Event Count Estimation

Laszlo Balazsi* Felix Chan† Laszlo Matyas‡

March 26, 2018

Abstract

This paper proposes a new estimation procedure called Event Count Estimator (ECE). The estimator is straightforward to implement and it is robust against outliers, censoring and excess zeros in the data. The paper establishes asymptotic properties of the new estimator and the theoretical results are supported by several Monte Carlo experiments. Monte Carlo experiments also show that the estimator has reasonable properties in moderate to large samples. As such, the cost of inefficiency for robustness is negligible from an applied viewpoint. The practical usefulness of the new estimator is demonstrated via an empirical application of Gravity Model of trade.

JEL: C13, C24, C55

Keywords: Big data, outliers, robust estimation, event count estimation, censoring, excess zeros.

*Central European University, Balazsi_Laszlo@alumni.ceu.edu.
†Curtin University, Felix.Chan@cbs.curtin.edu.au.
‡Central European University, matyas@ceu.edu.

Introduction

In classical econometrics and statistics, efficient estimation has had a central role. When data availability is limited, the standard errors of the parameter estimates of an econometrics model would generally be large to reflect the uncertainties due to small sample and the related limited amount of information available for estimation. In such cases, efficient estimation becomes important to obtain the most precise estimates. Efficient estimators often require sets of rather restrictive assumptions and hence, they are generally not robust against data features such as outliers, excess zeros or missing values. When data availability is not a binding issue, robustness becomes more important than efficiency for two reasons. First, aforementioned data features are common in 'Big Data'; and second, the different in standard errors between efficient and inefficient consistent estimators would generally becomes negligible as sample size grows. Therefore, consistent estimators that are robust against various data features would be desirable for 'Big Data' even if they are not efficient.

This paper proposes a conceptually simple technique called Event Count Estimator (hereafter EC Estimator or ECE), which trade optimality and efficiency for robustness against several data related problems. While optimal estimation methods (like Least Squares, Maximum Likelihood, etc.) are more efficient than the ECE under their respective sets of assumptions, the slight gained in precision is quickly offset by the biases induced from the violation of their assumptions.

The paper is organised as follows. Section 1 provides a simple example to motivate the concept of ECE. This is followed by four variants of the EC Estimator in Sections 2. Asymptotic properties can be found in Section 3 and Section 4 evaluates the finite sample performance of these estimators via Monte Carlo simulation. An empirical demonstration

is presented in Section 5 and Section 6 contains some concluding remarks.

1 Motivational Example

Let us motivate the concept of Event Count Estimation by providing a simple example. Consider the following data generating process for a simple location model:

$$y_i^* = \begin{cases} y_i & \text{if } u_i > \delta \\ S_i & \text{if } u_i \leq \delta \end{cases} \tag{1}$$

with

$$y_i = m_0 + e_i,$$

where e_i is independently and identically distributed following a symmetric distribution with mean 0 and variance σ^2, $u_i \sim U(0,1)$ and $S_i \sim U(a,b)$ with $a >> m_0$. For simplicity, let us assume all random variables are independent from each other.

Model (1) represents a situation where a fraction of the data have been contaminated. Neither the number of contaminated observations nor which observations were contaminated is known as they are both functions of the random variable u_i. The values of the contaminated observations are also unknown except the minimum bound of the contamination is much larger than the unknown mean m_0. The basic problem here is of course the best estimation strategy to estimate the location parameter, m_0.

Simple average is clearly biased, since $N^{-1}\sum_{i=1}^{N} y_i \xrightarrow{p} \mathbb{E}(y_i^*) = m_0(1-\delta) + \frac{\delta}{2}(b-a)$. A plausible solution is to 'trim' the sample by removing any observations with absolute value greater than some limit, say B. The problem with this approach is that the choice of B seems arbitrary and it removes more observations than required. Another approach

is to consider the following estimator:

$$\hat{m} = \arg\max_{m} \Pr\left[|y_i - m| < B\right]. \tag{2}$$

To see how this may work, consider

$$\begin{aligned}
\Pr\left[|y_i - m| < B\right] =& \Pr\left[|y_i^* - m| < B\right] \Pr[y_i = y_i^*] + \Pr\left[|S_i - m| < B\right] \Pr\left[y_i = S_i\right] \\
=& (1-\delta) \Pr\left[|y_i^* - m| < B\right] + \delta \Pr\left[|S_i - m| < B\right] \\
=& (1-\delta) \Pr\left[|\Delta m + e_i| < B\right] + \delta \Pr\left[|S_i - m| < B\right] \\
=& (1-\delta) \int_{-B-\Delta m}^{B-\Delta m} g(e)de + \delta \int_{-B-m}^{B-m} \frac{ds}{b-a},
\end{aligned}$$

where $\Delta m = m_0 - m$ with $g(.)$ denote the density function of e_i. Differentiate the last line with respect to m and set it to 0 gives:

$$(1-\delta)\left[g(B - \Delta m) - g(-B - \Delta m)\right] = 0.$$

Since the distribution of e_i is symmetric, $\Delta m = 0$, which implies $m = m_0$. Note that this approach is less sensitive to the choice of B because unlike the trimming approach, this approach does not remove any observation. In fact, all observations must be included in order to seek the value that maximises the probability of the event $|y_i - m| < B$.

The next section will formalise this approach in the context of linear regression and introduce several estimators which are straightforward to implement in practice.

2 Family of Estimators

Let us formulate first the concept of Event Count Estimator (ECE). Consider a simple linear regression model:

$$y_i = x_i'\beta_0 + \varepsilon_i, \quad i = 1 \dots N, \tag{3}$$

where x_i is a $k \times 1$ vector of explanatory variables, β is the $k \times 1$ unknown parameter vector and ε_i is the error term with zero first moment. The estimation aims at finding $\hat{\beta}_{ECE}$ for which the number of "successes", that is the number of observations that satisfy

$$d(y_i - x_i'\hat{\beta}_{ECE}) \leq B \tag{4}$$

is maximized, where $B > 0$ is some finite upper boundary, $d(.) \geq 0$ is some distance measure (metric), with $\hat{\beta}_{ECE}$ denotes the EC estimator. In practice, two of the most common metrics are the quadratic norm $d(x) = x^2$ or the absolute norm $d(x) = |x|$ but obviously other metrics are also possible. Formally, let

$$d_i(\beta, B) = \begin{cases} 1 & \text{if } d(y_i - x_i'\beta) \leq B \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This provides the foundation for the following ECE estimators.

2.1 Unrestricted ECE

Consider the average number of residuals that fall within the boundaries given β:

$$Q_N(\beta, B) = N^{-1} \sum_{i=1}^{N} d_i(\beta, B) \equiv N^{-1} k(\beta, B). \tag{6}$$

This leads to the first EC estimator called the *Unrestricted ECE*:

$$\hat{\beta}_{ECE_{UR}} = \arg\max_{\beta} Q_N(\beta, B). \tag{7}$$

The rationale of this estimator is to strike a balance between maintaining the information from extreme observations while minimising the impacts of outliers. Since the distribution of ε_i is unknown, extreme observations which are drawn from the distribution of ε_i are generally not distinguishable from the outliers that come from a different distribution. The idea here is to seek the parameter vector in such a way that maximises the inclusions of all observations that centred around the estimated conditional mean, subject to the boundary point B. This reduces the influence of observations outside the boundary points. However, since the decision of inclusion depends on the parameter vector, which is determined by all observations, information from observations outside the boundary point are not discarded.

2.2 Restricted Estimator

Under equation (3), the probability that $d_i = 1$ is

$$\Pr\left[d(y_i - x_i'\beta_0) \leq B\right] = \Pr\left[d(\varepsilon_i) \leq B\right] = p_i.$$

Under the assumption that ε_i are i.i.d., d_i follows a Bernoulli distribution, the sum of the d_i thus follows a Binomial distribution with parameter $k(\beta, B)$ and the probability

estimated by $p = k(\beta, B)/N$. This gives the following Likelihood function

$$\begin{aligned} L(\beta, B) &= \begin{pmatrix} N \\ k \end{pmatrix} p^k (1-p)^{N-k} \\ &= \begin{pmatrix} N \\ k \end{pmatrix} \left(\tfrac{k}{N}\right)^k \left(\tfrac{N-k}{N}\right)^{N-k} \end{aligned} \tag{8}$$

and the log-likelihood

$$\begin{aligned} l(\beta, B) = \log L(\beta, B) &= \log N! - \log k! - \log (N-k)! \\ &\quad - N \cdot \log N + k \cdot \log k + (N-k) \cdot \log (N-k). \end{aligned} \tag{9}$$

The *Restricted ECE* is obtained by maximising the log-likelihood function. That is,

$$\hat{\beta}_{ECE_R} = \arg\max_{\beta} l(\beta, B). \tag{10}$$

It is restricted in the sense that ε_i is restricted to be i.i.d. for all i.

The maximisation of equation (9) requires further discussion. By the chain rule and using the fact $\Gamma'(n+1) = n! \left(-\gamma + \sum_{j=1}^{k} j^{-1}\right)$ where γ denotes the *Euler-Mascheroni* constant and $\Gamma(.)$ denotes the Gamma function, the first order conditions are:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \frac{\partial l}{\partial k} \frac{\partial k}{\partial \beta} \\ &= \left[-\left(-\gamma + \sum_{j=1}^{k} \frac{1}{j}\right) + \left(-\gamma + \sum_{j=1}^{N-k} \frac{1}{j}\right) + \log k - \log (N-k) \right] \frac{\partial k}{\partial \beta} \\ &= \left[-\sum_{j=1}^{k} \frac{1}{j} + \sum_{j=1}^{N-k} \frac{1}{j} + \log \frac{k}{N-k} \right] \frac{\partial k}{\partial \beta} \\ &= 0. \end{aligned} \tag{11}$$

Note that equation (11) holds when $\frac{\partial l}{\partial k} = 0$ or $\frac{\partial l}{\partial \beta} = 0$. The latter is the first order condition of equation (7), which implies the Restricted and Unrestricted ECE are algebraically equivalent when this holds. However, additional solutions of equation (11) emerge when $\frac{\partial l}{\partial k} = 0$.

Case 1. $k = N - k$. $\log \frac{k}{N-k} = 0$ and the two sums cancel out: the first order condition trivially holds.

Case 2. $k > N - k$.

$$-\sum_{j=N-k+1}^{k} \frac{1}{j} + \log \frac{k}{N-k} = 0$$

gives a solution.

Case 3. $N - k > k$.

$$\sum_{j=k+1}^{N-k} \frac{1}{j} + \log \frac{k}{N-k} = 0$$

is a solution.

Given these additional solutions, some cautions are required when using the restricted ECE. These solutions, while satisfy the first order condition, do not necessarily represent an optimum. Thus, it is important to compare $\hat{\beta}_R$ and $\hat{\beta}_{ECE_{UR}}$ and see which should be ruled out.

2.3 Conditional ECE

The differentiability of the objective functions for both Restricted and Unrestricted ECE can create numerical difficulties in practices. As such, this paper proposes a third estimator called *Conditional ECE*. Unlike the other two estimators, the objective function of the conditional ECE is smooth and differentiable. As such, it is not only easier to implement in practice, it is also straightforward to derive consistency and asymptotic normality which

will facilitate valid inferences.

In order to establish the conditional ECE estimator, first define

$$u_i(\beta) = y_i - x_i'\beta \tag{12}$$

and the Law of Iterated Expectation suggests that

$$\mathbb{E}\left[d_i(u_i)\right] = \mathbb{E}\left[\mathbb{E}\left[d_i(u_i)|x_i\right]\right]. \tag{13}$$

Let $J_i(\beta) = \mathbb{E}\left[d_i(u_i)|x_i\right]$ with $P_N = N^{-1}\sum_{i=1}^{N} J_i(\beta)$ and define $Q = \lim_{N\to\infty} Q_N$, then under assumptions of the Weak Law of Large Number (WLLN) $P_N - Q = o_p(1)$. This means β can be estimated by seeking a parameter vector that maximises P_N. The function $J_i(\beta)$ can be derived as follows. First, noting that the distance function $d(u)$ can be decomposed as

$$d(u) = d_+(u) + d_-(u), \tag{14}$$

where

$$d_+(u) = \begin{cases} d(u) & u \geq 0 \\ 0 & u < 0 \end{cases} \qquad d_-(u) = \begin{cases} 0 & u \geq 0 \\ d(u) & u < 0. \end{cases} \tag{15}$$

So the function of "success", $d_i(u)$, can be written as

$$d_i(u_i) = \begin{cases} 1 & d_+(u_i) < B_u \text{ or } d_-(u_i) > -B_l \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

with the explicit assumption that $B_l < 0 < B_u$. Given the functions defined above, the

conditional expectation can be expressed as

$$\mathbb{E}\left[d_i(u_i)|x_i\right] = \Pr\left[d_+(u_i) < B_u|x_i\right] + \Pr\left[d_-(u_i) \leq -B_l|x_i\right].$$

Under the assumptions that $d_+(u)$ and $d_-(u)$ are one-to-one and onto under their specific domains, then their respective inverses exist.[1] Define the conditional density of u_i as $g(u_i|x_i)$, then

$$\begin{aligned}\Pr\left[d_+(u_i) < B_u|\, x_i\right] &= \Pr\left[u_i < d_+^{-1}(B_u)|x_i\right] \\ &= \left[\int_0^{\infty} g(u_i|x_i)du_i\right]^{-1} \int_0^{d_+^{-1}(B_u)} g(u_i|x_i)du_i. \end{aligned} \tag{17}$$

Similarly for the negative case:

$$\begin{aligned}\Pr\left[d_-(u_i) < -B_l|x_i\right] &= \Pr\left[u_i < d_-^{-1}(-B_l)|x_i\right] \\ &= \left[\int_{-\infty}^{0} g(u_i|x_i)du_i\right]^{-1} \int_{d_-^{-1}(-B_l)}^{0} g(u_i|x_i)du_i. \end{aligned} \tag{18}$$

The conditional expectation of $d_i(u_i)$ is therefore

$$\begin{aligned}\mathbb{E}\left[d_i(u_i)|x_i\right] = &\left[\int_0^{\infty} g(u_i|x_i)du_i\right]^{-1} \int_0^{d_+^{-1}(B_u)} g(u_i|x_i)du_i \\ &+ \left[\int_{-\infty}^{0} g(u_i|x_i)du_i\right]^{-1} \int_{d_-^{-1}(-B_l)}^{0} g(u_i|x_i)du_i. \end{aligned} \tag{19}$$

[1] This is not too restrictive as both quadratic and absolute loss functions satisfy this requirement.

By simple change of variable, it can be expressed in terms of the data:

$$\begin{aligned}\mathbb{E}\left[d_i(u_i)|x_i\right] \equiv & J_i(\beta) \\ = & \left[\int_{x_i'\beta}^{\infty} g(y_i|x_i)dy_i\right]^{-1} \int_{x_i'\beta}^{A_u+x_i'\beta} g(y_i|x_i)dy_i \\ & + \left[\int_{-\infty}^{x_;\beta} g(y_i|x_i)dy_i\right]^{-1} \int_{A_l+x_i'\beta}^{x_i'\beta} g(y_i|x_i)dy_i,\end{aligned} \tag{20}$$

where $A_u = d_+^{-1}(B_u)$ and $A_l = d_-^{-1}(-B_l)$. Define

$$P_N(\beta, B) = N^{-1}\sum_{i=1}^{N} J_i(\beta, B)$$

the conditional ECE is therefore

$$\hat{\beta}_{CECE} = \arg\max_{\beta} P_N(\beta). \tag{21}$$

The choice of the density function $g(\cdot)$ deserves further discussion. If $g(\cdot)$ is known, then estimator as defined in equation (21) can be readily implemented. If $g(\cdot)$ is not known, then there are two choices. The first is to assume a specific density and the second is to estimate the density function using non-parametric methods. For the first option, a convenient choice is the normal density, which is computationally simpler. Perhaps more importantly, it also leads to a consistent estimator even when the true underlying distribution is not normal as shown in Section 3. This is akin to the Quasi-Maximum Likelihood (QMLE) estimator and more generally, M-type estimator. For the second choice, any consistent non-parameter estimator for conditional density can be used to approximate $g(\cdot)$ as shown in Section 3.

Under normality, equation (20) becomes

$$J_i(\beta) = 2\left[\Phi\left(\frac{A_u}{\sigma_i}\right) - \Phi\left(\frac{A_l}{\sigma_i}\right)\right] \tag{22}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative function and $\sigma_i = |y_i - x_i'\beta|$.

3 Asymptotic Properties

This section presents several asymptotic properties of the family of ECE estimators introduced in Section 2.

3.1 Unrestricted ECE

For the Unrestricted ECE, consider the following assumptions:

A1. Let the data (y_i, x_i) be an outcome of a random variable W_i such that y_i and x_i follow equation (3).

A2. There exists a $k \times k$ non-singular matrix Σ_x such that $N^{-1}X'X - \Sigma_x = o_p(1)$ where $X = (x_1, \ldots, x_N)'$ and x_i is independent to ε_i for all i.

A3. Let $u_i(\beta) = y_i - x_i'\beta$ and denote Θ the compact parameter space such that $\beta_0, \beta \in \Theta$.

A4. The random variable W_i has the property that $u_i(\beta)$ is α-mixing $\forall \beta \in \Theta$ with size $-\dfrac{r}{r-1}$ for some $r > 1$.

A5. Let $g_i(\varepsilon)$ be the probability density function of ε_i such that $g_i(A_l) = g_i(A_u)$ where $A_l = d_-^{-1}(B_l)$ and $A_u = d_+^{-1}(B_u)$.

Some comments on these assumptions are warranted. Assumptions A1 and A3 are standard in the literature. Assumption A5 provides a condition to determine the boundary

points in case ε_i follows an asymmetric distribution. Obviously $B_u = -B_l$ if ε_i follows a symmetric distribution. Assumption A4 imposes a memory structure on W_i and allows the distribution of ε_i to be different across i. This covers the case where ε_i is stationary but serially correlated and heteroskedastic.

Proposition 1. *Under Assumptions A1 - A4,* $\hat{\beta}_{ECE_{UR}} - \beta_0 = o_p(1)$.

Proof. See Appendix A. □

3.2 Conditional ECE

The first set of asymptotic results concern with $J_i(\beta)$ as defined in equation (22). In addition to Assumptions A1-A3 and A4, consider the following:

B 1. $A_u = -A_l$.

B 2. $\mathbb{E}\left(\frac{A_u^2}{\varepsilon_i^2}\right) > \sqrt{2}$.

B 3. Let $p(\varepsilon)$ be the parent distribution of ε_i such that

$$\mathbb{E}\left[\phi\left(\frac{A_u}{\varepsilon_i}\right)\frac{A_u}{\varepsilon_i^2}\middle|\varepsilon_i > 0\right]\Pr\left(\varepsilon_i > 0\right) = \mathbb{E}\left[\phi\left(\frac{A_u}{\varepsilon_i}\right)\frac{A_u}{\varepsilon_i^2}\middle|\varepsilon_i < 0\right]\Pr\left(\varepsilon_i < 0\right), \qquad (23)$$

where $\phi(\cdot)$ denotes the standard normal density function.

B 4. Let $z_i = \frac{A_u}{|\varepsilon_i|}$ and $w_i = \phi(z_i)z_i$ then $\mathbb{E}\left|\frac{w_i}{\varepsilon_i}\right|^2 < \infty$.

B 5. The random variable $q_i = w_i x_i$ is α-mixing with size $-r/(r-2)$ or ϕ-mixing with size $-r/2(r-1)$ for some $r > 2$.

Assumption B1 is a simplifying assumption. Propositions 3 and 4 still hold with some modifications for asymmetric boundary points. There are two advantages of imposing symmetric boundaries. First, they are much easier to implement. Second, it is also easier

to derive a condition which will help to determine a valid boundary point. Along with Assumption B2, the following establishes a sufficient condition for A_u.

Proposition 2. *Let $A_u > \sigma_0\sqrt{2}$ where σ_0 denotes the standard deviation of ε_i then Assumption B2 holds.*

Proof. See Appendix A. □

While σ_0 is generally unknown, it can be replaced by the estimated standard deviations of the residuals from the OLS estimator.

Assumption B3 trivially holds if the distribution of ε_i is symmetric. Thus, it imposes the type of asymmetric distributions for ε_i in order to ensure consistency and asymptotic normality of conditional ECE. However, it is possible to relax B3 by removing Assumption B1. Under asymmetric boundary points, Assumption B3 becomes

$$\mathbb{E}\left[\phi\left(\frac{A_u}{\varepsilon_i}\right)\frac{A_u}{\varepsilon_i^2}\middle|\varepsilon_i > 0\right]\Pr\left(\varepsilon_i > 0\right) = \mathbb{E}\left[\phi\left(\frac{A_l}{\varepsilon_i}\right)\frac{A_l}{\varepsilon_i^2}\middle|\varepsilon_i < 0\right]\Pr\left(\varepsilon_i < 0\right). \tag{24}$$

Unlike Assumption B3, which only has one parameter, A_u, the condition above involves two parameters, A_u and A_l. Thus, the number of asymmetric distributions allowed is greater under asymmetric boundary than the symmetric boundary case.

Assumption B4 is a technical assumption to ensure all necessarily moments exists, while Assumption 5 imposes the type of dependence of ε_i, includes stationary serially correlated processes.

Proposition 3. *Consider the estimator as defined in equations (21) and (22), under Assumptions A1-A3 and B1 - B3, B5, $\hat{\beta}_{CECE} - \beta_0 = o_p(1)$.*

Proof. See Appendix A. □

Proposition 4. *Consider the estimator as defined in equations (21) and (22), under Assumptions A1-A3 and B1 - B5,* $\sqrt{N}\left(\hat{\beta}_{CECE}-\beta_0\right)\xrightarrow{d}N\left(0,c\Sigma_x^{-1}\right)$ *where*

$$c=\frac{\mathbb{E}\left[\varepsilon_i^{-2}\phi^2\left(A_u\varepsilon_i^{-1}\right)A_u^2\varepsilon_i^{-2}\right]}{\left\{\mathbb{E}\left[\varepsilon_i^{-2}(2-A_U^2\varepsilon_i^{-2})\phi(A_u\varepsilon_i^2)A_u|\varepsilon_i|^{-1}\right]\right\}^2}. \tag{25}$$

Proof. See Appendix A. □

The variance-covariance matrix, $c\Sigma_x^{-1}$, can be estimated consistently by $\hat{c}\hat{\Sigma}_x^{-1}$, where

$$\hat{\Sigma}_x=\frac{X'X}{N} \tag{26}$$

$$\hat{c}=N^{-1}\sum_{i=1}^{N}\frac{\varepsilon_i^{-2}\phi^2\left(A_u\varepsilon_i^{-1}\right)A_u^2\varepsilon_i^{-2}}{\{\varepsilon_i^{-2}(2-A_U^2\varepsilon_i^{-2})\phi(A_u\varepsilon_i^2)A_u|\varepsilon_i|^{-1}\}^2}. \tag{27}$$

Under Assumption A2 $\hat{\Sigma}_x-\Sigma_x=o_p(1)$ and under Assumptions B4 and B5, $\hat{c}-c=o_p(1)$, therefore $\hat{c}\hat{\Sigma}_x^{-1}-c\Sigma_x^{-1}=o_p(1)$ by the Continuous Mapping Theorem.

While the approach above is consistent and computationally straightforward, it is generally not efficient if ε_i is not normally distributed. A more efficient estimator can be obtained by estimating the conditional cumulative distribution function non-parametrically. Specifically, rewrite equation (20) as

$$J_i(\beta)=\frac{G(A_u+x_i'\beta)-G(x_i'\beta)}{1-G(x_i'\beta)}+\frac{G(x_i'\beta)-G(A_i+x_i'\beta)}{G(x_i'\beta)} \tag{28}$$

and define

$$\hat{J}_i(\beta)=\frac{\hat{G}(A_u+x_i'\beta)-\hat{G}(x_i'\beta)}{1-\hat{G}(x_i'\beta)}+\frac{\hat{G}(x_i'\beta)-\hat{G}(A_i+x_i'\beta)}{\hat{G}(x_i'\beta)},$$

where $\hat{G}(u)$ is a consistent non-parametric estimator of $G(u)$ such that $\sup_u|\hat{G}(u)-G(u)|=o_p(1)$, then by Continuous Mapping Theorem $\sup_\beta|\hat{J}_i(\beta)-J_i(\beta)|=o_p(1)$. Intuitively, as N becomes sufficiently large, $\hat{G}$ approaches G and therefore $\hat{\beta}_{CECE\_NP}-\hat{\beta}_{CECE\_NPA}=o_p(1)$,

where

$$\hat{\beta}_{CECE\_NP} = \arg\max_{\beta} N^{-1} \sum_{i=1}^{N} \hat{J}_i(\beta), \tag{29}$$

$$\hat{\beta}_{CECE\_NPA} = \arg\max_{\beta} N^{-1} \sum_{i=1}^{N} J_i(\beta), \tag{30}$$

with $J_i(\beta)$ as defined in equation (28). Thus, if $\hat{\beta}_{CECE\_NPA} - \beta_0 = o_p(1)$, then $\hat{\beta}_{CECE\_NP} - \beta_0 = o_p(1)$.

In addition to Assumptions A1-A4, consider the following assumptions:

C1. The conditional cumulative distribution function, $G(u)$, is twice differentiable with smooth first and second derivatives.

C2. There exist A_l and A_u such that $A_l < 0 < A_u$ and

$$[1 - G(0)]\,[g(0) - g(A_l)] = G(0)\,[g(A_u) - g(0)]\,. \tag{31}$$

C3. Let $g(u)$ denotes the derivative of $G(u)$, there exist consistent non-parametric estimators, $\hat{g}(u)$ and $\hat{G}(u)$, for $g(u)$ and $G(u)$, respectively, such that $\sup_u |\hat{g}(u) - g(u)| = o_p(1)$ and $\sup_u |\hat{G}(u) - G(u)| = o_p(1)$.

Proposition 5. *Under Assumptions A1-A4 and C1-C2,* $\hat{\beta}_{CECE\_NPA} - \beta_0 = o_p(1)$ *and* $\hat{\beta}_{CECE\_NP} - \beta_0 = o_p(1)$.

Proof. See Appendix A. □

4 Finite Sample Performance

This section examines the finite sample performances of the various ECE estimators under different scenarios via several Monte Carlo experiments. The performances will be

compared to Least Squares (LS) as it provides the standard benchmark. The choice of boundary points are based on Assumption A5 for the unrestricted estimator, Assumption B1 and Proposition 2 for the CECE estimator and Assumption C2 for the non-parametric CECE estimator. A program code is written to get the optimal boundaries and construct the EC estimator.[2]

4.1 When LS Assumptions are Satisfied

The first set of Monte Carlo experiments focus on the ideal case when all OLS assumptions are satisfied. This gives the benchmark performance of various ECE estimators relative to the OLS under different error distributions. The Data Generating Process (DGP) is

$$y_i = 0.5x_i + \varepsilon_i. \tag{32}$$

Unless otherwise stated, the regressor is drawn from a sequence of independent uniform distribution, $x_i \sim U(-10, 10)$. This paper considers three different distributions of ε_i, namely, $N(0, 1)$, $t(5)$ and the skewed normal distribution, $SN(0, 1, 0.8)$, as defined in Azzalini (1985). Each Monte Carlo experiment consists of 1000 replications.

Figure 1 captures a typical comparison of the unrestricted ECE and OLS estimates, while a more detailed Figure 9, with various sample sizes and DGPs is included in Appendix B.

[2]Codes are available at `https://www.dropbox.com/sh/68186kcos9jb35p/AABTlXLuaUTm9rja0ElNZKLpa?dl=0`.

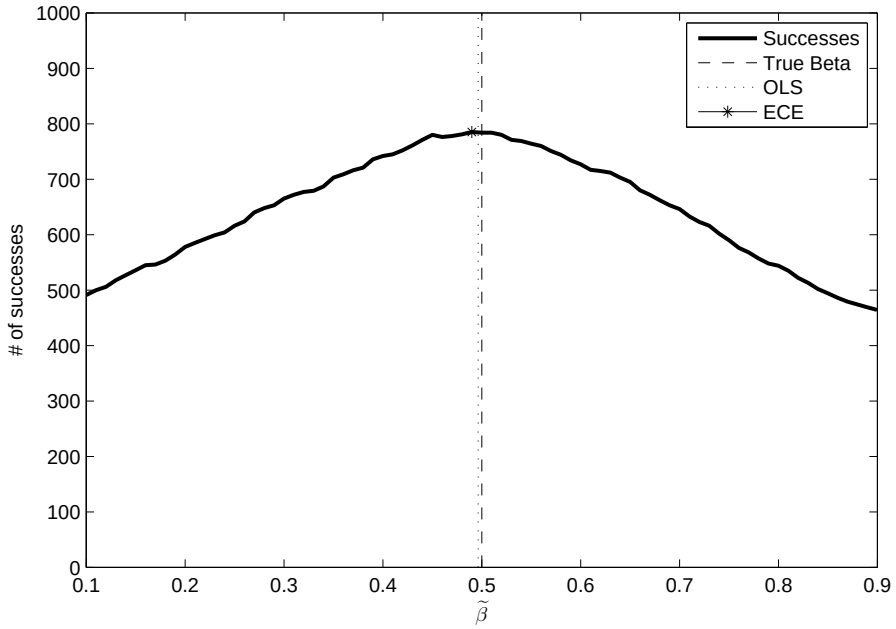


Figure 1: The dashed vertical line indicates the true parameter value 0.5, the dotted line the OLS estimate. The number of "successes" for each $\tilde{\beta}$ value is shown by the thick curve, and its maximum (indicated with black stars) are the respective unrestricted ECE estimates.

The dashed vertical line on Figure 1 indicates the true β value, while the dotted one is the OLS estimate. The ECE is constructed with a classic gridsearch. It starts from some initial β_0 value, then the number of "successes", $\sum_i d_i$, is counted for each $\hat{\beta}$. This is indicated by the thick curve. The maximum of the curve gives the unrestricted ECE estimator, indicated by a black star.

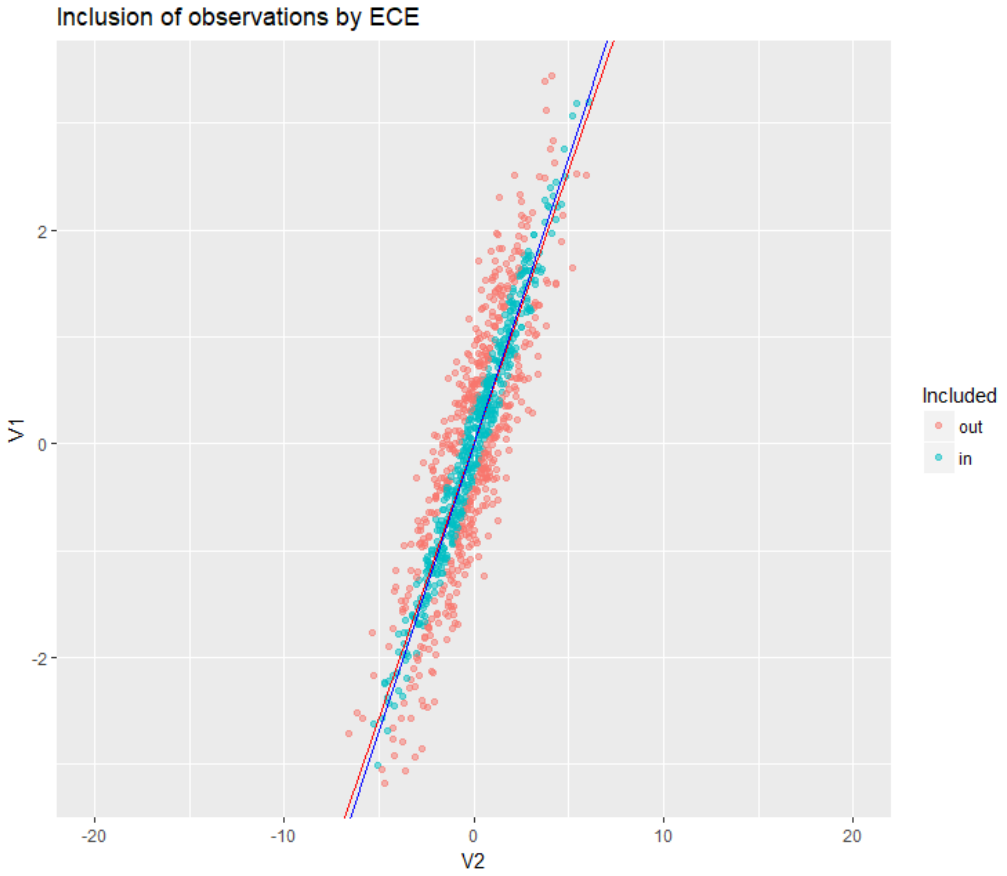


Figure 2: The used and unused observations by ECE when all LS assumptions are satisfied. The blue line shows the ECE estimator, the red the LS, which uses all (blue and red) observations.

As shown in Figure 1, the unrestricted ECE estimator performs considerably well when compared to OLS, even when all OLS assumptions hold. From Figure 9 it is clear that, even for small sample sizes the EC estimate often falls as close to the true β as the OLS. Furthermore, any remaining difference vanishes asymptotically, as the sample size grows. As the *y-axis* of Figure 1 suggests, around 50% of the total observations are used when constructing the ECE (mimicking how the boundary was chosen), meaning that around 50% of the observations lie inside the $|B_u - B_l|$-wide interval around $\hat{\beta}_{ECE}$. The concept of observation falling in and -out when constructing the ECE estimator is depicted nicely in Figure 2, where the red observations were those excluded from the unrestricted ECE

estimator.

Table 1: Monte Carlo Results: All LS assumptions are satisfied

Case 1: $\varepsilon_t \sim N(0,1)$**,** $x_t \sim U(-10,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4990	1.136E-3	0.4993	2.178E-4	0.5004	4.347E-5
$\hat{\beta}_{CECE}$	0.4982	2.101E-3	0.5004	9.210E-4	0.5003	3.789E-4
$\hat{\beta}_{CECE\_NP}$	0.4985	2.863E-3	0.4933	9.326E-4	0.4961	3.638E-4
$\hat{\beta}_{OLS}$	0.4994	3.238E-4	0.4999	2.761E-5	0.5000	2.771E-6
Case 2: $\varepsilon_t \sim t(5)$**,** $x_t \sim U(-10,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4982	2.329E-3	0.5011	4.876E-4	0.4995	8.280E-5
$\hat{\beta}_{CECE}$	0.5007	1.939E-3	0.5013	6.480E-4	0.5008	2.770E-5
$\hat{\beta}_{CECE\_NP}$	0.5006	1.394E-3	0.4921	4.594E-4	0.4960	1.827E-5
$\hat{\beta}_{OLS}$	0.5005	5.381E-4	0.5000	5.043E-5	0.5001	4.956E-6
Case 3: $\varepsilon_t \sim SN(0,1,0.8)$**,** $x_t \sim U(-10,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4982	2.050E-3	0.5006	3.578E-4	0.5001	9.014E-5
$\hat{\beta}_{CECE}$	0.5007	3.086E-3	0.4996	3.890E-4	0.4999	8.163E-5
$\hat{\beta}_{CECE\_NP}$	0.5004	1.025E-5	0.4996	2.817E-4	0.5001	6.741E-5
$\hat{\beta}_{OLS}$	0.5005	5.417E-4	0.5002	7.854E-5	0.5000	8.819E-6

Table 1 contains the results from all ECE estimators, which gives a more complete picture about the performance of ECE estimators under the classical linear regression assumptions. As shown in Table 1, all ECE estimates converged to the true value. While the OLS remains superior in terms of efficiency with the smallest variance at every sample size, the ECE estimators did follow closely with the non-parametric CECE being the most efficient. This is not surprising since OLS should be the most efficient under these ideal conditions, where as the non-parametric CECE gain additional efficiency over the other ECE estimators by approximating the underlying distribution. This is also supported by the fact the variances of the non-parametric CECE estimate are extremely close to the CECE estimate when $\varepsilon_t \sim N(0,1)$ for each sample size.

4.2 The Presence of Outliers

The problem of outliers has long been recognized to have serious effects in econometric analysis (Chatterjee and Hadi, 1986; Osborne and Overbay, 2004). The most frequently used methods in empirical studies to minimise the impact of outliers is the Least Absolute Deviation (LAD) estimator (Rice and White, 1964; Ellis, 2000; Mishra and Dasgupta, 2003), but this is not particularly satisfactory. There were other more sophisticated methods developed in the literature, but these methods were not popular in empirical studies, as they were usually complicated and difficult to implement. They also often relied on strong assumptions see, for example, the trimmed least squares (Honoré, 1992; Ruppert and Carroll, 1980) and its generalizations like Bayesian weight trimming (Elliott and Little, 2000; Little, 1991), to name a few.

While the presence of outliers does not violate LS assumptions, many empirical studies demonstrated that OLS can be extremely sensitive to them in finite samples. To produce such scenario in the next Monte Carlo experiment, we assume that 1% (or 1 if $N \leq 100$) of the observations are outliers, and so their new values are set 10 times larger than the values from their original draws. We then compare the performances of the EC and OLS estimators. Figure 3 shows a typical unrestricted EC and OLS estimates in the presence of outliers. Further results with different sample sizes and underlying DGPs (but similar qualitative outcomes) are collected in Figure 10 in Appendix B.

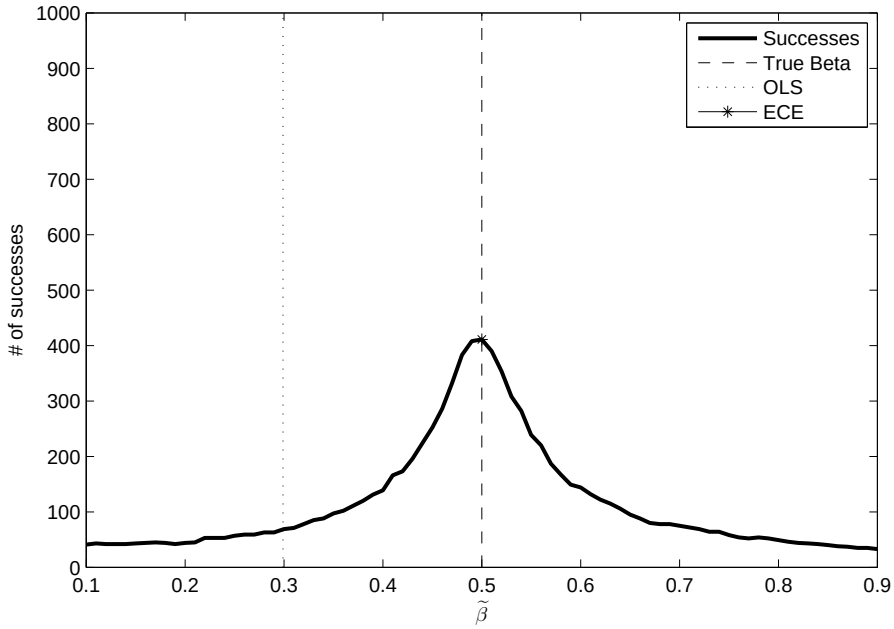


Figure 3: OLS vs. unrestricted ECE in presence of outliers. The dashed vertical line indicates the true parameter value 0.5, the dotted line the OLS estimate. The number of "successes" for each $\tilde{\beta}$ value is shown by the thick curve, and its maximum (indicated with black stars) are the respective unrestricted ECE estimates.

As seen from Figure 3, the OLS estimate is far from the true β parameter, while the unrestricted ECE is much closer to the true value. Intuitively, ECE is able to exclude the 1% outliers from the sample, which leads to more precise estimates. This concept can be illustrated more clearly in Figure 4, where the slopes of the unrestricted ECE and the LS are considerably different.

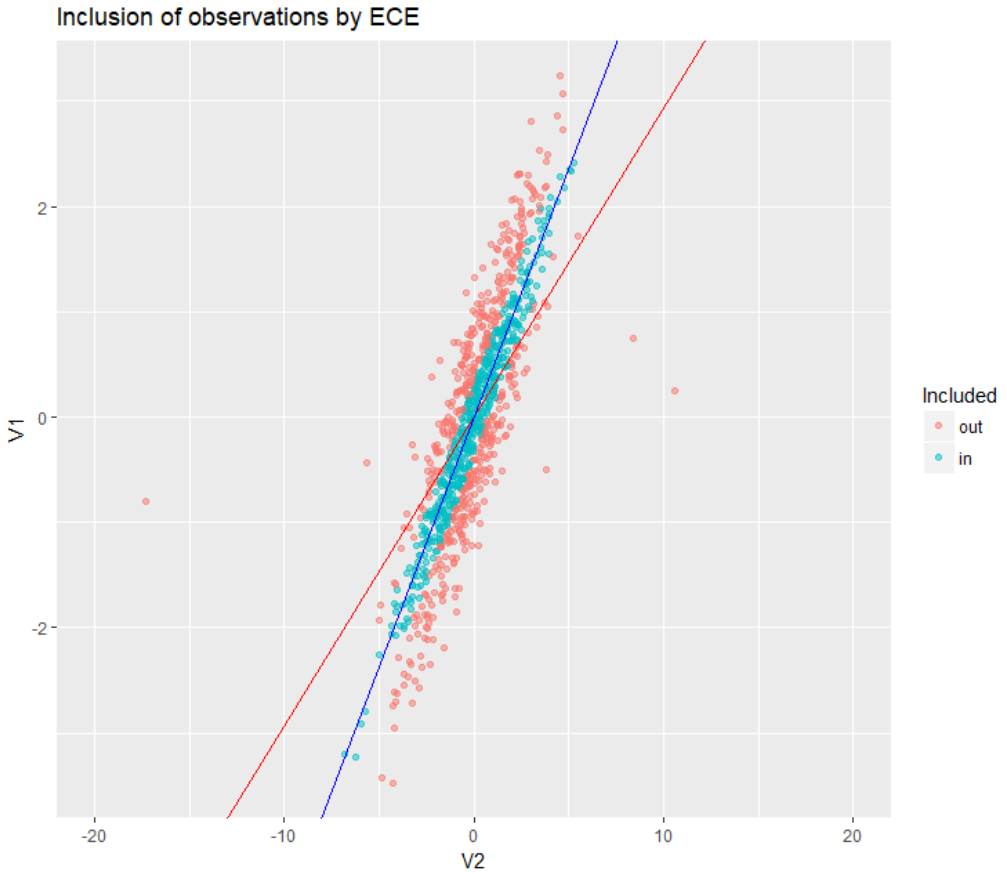


Figure 4: The used and unused observations by ECE in case of outliers. The blue line shows the ECE estimator, the red the LS, which uses all (blue and red) observations.

While this highlights the advantage of unrestricted ECE, it is important to note that this setting assumes the *fraction* of outliers (1%) is fixed for all sample sizes. In other words, the number of outliers grows as sample size increases. As this is probably a strong assumption, we have also experimented with fixing the *number* of outliers and increased the sample size. Not surprisingly, the "weight" of the outliers in computing the OLS estimates shrank, while these observations kept being excluded by the unrestricted ECE estimates, resulting in the convergence of the two respective estimators around the true β value. Nevertheless, it is clear from the results, that the unrestricted ECE outperforms OLS when the presence of outliers is non-negligible.

Table 2: Monte Carlo Results: Additional outliers

Case 1: $\varepsilon_t \sim N(0,1)$, $x_t \sim U(-10,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.5308	2.371E-3	0.5180	5.018E-5	0.5001	6.502E-6
$\hat{\beta}_{CECE}$	0.5258	2.272E-3	0.4999	4.058E-5	0.5000	5.591E-6
$\hat{\beta}_{CECE\_NP}$	0.4407	8.047E-3	0.4638	5.027E-5	0.4787	4.715E-6
$\hat{\beta}_{OLS}$	0.3475	8.007E-3	0.3511	8.563E-4	0.3500	7.571E-5
Case 2: $\varepsilon_t \sim t(5)$, $x_t \sim U(-10,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4991	7.247E-3	0.5047	5.984E-5	0.4964	1.849E-6
$\hat{\beta}_{CECE}$	0.4994	7.902E-3	0.5004	6.836E-5	0.5001	5.927E-6
$\hat{\beta}_{CECE\_NP}$	0.4378	9.732E-3	0.464	7.139E-5	0.4692	7.021E-6
$\hat{\beta}_{OLS}$	0.3557	7.601E-3	0.3514	7.991E-4	0.3502	8.327E-5
Case 3: $\varepsilon_t \sim SN(0,1,0.8)$, $x_t \sim U(-10,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4991	8.403E-3	0.4999	6.781E-4	0.5001	1.764E-5
$\hat{\beta}_{CECE}$	0.4994	2.689E-3	0.5001	7.891E-4	0.4999	2.417E-5
$\hat{\beta}_{CECE\_NP}$	0.4378	1.117E-3	0.4701	8.679E-4	0.4799	3.134E-5
$\hat{\beta}_{OLS}$	0.3557	8.135E-3	0.3541	8.021E-4	0.3551	9.812E-5

Additional set of Monte Carlo experiments to examine the performance of all the ECE estimators have also been conducted. The results can be found in Table 2. As shown in the table, the performance of non-parametric CECE was not as impressive as the other ECE variants. In fact, both unrestricted and conditional ECE have performed remarkably well in the presence of outliers, while the non-parametric CECE performed relatively poorly. However, it does show sign of convergence to the true value and it appears to be better than OLS. The reason behind this poor performance may be due to the difficulties in estimating the underlying distribution non-parametrically in the presence of outliers. Clearly, if the non-parametric estimate of the distribution is poor, then the non-parametric CECE will be severely affected. However, both unrestricted and conditional CECE remained robust in this case as they are far less sensitive to the underlying distribution.

4.3 The Failure of Random Sampling

The third set of Monte Carlo experiments concerns with the following data generating process:

$$y^* = X\beta + \varepsilon \quad \text{where} \quad y^* = \begin{cases} y & \text{if } y \geq a \\ a & \text{if } y < a, \end{cases}$$

with some fixed a, that is, the dependent variable is left censored. By assigning $a = -2.5$, Figure 5 shows typical estimates from the unrestricted ECE and OLS while Figure 6 shows how the proper censored values get disregarded by the unrestricted ECE.

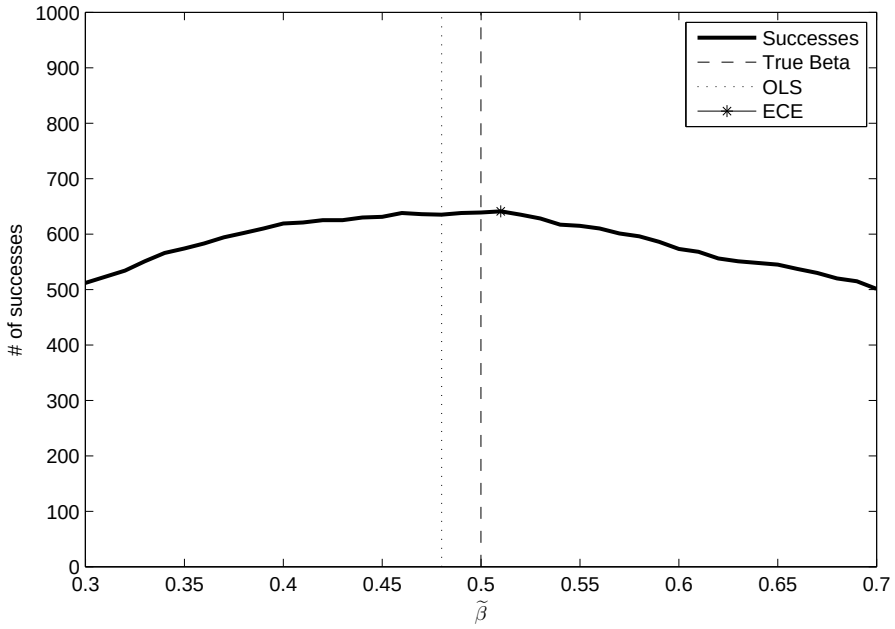


Figure 5: OLS and ECE estimates in case of censored data. $X \sim N(0, 4)$ and $N = 1000$. The dashed vertical line indicates the true parameter value 0.5, the dotted line the OLS estimate. The number of "successes" for each $\tilde{\beta}$ value is shown by the thick curve, and its maximum (indicated with black stars) are the respective ECE estimates.

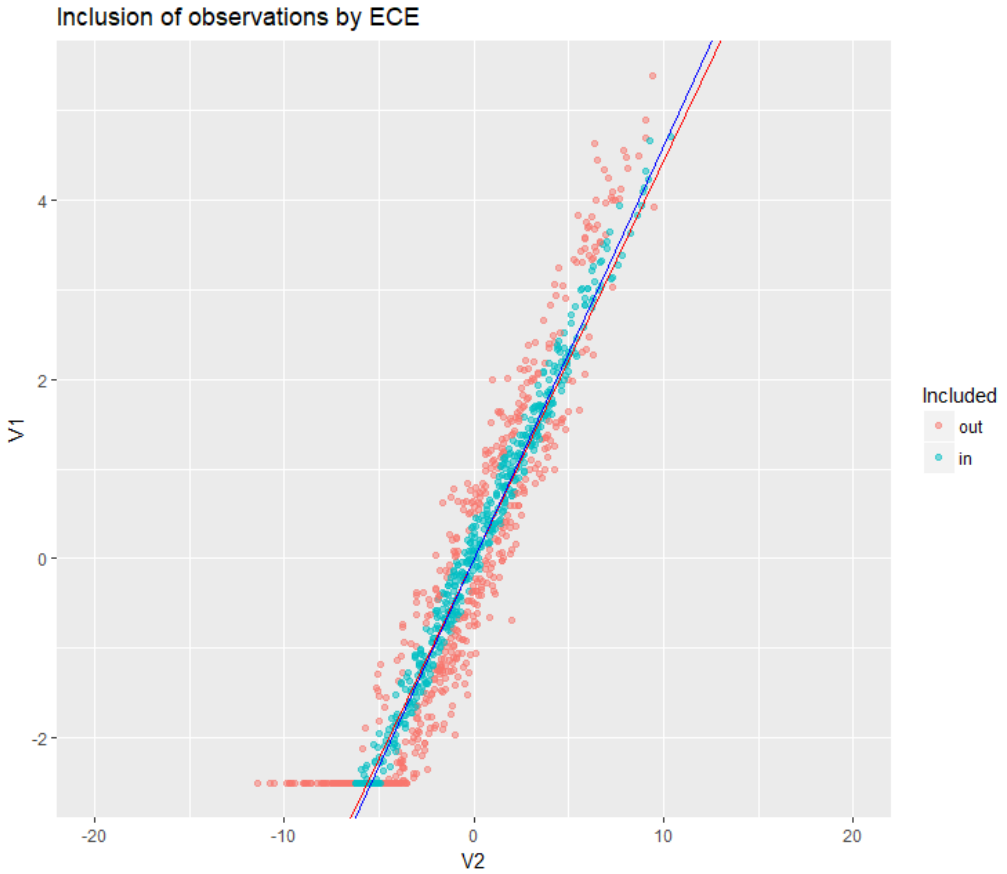


Figure 6: The used and unused observations by ECE with censored data. The blue line shows the ECE estimator, the red the LS, which uses all (blue and red) observations.

From Figure 5, it is clear that both the ECE and the OLS give imprecise estimates of the true parameter value, but the unrestricted ECE systematically provides estimate closer to the truth, which difference virtually vanishes as the sample size grows.

An interesting type of censoring is the phenomenon of *excess zeros* (also called inflated zeros, see Linnemann, 1966; Wang and Winters, 1991; Eichengreen and Irwin, 1995), that is, observations under a certain threshold can not be distinguished from (and so are treated as) zero. For example when modelling trade, it is often the case that the observed zero trade flows correspond to small, but positive flows of goods, which in turn downward biases the estimators, underestimating the true effects.

Given the practical importance of this scenario, we are going to focus the remaining Monte Carlo experiments to investigate the performance of ECE estimators in this case. To model the case of excess zeros, and to stimulate a somewhat trade-like data, we assume, that $X \sim U[0, 10]$ (for example GDP), and that the true parameter $\beta = 0.5$. The latent variable y is then generated with an X, the true β with ε follows three different distributions, namely, $N(0, 1)$, $t(5)$ and $SN(0, 1, 0.8)$. The estimator is based on the regression between X and y^*, where

$$y^* = \begin{cases} y & \text{if } y \geq 2 \\ 0 & \text{if } y < 2\,. \end{cases}$$

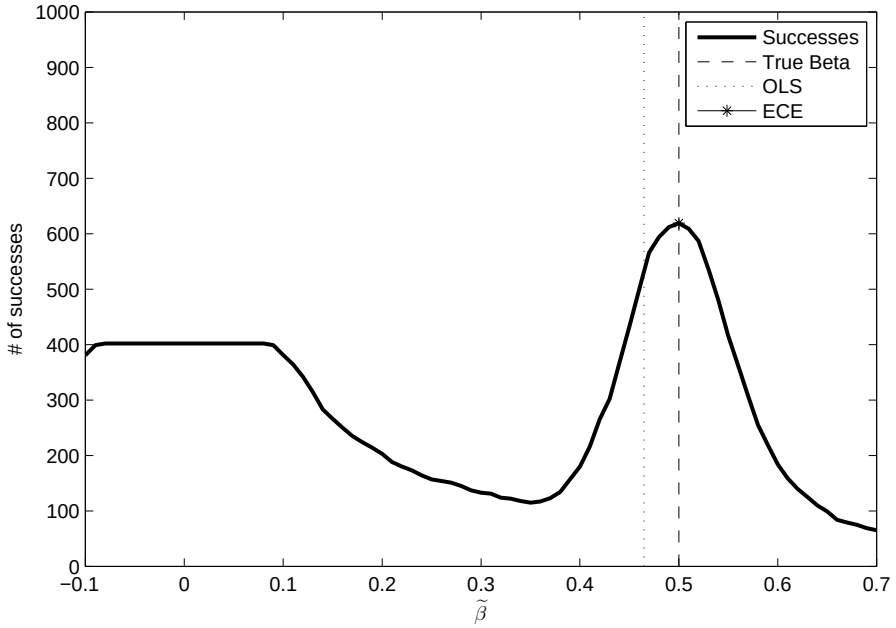


Figure 7: ECE vs. OLS in the presence of excess zeros. $X \sim U[0, 10]$ and $N = 1000$. The dashed vertical line indicates the true parameter value 0.5, the dotted line the OLS estimate. The number of "successes" for each $\tilde{\beta}$ value is shown by the thick curve, and its maximum (indicated with black stars) are the respective ECE estimates.

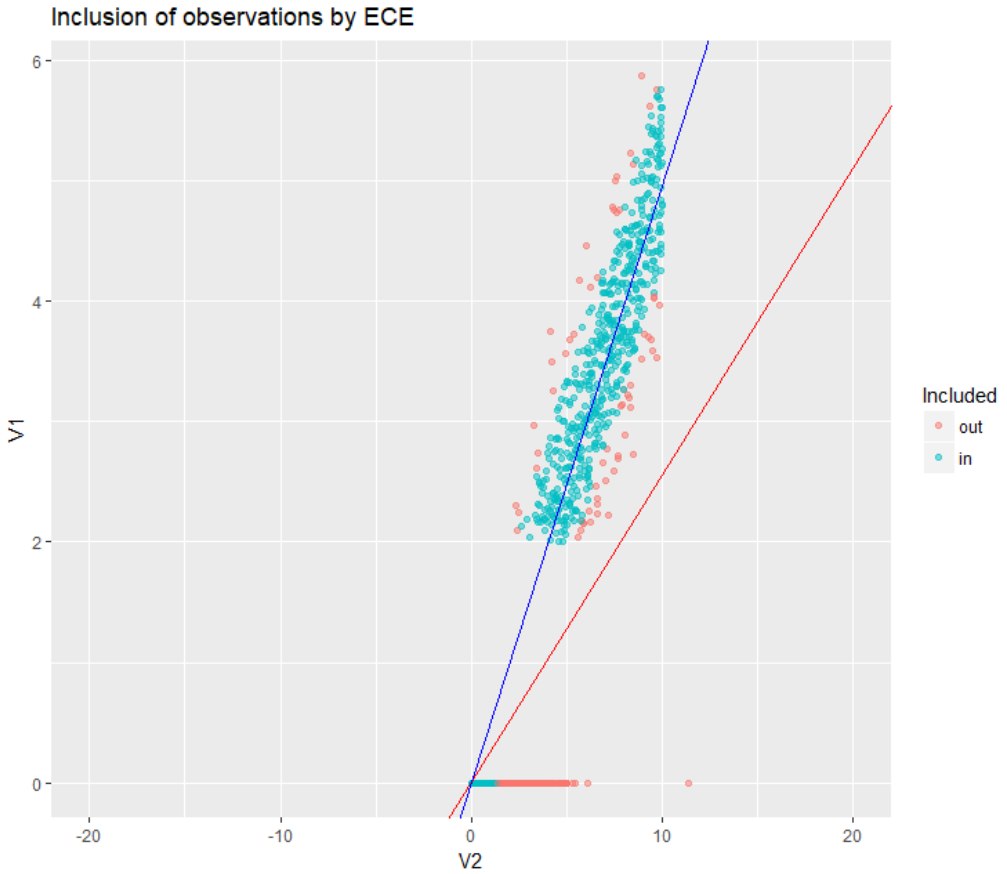


Figure 8: The used and unused observations by ECE with excess zeros. The blue line shows the ECE estimator, the red the LS, which uses all (blue and red) observations.

Figure 7 shows how the OLS estimate (dotted) is biased downwards (and falls closer to zero), but how the unrestricted ECE (black) is unaffected by the excessive zero observations in the data. Figure 8 contains a similar graph which shows the underlying concept of disregarding "suspicious" observations by the unrestricted ECE.

Table 3: Monte Carlo Results: Excess Zeros

Case 1: $\varepsilon_t \sim N(0,1)$, $x_t \sim U(0,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4631	2.277E-3	0.4639	6.561E-3	0.4936	2.401E-4
$\hat{\beta}_{CECE}$	0.4635	2.076E-3	0.4931	5.845E-3	0.5023	2.803E-4
$\hat{\beta}_{CECE\_NP}$	0.5872	2.156E-3	0.5732	7.778E-4	0.5689	6.62E-5
$\hat{\beta}_{OLS}$	0.4683	4.724E-4	0.4682	4.562E-5	0.4681	4.661E-6
Case 2: $\varepsilon_t \sim t(5)$, $x_t \sim U(0,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4571	3.541E-3	0.4694	9.268E-3	0.4889	2.854E-4
$\hat{\beta}_{CECE}$	0.4753	1.729E-3	0.4971	3.214E-3	0.5028	2.298E-4
$\hat{\beta}_{CECE\_NP}$	0.5401	5.006E-3	0.536	3.948E-4	0.5354	3.311E-5
$\hat{\beta}_{OLS}$	0.4716	6.649E-4	0.4705	6.459E-5	0.4706	6.967E-6
Case 3: $\varepsilon_t \sim SN(0,1,0.8)$, $x_t \sim U(0,10)$						
Estimators	$N = 100$		$N = 1000$		$N = 10000$	
	Mean	Variance	Mean	Variance	Mean	Variance
$\hat{\beta}_{UR}$	0.4693	3.079E-3	0.4701	9.880E-4	0.4866	8.901E-4
$\hat{\beta}_{CECE}$	0.5116	1.079E-3	0.5016	9.9901E-4	0.5020	9.801E-4
$\hat{\beta}_{CECE\_NP}$	0.5612	1.084E-3	0.5490	8.441E-4	0.5501	8.321E-4
$\hat{\beta}_{OLS}$	0.4709	7.644E-4	0.4712	8.901E-5	0.4708	7.013E-6

Further results based on the Monte Carlo experiments are collected in Table 3. The results indicate that both unrestricted and CECE performed extremely well in large sample whereas both non-parametric CECE and OLS perform relatively poorly. In fact, OLS often outperformed non-parametric CECE in this case. This, again, may be due to the difficulties in obtaining a consistent non-parametric estimate of the underlying distribution which lead to severely biased estimate.

5 Empirical Illustration

In this section we apply the theoretical results in practice to show how the unrestricted ECE can be used to get fine estimates in the presence of excess zeros. To illustrate this we compare unrestricted ECE to Poisson Pseudo Maximum Likelihood (PPML, proposed by

Santos Silva and Tenreyro, 2006) and to Least Squares on a bilateral trade data set. The panel comprises 180 trading countries observed annually over 53 years (for the period 1960-2012), giving nearly 1.1 million exporterimportertime data points. Raw net import-export data were collected from IMF's Direction of Trade Statistics Yearbook, and were deflated to 2000 US $ using US CPI from IMF's International Financial Statistics Yearbook. Population and GDP measures were obtained from the World Bank's World Development Indicators. Other country- and country-pair specific demographics were collected from the World Trade Organization, CIA's Factbook and Wikipedia. We seek to identify the factors exerting trade flows, like bi- or multilateral trade agreements, GDP, or distance of the trading parties.

By observing the data the presence of excess zeros, as is often the case for trade data, becomes obvious: over 315 000 of the 818 000 observed trade flows are nil. We estimate a standard gravity model with fixed effects and additive disturbances

$$y_{ijt} = x'_{ijt}\beta + \alpha_i + \gamma_j + \lambda_t + \varepsilon_{ijt}\,, \tag{33}$$

proposed by Matyas (1997), where the left and right hand variables are in logs, not measured in levels. The employed variables are similar to Rose (2004) and Konya et al. (2011): y is the natural logarithm of real export from country i to country j in year t, x' comprises all country-pair, country-time and country-specific characteristics,[3] α_i and γ_j are exporter and importer country fixed effects, finally, λ_t is the year fixed effect.

Along with the unrestricted ECE, multiple reference estimators are constructed. First, as the logarithm of zero is not defined, we can simply omit the zero observations (Brada and

[3]The regressors used are: BOTHIN (GATT/WTO membership dummy); rGDP (log real GDP measure); rGDP/POP (log real GDP per capita); DIST (log great circle distance); LAND (log land area of the country); CLANG (common language dummy); CBORD (common border dummy); LLOCK (landlocked dummy); ISLAND (island dummy); EVCOL (ever colonized dummy); COMCOL (common colonizer dummy); MUNI (same monetary union dummy); TA (common trade agreement dummy).

Mendez, 1985; Bikker, 1987), and estimate the log-linearized model with Least Squares. This obviously ignores a large, non-random part of the data set (more distant or smaller country-pairs exhibiting zero trade), which might suggest that the captured effect of trade membership on trade activity overestimates the true one (if zero trade flows are not uncommon between member countries), or in fact underestimates it (if zero trade flows are less typical between member countries).

To account for zero flows as well, common practice suggests (see Linnemann, 1966; Wang and Winters, 1991; Eichengreen and Irwin, 1995; Frankel, 1997) to add a nuance, say 1, to the observations (still measured in levels). In this way changes in large trade flow values are non-noticeable, especially after taking logarithms, and zero trades remain zeros if measured in logs. Estimating the model with Least Squares obviously alleviates the selectivity issue arisen from leaving out small countries systematically, for example, yet still suffers from at least one issue. Santos Silva and Tenreyro (2006), also detailed in Baltagi et al. (2017), draw attention to the misleading inferences drawn from a log-linearized gravity model estimated by Least Squares in the presence of excess zeros or heteroscedasticity. They argue, that OLS estimates can be problematic and heavily biased on a log-linear model as (i) log-linearization of a constant elasticity model generally results in an error term whose mean depends on the covariates (ii) rounding-down of flows might occur, thus zero trades from small ones can not be distinguished, leading to possibly severe measurement errors.[4] As a consequence, Santos Silva and Tenreyro (2006) suggest the Poisson Pseudo Maximum Likelihood (PPML) estimator to estimate gravity-type equations, which is consistent even under the data and model formulational problems listed above, thus capable to uncover the true effects various economic factors exert on trade activity. Table 4 collects the estimates together with the t-values.

[4]More generally, estimation problems with log-linear models are mere implications of Jensen's inequality: $\mathbb{E}(\log y) \neq \log \mathbb{E}(y)$.

Table 4: The ECE, Least Squares and PPML estimators for a gravity equation

Variable	ECE		PPML		OLS (with zeros)		OLS (w/o zeros)	
	$\hat{\beta}$	t-value	$\hat{\beta}$	t-value	$\hat{\beta}$	t-value	$\hat{\beta}$	t-value
BOTHIN	0.778	24.29	0.951	19.77	0.758	14.58	0.6415	22.12
rGDP1	1.849	136.96	0.943	33.26	1.301	59.14	0.817	67.30
rGDP2	1.510	117.71	0.737	30.78	1.486	67.55	0.708	60.31
rGDPPOP1	-0.070	6.74	0.022	5.62	-0.100	10	-0.006	1.17
rGDPPOP2	-0.192	15.88	0.120	6.80	-0.181	16.45	-0.02	3.55
DIST	-2.278	265.71	-0.682	-96.02	-2.256	225.6	-1.086	222.65
LAND1	-	-	0.062	3.79	-	-	-	-
LAND2	-	-	0.148	7.30	-	-	-	-
CLANG	0.904	63.42	0.143	12.18	0.896	50.27	0.424	50.26
CBORD	1.145	31.00	0.560	40.49	1.161	27.64	0.816	45.61
LLOCK1	-	-	2.490	12.72	-	-	-	-
LLOCK2	-	-	0.139	1.19	-	-	-	-
ISLAND1	-	-	-0.717	-3.03	-	-	-	-
ISLAND2	-	-	-0.941	-4.45	-	-	-	-
EVCOL	1.506	28.22	0.270	9.34	1.331	18.75	1.067	36.67
COMCOL	1.283	65.13	0.217	11.79	1.277	58.05	0.403	36.98
MUNI	0.315	19.11	0.211	14.33	0.576	24	0.461	40.51
TA	1.531	77.32	0.401	29.51	1.076	59.78	0.516	60.26
R^2	0.5579		0.883		0.5952		0.9845	
No. obs.	697 938		813 234		813 234		499 708	
No. par.	413		413		413		413	

The dependent variable is real export measured in logs, except for the PPML estimates, where is measured in levels. Similar results are obtained if import is used as a proxy for real trade flows, therefore they are not reported. The ECE estimates are obtained by applying ECE on a transformed data set, where all fixed effects are already removed.

Several regularities stand out from Table 4. First, the OLS estimates with the inclusion of zero observations are much larger in magnitude than the ones without taking zeros into account. This sheds some light on the conjecture that, omission of observations (in this case pairs of smaller countries who are quite possibly not members of the same trade union) downward biases the true effect. Precisely, what we see is that country-pairs with no export are more distant, have smaller GDPs, and are less likely to be members of the same trade agreement (trade union), than their trading counterparts (Table 5). It is then expected that including zero trade flows will further push-up in magnitude the effects various economic factors have on real trade activity.

Table 5: Zero vs. non-zero flows

Variable	Non-zero obs.	Zero obs.
Same trading agreement (prob.)	0.1984	0.1286
Log GDP of exporter (log $)	23.38	22.38
Log GDP of importer (log $)	23.31	22.58
Distance (km)	4209.35	4691.81

Least Squares nevertheless is expected to lead to inconsistent estimates so long the mean of ε is a function of the observables. The PPML, which estimates model (33) in its multiplicative form, and yields consistent estimators, produces results somewhat comparable to OLS. The parameter estimates in general are closer to zero economically, the Land Locked dummy of the importer country does not even differ statistically from zero. The major drivers of trade activity, however, exert similar effects on real export (trade activity) to what Least Squares or ECE suggest. Those PPML estimates coincide with Santos Silva and Tenreyro (2006) who raise doubts on the strong explanatory powers of common country-pair characteristics (border, linguistics), or a larger than one elasticity estimate for GDP.

The unrestricted ECE, as being performed on the log-linearized gravity model, falls close to the OLS estimates, yet are somewhat higher. This may be the result of the ECE excluding some of the observations with zero flows, those which would artificially mitigate the effect of economic factors on trade, like censored values. Simulated standard errors are constructed to enable drawing inferences on the results.

Following Assumption A5 we set $B = 8$ which seemed to be the most reliable in terms of the number of observations included in the estimation. A non-proper choice of B means that either too many or too few observations are used to construct the estimator. In the first case, if nearly the whole sample is used (B is too wide), slightly different B can give fundamentally different estimates, especially if optimization starts from different initials.

This is so as now multiple $\tilde{\beta}$ have a 2B-wide interval around which encompasses the high number of observations. On the contrary, when B is too narrow, too few observations, say M, are used to construct the ECE, consequently multiple $\tilde{\beta}$, may be far apart, are suitable to formulate an ECE with M observations. This chain of thoughts implies that ECE is "stable" and produced good estimates if similar B values lead to similar estimates, and that the optimizer converges to the same B from different starting points.

6 Conclusion

This paper introduced a new estimation strategy called Event Count Estimator which have several variants, namely unrestricted ECE, conditional ECE (CECE) and Non-parametric CECE. In addition to derive several statistical properties of the proposed estimators, we also showed that this new family of estimators outperformed OLS in the presence of outliers and several other data related problems. Although ECE in theory is less efficient than the OLS as it uses less information from the data, the effect is significant only in small sample. In moderate and large sample, Monte Carlo simulations showed that the unrestricted ECE and CECE performed substantially better than the OLS estimator in the presence of outliers and excess zeros. This is due to their abilities to exclude potentially problematic observations which lead to more robust estimates.

The practical usefulness of the unrestricted ECE estimator was demonstrated through an empirical application that contains data problems such as outliers and excess zeros. Specifically, we estimated a standard gravity model of trade with fixed effects using three different estimators, namely, the unrestricted ECE, Poisson Pseudo Maximum Likelihood (PPML) and Least Squares. The results suggested that the LS and PPML are highly sensitive to the underlying assumptions of the data generating process whereas the unrestricted ECE appeared to be much more robust.

Appendix A: Technical Proofs

The Lemmas below is useful for proving Proposition 1.

Lemma 1. *There exists* $r > 1$ *and* $0 < \delta \leq r$ *such that*

$$\lim_{N\to\infty} \sum_{i=1}^{N} \left\{ \frac{\mathbb{E}\left|d_i(\beta, B) - \mathbb{E}\left[d_i(\beta, B)\right]\right|^{r+\delta}}{i^{r+\delta}} \right\}^{1/r} < \infty. \tag{A.1}$$

Proof: Recall $\mathbb{E}\left[d_i(\beta, B)\right] = \Pr\left[B_i < u_i(\beta) < B_u\right]$ and therefore $\mathbb{E}\left|d_i(\beta, B) - \mathbb{E}\left[d_i(\beta, B)\right]\right| < 1$ and hence

$$\begin{aligned}\lim_{N\to\infty} \sum_{i=1}^{N} \left\{ \frac{\mathbb{E}\left|d_i(\beta, B) - \mathbb{E}\left[d_i(\beta, B)\right]\right|^{r+\delta}}{i^{r+\delta}} \right\}^{1/r} &< \lim_{N\to\infty} \sum_{i=1}^{N} i^{-\frac{r+\delta}{r}} \\ &\equiv \zeta\left(\frac{r+\delta}{r}\right),\end{aligned}$$

where $\zeta(.)$ denotes the Riemann zeta function, which has finite value when the argument is greater than 1. This means the lemma holds for any $r \geq 1$ and $0 < \delta \leq r$ that satisfy

$$\frac{r+\delta}{r} > 1.$$

This completes the proof. ■

Lemma 2. *Define*

$$\bar{Q}_N(\beta, B) = N^{-1} \sum_{i=1}^{N} \Pr\left(A_l < u_i(\beta) < A_u\right), \tag{A.2}$$

then under Assumption (4) $Q_N(\beta, B) - \bar{Q}_N(\beta, B) = o_p(1)$.

Proof: Under Assumption A4 and Lemma 1, $Q_N - \bar{Q}_N = o_p(1)$ by Theorem 2.10 in McLeish (1975) (See also Theorem 3.47 in White (1999)). This completes the proof. ■

Proof of Proposition 1: First note

$$\begin{aligned}\bar{Q}_N &= \sum_{i=1}^{N} \Pr\left(A_l < u_i(\beta) < A_u\right) \\ &= \sum_{i=1}^{N} \Pr\left(A_l < \varepsilon_i + x_i'(\beta_0 - \beta) < A_u\right) \\ &= \sum_{i=1}^{N} \Pr\left(A_l - x_i'(\beta_0 - \beta) < \varepsilon_i < A_u - x_i'(\beta_0 - \beta)\right) \\ &= \sum_{i=1}^{N} \int_{A_l - x_i'\Delta\beta}^{A_u - x_i'\Delta\beta} g_i(\varepsilon) d\varepsilon,\end{aligned}$$

where $\Delta\beta = \beta_0 - \beta$. Differentiate the last line with respect to β gives:

$$\frac{\partial \bar{Q}_N}{\partial \beta} = \sum_{i=1}^{N} \left[g(A_u - x_i'\Delta\beta) - g(A_l - x_i'\Delta\beta)\right] x_i,$$

which equals to the null vector only when $\Delta\beta = 0$ under Assumption A5. Moreover, the density function is concave as implied by Assumption A5 and therefore $\bar{Q}_N$ has a unique global maximum at β_0. Now, Lemma 2 showed that $Q_N(\beta, B)$ converges to a non-stochastic function $\bar{Q}_N$ which has a unique global maximum. Furthermore, Assumption A3 ensures the compactness of the parameter space and that $Q_N(\beta, B)$ is a measurable function of y_i. Therefore $\hat{\beta}_{ECE_{UR}} - \beta_0 = o_p(1)$ by Theorem 4.1.1 in Amemiya (1985). This completes the proof. ■

Proof of Proposition 2: By Jensen's inequality

$$\mathbb{E}\left(\frac{A_u^2}{\varepsilon_i^2}\right) > \frac{A_u^2}{\mathbb{E}(\varepsilon_i^2)} \tag{A.3}$$

and simple algebra manipulation leads to the result. This completes the proof. ■

Proof of Proposition 3: It is straightforward to show that

$$\frac{\partial J_i}{\partial \beta} = 4\text{sgn}\left(u_i\right)\phi(z_i)z_i x_i, \tag{A.4}$$

where $\text{sgn}(x) = 1$ if $x \geq 0$ and $\text{sgn}(x) = -1$ if $x < 0$. Note that $\sigma_i = |u_i| = |\varepsilon_i + x_i'\Delta\beta|$, therefore, when $\Delta\beta = 0$, i.e. $\beta = \beta_0$, equation (A.4) becomes

$$\frac{\partial J_i}{\partial \beta} = \frac{4}{|\varepsilon_i|}\phi\left(\frac{A_u}{\varepsilon_i}\right)\frac{A_u}{\varepsilon_i}x_i.$$

By construction, the expectation of $P_N(\beta, B)$ as defined by equation (22) exists and finite. Under Assumption B4 and by WLLN, $P_N(\beta)$ converges to a non-stochastic function in β. Under Assumption B2, $P_N(\beta)$ is a concave function in β with the partial derivatives of $P_N(\beta)$ also converges to 0 when $\beta = \beta_0$. Thus, $P_N(\beta)$ is uniquely maximised at β_0 as $N \to \infty$. Along with Assumptions A1 and A3, $\hat{\beta}_{CECE} - \beta_0 = o_p(1)$ by Theorem 4.1.1 in Amemiya (1985). This completes the proof. ■

Proof of Proposition 4: First note that

$$\frac{\partial^2 J_i}{\partial \beta \partial \beta'} = 4\left(2 - \frac{A_u^2}{u_i^2}\right)\phi\left(\frac{A_u}{|u_i|}\right)\frac{A_u}{|u_i|}\frac{x_i x_i'}{u_i^2}, \tag{A.5}$$

which exists and continuous in β with finite first moment under Assumption B4. Consider a sequence β_N such that $\beta_N - \beta_0 = o_p(1)$, then equation (A.5) converges to

$$c_0 \equiv \mathbb{E}\frac{\partial^2 J_i}{\partial \beta \partial \beta'}\bigg|_{\beta=\beta_0} = 4\mathbb{E}\left[(2 - z_i^2)\phi(z_i)\frac{z_i}{\varepsilon_i^2}\right]\Sigma_x \tag{A.6}$$

by Continuous Mapping Theorem. Under Assumptions A2 and B4, c_0 is non-singular. Now, Assumptions A2, B4 and B5 imply

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial J_i}{\partial \beta}\xrightarrow{d} N(0,c_1),$$

where $c_1 = 16\mathbb{E}\left[\varepsilon_i^{-2}\phi^2(z_i)z_i^2\right]\Sigma_x$ by Theorem 5.20 in White (1999). The result then followed by Theorem 4.1.3 in Amemiya (1985). This completes the proof. ■

Proof of Proposition of 5: Under Assumption A3-A2 and C1, β_0 belongs to a compact Euclidean subset of $\mathbb{R}^k$, $J_i(\beta)$ is continuous in β and measurable with respect to y for all i. Moreover, $N^{-1}\sum_{i=1}^{N} J_i(\beta)$ converges to $\mathbb{E}\left[J_i(\beta)\right]$ in probability by WLLN. Note that

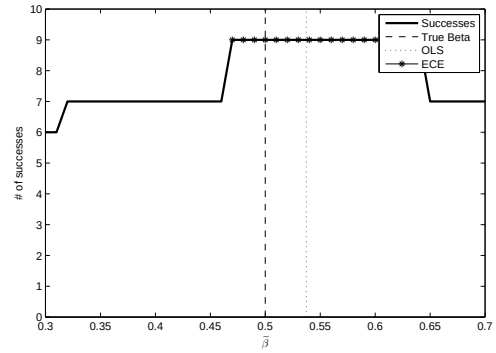
$$\begin{aligned}
\left.\frac{\partial J_i}{\partial \beta}\right|_{\beta=\beta_0} &= \left[1-G(0)\right]^{-2}\left\{\left[1-G(0)\right]\left[g(A_u)-g(0)\right]+g(0)\left[G(A_u)-G(0)\right]\right\}x_i \\
&\quad + \left[G(0)\right]^{-2}\left\{G(0)\left[g(0)-g(A_l)\right]-g(0)\left[G(0)-G(A_l)\right]\right\}x_i \\
&= \left\{G(0)\left[1-G(0)\right]\right\}^{-1}\left\{G(0)\left[g(A_u)-g(0)\right]+\left[1-G(0)\right]\left[g(0)-g(A_l)\right]\right\}x_i \\
&\quad + g(0)\left\{G(0)\left[1-G(0)\right]\right\}^{-2}\times \\
&\quad \left\{G^2(0)\left[G(A_u)-G(0)\right]-\left[1-G(0)\right]^2\left[G(0)-G(A_l)\right]\right\}x_i.
\end{aligned}$$

For symmetric distribution, set $A_l = -A_u$ and note that $G(0) = 1-G(0)$, $G(A_u) = 1-G(A_l)$ and $g(A_u) = g(A_l)$ then straightforward algebra shows that $\left.\frac{\partial J_i}{\partial \beta}\right|_{\beta=\beta_0} = 0$. If g is not even, then select A_u and A_l such that Assumption C2 holds, then straightforward algebra shows $\left.\frac{\partial J_i}{\partial \beta}\right|_{\beta=\beta_0} = 0$.

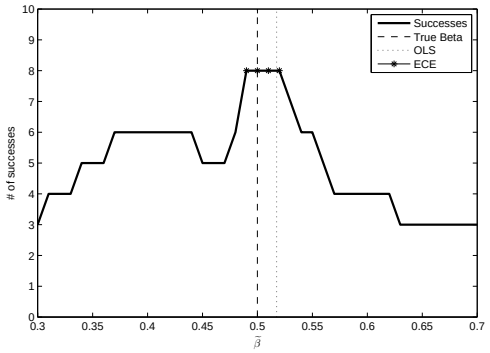
β_0 maximises $\mathbb{E}\left[J_i(\beta)\right]$ uniquely and therefore $\hat{\beta}_{CECE\_NPA} - \beta_0 = o_p(1)$ by Theorem 4.1.1 in Amemiya (1985). Under Assumption C3, the $\hat{G}$ approaches G as N grows suffi-

ciently large, therefore the vector that maximises $\hat{G}$ approaches the vector that maximises G as $N \to \infty$. Hence, $\hat{\beta}_{CECE\_NP} - \beta_0 = o_p(1)$. This completes the proof. ■

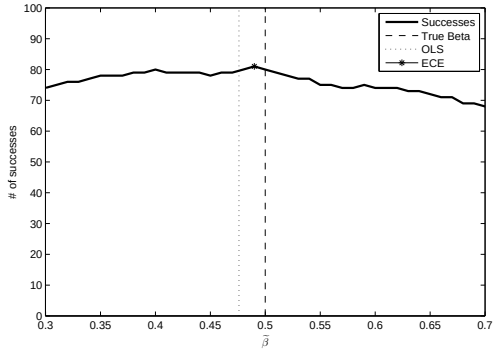
Appendix B: Additional Tables and Figures



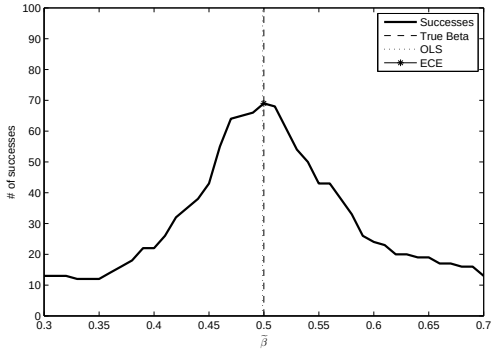
(a) $X \sim N(0,1)$ and $N = 10$



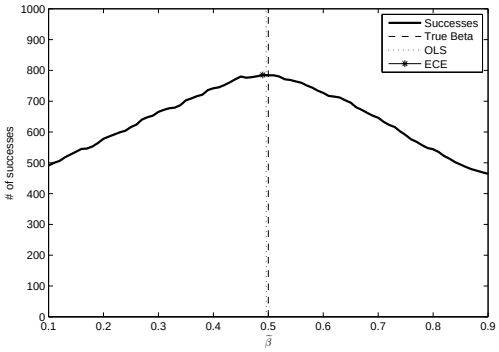
(b) $X \sim U[-10,10]$ and $N = 10$



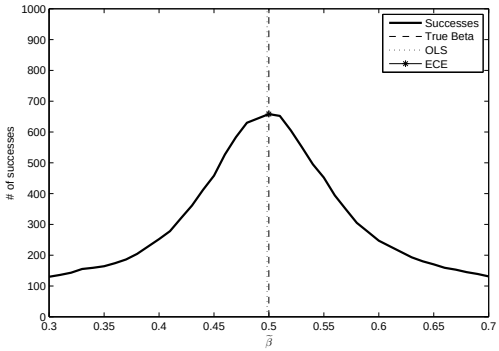
(c) $X \sim N(0,1)$ and $N = 100$



(d) $X \sim U[-10,10]$ and $N = 100$

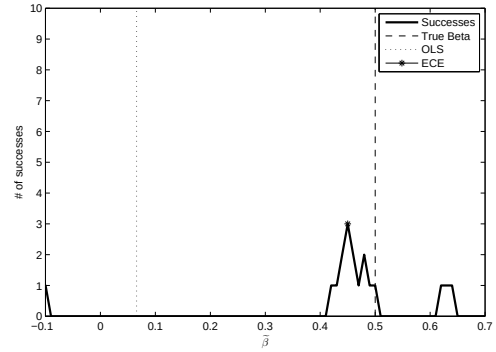


(e) $X \sim N(0,1)$ and $N = 1000$

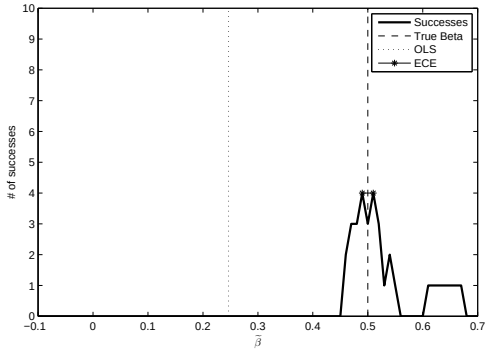


(f) $X \sim U[-10,10]$ and $N = 1000$

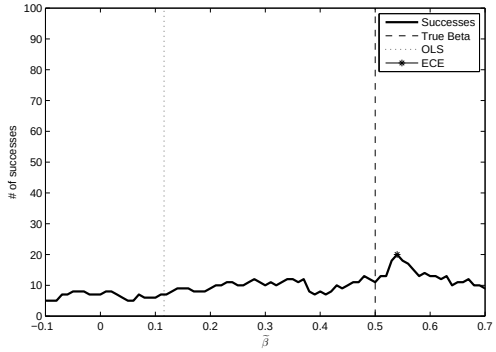
Figure 9: Comparison of ECE and OLS estimates when all OLS assumptions are satisfied. The dashed vertical line indicates the true parameter value 0.5, the dotted line the OLS estimate. The number of "successes" for each $\tilde{\beta}$ value is shown by the thick curve, and its maximum (indicated with black stars) are the respective ECE estimates.



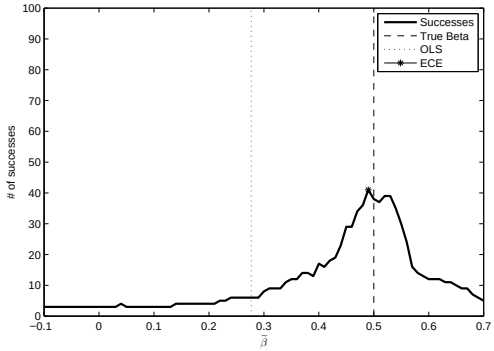
(a) $X \sim N(0,1)$ and $N = 10$



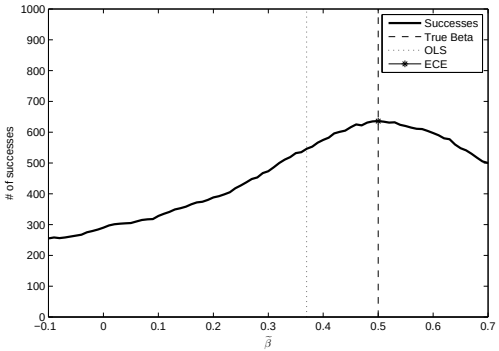
(b) $X \sim U[-10,10]$ and $N = 10$



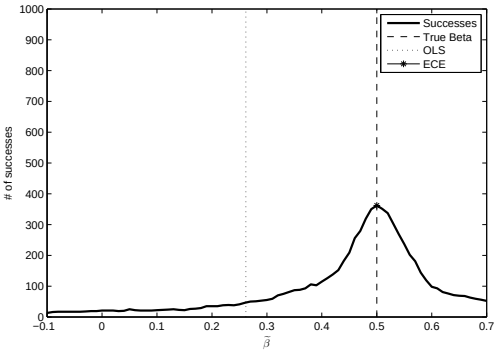
(c) $X \sim N(0,1)$ and $N = 100$



(d) $X \sim U[-10,10]$ and $N = 100$



(e) $X \sim N(0,1)$ and $N = 1000$



(f) $X \sim U[-10,10]$ and $N = 1000$

Figure 10: Typical ECE vs. OLS estimates with 1% outliers, with different sample sizes and for different DGPs.

References

Amemiya, T. 1985. *Advanced Econometrics*. Harvard University Press.

Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, **12**, 171–178.

Baltagi, B. H., Egger, P. H., and Erhardt, K. 2017. The Estimation of Gravity Models in International Trade. In: Matyas, L. (ed), *The Econometrics of Multi-dimensional Panels - Theory and Applications*. Springer Verlag (forthcoming).

Bikker, J. 1987. An International Trade Flow Model with Substitution: An Extension of the Gravity Model. *Kyklos*, **40**(3), 315–337.

Brada, J., and Mendez, J. 1985. Economic Integration among Developed, Developing and Centrally Planned Economies: A Comparative Analysis. *The Review of Economics and Statistics*, **67**(4), 549–556.

Chatterjee, S., and Hadi, A. 1986. Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, **1**(3), 379–393.

Eichengreen, B., and Irwin, D. A. 1995. Trade Blocs, Currency Blocs and the Reorientation of World Trade in the 1930s. *Journal of International Economics*, **38**, 1–24.

Elliott, M. R., and Little, R. J. A. 2000. Model-based Approaches to Weight Trimming. *Journal of Official Statistics*, **16**, 191–210.

Ellis, S. P. 2000. Singularity and Outliers in Linear Regression with Application to Least Squares, Least Absolute Deviation, and Least Median of Squares Linear Regression. *Metron*, **58**(1).

Frankel, J. 1997. Regional Trading Blocs in the World Economic System. *Washington, DC: Institute for International Economics.*

Honoré, B. 1992. Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects. *Econometrica*, **60**(3).

Konya, L., Matyas, L., and Harris, M. 2011. GATT/WTO Membership Does Promote International Trade After All – Some New Empirical Evidence. *Working paper, Munich Personal RePEc Archive, 34978, online at http://mpra.uni-muenchen.de/34978/.*

Linnemann, H. 1966. *An Econometric Study of International Trade Flows.* Amsterdam: NorthHolland.

Little, R. J. A. 1991. Inference with Survey Weights. *Journal of Official Statistics*, **7**, 405–424.

Matyas, L. 1997. Proper Econometric Specification of the Gravity Model. *The World Economy*, **20**, 363–369.

McLeish, D.L. 1975. A Maximal Inequality and Dependent Strong Laws. *Annals of Probability*, **3**, 826–836.

Mishra, S. K., and Dasgupta, M. 2003. Least Absolute Deviation Estimation of Multi-Equation Linear Econometric Models: A Study Based on Monte Carlo Experiments. *NEHU Economics Working Paper No. skm/02.*

Osborne, J. W., and Overbay, A. 2004. The Power of Outliers (and Why Researchers Should Always Check for Them). *Practical Assessment, Research & Evaluation*, **9**, 1–12.

Rice, J. R., and White, J. S. 1964. Norms for Smoothing and Estimation. *SIAM Review*, **6**, 243–256.

Rose, A. K. 2004. Do We Really Know That the WTO Increases Trade? *American Economic Review*, **94**, 98–114.

Ruppert, D., and Carroll, R. J. 1980. Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association*, **75**(372), 828–838.

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

Wang, Z. K., and Winters, L. 1991. The Trading Potential of Eastern Europe. *No. 610, CEPR Discussion Papers.*

White, H. 1999. *Asymptotic Theory for Econometricians,* Revised edition. Academic Press.