4)$	$(A5)$
	1	52%	25%	0%	0%	23%
	2	24%	67%	4%	0%	5%
1000	3	18%	60%	20%	0%	3%
	4	25%	56%	19%	0%	0%
	5	14%	14%	63%	0%	9%
	1	93%	0%	0%	4%	2%
	2	5%	95%	0%	0%	0%
10,000	3	0%	31%	67%	2%	0%
	4	1%	67%	19%	13%	0%
	5	41%	14%	3%	0%	41%

Notes: This table shows the selection frequency of each identifying restriction for 10,000 simulations. These simulations vary by sample size (as reported by column "n") and by Data Generating Process (as reported in column "DGP"). Selection is done assuming the probability model to be logistic. Interpretation: when the sample size is n=1000 and generated by DGP1, a model with moments (A1) is chosen 51% of the time.

6 Application

We now revisit three empirical studies published in top-tier economic journals where the log of zero had to be addressed. First, Santos Silva and Tenreyro (2006) compare various estimators to estimate gravity models of trade. Second, Michalopoulos and Papaioannou (2013) use the popular fix to examine the role of pre-colonial ethnic institutions on economic development. Third, Card and DellaVigna (2020) investigate the preferences of academic journal editors with the IHS transformation in the context of endogenous regressors.

For brevity, we report only the main estimates for the most relevant estimators. We focus on the data-driven selected δ for iOLS$_\delta$, along with iOLS$_U$, PPML and IHS. Standard errors, as reported in parenthesis, are obtained using 300 pairs bootstrap. Comprehensive results for all estimators and tests discussed in the paper are given in the Appendix.

6.1 Santos Silva and Tenreyro (2006)

First, we study the gravity model of Santos Silva and Tenreyro (2006). Their Table 3 reports models of bilateral trade for data covering 136 countries in 1990. For importer (I) and exporter (X) countries, they control for log(GDP), log(GDP per capita), log(Distance), along with dummies for contiguity, shared language, colonial ties, access to the oceans, remoteness (which measures the access to other trading partners), free trade agreement (FTA), and the existence of a preferential trade agreement. They advocate for PPML over a log-linear model arguing that the latter is biased in presence of heteroskedastic errors.

We report the t-statistic and main estimates in Table 7 in order to compare the different approaches.[36] The tests provide evidence against iOLS estimators but fail to reject PPML. Although the test for PPML0 fail to reject that zeros can be discarded, the test for Heckman provides evidence of the opposite. Therefore, dropping zeros is not recommended and PPML should be preferred when using the parametric test with logit probability. However, it should be noted that using a nonparametric

[36]The associated $\hat{\lambda}$ statistics for all specifications are provided in Table 22 in the Appendix.

approach (kNN) to estimate the conditional probability of observing zeros for the test provides a different conclusion. PPML is no longer the preferred model and is rejected in favor of iOLS$_{\delta=100}$ and iOLS$_U$ (see Table 22 in the Appendix).

Table 7: Estimates from Santos Silva and Tenreyro (2006)'s Table 3

	iOLS$_{\delta=100}$	PPML*	iOLS$_U$	Heckman	PPML0
Log(Distance)	-1.52	**-0.78**	-1.48	-1.26	-0.78
	(0.08)	**(0.06)**	(0.08)	(0.04)	(0.06)
Contiguity	0.13	**0.19**	0.20	0.15	0.20
	(0.38)	**(0.10)**	(0.38)	(0.13)	(0.10)
Language	0.76	**0.75**	0.69	0.77	0.75
	(0.13)	**(0.14)**	(0.13)	(0.07)	(0.14)
Colonial	0.41	**0.03**	0.39	0.44	0.02
	(0.15)	**(0.15)**	(0.15)	(0.07)	(0.15)
FTA	1.45	**0.18**	1.55	0.46	0.18
	(0.49)	**(0.10)**	(0.57)	(0.11)	(0.10)
$\hat{\lambda}$	0.46	**1.26**	0.44	0.81	-0.22
	(0.06)	**(0.39)**	(0.06)	(0.09)	(0.37)
t-Stat.	[-9.82]	**[0.68]**	[-9.24]	[8.75]	[-0.59]

Notes: This table displays the main coefficients, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for several models of trade gravity, based on using a logistic probability model. The full list of control variables is provided in Section 6.1. $iOLS_{\delta=100}$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. The symbol ∗denotes the specification recommended by the authors in their original article. Our preferred specification (i.e. with the smallest t-stat) is in bold.

6.2 Michalopoulos and Papaioannou (2013)

Michalopoulos and Papaioannou (2013) 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": $\log(0.01+Y_i)$. The authors focuses on the coefficient associated with Murdock's 1967 index of jurisdictional hierarchy.[37] The cross-sectional unit is ethnicity-by-country. They control cumulatively for population

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

density, location, and geography,[38] and find positive and significant estimates.

Table 8: Estimates from Michalopoulos and Papaioannou (2013)'s Table 2

	PF*	$iOLS_{\delta=0.05}$	$iOLS_{\delta=100}$	PPML	$iOLS_U$
Pop.					
$\hat{\beta}$	0.35	**0.53**	0.44	0.29	0.41
	(0.07)	**(0.11)**	(0.13)	(0.12)	(0.14)
$\hat{\lambda}$	-0.88	**1.01**	1.08	2.62	1.09
	(0.29)	**(0.02)**	(0.06)	(0.57)	(0.07)
t-Stat.	[-6.40]	**[0.67]**	[1.36]	[2.85]	[1.30]
Pop. & Loc.					
$\hat{\beta}$	0.32	**0.40**	0.36	0.14	0.35
	(0.06)	**(0.09)**	(0.09)	(0.11)	(0.10)
$\hat{\lambda}$	-0.56	**1.00**	1.03	3.68	1.03
	(0.24)	**(0.04)**	(0.05)	(1.24)	(0.05)
t-Stat.	[-6.58]	**[-0.04]**	[0.58]	[2.16]	[0.55]
Pop. & Loc. & Geo.					
$\hat{\beta}$	0.19	0.09	0.11	0.00	**0.10**
	(0.05)	(0.11)	(0.09)	(0.10)	**(0.09)**
$\hat{\lambda}$	-0.18	0.82	0.85	1.94	**0.85**
	(0.12)	(0.19)	(0.21)	(0.91)	**(0.22)**
t-Stat.	[-9.46]	[-0.91]	[-0.69]	[1.04]	**[-0.68]**

Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. These tests are based on using a logistic probability model. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). Three specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). Full estimates are available in Table 26 of the Appendix. The symbol ∗denotes the specification used by the authors in their original article. Our preferred specifications (i.e. with the smallest t-stat) are in bold.

The results are reported in Table 8. The popular fix provides overly precise estimates but is always rejected by our tests with λ far from 1. In comparison, iOLS yields values of λ fairly close to 1. Adding geographic controls lowers the estimated β in all cases to the point where it is not statistically significant anymore. We fail

[38]We focus on columns (2)-(4) of their Table 2. Full results along with a replication of their Table 3 (Panel A, column (1)-(4)), which includes additional country fixed effects, are provided in the Appendix. The same conclusions apply.

to reject PPML in this specification although with a fairly imprecise estimate of λ. Those results reveal a much weaker statistical relationship between the variables under study than considered in the original paper.[39] Interestingly, the best model suggested by the test differs depending on the explanatory variables included in the specification. Indeed, the exogeneity condition directly depends on X, which implies that the right model to use for the estimation of the effects can change from one specification to another, simply by adding a new variable in X, even though these specifications are very close.[40]

6.3 Card and DellaVigna (2020)

Finally, we revisit Card and DellaVigna (2020)'s study of journal editors. The data contains submission-level information from four leading economics journal matched to Google Scholar citations. The authors address the log of zero with the IHS transformation and study the role of many control variables. For simplicity, we focus our attention on measuring the impact of receiving an invitation to *Revise & Resubmit* on the number of citations, which is considered an endogenous variable. The authors take a control function approach, instrumented using the "leave-out mean R&R rate of the editor".[41]

Table 9 reports estimates in three cases: without correcting for the endogeneity, using the control function approach, and with instrumental variables. In all cases, $iOLS_{\delta=50}$ is selected and PPML is rejected. The IHS is also rejected although with a λ fairly close to 1.[42] Accounting for the endogeneity of this variable yields a lower

[39]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 the Appendix). In contrast, we find an absence of significance in terms of pre-ethnic political hierarchy and economic development.

[40]As seen in Table 25 in the Appendix, the nonparametric probability model used for the tests yields qualitatively similar results for our main specification of interest (i.e, with population, location, and geographical controls). Indeed, $\text{iOLS}_{\delta=100}$ and iOLS_U remain our favored specification, associated in both cases with λ close to one.

[41]This instrument measures the frequency with which the same editor has invited *other* authors to revise their manuscript before reassessment.

[42]The kNN nonparametric probability model favors $\text{iOLS}_{\delta=50}$ in the instrumental variable case but rejects all models in the other cases, as shown in Table 32 in the Appendix.

estimate, even negative when using iOLS. Yet, it is never statistically significant.

This effect is interpreted as the mechanical publication effect and has a positive sign in the original paper: an invitation for R&R should yield additional citations. Although not statistically significant, its sign changes when using $iOLS_\delta$ and $iOLS_U$ or using 2SLS instead of the control function. We interpret this negative sign as follows. An editor which is more likely to offer R&R will mechanically do so for papers with lesser potential to attract citations. Specification tests presented in the Appendix point in favor of $iOLS_{\delta=50}$ in all specifications, and reject all other estimators based on the Poisson condition or discarding zeros.

Table 9: Estimates from Card and DellaVigna (2020)

	IHS*	iOLS$_{\delta=50}$	PPML	iOLS$_U$
No correction for Endogeneity				
$\hat{\beta}$	0.57	**0.48**	0.53	0.48
	(0.05)	**(0.04)**	(0.04)	(0.04)
$\hat{\lambda}$	0.97	**1.00**	1.12	1.00
	(0.00)	**(0.00)**	(0.01)	(0.00)
t-Stat.	[-6.64]	**[0.73]**	[8.63]	[-3.06]
Control Function				
$\hat{\beta}$	0.07	**-0.09**	0.11	-0.08
	(0.14)	**(0.13)**	(0.13)	(0.13)
$\hat{\lambda}$	0.97	**1.00**	1.13	1.00
	(0.00)	**(0.00)**	(0.02)	(0.00)
t-Stat.	[-5.84]	**[0.83]**	[8.36]	[-2.93]
Instrumental Variables				
$\hat{\beta}$	-1.77	**-1.20**	-1.36	-1.11
	(1.32)	**(1.86)**	(0.91)	(1.76)
$\hat{\lambda}$	0.97	**1.00**	1.14	1.00
	(0.01)	**(0.10)**	(0.02)	(0.00)
t-Stat.	[-3.53]	**[-0.04]**	[7.54]	[-0.88]

Notes: This table displays the coefficient associated with an invitation to revise & resubmit (R&R), standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of citations. $iOLS_\delta$ and $iOLS_U$ are defined in Section 3. Three specifications are presented: no correction for endogeneity contrasts with control function and instrumental variables which rely on the Editor leave-out mean R&R rate for identification. The symbol $*$denotes the specification used by the authors in their original article. Our preferred specifications (i.e. with the smallest t-stat) are in bold.

7 Conclusion

This paper developed multiple contributions to address a common yet unresolved issue faced in empirical research: the log of zero. First, we have attempted to clarify some issues and misconceptions with respect to existing practices, such as adding an arbitrary constant to the dependent variable. Second, we have derived a new family of estimators to estimate log-linear models when the dependent variable can take non-positive values. Those estimators have 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. Third, we have developed testing procedures to verify the underlying exogeneity restrictions imposed by our estimators and other well-known approaches, such as PPML or IHS. We show how these tests can be helpful to guide empirical research. Fourth, all methods are illustrated through numerical simulations and replications of recent publications in top-tier economics journals. We find that the exogeneity restrictions used by our estimators are rarely rejected and often selected as the best solution in those applications.

The main takeaway from our research should be that no single method works for all settings, hence different methods can lead to different conclusions. Hopefully, empirical researchers are now better equipped to substantiate their preferred method in any given setting. 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, including semi-parametric models of sample selection and regularized models like the lasso, which we leave for future research.

References

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

Battese, G. E. 1997. "A Note on the Estimation of Cobb-Douglas Production Functions When Some Explanatory Variables Have Zero Values." *Journal of Agricultural Economics*, 48(13): 250–252.

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.

Blundell, Richard, and Jean Marc Robin. 1999. "Estimation in Large and Disaggregated Demand Systems: An Estimator for Conditionally Linear Systems." *Journal of Applied Econometrics*, 14(3): 209–232.

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.

Cameron, A., and Pravin Trivedi. 2005. *Microeconometrics.* Cambridge University Press.

Card, D., and S. DellaVigna. 2020. "What Do Editors Maximize? Evidence from Four Economics Journals." *The Review of Economics and Statistics*, 102(1): 195–217.

Carroll, Raymond J., and David Ruppert. 1984. "Power Transformations when Fitting Theoretical Models to Data." *Journal of the American Statistical Association*, 79(386): 321–328.

Chamberlain, Gary. 1987. "Asymptotic efficiency in estimation with conditional moment restrictions." *Journal of econometrics*, 34(3): 305–334.

Ciani, Emanuele, and Paul Fisher. 2018. "Dif-in-Dif Estimators of Multiplicative Treatment Effects." *Journal of Econometric Methods*, 8(1): 20160011.

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.

Dembinski, Hans, Michael Schmelling, and Roland Waldi. 2019. "Application of the iterated weighted least-squares fit to counting experiments." *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 940: 135141.

De Vos, Kristof. 2015. "Negative Wholesale Electricity Prices in the German, French and Belgian Day-Ahead, Intra-Day and Real-Time Markets." *The Electricity Journal*, 28(4): 36–50.

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.

Draca, Mirko, Stephen Machin, and John Van Reenen. 2011. "Minimum Wages and Firm Profitability." *American Economic Journal: Applied Economics*, 3(1): 129–51.

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.

Goldberger, Arthur S. 1968. "The Interpretation and Estimation of Cobb-Douglas Functions." *Econometrica*, 36(3/4): 464–472.

Gourieroux, C., A. Monfort, and A. 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, K., and T. Mayer. 2019. "Brands in Motion, How Frictions Shape Multinational Production." *American Economic Review*, forthcoming.

Head, Keith, and Thierry Mayer. 2014. "Gravity Equations: Workhorse,Toolkit, and Cookbook." In *Handbook of International Economics*. Vol. 4 of *Handbook of International Economics*, , ed. G. Gopinath, . Helpman and K. Rogoff, Chapter 0, 131–195. Elsevier.

Heckman, James J. 1979. "Sample selection bias as a specification error." *Econometrica: Journal of the econometric society*, 153–161.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein. 2008. "Estimating Trade Flows: Trading Partners and Trading Volumes*." *The Quarterly Journal of Economics*, 123(2): 441–487.

Huber, Kilian. 2018. "Disentangling the Effects of a Banking Crisis: Evidence from German Firms and Counties." *American Economic Review*, 108(3): 868–98.

Johnson, N. L. 1949. "Systems of Frequency Curves Generated by Methods of Translation." *Biometrika*, 36(1-2): 149–176.

Johnson, S. R., and G. C. Rausser. 1971. "Effects of Misspecifications of Log-Linear Functions When Sample Values Are Zero or Negative." *American Journal of Agricultural Economics*, 53(1): 120–124.

MacKinnon, James G, and Lonnie Magee. 1990. "Transforming the dependent variable in regression models." *International Economic Review*, 315–339.

MaCurdy, T. E., and J. H. Pencavel. 1986. "Testing between Competing Models of Wage and Employment Determination in Unionized Markets." *Journal of Political Economy*, 94(3): S3–S39.

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

Santos Silva, J.M.C., and S. Tenreyro. 2010. "On the existence of the maximum likelihood estimates in Poisson regression." *Economics Letters*, 107(2): 310 – 312.

Santos Silva, J.M.C., and S. Tenreyro. 2011. "poisson: Some convergence issues." *The Stata Journal*, 11(2): 207 – 212.

Santos Silva, Joo 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.

Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica*, 26(1): 24–36.

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

Wooldridge, JW. 1997. "Quasi-likelihood methods for count data." *Handbook of applied econometrics*, 2: 352–406.

Worrall, John L., and Tomislav V. Kovandzic. 2010. "Police levels and crime rates: An instrumental variables approach." *Social Science Research*, 39(3): 506–516.

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.

A Mathematical Appendix

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

$$\beta = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\beta)\right], \tag{37}$$

where $\tilde{Y}_i(\beta) = \log(Y_i + \exp(X_i'\beta)) - c(\beta)$ 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

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

The sample counterpart of this mapping is given by

$$\hat{M}_n(\phi) = [X'X]^{-1} X' \hat{\tilde{Y}}_i(\phi), \tag{39}$$

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) who develop a convergence theory for iterative estimators. Following their theory, the convergence of iOLS requires that $M(\cdot)$ and $\hat{M}_n(\cdot)$ be contraction mappings, asymptotically.[43]

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 Dominitz and Sherman (2005), 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 $|\hat{\beta}_0 - \beta|$;[44]

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

[43] The reader is referred to Dominitz and Sherman (2005) for a formal definition of an asymptotic contraction mapping.

[44] Note that Dominitz and Sherman (2005)'s condition (i) is about $M(\cdot)$ and not $\hat{M}_n(\cdot)$. However those conditions imply each other under conditions (iii) and (iv) by applying their Lemma 3 with trivial modifications.

(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* $n\to\infty$.

For condition (i), we adapt the proof of Lemma 5 in Dominitz and Sherman (2005) 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. From a multivariate Taylor expansion argument, Dominitz and Sherman (2005) 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 that unity as* $n\to\infty$*. Note that we have*

$$\begin{aligned}\hat{V}_n(\phi) &= [X'X]^{-1}X'\nabla_\phi\hat{\hat{Y}}(\phi)\\ &= n^{-1}X'\nabla_\phi\hat{\hat{Y}}(\phi),\end{aligned} \tag{40}$$

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

$$\left[\nabla_\phi\hat{\hat{Y}}(\phi)\right]_{i,k} = \frac{\exp(X_i'\phi)X_{ki}}{Y_i+\exp(X_i'\phi)} - \frac{\partial\hat{c}(\phi)}{\partial\phi_k}. \tag{41}$$

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* $k>1$.

$$\frac{\partial\hat{c}(\phi)}{\partial\phi_k} = n^{-1}\sum_{i=1}^n\frac{\exp(X_i^{r\prime}\phi^r+\hat{\phi}^1)(\frac{\partial\hat{\phi}^1}{\partial\phi_k}+X_{ki})}{Y_i+\exp(X_i^{r\prime}\phi^r+\hat{\phi}^1)} - n^{-1}\sum_{i=1}^n(\frac{\partial\hat{\phi}^1}{\partial\phi_k}+X_{ki}), \tag{42}$$

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

$$\frac{\partial\hat{c}(\beta)}{\partial\phi_k} = n^{-1}\sum_{i=1}^n\frac{X_{ki}}{1+U_i}+O_p(1), \tag{43}$$

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*

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n}\frac{X_{ki}X_{li}}{1+U_i} - n^{-1}\sum_{i=1}^{n}X_{ki}n^{-1}\sum_{j=1}^{n}\frac{X_{lj}}{1+U_j}, \tag{44}$$

for $l > 1$ and

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

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

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

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

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'WX, \tag{47}$$

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

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'W^{1/2}W^{1/2}X. \tag{48}$$

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

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

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

$$\max_{|a|=1} a' \left[\hat{V}_n(\beta)\right]_{k,l>1} a = \max_{|a|=1} 1 - a'X'D^{1/2}D^{1/2}Xa. \tag{50}$$

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.

Let us turn to condition (ii). Following Dominitz and Sherman (2005), a sufficient condition to satisfy (ii) 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.

For condition (iii), we want to show that $n^{1/2}\sup_{\phi\in B_0}|\hat{M}_n(\phi) - M(\phi)| = O_p(1)$ 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 $\phi \in B_0$, and thus in particular

$$n^{1/2}\sup_{\phi\in B_0}|\hat{M}_n(\phi) - M(\phi)| = O_p(1). \tag{51}$$

For 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. Applying Theorem 1 in Dominitz and Sherman (2005) yields the desired convergence result.

Proof 2 (Proof of Theorem 1: Normality) *We now make use of Theorem 4 in Dominitz and Sherman (2005) 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

$$\hat{c}(\beta) = n^{-1}\sum_{i=1}^{n}\log(n^{-1}\sum_{j=1}^{n}U_j + U_i) - \log(n^{-1}\sum_{j=1}^{n}U_j) \xrightarrow{p} E[\log(1+U_i)] = c, \quad (52)$$

as $n \to \infty$, and

$$\hat{\tilde{Y}}_i(\beta) = \log(1+U_i) + X_i'\beta - \hat{c}(\beta), \quad (53)$$

so that

$$\sqrt{n}[X'X]^{-1}X'\hat{\tilde{Y}}_i(\beta) = \sqrt{n}\left(\beta + [X'X]^{-1}X'(\log(1+U) - \hat{c}(\beta))\right). \quad (54)$$

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

$$\sqrt{n}\left(\hat{M}_n(\beta) - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad (55)$$

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 the following result

$$\sqrt{n}\left(\hat{\beta}_{i(n)} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Omega^{-1}), \quad (56)$$

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

$$V(\beta) = E[X_iX_i']^{-1}E[\frac{X_iX_i'}{1+U_i}], \quad (57)$$

of which each element is strictly below 1. Therefore sandwich-type covariance estimators are changed from the classical expression

$$\hat{\Sigma} = (\frac{1}{n}X'X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'X)^{-1} \quad (58)$$

to

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{59}$$

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* $\delta \neq 1$.

An approximate solution consists in evaluating U_i *at its mean so that a simple (though biased) estimator is given by*

$$\hat{\Omega} = \frac{1+\delta}{\delta} \times \hat{\Sigma}, \tag{60}$$

where $\hat{\Sigma}$ *is the estimated covariance matrix (robust or not) of the OLS estimator in the last iteration. For instance, this approximation yields standard errors twice as large as those of the OLS procedure for* $\delta = 1$.

Proof 3 (Proof of Theorem 2: iOLS$_U$) *This proof is similar to that of the previous theorem, with small modifications.*

Let us now consider

$$\begin{aligned}\hat{V}_n(\phi) &= [X'X]^{-1}X'\nabla_\phi \hat{\tilde{Y}}(\phi) \\ &= n^{-1}X'\nabla_\phi \hat{\tilde{Y}}(\phi),\end{aligned} \tag{61}$$

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

$$\left[\nabla_\phi \hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\delta \exp(X_i'\phi)X_{ki}}{Y_i + \delta \exp(X_i'\phi)} + \frac{\partial \hat{U}_i(\phi)}{\partial \phi_k}\left(\frac{1}{1+\delta} - \frac{1}{\hat{U}_i(\phi)+\delta}\right). \tag{62}$$

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

$$\left[\nabla_\beta \hat{\tilde{Y}}(\beta)\right]_{i,k} = X_{ki}\left(1 - \frac{U_i}{1+\delta}\right), \tag{63}$$

which yields

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} X_{ki}X_{li}\left(1 - \frac{U_i}{1+\delta}\right). \tag{64}$$

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

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{65}$$

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

Proof 4 (Proof of Theorem 2: iOLS$_\epsilon$)) *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*

$$\left[\nabla_\phi \hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\delta\exp(X_i'\phi)X_{ki}}{Y_i + \delta\exp(X_i'\phi)} - \frac{1}{\hat{U}_i(\phi)+\delta}\frac{\partial \hat{U}_i(\phi)}{\partial \phi_k} + \frac{1}{1+\delta}\frac{\partial(Y_i - \exp(X_i'\phi))}{\partial \phi_k}. \tag{66}$$

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

$$\left[\nabla_\beta \hat{\tilde{Y}}(\beta)\right]_{i,k} = X_{ki}\left(1 - \frac{\exp(X_i'\beta)}{1+\delta}\right), \tag{67}$$

which yields

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} X_{ki}X_{li}\left(1 - \frac{\exp(X_i'\beta)}{1+\delta}\right). \tag{68}$$

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

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{69}$$

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.

Proof 5 (Proof of Theorem 3 : Instrumental Variables Consistency) *Recall that the parameter $\beta \in \mathbb{R}^K$ is characterized by the fixed-point equation*

$$\beta^{IV} = E[\breve{X}_i\breve{X}_i{}']^{-1}E\left[\breve{X}_i\tilde{Y}_i(\beta)\right], \tag{70}$$

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

$$M^{IV}(\phi) = E[\breve{X}_i\breve{X}_i{}']^{-1}E\left[\breve{X}_i\tilde{Y}_i(\phi)\right]. \tag{71}$$

The sample counterpart of this mapping is given by

$$\hat{M}_n^{IV}(\phi) = [\breve{X}_i{}'\breve{X}_i]^{-1}\breve{X}_i{}'\hat{\tilde{Y}}_i(\phi), \tag{72}$$

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

Our proof is very similar to the one used to show Theorem 1. 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

$$\begin{aligned}\hat{V}_n^{IV}(\phi) &= [\breve{X}'\breve{X}]^{-1}\breve{X}'\nabla_\phi\hat{\tilde{Y}}(\phi)\\ &= n^{-1}\breve{X}'\nabla_\phi\hat{\tilde{Y}}(\phi),\end{aligned} \tag{73}$$

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

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

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

$$\frac{\partial \hat{c}(\phi)}{\partial \phi_k} = n^{-1}\sum_{i=1}^n \frac{\exp(X_i^{r\prime}\phi^r + \hat{\phi}^1)(\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki})}{Y_i + \exp(X_i^{r\prime}\phi^r + \hat{\phi}^1)} - n^{-1}\sum_{i=1}^n (\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki}), \tag{75}$$

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

$$\begin{aligned}\frac{\partial \hat{c}(\beta)}{\partial \phi_k} =& n^{-1}\sum_{i=1}^n \frac{X_{ki}}{1+U_i} - n^{-1}\sum_{i=1}^n X_{ki} + O_p(1)\\ =& n^{-1}\sum_{i=1}^n \frac{X_{ki}U_i}{1+U_i} + O_p(1),\end{aligned} \tag{76}$$

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

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{\breve{X}_{ki}X_{li}}{1+U_i} - (n^{-1}\sum_{i=1}^n \breve{X}_{ki})(n^{-1}\sum_{j=1}^n \frac{X_{lj}U_j}{1+U_j}), \tag{77}$$

for $l > 1$ and

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{\breve{X}_{ki}}{1+U_i}, \tag{78}$$

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

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n}\frac{\breve{X}_{ki}X_{li}}{1+U_i}. \tag{79}$$

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

$$\left[\hat{V}_n^{IV}(\beta^{IV})\right] = n^{-1}X'P_zWX, \tag{80}$$

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

$$\begin{aligned}\left[\hat{V}_n^{IV}(\beta)\right] &= n^{-1}X'P_z(I_n - D)X \\ &= X'P_zX - n^{-1}X'P_zDX \\ &= I_K - n^{-1}X'P_zDX,\end{aligned} \tag{81}$$

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

$$\max_{|a|=1} a'\left[\hat{V}_n^{IV}(\beta)\right]a = \max_{|a|=1} 1 - a'X'\frac{1}{2}(P_zD + D'P_z)Xa. \tag{82}$$

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 6 (Proof of Theorem 3: 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*

$$\hat{c}(\beta) \xrightarrow{p} E[\log(1 + U_i)] = c, \tag{83}$$

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

$$\hat{\tilde{Y}}_i(\beta) = \log(1 + U_i) + X_i'\beta - \hat{c}(\beta), \tag{84}$$

so that

$$\sqrt{n}[\breve{X}'\breve{X}]^{-1}\breve{X}'\hat{\tilde{Y}}_i(\beta) = \sqrt{n}\left(\beta + [\breve{X}'\breve{X}]^{-1}\breve{X}'(\log(1 + U) - \hat{c}(\beta))\right). \tag{85}$$

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*

$$\sqrt{n}\left(\hat{M}_n^{IV}(\beta) - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma), \tag{86}$$

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*

$$\sqrt{n}\left(\hat{\beta}_{i(n)}^{IV} - \beta^{IV}\right) \xrightarrow{d} \mathcal{N}(0, [\Omega^{IV}]^{-1}), \tag{87}$$

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*

$$V(\beta) = E[\breve{X}_i\breve{X}_i']^{-1}E[\frac{\breve{X}_i X_i'}{1 + U_i}]. \tag{88}$$

Therefore sandwich-type covariance estimators are given by

$$\tilde{\Sigma} = (\frac{1}{n}X'\frac{1}{2}(P_z(I-W)+(I-W)P_z)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'\frac{1}{2}(P_z(I-W)+(I-W)P_z)X)^{-1}, \tag{89}$$

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 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). For example, Worrall and Kovandzic (2010) relies on exogenous variation in federal funding laws to instrument the size of the police force.

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, as obtained for example with the 2SLS "popular fix" estimator $\hat{\beta}^{2SLSPF} = [X'P_zX]^{-1}X'P_z \log(Y+\Delta) \in \mathbb{R}^K$, 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 $\hat{\beta}^{2SLS}_{t+1} = (X'P_zX)^{-1}(X'P_z\hat{\tilde{Y}}(\hat{\beta}_t))$, and update t to $t+1$;*

4. *Iterate steps 2 and 3 until $\hat{\beta}^{2SLS}_t$ 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 in the following theorem.

Theorem 3 (Consistency and Asymptotic Normality) *Under the above assumptions, the i2SLS estimator is consistent and achieves the parametric rate of convergence $n^{-1/2}$. Formally, we have*

$$n^{1/2}|\hat{\beta}^{IV}_{t(n)} - \beta| = O_p(1) \tag{90}$$

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

$$\sqrt{n}\left(\hat{\beta}^{IV}_{t(n)} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Omega^{IV}), \tag{91}$$

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, from a computational standpoint, when identified on the basis of an instrumental variable.[45] 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.

Finally, the specification tests developed for iOLS are easily adapted for situations

[45]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.

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.2 Dispensable zeros (iOLS$_S$)

In some circumstances, zeros can be dropped without consequence for identification. However, doing so comes at the cost of a loss of efficiency. The condition for zeros to be "dispensable" in the PPML framework is

$$E(U_i|X_i, U_i > 0) = c, \tag{92}$$

where c is a constant. This condition holds either when both $E(U_i|X_i)$ and $Pr(U_i > 0|X_i)$ are constant in X_i or when both vary with X_i. In the former case, whether one chooses to discard the zeros before estimation has no consequence for identification but will affect precision. In the (somewhat more realistic) latter case, dropping zeros is required for identification unless the term $Pr(U_i > 0|X_i)$ is explicitly modelled.

We now show that the iOLS estimators can accommodate this latter situation without loss of efficiency even when zeros are dispensable. Without loss of generality, we will focus on iOLS$_U$ and assume $E(U_i|X_i, U_i > 0) = c$, although similar conditions could be considered for iOLS$_\delta$. We propose to keep all observations but introduce a correction such that the conditional expectation $E(U_i|X_i, U_i > 0) = \textit{constant}$ is respected. Let $\tilde{Y}_i$ denote the transformed dependent variable in (24), and take the conditional expectation on both sides to obtain

$$E[\tilde{Y}_i|X_i] = X_i'\beta + \frac{1}{1+\delta}\left(E[U_i|X_i] - 1\right). \tag{93}$$

Substituting $E(U_i|X_i) = cPr(U_i > 0|X_i)$, which holds by definition, yields

$$E[\tilde{Y}_i|X_i] = X_i'\beta + \frac{1}{1+\delta}\left(cPr(U_i > 0|X_i) - 1\right). \tag{94}$$

Let us further assume that $c = 1/Pr(U_i > 0)$, without loss of generality, and rearrange the above expression into

$$E[\tilde{Y}_i|X_i] = X_i'\beta + \frac{1}{1+\delta}\left(\frac{Pr(U_i > 0|X_i)}{Pr(U_i > 0)} - 1\right), \tag{95}$$

which is equivalent to

$$E[\tilde{Y}_i|X_i] - \frac{1}{1+\delta}\left(\frac{Pr(U_i > 0|X_i) - Pr(U_i > 0)}{Pr(U_i > 0)}\right) = X_i'\beta. \tag{96}$$

Therefore, we can define a new transformation of the dependent variable

$$\tilde{Y}_i(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c_i(\beta), \tag{97}$$

where

$$\begin{aligned} c_i(\beta) = &\log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i \exp(-X_i'\beta) - 1) \\ &- \frac{1}{1+\delta}\left(\frac{Pr(U_i > 0|X_i) - Pr(U_i > 0)}{Pr(U_i > 0)} - 1\right), \end{aligned} \tag{98}$$

is such that the exogeneity condition holds in the linear model because the conditional expectation of the new transformed dependent variable has the correct mean.

We denote iOLS$_S$ the iOLS estimator based on this transformation. Before applying the iterative procedure, one needs to estimate a probability model to obtain predictions of $Pr(U_i > 0|X_i)$. In our practical implementation, we specify a logistic probability model to remain simple. $Pr(U_i > 0)$ is given by the average across observations.

The asymptotic properties of iOLS$_S$ depends on those of the estimator of $Pr(U_i > 0|X_i)$. Proving the consistency of iOLS$_S$ directly follows from that of iOLS$_U$, where the new added term in c_i does not depend on β but only on $Pr(U_i > 0|X_i)$. Therefore, $\sqrt{n}$-consistency is achieved if one has a $\sqrt{n}$-consistent estimator of that conditional probability. A nonparametric estimator will hence yield a slower convergence rate. Besides, this two-step estimation procedure requires one to correct the estimates' standard errors. A simple yet more computationally demanding approach is to use a bootstrap procedure.

B.3 Negative values

It sometimes occur that the dependent variable of interest take negative values in some instances. For example, wholesale hourly electricity prices can be negative for some observations (De Vos, 2015) or firms' profits can turn negative (Draca, Machin and Van Reenen, 2011). This prevents the use of a log-transformation, or requires one to discard observations with negative values.

Our estimator extends to dependent variables taking negative values. However, one needs to specify a model for negative values. For simplicity, we consider model (1) and assume that U_i can now take both positive and negative values. The "popular fix" counterpart in this context would be to add a constant plus the minimum of Y_i in absolute terms. We consider, instead, the following model

$$\log(Y_i + \rho \exp(X_i'\beta)) = X_i'\beta + \log(\rho + U_i) \tag{99}$$

where ρ must be chosen such that $Y_i + \rho \exp(X_i'\beta) > 0 \ \forall i$. A necessary identifying restriction is then given by $E[X_i'(\log(\rho + U_i) - c)] = 0$.

This transformation means that the log function's vertical asymptote at zero is shifted leftwise towards the minimum value of Y. Therefore, this approach is fundamentally different from the IHS, which imposes a S-shape transformation around zero.

There are three possibilities to choose ρ. First, the error U_i is known to be bounded below so that $U_i \geq \underline{\mathrm{U}}$. One can simply choose $\rho = \delta + |\underline{\mathrm{U}}|$, where the choice of δ follows the same argument as in the non-negative setting. The rest of the procedure remains unchanged compared to iOLS$_\delta$.

Second, the error U_i is known to be bounded below, but $\underline{\mathrm{U}}$ is unknown. It can be estimated by taking the first-order statistic $\hat{\underline{\mathrm{U}}} = \min_i \frac{Y_i}{\exp(X_i'\beta)}$. In this case, $\hat{\rho} = \delta + |\hat{\underline{\mathrm{U}}}|$ is estimated from the data. It is akin to the popular fix for negative data in that it also consists in adding an order statistic to the dependent variable. Here, the convergence rate of $\hat{\underline{\mathrm{U}}}$ is crucial to determine that of the iOLS estimator. For instance, if U_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 $n^{-1/2}$.

Finally, the error U_i could be unbounded. It is then unclear what is the appropriate exogeneity restriction. For instance, the first-order statistic of a Gaussian error will go to $-\infty$ at rate $\log(n)$. The parameter ρ will slowly decrease with the sample size and never converge. iOLS consistency would require the identifying restriction to depend on the sample size, which may not be meaningful in empirical applications. This case can be addressed by imposing the same restriction for all sample sizes, say $E[U_i|X_i] = 1$ for instance, and consider the approach detailed in the previous paragraph as a reasonable approximation.

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

$$Y_i = X_i^{\beta} U_i, \tag{100}$$

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

$$\log(Y_i + \delta X_i^{\beta}) = \beta \log(X_i) + \eta_i, \tag{101}$$

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

$$E[\log(X_i)\eta_i|X_i > 0]Pr(X_i > 0) + \lim_{\epsilon \to 0} E[\log(\epsilon)\eta_i|X_i = 0]Pr(X_i = 0) = 0, \tag{102}$$

which can be rewritten into

$$E[\log(X_i)\eta_i|X_i > 0]Pr(X_i > 0) + \lim_{\epsilon \to 0} \log(\epsilon)E[\mathbb{1}_{(X_i=0)}\eta_i]Pr(X_i = 0) = 0. \quad (103)$$

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 model,[46] as in

$$\log(Y_i + X_i^{\beta}) = \beta_0 \mathbb{1}_{(X_i=0)} + \beta\tilde{X}_i + \eta_i, \quad (104)$$

where $\tilde{X}_i = \log(X_i)$ for $X_i > 0$ and is equal to 0 otherwise, and the two parameters β_0 and β should be equal in principle. 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 Incidental parameter problem

In non-linear panel data models, individual fixed-effects are not 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. It is a well-known issue with maximum likelihood estimators, but is not a problem for linear estimators because the randomness of individual fixed-effects is "averaged out" and the parameters of interest are hence consistently estimated.

Our estimators do not suffer from this problem as soon as fixed-effects are averaged out at each iteration. A modified version of the iOLS algorithm can be used to

[46] See also Battese (1997) for a very similar approach.

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 fasten computations exactly like above. Their approach bears some similarities with our approach for $iOLS_\epsilon$ (additive poisson), except that ours involves less matrix inversions.

B.6 The log of a ratio

Researchers are sometimes willing to estimate equations of the form

$$\log(Y_{i1}/Y_{i2}) = X_i'\beta + \varepsilon_i, \tag{105}$$

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

$$\log((Y_{i1} + \Delta)/(Y_{i2} + \Delta)) = X_i'\beta + \omega_i, \tag{106}$$

for some $\Delta > 0$.[47] 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

$$\begin{aligned} \log(Y_{i1} + \Delta) &= X_i'\beta_1 + \varepsilon_{1i} \\ \log(Y_{i2} + \Delta) &= X_i'\beta_2 + \varepsilon_{2i}, \end{aligned} \tag{107}$$

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 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 Enforcing the log-linear model's exogeneity condition

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

[47] Alternatively, in the context of a growth rate, Huber (2018) suggests using the "symmetric growth". This fix consists in using $2\frac{y_t - y_{t-1}}{y_t + y_{t+1}}$ as "a second-order approximation to the ln growth rate. This measure is bounded in the interval [2, 2]. It has become standard in the establishment-level literature because it naturally accommodates zeros in the outcome variable, for example due to zero household debt or firm exit".

equal to 1 otherwise. Therefore, the iOLS transform becomes

$$\log(Y_i + (1 - \xi_i)\delta \exp(X_i'\beta)) = X'\beta + \log((1 - \xi_i)\delta + U_i). \quad (108)$$

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

$$E(\log((1 - \xi_i)\delta + \exp(\varepsilon_i)\xi_i)|X) = E(\varepsilon_i|\xi_i = 1, X)P(X) + \log(\delta)(1 - P(X)) \quad (109)$$

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

$$E(\varepsilon_i|X) = E(\varepsilon_i|\xi_i = 1, X)P(X) + E(\varepsilon_i|\xi_i = 0, X)(1 - P(X)). \quad (110)$$

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

$$E(\varepsilon_i|\xi_i = 0, X) = \log(\delta), \quad (111)$$

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

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

$$E(\varepsilon_i|\xi_i = 1, X) = -E(\varepsilon_i|\xi_i = 0, X)(1 - P(X))P(X)^{-1}. \quad (112)$$

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

$$E(\varepsilon_i|\xi_i = 1, X) = -\log(\delta(X))(1 - P(X))P(X)^{-1}. \quad (113)$$

This approach can be helpful although the choice of the candidate functions for $\delta(\cdot)$ to be tested is beyond the scope of this paper. Numerical simulations reveal that the algorithm has similar performance than iOLS$_\delta$.

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

$$E[U_i|Z_i] = E[U_i|Z_i, U_i > 0]Pr(U_i > 0|Z_i) = E(U_i), \tag{114}$$

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

$$U_i = \lambda E[U]Pr(U_i > 0|Z_i)^{-1} + \nu_i \tag{115}$$

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

$$H_0 : E[\log(\delta + U_i)|Z_i, U_i > 0] - \log(\delta) = \frac{c - \log(\delta)}{Pr(U_i > 0|Z_i)}, \tag{116}$$

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

$$E[\omega_i|Z_i, U_i > 0]P(Z_i) + E[\omega_i|Z_i, U_i = 0](1 - P(Z_i)) = 0 \tag{117}$$

which can be rearranged into

$$E[\omega_i|Z_i, U_i > 0] = -E[\omega_i|Z_i, U_i = 0](1 - P(Z_i))P(Z_i)^{-1}. \tag{118}$$

For the popular fix estimator, substituting the expression of ω_i on the right-hand-side gives

$$E[\omega_i|Z_i, U_i > 0] = -(\log(\Delta) - E(X'\beta|Z, U > 0))(1 - P(Z_i))P(Z_i)^{-1}, \qquad (119)$$

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

$$E[\omega_i|Z_i, U_i > 0] = E(X'\beta|Z, U > 0)(1 - P(Z_i))P(Z_i)^{-1}. \qquad (120)$$

C Additional Simulations

There are six DGPs specified as follows:

- DGP 1: (A1) $E[X_i'(\log(1+U_i)-c)] = 0$. This DGP is useful to illustrate iOLS$_\delta$. Let us assume that $\log(1+\varepsilon_i)$ is uniformly distributed as $U[\frac{c}{2P(X_i)}, \frac{3c}{2P(X_i)}]$ with X_{1i} and X_{1i} also uniformly distributed as $U[-1,2]$. Choosing $c = 0.41512$ yields the desired condition $E[X_i'(\log(1+U_i)-c)] = 0$ with $E(U_i) = 1$.

- DGP 2 (A2): $E[U_i|X_i] = 1$. This DGP is aimed at comparing the alternative modelling approaches to PPML. We assume that $(X_{1i}, X_{2i})'$ is bivariate normal with mean zero, variance $\sigma^2_{X1} = \sigma^2_{X2} = 1$ and covariance $\sigma_{X1X2} = -0.3$. We further assume that ε_i is Gaussian with mean $-\log(P(X_i)) - 1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(X_i)$.

- DGP 3 (A3): $E[U_i|U_i > 0, X_i] = 1$. This DGP is such that discarding zeros and using PPML yields consistent estimates. $(X_{1i}, X_{2i})'$ is distributed as in DGP 2, but now we assume $\exp(\varepsilon_i) \sim U[1 - \min(1, \frac{|X_{1i}+X_{2i}|}{2}), 1 + \min(1, \frac{|X_{1i}+X_{2i}|}{2})]$. The purpose is to have a multiplicative error with mean 1 and with variance as a function of X_i, as for DGP 1 and 2 but not through $P(X)$.

- DGP 4 (A4): Heckit. This DGP is such that Heckman's model applies. Let $(X_{1i}, X_{2i})'$ be distributed as in DGP 2 and 3. In addition, assume that $\xi_i = 1$ if $X_i'\gamma + \nu_i > 0$ and $\xi_i = 0$ otherwise. We further assume $(\varepsilon_i, \nu_i)'$ to be iid joint Gaussian with variances $\sigma^2_\varepsilon = \sigma^2_\nu = 3$ and covariance $\sigma^2_{\varepsilon\nu} = -2.7$.

- DGP 5 (A5): Inverse Hyperbolic Sine. This DGP is designed so that applying the IHS transform yields consistent OLS estimates. Let $(X_{1i}, X_{2i})'$ be iid uniform draws in $[-0.5, 0.5]$ and ϵ_i be iid uniform in $[-X_i'\beta, X_i'\beta + 2X_i'\beta(1-P(X))P(X)^{-1}]$. The model is $\log(Y_i + \sqrt{Y_i^2+1}) = X_i'\beta + \omega_i$, with $\omega_i = \xi_i\epsilon_i - (1-\xi_i)X_i'\beta$.

- DGP 6 (IV): $E[U_i|X_i] \neq 1$ but $E[U_i|Z_i] = 1$. We finally turn to IV regression. 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. 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 $X_{ik} = 0.8Z_{ik} + 0.2\varepsilon_i$, for $k = 1, 2$.

Table 10 reports the results for DGP 3, where we can drop zeros because $E[U_i|U_i > 0, X_i] = 1$ is the right identifying restriction. We first observe in this case that only iOLS$_S$ and PPML0 provide consistent estimates of β_1 and β_2. Second, we note that iOLS$_S$ provides standard errors which are several times smaller than those corresponding to PPML0. Third, we report that OLS without zeros and the Heckman correction provide estimates with some bias. This bias is more accentuated for the popular fix, the inverse hyperbolic sine transformation and also applies to the remaining models. These results suggest that ignoring to drop zeros when they are dispensable can lead to biased estimates.

Table 11 reports the results for DGP 4 which relies on the joint normality of the error terms in the selection and outcome equations. As expected, the Heckman model provides the right estimates. The standard errors of the associated parameters are large nonetheless. We also observe that other models provide biased estimate, suggesting that ignoring the selection process which governs the zeros can lead to misleading conclusions.

Finally, Table 13 reports the results for DGP 5, where the regressors are endogenous and requires the use of instrumental variables to achieve identification under the assumption that $E[U_i|Z_i] = 1$. First, we observe that only i2SLS$_U$ provides consistent estimates. Second, however and as expected, i2SLS$_{\delta=100}$ provides similar results to i2SLS$_U$ with a slight bias. This bias is more accentuated in i2SLS$_{\delta=1}$ and is even more severe for the other models. In particular, the instrumental variable popular fix reports the wrong sign for β_2 further demonstrating its invalidity.

Table 10: Simulations: DGP 3 (A3)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	0.10	1.51	0.41	0.11	1.51	0.41
		(0.15)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)
	$iOLS_{\delta=100}$	−0.04	1.32	0.68	−0.04	1.31	0.68
		(0.10)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)
(A2)	$iOLS_U$	−0.04	1.31	0.69	−0.04	1.30	0.69
		(0.10)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)
	$iOLS_\varepsilon$	−0.02	1.26	0.72	−0.03	1.27	0.73
		(0.30)	(0.11)	(0.12)	(0.11)	(0.04)	(0.04)
	$PPML$	−0.02	1.26	0.72	−0.03	1.27	0.73
		(0.30)	(0.11)	(0.12)	(0.11)	(0.04)	(0.04)
(A3)	$iOLS_S$	0.06	1.00	1.00	0.06	1.00	1.00
		(0.05)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
	OLS	0.98	0.91	0.91	0.98	0.91	0.91
		(0.05)	(0.04)	(0.04)	(0.02)	(0.01)	(0.01)
	$PPML0$	1.00	1.00	1.00	1.00	1.00	1.00
		(0.17)	(0.07)	(0.07)	(0.07)	(0.03)	(0.03)
(A4)	$Heckman$	0.97	0.91	0.91	0.98	0.91	0.91
		(1.38)	(0.26)	(0.26)	(0.41)	(0.08)	(0.08)
Others	$IHST$	0.67	0.72	−0.02	0.67	0.72	−0.02
		(0.09)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)
	PF	0.24	0.68	−0.00	0.24	0.68	−0.00
		(0.08)	(0.06)	(0.05)	(0.03)	(0.02)	(0.02)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP3, as described in Section **??**. The column “Cond.” identifies the family of identifying condition on which the models in column “Estim.” rely. 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.

Table 11: Simulations: DGP 4 (A4)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	-0.98	1.86	0.14	-0.97	1.85	0.15
		(0.19)	(0.11)	(0.11)	(0.06)	(0.04)	(0.03)
	$iOLS_{\delta=100}$	-1.03	1.67	0.33	-1.02	1.66	0.34
		(0.16)	(0.09)	(0.09)	(0.05)	(0.03)	(0.03)
(A2)	$iOLS_U$	-1.03	1.66	0.35	-1.02	1.65	0.35
		(0.16)	(0.09)	(0.09)	(0.05)	(0.03)	(0.03)
	$iOLS_\varepsilon$	-0.93	1.53	0.44	-0.94	1.54	0.45
		(0.47)	(0.19)	(0.16)	(0.19)	(0.08)	(0.06)
	$PPML$	-0.93	1.53	0.44	-0.94	1.54	0.45
		(0.47)	(0.19)	(0.16)	(0.20)	(0.08)	(0.06)
(A3)	$iOLS_S$	-0.92	1.34	0.66	-0.92	1.35	0.65
		(0.13)	(0.07)	(0.07)	(0.04)	(0.02)	(0.02)
	OLS	-0.57	1.29	0.71	-0.56	1.29	0.71
		(0.12)	(0.06)	(0.07)	(0.04)	(0.02)	(0.02)
	$PPML0$	0.04	1.31	0.66	0.03	1.33	0.67
		(0.41)	(0.18)	(0.14)	(0.18)	(0.07)	(0.05)
(A4)	$Heckman$	1.02	1.00	1.00	1.00	1.00	1.00
		(2.41)	(0.45)	(0.45)	(0.73)	(0.14)	(0.14)
Others	$IHST$	0.29	0.68	0.00	0.29	0.68	0.00
		(0.07)	(0.05)	(0.04)	(0.02)	(0.02)	(0.01)
	PF	-0.08	0.63	0.00	-0.08	0.63	0.00
		(0.06)	(0.05)	(0.04)	(0.02)	(0.02)	(0.01)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP4, as described in Section **??**. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. 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.

Table 12: Simulations: DGP 5 (A5)

Cond.	Estim.	n=1000 β_0	n=1000 β_1	n=1000 β_2	n=10,000 β_0	n=10,000 β_1	n=10,000 β_2
(A1)	$iOLS_{\delta=1}$	2.12	3.23	4.33	2.13	3.23	4.35
		(0.09)	(0.41)	(0.43)	(0.03)	(0.13)	(0.14)
	$iOLS_{\delta=100}$	2.06	3.34	5.42	2.07	3.34	5.45
		(0.09)	(0.31)	(0.32)	(0.03)	(0.09)	(0.10)
(A2)	$iOLS_U$	2.06	3.35	5.49	2.07	3.35	5.52
		(0.09)	(0.30)	(0.31)	(0.03)	(0.09)	(0.10)
	$iOLS_\varepsilon$	1.89	2.96	6.24	1.93	2.99	6.30
		(0.37)	(0.87)	(1.26)	(0.11)	(0.27)	(0.39)
	$PPML$	1.89	2.96	6.24	1.93	2.99	6.29
		(0.37)	(0.87)	(1.26)	(0.11)	(0.27)	(0.40)
(A3)	$iOLS_S$	2.06	3.02	5.75	2.07	3.03	5.78
		(0.09)	(0.25)	(0.27)	(0.03)	(0.08)	(0.08)
	OLS	1.78	2.08	3.81	1.78	2.08	3.81
		(0.10)	(0.34)	(0.37)	(0.03)	(0.11)	(0.12)
	$PPML0$	2.88	2.65	6.55	2.92	2.67	6.60
		(0.34)	(0.81)	(1.17)	(0.10)	(0.25)	(0.37)
(A4)	$Heckman$	−32.86	7.64	−4.78	−39.49	10.59	−4.94
		(436.74)	(18.21)	(18.86)	(21.66)	(4.00)	(4.54)
Others	$IHST$	1.00	1.00	1.00	1.00	1.00	1.00
		(0.05)	(0.19)	(0.20)	(0.02)	(0.06)	(0.06)
	PF	0.58	0.93	0.94	0.58	0.93	0.94
		(0.05)	(0.18)	(0.19)	(0.02)	(0.06)	(0.06)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP5, as described in Section **??**. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. 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.

Table 13: Simulations: DGP 5 (IV-A2)

	n=1000			n=10,000		
Estim.	β_0	β_1	β_2	β_0	β_1	β_2
$i2SLS_{\delta=1}$	1.07	1.32	0.68	1.07	1.31	0.69
	(0.25)	(0.16)	(0.13)	(0.08)	(0.05)	(0.04)
$i2SLS_{\delta=100}$	0.99	1.03	0.97	0.99	1.02	0.98
	(0.22)	(0.13)	(0.11)	(0.06)	(0.04)	(0.03)
$i2SLS_U$	0.99	1.01	0.98	1.00	1.00	1.00
	(0.22)	(0.13)	(0.11)	(0.07)	(0.04)	(0.04)
$i2SLS_\varepsilon$	1.42	0.99	0.95	1.33	1.02	1.01
	(0.58)	(0.27)	(0.27)	(0.32)	(0.12)	(0.15)
$PPML$	0.29	1.42	1.26	0.33	1.39	1.27
	(0.67)	(0.27)	(0.29)	(0.34)	(0.13)	(0.15)
OLS	0.83	1.10	1.66	0.91	1.05	1.63
	(0.10)	(0.06)	(0.06)	(0.09)	(0.06)	(0.04)
$Heckman$	−1.66	1.69	1.07	−1.65	1.67	1.01
	(1.73)	(0.40)	(0.40)	(0.58)	(0.14)	(0.16)
$2SLS$	1.53	0.66	1.33	1.52	0.66	1.34
	(0.11)	(0.07)	(0.06)	(0.03)	(0.02)	(0.02)
$IVPF$	−2.10	1.25	−0.47	−2.10	1.26	−0.48
	(0.23)	(0.16)	(0.13)	(0.07)	(0.05)	(0.04)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP5, as described in Section **??**. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. 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.

Table 14: Simulations: Data-driven selection of δ (iOLS$_\delta$, kNN)

		δ						
n	DGP	0.1	0.5	1	5	10	50	100
1000	1	6%	19%	21%	21%	10%	7%	16%
	2	0%	2%	4%	9%	8%	11%	65%
10,000	1	0%	17%	62%	21%	1%	0%	0%
	2	0%	0%	0%	2%	7%	12%	79%

Notes: This table shows the relative frequency with which a given δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ was chosen on the basis of the 10,000 simulations. These simulations vary by sample size n and by DGP. Interpretation: when the sample size is n=10,000 and the data was generated using DGP1, $\delta = 1$ was selected 50% of the time.

Table 15: Simulations: Data-driven selection of δ (iOLS$_\delta$, Probit)

		δ						
n	DGP	0.1	0.5	1	5	10	50	100
1000	1	19%	17%	20%	19%	11%	6%	9%
	2	8%	7%	7%	8%	6%	5%	58%
10,000	1	0%	11%	57%	30%	1%	0%	0%
	2	0%	0%	0%	1%	3%	7%	89%

Notes: This table shows the relative frequency with which a given δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ was chosen on the basis of the 10,000 simulations. These simulations vary by sample size n and by DGP. Interpretation: when the sample size is n=10,000 and the data was generated using DGP1, $\delta = 1$ was selected 50% of the time.

Table 16: Simulations: Specification testing (Probit)

n	DGP	$t_{\delta=1}$	$t_{\delta=100}$	t_U	t_ε	t_{PPML}	t_{PF}	t_{IHST}	t_{HECK}	t_{PPML0}
1000	1	6%	23%	21%	63%	63%	21%	21%	83%	57%
	2	9%	6%	6%	4%	4%	11%	10%	32%	8%
	3	11%	8%	9%	7%	7%	12%	11%	5%	6%
	4	9%	7%	8%	10%	10%	29%	29%	10%	7%
	5	6%	7%	7%	19%	19%	5%	5%	0%	0%
10,000	1	13%	100%	100%	100%	100%	91%	92%	100%	100%
	2	33%	8%	7%	4%	4%	54%	49%	100%	41%
	3	51%	27%	30%	34%	34%	40%	36%	5%	5%
	4	46%	25%	28%	23%	23%	99%	99%	56%	11%
	5	6%	22%	25%	94%	94%	5%	5%	43%	7%

Notes: This table shows the relative rejection frequency of each null hypothesis for 10,000 simulations. These simulations vary by sample size (as reported by the column "n") and by Data Generating Process (as reported in the column "DGP"). Interpretation: when the sample size is n=1000 and the data was generated using DGP1, $t_{\delta=1}$ was rejected 6% of the time.

Table 17: Simulations: Model selection (kNN)

n	DGP	(*A*1)	(*A*2)	(*A*3)	(*A*4)	(*A*5)
1000	1	33%	41%	0%	0%	26%
	2	7%	62%	5%	0%	25%
	3	1%	48%	34%	0%	17%
	4	2%	65%	28%	0%	4%
	5	15%	39%	24%	0%	22%
10,000	1	62%	5%	0%	0%	32%
	2	0%	94%	0%	1%	5%
	3	0%	21%	74%	5%	0%
	4	0%	68%	19%	13%	0%
	5	22%	46%	2%	0%	30%

Notes: This table shows the selection frequency of each identifying restriction for 10,000 simulations. These simulations vary by sample size (as reported by column "n") and by Data Generating Process (as reported in column "DGP"). Interpretation: when the sample size is n=1000 and generated by DGP1, a model with moments (A1) is chosen 51% of the time.

Table 18: Simulations: Model selection (Probit)

n	DGP	(*A*1)	(*A*2)	(*A*3)	(*A*4)	(*A*5)
1000	1	54%	26%	0%	0%	19%
	2	18%	70%	5%	0%	6%
	3	11%	66%	18%	0%	5%
	4	21%	61%	17%	0%	0%
	5	5%	6%	87%	0%	2%
10,000	1	87%	0%	0%	12%	1%
	2	1%	98%	0%	1%	0%
	3	0%	37%	61%	1%	1%
	4	0%	79%	18%	3%	0%
	5	43%	14%	3%	0%	41%

Notes: This table shows the selection frequency of each identifying restriction for 10,000 simulations. These simulations vary by sample size (as reported by column "n") and by Data Generating Process (as reported in column "DGP"). Interpretation: when the sample size is n=1000 and generated by DGP1, a model with moments (A1) is chosen 51% of the time.

D Data Appendix

D.1 American Economic Review (2016-2020)

Table 19: 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 20: 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

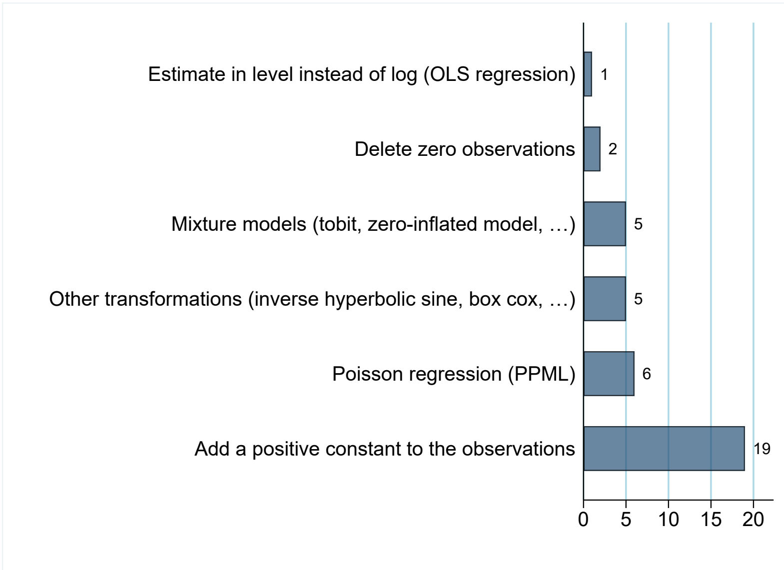

Figure 4: Proposed solutions by category on ResearchGate (November 2018)

D.3 Wooclap Survey

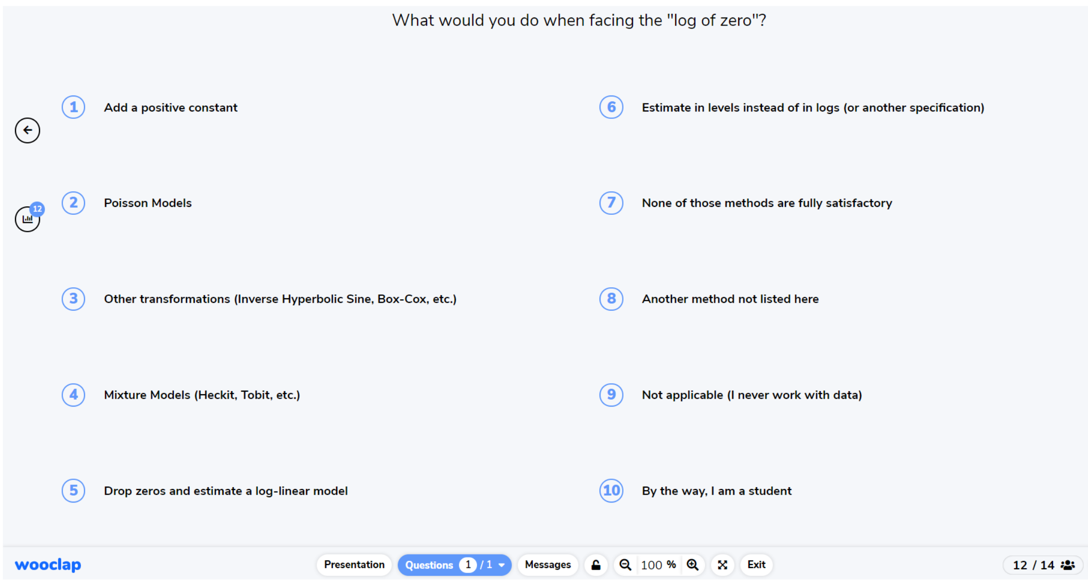

Figure 5: Wooclap Survey

Description. The survey was implemented during 3 seminars (CREST, HEC Montral, and University of Montral) in 2021, before the speaker clarified the different approaches. The attendees could provide multiple answers to the questions displayed in Figure 5 and were invited to indicate if they were a student. Results are presented in Table 21.

Table 21: 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 %
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 Santos Silva and Tenreyro (2006)

Table 22: Tests for Santos Silva and Tenreyro (2006)'s Table 3

	PF	IHST	$t_{\delta=100}$	t_{PPML}	t_{iOLS_U}	t_{HECK}	t_{PPML0}
Logit Model							
$\hat{\lambda}$	0.04	0.03	0.46	1.26	0.44	0.81	-0.22
(s.e)	(0.02)	(0.02)	(0.06)	(0.39)	(0.06)	(0.09)	(0.37)
t-Stat.	[-53.45]	[-57.90]	[-9.82]	[0.68]	[-9.24]	[8.75]	[-0.59]
KNN Model							
$\hat{\lambda}$	0.27	0.21	0.93	1.72	0.92	0.81	0.75
(s.e)	(0.01)	(0.01)	(0.05)	(0.25)	(0.05)	(0.10)	(0.37)
t-Stat.	[-49.85]	[-66.82]	[-1.55]	[2.86]	[-1.51]	[8.47]	[2.00]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for several models of trade gravity presented in Table 7. $iOLS_{\delta=100}$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. $\hat{\lambda}$ and t-tests are defined in Section 4.1.

Table 23: Logit Model for Santos Silva and Tenreyro (2006)'s Table 3

	$\mathbb{1}(Trade > 0)$	
	Coef.	*s.e.*
Log-GDP (Exp.)	0.82	(0.02)
Log-GDP (Imp.)	0.59	(0.01)
Log-GDP per Capita (Exp.)	0.18	(0.02)
Log-GDP per Capita (Imp.)	0.20	(0.02)
Log(Distance)	-0.84	(0.04)
Contiguity	-0.84	(0.18)
Language	0.63	(0.07)
Colonial	0.26	(0.07)
LandLocked (Exp.)	0.10	(0.06)
LandLocker (Imp.)	-0.13	(0.06)
Remote (Exp.)	0.24	(0.09)
Remote (Imp.)	-0.12	(0.09)
Free Trade Agreement	2.27	(0.26)
Openness	0.49	(0.05)

Notes: This table displays the logit estimates and standard errors (s.e) based on 300 pairs bootstrap used to calculate the various t-statistics of Tables 7 and 22.

D.5 Michalopoulos and Papaioannou (2013)

Table 24: Tests for Michalopoulos and Papaioannou (2013)'s Table 2 (Logit)

	PF	$iOLS_{\delta=0.1}$	$iOLS_{\delta=0.5}$	$iOLS_{\delta=100}$	PPML	$iOLS_U$	HECK	PPML0
No Controls								
$\hat{\lambda}$	-2.95	1.00	1.00	1.01	1.04	1.00	65.16	30.07
(s.e)	(0.33)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(16782)	(7049)
t-Stat.	[-11.98]	[1.19]	[1.17]	[2.43]	[2.11]	[0.75]	[0.00]	[0.00]
Pop. Controls								
$\hat{\lambda}$	-0.88	1.00	1.01	1.08	2.62	1.09	19.38	11.12
(s.e)	(0.29)	(0.02)	(0.02)	(0.06)	(0.57)	(0.07)	(5.50)	(4.64)
t-Stat.	[-6.40]	[0.03]	[0.67]	[1.36]	[2.85]	[1.30]	[3.34]	[2.18]
Pop. & Loc. Controls								
$\hat{\lambda}$	-0.56	0.98	1.00	1.03	3.68	1.03	7.02	0.51
(s.e)	(0.24)	(0.05)	(0.04)	(0.05)	(1.24)	(0.05)	(1.91)	(2.45)
t-Stat.	[-6.58]	[-0.41]	[-0.04]	[0.58]	[2.16]	[0.55]	[3.15]	[-0.20]
Pop. & Loc. & Geo. Controls								
$\hat{\lambda}$	-0.18	0.80	0.82	0.85	1.94	0.85	2.84	0.80
(s.e)	(0.12)	(0.19)	(0.19)	(0.21)	(0.91)	(0.22)	(0.55)	(0.65)
t-Stat.	[-9.46]	[-1.06]	[-0.91]	[-0.69]	[1.04]	[-0.68]	[3.35]	[-0.30]
Pop. & Loc. & Geo. Controls with Rule of Law Index								
$\hat{\lambda}$	-0.10	0.60	0.62	0.62	1.96	0.62	2.56	0.47
(s.e)	(0.12)	(0.22)	(0.22)	(0.24)	(1.11)	(0.24)	(0.56)	(0.66)
t-Stat.	[-9.13]	[-1.83]	[-1.72]	[-1.57]	[0.86]	[-1.56]	[2.78]	[-0.81]
Pop. & Loc. & Geo. Controls with Log(GDP/Capita)								
$\hat{\lambda}$	-0.06	0.40	0.42	0.45	1.76	0.46	2.37	0.41
(s.e)	(0.10)	(0.26)	(0.27)	(0.28)	(1.54)	(0.28)	(0.47)	(0.73)
t-Stat.	[-10.50]	[-2.29]	[-2.17]	[-1.95]	[0.49]	[-1.93]	[2.90]	[-0.81]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night, and presented in Tables 8 and 26. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 4.1.This table relies on the logit procedure for these tests. Six specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). The last two controls for, respectively, the quality of the legal system (in 2007) and GDP per Capita (in 2007).

Table 25: Tests for Michalopoulos and Papaioannou (2013)'s Table 2 (KNN)

	PF	$iOLS_{\delta=0.1}$	$iOLS_{\delta=0.5}$	$iOLS_{\delta=100}$	PPML	$iOLS_U$	HECK	PPML0
No Controls								
$\hat{\lambda}$	-3.61	1.23	1.24	1.26	2.30	1.31	65.16	0.00
(s.e)	(1.22)	(0.20)	(0.21)	(0.30)	(0.53)	(0.32)	(190.65)	(39.73)
t-Stat.	[-3.78]	[1.16]	[1.13]	[0.86]	[2.46]	[0.96]	[0.34]	[-0.03]
Pop. Controls								
$\hat{\lambda}$	-2.14	0.91	0.89	0.81	0.02	0.79	19.38	1.52
(s.e)	(0.33)	(0.04)	(0.05)	(0.11)	(0.72)	(0.12)	(5.58)	(1.42)
t-Stat.	[-9.39]	[-2.23]	[-2.10]	[-1.73]	[-1.36]	[-1.82]	[3.29]	[0.37]
Pop. & Loc. Controls								
$\hat{\lambda}$	-2.10	0.94	0.92	0.87	1.62	0.87	7.02	-1.81
(s.e)	(0.27)	(0.03)	(0.04)	(0.07)	(0.70)	(0.07)	(2.14)	(3.01)
t-Stat.	[-11.34]	[-1.96]	[-2.02]	[-1.84]	[0.88]	[-1.85]	[2.82]	[-0.93]
Pop. & Loc. & Geo. Controls								
$\hat{\lambda}$	-1.83	1.08	1.05	1.01	1.56	1.01	2.84	-1.55
(s.e)	(0.30)	(0.04)	(0.04)	(0.06)	(0.24)	(0.06)	(0.64)	(1.57)
t-Stat.	[-9.43]	[1.91]	[1.19]	[0.12]	[2.33]	[0.13]	[2.86]	[-1.63]
Pop. & Loc. & Geo. Controls with Rule of Law Index								
$\hat{\lambda}$	-1.24	1.08	1.06	1.02	1.73	1.02	2.56	-0.46
(s.e)	(0.28)	(0.06)	(0.07)	(0.09)	(0.57)	(0.09)	(0.56)	(0.87)
t-Stat.	[-7.98]	[1.39]	[0.89]	[0.23]	[1.29]	[0.25]	[2.77]	[-1.68]
Pop. & Loc. & Geo. Controls with Log(GDP/Capita)								
$\hat{\lambda}$	-1.61	1.17	1.14	1.11	2.15	1.11	2.37	0.61
(s.e)	(0.26)	(0.08)	(0.09)	(0.13)	(0.97)	(0.13)	(0.45)	(0.70)
t-Stat.	[-9.96]	[1.97]	[1.51]	[0.83]	[1.18]	[0.84]	[3.02]	[-0.56]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night, and presented in Tables 8 and 26. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 4.1. This table relies on the KNN procedure for these tests. Six specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). The last two controls for, respectively, the quality of the legal system (in 2007) and GDP per Capita (in 2007).

Table 26: Estimates from Michalopoulos and Papaioannou (2013)'s Table 2

Coefficient estimate on *Jurisdictional Hierarchy*

	PF	$iOLS_{\delta=0.1}$	$iOLS_{\delta=0.5}$	$iOLS_{\delta=100}$	PPML	$iOLS_U$	HECK	PPML0
No Controls								
β	0.41	0.66	0.66	0.44	0.50	0.38	5.65	0.43
(s.e)	(0.07)	(0.11)	(0.10)	(0.18)	(0.21)	(0.20)	(23.06)	(0.21)
t-Stat.	[-11.98]	[1.19]	[1.17]	[2.43]	[2.11]	[0.75]	[0.00]	[0.00]
Pop. Controls								
β	0.35	0.53	0.53	0.44	0.29	0.41	1.59	0.30
(s.e)	(0.07)	(0.11)	(0.11)	(0.13)	(0.12)	(0.14)	(0.42)	(0.11)
t-Stat.	[-6.40]	[0.03]	[0.67]	[1.36]	[2.85]	[1.30]	[3.52]	[2.39]
Pop. & Loc. Controls								
β	0.32	0.42	0.40	0.36	0.14	0.35	0.72	0.12
(s.e)	(0.06)	(0.09)	(0.08)	(0.09)	(0.11)	(0.10)	(0.18)	(0.11)
t-Stat.	[-6.58]	[-0.41]	[-0.04]	[0.58]	[2.16]	[0.55]	[3.67]	[0.21]
Pop. & Loc. & Geo. Controls								
β	0.19	0.05	0.09	0.11	0.00	0.10	0.18	0.01
(s.e)	(0.05)	(0.11)	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)
t-Stat.	[-9.46]	[-1.06]	[-0.91]	[-0.69]	[1.04]	[-0.68]	[5.16]	[1.23]
Pop. & Loc. & Geo. Controls with Rule of Law Index								
β	0.16	0.01	0.05	0.07	-0.04	0.07	0.12	-0.02
(s.e)	(0.06)	(0.12)	(0.11)	(0.10)	(0.11)	(0.10)	(0.09)	(0.11)
t-Stat.	[-9.13]	[-1.83]	[-1.72]	[-1.57]	[0.86]	[-1.56]	[4.57]	[0.71]
Pop. & Loc. & Geo. Controls with Log(GDP/Capita)								
β	0.20	0.01	0.04	0.04	-0.11	0.04	0.16	-0.09
(s.e)	(0.05)	(0.12)	(0.12)	(0.10)	(0.10)	(0.10)	(0.08)	(0.10)
t-Stat.	[-10.50]	[-2.29]	[-2.17]	[-1.95]	[0.49]	[-1.93]	[5.01]	[0.56]

Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. The t-Stats rely on the logit probability model and procedure. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). Six specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). The last two controls for, respectively, the quality of the legal system (in 2007) and GDP per Capita (in 2007).

Table 27: Tests for Michalopoulos and Papaioannou (2013)'s Table 3 (Logit) (Panel A, columns (1)-(4))

	PF	$\text{iOLS}_{\delta=100}$	PPML	iOLS_U	HECK	PPML0
Country Fixed Effects Only						
$\hat{\lambda}$	-1.50	1.00	1.26	1.00	5.56	6.47
(s.e)	(0.22)	(0.01)	(0.14)	(0.01)	(3.47)	(4.29)
t-Stat.	[-11.42]	[0.31]	[1.86]	[0.27]	[1.31]	[1.27]
with Loc. & Geo. Controls						
$\hat{\lambda}$	-0.11	0.75	1.39	0.76	1.32	1.12
(s.e)	(0.16)	(0.22)	(1.02)	(0.22)	(0.62)	(0.66)
t-Stat.	[-6.86]	[-1.12]	[0.39]	[-1.08]	[0.52]	[0.18]
with Pop. Controls						
$\hat{\lambda}$	-0.05	0.55	4.60	0.55	0.65	-7.17
(s.e)	(0.21)	(0.26)	(1.78)	(0.26)	(1.41)	(3.26)
t-Stat.	[-5.07]	[-1.74]	[2.03]	[-1.72]	[-0.25]	[-2.51]
with Pop. & Loc. & Geo. Controls						
$\hat{\lambda}$	-0.01	0.16	0.22	0.16	0.96	-0.86
(s.e)	(0.14)	(0.36)	(1.75)	(0.36)	(0.52)	(0.95)
t-Stat.	[-7.10]	[-2.32]	[-0.44]	[-2.31]	[-0.08]	[-1.97]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 4.1. This table relies on the Logit procedure for these tests. Four specifications are presented, controlling for different combinations of population density (Pop.), Location (Loc.), and Geography (Geo.) along with country fixed effects.

Table 28: Tests for Michalopoulos and Papaioannou (2013)'s Table 3 (KNN) (Panel A, columns (1)-(4))

	PF	$\text{iOLS}_{\delta=100}$	PPML	iOLS_U	HECK	PPML0
Country Fixed Effects Only						
$\hat{\lambda}$	-2.66	1.02	1.14	1.02	5.56	-5.85
(s.e)	(0.59)	(0.07)	(0.17)	(0.07)	(3.91)	(4.93)
t-Stat.	[-6.25]	[0.30]	[0.81]	[0.29]	[1.17]	[-1.39]
with Loc. & Geo. Controls						
$\hat{\lambda}$	-1.70	0.95	0.99	0.95	1.32	-0.38
(s.e)	(0.40)	(0.10)	(0.43)	(0.10)	(0.63)	(1.49)
t-Stat.	[-6.80]	[-0.52]	[-0.01]	[-0.53]	[0.51]	[-0.93]
with Pop. Controls						
$\hat{\lambda}$	-2.07	0.95	1.70	0.95	0.65	0.58
(s.e)	(0.42)	(0.06)	(0.72)	(0.06)	(1.29)	(2.57)
t-Stat.	[-7.25]	[-0.86]	[0.97]	[-0.85]	[-0.27]	[-0.16]
with Pop. & Loc. & Geo. Controls						
$\hat{\lambda}$	-1.81	1.03	1.68	1.03	0.96	-0.70
(s.e)	(0.34)	(0.07)	(0.53)	(0.07)	(0.52)	(0.95)
t-Stat.	[-8.33]	[0.46]	[1.28]	[0.45]	[-0.08]	[-1.80]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 4.1. This table relies on the KNN procedure for these tests. Four specifications are presented, controlling for different combinations of population density (Pop.), Location (Loc.), and Geography (Geo.) along with country fixed effects.

Table 29: Estimates from Michalopoulos and Papaioannou (2013)'s Table 3 (Panel A, columns (1)-(4))

Coefficient estimate on *Jurisdictional Hierarchy*

	PF	$iOLS_{\delta=100}$	PPML	$iOLS_U$	HECK	PPML0
Country Fixed Effects Only						
β	0.33	0.38	0.34	0.38	0.71	0.29
(s.e)	(0.07)	(0.09)	(0.19)	(0.09)	(3.47)	(0.20)
t-Stat.	[-11.42]	[0.31]	[1.86]	[0.27]	[1.60]	[1.51]
with Loc. & Geo. Controls						
β	0.28	0.43	0.03	0.44	0.38	0.02
(s.e)	(0.07)	(0.12)	(0.18)	(0.13)	(0.62)	(0.17)
t-Stat.	[-6.86]	[-1.12]	[0.39]	[-1.08]	[2.14]	[1.70]
with Pop. Controls						
β	0.21	0.25	-0.02	0.25	0.19	-0.03
(s.e)	(0.05)	(0.06)	(0.11)	(0.06)	(1.41)	(0.11)
t-Stat.	[-5.07]	[-1.74]	[2.03]	[-1.72]	[0.46]	[-2.20]
with Pop. & Loc. & Geo. Controls						
β	0.18	0.15	-0.11	0.15	0.14	-0.09
(s.e)	(0.04)	(0.08)	(0.09)	(0.08)	(0.52)	(0.10)
t-Stat.	[-7.10]	[-2.32]	[-0.44]	[-2.31]	[1.86]	[-0.91]

Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. The t-Stats rely on the logit probability model and procedure. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). Four specifications are presented, controlling for different combinations of population density (Pop.), Location (Loc.), and Geography (Geo.) along with country fixed effects.

Table 30: Logit Estimates for Michalopoulos and Papaioannou (2013)'s Table 2

	$\mathbb{1}(Light > 0)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Jurisdictional Hierarchy	0.85	0.68	0.68	0.62	0.55	0.41
	(0.16)	(0.16)	(0.20)	(0.23)	(0.23)	(0.26)
Population Density	No	Yes	Yes	Yes	Yes	Yes
Location Controls	No	No	Yes	Yes	Yes	Yes
Geographic Controls	No	No	No	Yes	Yes	Yes
Rule of Law Controls	No	No	No	No	Yes	No
Log(GDP per capita (2007))	No	No	No	No	No	Yes

Notes: This table displays the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 8, 24, and 26.

Table 31: Logit Estimates for Michalopoulos and Papaioannou (2013)'s Table 3 (Panel A, columns (1)-(4))

	$\mathbb{1}(Light > 0)$			
	(1)	(2)	(3)	(4)
Jurisdictional Hierarchy	0.85	0.74	0.68	0.62
	(0.16)	(0.20)	(0.16)	(0.23)
Country Fixed Effects	Yes	Yes	Yes	Yes
Population Density	No	No	Yes	Yes
Geographic Controls	No	Yes	No	Yes
Location Controls	No	Yes	No	Yes

Notes: This table displays the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 28 and 29.

D.6 Card and DellaVigna (2020)

Table 32: Tests for Card and DellaVigna (2020) (KNN)

	IHS	$\text{iOLS}_{\delta=50}$	PPML	iOLS_U
No correction for Endogeneity				
$\hat{\lambda}$	0.64	0.97	0.89	0.97
(s.e)	(0.01)	(0.01)	(0.02)	(0.01)
t-Stat.	[-27.35]	[-4.41]	[-4.50]	[-5.10]
Control Function				
$\hat{\lambda}$	0.60	0.96	0.87	0.95
(s.e)	(0.01)	(0.01)	(0.03)	(0.01)
t-Stat.	[-28.58]	[-6.12]	[-4.22]	[-6.61]
Instrumental Variables				
$\hat{\lambda}$	0.57	0.96	0.91	0.96
(s.e)	(0.02)	(0.08)	(0.04)	(0.01)
t-Stat.	[-25.79]	[-0.48]	[-2.09]	[-3.20]

Notes: This table displays the coefficient associated with an invitation to revise & resubmit (R&R), standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of citations based on the KNN procedure. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 3 and 4.1. Three specifications are presented: no correction for endogeneity (OLS) contrasts with control function (CF) and instrumental variables (IV) which rely on the Editor leave-out mean R&R rate for identification.

Table 33: First-Stage Estimates Based on Card and DellaVigna (2020)

	Revise & Resubmit	
	Coef.	*s.e*
Editor leave-out-mean R&R rate	0.38	(0.08)
Fractions of referee recommendations		
Reject	-0.00	(0.01)
No Recommendation	0.21	(0.02)
Weak R&R	0.29	(0.01)
R&R	0.69	(0.02)
Strong R&R	0.96	(0.03)
Accept	0.91	(0.03)
Author Publications in 35 high-impact journal		
One Publication	-0.00	(0.01)
Two Publications	0.01	(0.01)
Three Publications	0.02	(0.01)
Four or Five Publications	0.03	(0.01)
Six or More Publications	0.05	(0.01)
Number of authors		
Two Author	-0.01	(0.01)
Three Authors	-0.00	(0.01)
Four Authors	0.01	(0.01)

Notes: This table provides the first stage estimates used for the instrumental variable estimates provided in Table 9 and based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable before transformation is a dummy equal to one if the authors were invited to resubmit. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. The editor leave-out-mean R&R rate is the main variable of interest and is considered as an exogenous instrument, measuring the proclivity with which an editor invites other authors to revise and resubmit their research. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Table 34: Logit Estimates based on Card and DellaVigna (2020) (OLS)

	$\mathbb{1}(Citations > 0)$	
	Coef.	*s.e*
Revise & Resubmit	0.40	(0.10)
Fractions of referee recommendations		
Reject	0.76	(0.08)
No Recommendation	0.77	(0.16)
Weak R&R	1.44	(0.14)
R&R	1.86	(0.16)
Strong R&R	2.00	(0.25)
Accept	2.24	(0.28)
Author Publications in 35 high-impact journal		
One Publication	0.30	(0.06)
Two Publications	0.56	(0.08)
Three Publications	0.75	(0.08)
Four or Five Publications	1.00	(0.09)
Six or More Publications	0.88	(0.09)
Number of Authors		
Two Authors	0.27	(0.05)
Three Authors	0.36	(0.07)
Four Authors	0.56	(0.13)

Notes: This table provides the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 9 and **??** (in the OLS case), based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable is a dummy equal to one if the authors obtained at least one citation. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Table 35: Logit Estimates Based on Card and DellaVigna (2020) (Control Function)

	$\mathbb{1}(Citations > 0)$	
	Coef.	*s.e*
Revise & Resubmit	-0.15	(0.26)
Control Function	0.35	(0.16)
Fractions of referee recommendations		
Reject	0.76	(0.08)
No Recommendation	0.88	(0.16)
Weak R&R	1.57	(0.15)
R&R	2.18	(0.22)
Strong R&R	2.48	(0.33)
Accept	2.69	(0.34)
Author Publications in 35 high-impact journal		
One Publication	0.31	(0.06)
Two Publications	0.56	(0.08)
Three Publications	0.76	(0.09)
Four or Five Publications	1.01	(0.09)
Six or More Publications	0.90	(0.09)
Number of Authors		
Two Authors	0.27	(0.05)
Three Authors	0.36	(0.07)
Four Authors	0.56	(0.13)

Notes: This table provides the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 9 and **??** (in the Control Function (CF) case), based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable is a dummy equal to one if the authors obtained at least one citation. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. The editor leave-out-mean R&R rate is used to form a control function for the invitation to revise and resubmit the manuscript. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Table 36: Logit Estimates Based on Card and DellaVigna (2020) (Instrumental Variable)

	$\mathbb{1}(Citations > 0)$	
	Coef.	*s.e*
Editor leave-out mean R&R rate	-1.06	(0.71)
Fractions of referee recommendations		
Reject	0.76	(0.08)
No Recommendation	0.86	(0.16)
Weak R&R	1.52	(0.14)
R&R	2.08	(0.15)
Strong R&R	2.33	(0.23)
Accept	2.54	(0.27)
Author Publications in 35 high-impact journal		
One Publication	0.31	(0.06)
Two Publications	0.56	(0.08)
Three Publications	0.76	(0.08)
Four or Five Publications	1.01	(0.09)
Six or More Publications	0.90	(0.09)
Number of authors		
Two Author	0.27	(0.05)
Three Authors	0.36	(0.07)
Four Authors	0.56	(0.13)

Notes: This table provides the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 9 (in the Instrumental Variable (IV) case), based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable is a dummy equal to one if the authors obtained at least one citation. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. The editor leave-out-mean R&R rate is the main variable of interest and is considered as an exogenous instrument, measuring the proclivity with which an editor invites other authors to revise and resubmit their research. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.