On Uniform Inference in Nonlinear Models with Endogeneity

Shakeeb Khan* Boston College

Denis Nekipelov* University of Virginia

First Version: October 2014

This Version: September 2019

Abstract

This paper explores the uniformity of inference for parameters of interest in nonlinear econometric models with endogeneity. Here the notion of uniformity arises because the behavior of estimators of parameters is shown to vary with where they lie in the parameter space. As a result, inference becomes nonstandard in a fashion that is loosely analogous to inference complications found in the unit root and weak instruments literature, as well as the models recently studied in Andrews and Cheng (2012), Chen, Ponomareva, and Tamer (2014), Han and McCloskey (2019). Our main illustrative example is the standard sample selection model, where the parameter is the intercept term as in Heckman (1990), Andrews and Schafgans (1998) and Lewbel (2007). We show here there is a *discontinuity* in the limiting distribution for an estimator despite it being uniformly (across degrees of selection) consistent. This discontinuity prevents standard inference procedures from being uniformly valid, and motivates the development of new methods, for which we establish asymptotic properties. Finite sample properties of the procedure is explored through a simulation study and an empirical illustration using the Mroz (1987) data set as in Newey, Powell, and Walker (1990).

JEL Classification: C12,C13,C14,C15.

Keywords: Selection on observables and unobservables, uniform inference, fixed and drifting sequences of parameters.

*We are grateful to seminar participants at UW-Madison, Penn State, Harvard/MIT, HKUST, SHUFE, SMU, NUS, Columbia, LSE, University of British Columbia, University of Colorado Boulder, University of Southern California, University of Toronto, University of Washington, and conference participants at CEME in Cornell, and the LAMES in Mexico City for helpful comments. Support from the NSF is gratefully acknowledged.

1 Introduction

Endogeneity and sample selectivity are frequently encountered in econometric models, and failure to correct for them appropriately can result in incorrect inference. In linear models, with the availability of appropriate instruments, two-stage least squares (2SLS) yields consistent estimates without the need for making parametric assumptions on the error disturbances. This is not the case in non-linear models, as the consistency of 2SLS depends critically upon the orthogonality conditions that arise in the linear-regression context.

One approach for handling endogeneity in many non-linear models has required parametric specification of the error disturbances. A more recent literature in econometrics has developed methods that do not require parametric distributional assumptions, which is more in line with the 2SLS approach in linear models. This semiparametric, or "distribution-free" approach can be roughly divided into two groups , depending on the source of endogeneity that arises in the model.

In one case the source of endogeneity is that the data available to the econometrician is selected nonrandomly, resulting in what is now well known as *sample selection bias* (Gronau (1973), Heckman (1974)). Distribution free methods were proposed in Powell (1986),Ahn and Powell (1993), Lewbel (2007), Newey (2009), and recent work in D'Hautefoeuille, Maurel, and Zhang (2018) and Honoré and Hu (2019). In the other case, the source of endogeneity is the explanatory variables themselves. Recent work such as Blundell and Powell (2003), Vytlacil and Yildiz (2007), Khan and Nekipelov (2018) proposed semiparametric, distribution free methods which are robust to misspecification of the distribution of the unobserved components of the model.

In this paper we point out that the inference problem in these models with endogeneity has not yet been adequately solved since they have yet to propose an inference method that is *uniformly valid* in the parameters of the model. By this we mean that the large sample properties of estimators for these models will vary depending on the values of the unknown parameters of the model. Furthermore this often results in a limiting distribution theory that will be *discontinuous* in these parameters. It is this discontinuity which motivates the new inference procedures we propose.

The rest of the paper is organized as follows. The next section illustrates the main difficulty with conducting inference by reconsidering the semiparametric[1] sample selection model Ahn and Powell (1993), Heckman (1990), Lewbel (2007), Andrews and Schafgans (1998)),Newey

[1]Nonparametric sample selection models were considered in Das, Newey, and Vella (2003).

(2009). In Sections 2 and 3 we show that the large sample behavior of existing inference methods vary discontinuously with the degree of selection on unobserved variables, with one extreme case being when selection is on observed variables only. As we will show this discontinuity results in impossibility results for valid uniform inference, and motivates our new inference procedures. Sections 4 and 5 explores the finite sample properties of our new inference methods in two ways. In Section 4 we consider a simulation study, and in Section 5, we apply the new inference method proposed in Section 2 to study the slope coefficients in a female labor supply curve, using the data set introduced in Mroz (1987). Section 6 explores how our proposed inference models can be used to conduct valid inference for parameters of interest in other nonlinear models with endogeneity, such as triangular and non triangular systems of discrete variable equations often explored in labor economics and empirical industrial organization.

Section 7 concludes by summarizing our results and suggesting areas for future research that will aim to primarily address the unresolved issues in this paper. An Appendix collects all the proofs of the main theorems in the paper.

2 Identification and Inference in the Sample Selection model

In this section we illustrate the complications that can arise when conducting inference in the sample selection model, which has been of widespread interest in both theoretical and applied econometrics. This is because estimation of economic models is often confronted with the problem of sample selectivity, which is well known to lead to specification bias if not properly accounted for. Sample selectivity arises from nonrandomly drawn samples which can be due to either self-selection by the economic agents under investigation, or by the selection rules established by the econometrician. In labor economics, the most studied example of sample selectivity is the estimation of the labor supply curve, where hours worked are only observed for agents who decide to participate in the labor force. Examples include the seminal works of Gronau (1973) and Heckman (1974). It is well known that the failure to account for the presence of sample selection in the data may lead to inconsistent estimation of the parameters aimed at capturing the behavioral relation between the variables of interest.

Econometricians typically account for the presence of sample selectivity by estimating a bivariate equation model known as the sample selection model (or using the terminology of Amemiya (1985), the Type 2 Tobit model). The first equation, typically referred to as the "selection" equation, relates the binary selection rule to a set of regressors. The

second equation, referred to as the "outcome" equation, relates a continuous dependent variable, which is only observed when the selection variable is 1, to a set of possibly different regressors.

We express mathematically with the following model:

$$\begin{aligned} D &= \mathbf{1}\{Z - V \geq 0\} \\ Y &= DY^* = D \cdot (\theta_0 + U) \end{aligned} \tag{2.1}$$

Where $\theta_0 \in \mathbf{R}$ is the unknown parameter of interest, Z is the observed instrumental variable, and U, V are unobserved disturbances, which are independent of the instrument, but not necessarily independent of each other. The observed dependent variable D in the selection equation is binary, with $\mathbf{1}\{\cdot\}$ denoting the usual indicator function, and the dependent variable of the outcome equation, Y^*, is only observed when $D = 1$.

The above model is in one sense a condensed version what is often estimated in practice. The standard setup usually includes additional covariates, denoted by the observed random vector X in the second equation, where Y^* would be expressed as

$$Y^* = \theta_0 + X'\beta_0 + U$$

in which case Z would also be a vector whose dimension would usually exceed that of the dimension of X, and β_0 would also be a parameter to conduct inference on- see, e.g. Ahn and Powell (1993)

Our focus is on the condensed model and θ_0 only, for the following reasons. First, θ_0 is the parameter of interest in much of the treatment effects literature as it relates to the average treatment effect- see, e.g. Heckman (1990) and Andrews and Schafgans (1998). As discussed there the economic interpretation of an sample selection model makes inference on the intercept particularly important. It is required for the evaluation of the *wage gap* between unionized and nonunionized workers or between two different socioeconomic groups,- see, e.g. Oaxaca (1973), Smith and Welch (1986), Baker, Benjamin, Cegep, and Grant (1995). In the program evaluation literature the intercept permits evaluation of the net benefit of a social program by permitting comparisons of the actual outcome of participants with the expected outcome had they not chosen to participate..

Second, θ_0 is the parameter for which the difficulty in conducting inference can vary with the degree of selection, measured by the correlation between U and V. This is generally not the case for inference on β_0 for which inference on can be handled by existing methods.

What complicates inference on for θ_0 is that how well one can estimate θ_0 depends on the type of selection in the model, something which is unknown to the econometrician. For example, if the selection in the model is on *observables only*, which corresponds to U, V being uncorrelated with each other, than θ_0 can be consistently estimated at the standard parametric rate by, for example OLS or WLS only using the observations where $D = 1$. However, both OLS and WLS will be *inconsistent* if there is any amount of *selection on unobservables*. An alternative estimator would be to take into account selection on unobservables. One such estimator is proposed in Heckman (1990) and Andrews and Schafgans (1998). We propose a different one in this paper that will be the basis of our inference procedure.

Neither the Andrews and Schafgans (1998) (AS) estimator nor the new estimator (KN) we propose will have standard asymptotic properties (i.e parametric rates of convergence, limiting Gaussian distributions). These nonstandard properties will continue to hold even in the case when selection on observables only. The comparison of both the AS and KN estimators to the standard OLS and WLS estimators represents the classical robustness-efficiency tradeoff; OLS,WLS is not robust to selection on unobservables, but is more efficient than AS or KN if selection is on observables only.

To introduce an inference procedure that allows for both types of selection we consider the behavior of the KN estimator under drifting parameter sequences. For the problem at hand, one way to interpret these sequences would the correlation between U and V converging to 0, so that in the limit, the selection is on observables only.

To facilitate this discussion in the remainder of this section, we will distinguish between realizations of the random variables from a random sample and the random variables themselves. Our notation will be conventional in the sense that lower case letters with a subscript i will denote realizations from a random sample of n observations, and capitalized letters will denote the random variables themselves. So for example, in the above base model described, d_i, z_i, v_i, y_i, u_i will denote realizations of draws from the random variables D, Z, V, Y, U.

One of the main complications for estimation and inference procedure in this sample selection model is the unknown joint distribution of U and V. In this case, one may be inclined to pre-test for the correlation between the error terms in the two equations, and if it becomes clear that the error terms are uncorrelated, one may use the mean of the outcome whenever the dummy D is not equal to zero, as an estimate for θ_0. By the standard CLT, this mean will converge to expectation at a parametric rate. However, if one establishes

that U and V are correlated, than the full distribution of U and V needs to be explored and thus the estimator for θ_0 may need to employ an estimated unknown function leading to a slow rate of its convergence.

As we will show, a cause of behavior of the estimator is the tail structure of the distribution of U and V. It turns out that we can find two joint distributions of U and V which will be arbitrarily close to each other, yet the corresponding estimator for the parameter of interest θ_0 may have drastically different performance both in the rate of convergence and in the the asymptotic distribution. In practical terms this implies that a small amount of contamination in the data leading to a small correlation between U and V may have a substantial impact on the properties of the estimator for the parameter of interest.

We focus our discussion by analyzing the estimators arising in the two cases: when U and V are correlated and when they are not and then we design the procedure that bridges the gap between the two distributions.

Before starting the formal analysis we present the general assumptions that we impose on the structure of the distribution of error terms and the covariates.

ASSUMPTION 1 *(i) Z has a full support on $\mathbb{R}$ with the density $f_Z(\cdot)$ that is absolutely continuous and such that $1/f_Z(\cdot)$ is absolutely integrable on any bounded subset of $\mathbb{R}$.*

(ii) U and V have absolutely continuous strictly positive joint density supported on $\mathbb{R} \times \mathbb{R}$ such that $(U, V) \perp Z$ and $E\left[|U|^2 \,|\, V = v\right] < \infty$ uniformly over $v \in \mathbb{R}$.

(iii) The conditional density $f_{U|V}(\cdot \,|\, v)$ is well defined for each $v \in \mathbb{R}$, it is bounded for each v.

First, we establish the general identification result for the parameter of interest. We note that our only normalization is $E[U] = 0$. In this case the expectation of the "combined" error term U, D in the main equation of the selection model is not equal to zero. Although no information is available regarding the structure of correlation between U and V, the marginal distribution of V may be recovered from the selection equation. But this is not informative for the conditional distribution of U given V.

In this case the identification argument works only in the limit. In fact, we note that

$$E\left[Y \,|\, D = 1, Z = z\right] = \theta_0 + E\left[U \,|\, V \leq z\right]$$

Alternatively, we can write

$$\theta_0 = \frac{E\left[Y \mid Z = z\right]}{P(z)} - E\left[U \mid V \leq z\right],$$

where $P(z) = E[D|Z = z]$. Then given the assumption that support of Z is large, we can see that $\lim_{z \to +\infty} P(z) = 1$ and $\lim_{z \to \infty} E\left[U \mid V \leq z\right] = E[U] = 0$, therefore

$$\theta = \lim_{z \to +\infty} E\left[Y \mid Z = z\right].$$

We note that this expresses the parameter of interest in terms of the observable conditional expectation $E[Y|Z]$. Thus, this demonstrates the identification of this parameter under Assumption 1 which does not require the knowledge of any features of the joint distribution of (U, V). However, without further assumptions the identification is based on the limiting values of the "instrument" Z. This is why parameters identified in this manner are frequently referred to as *identified at infinity*.

Heckman (1990) and Andrews and Schafgans (1998) develop semiparametric inference procedures for the intercept parameter in the selection model.

Here we will consider inference based on a closed form estimator that has a similar structure to the estimator considered by Lewbel (1998) who studied the estimation of the intercept of the binary choice model under mean restriction imposed on the error term. We work with this estimator because its closed form facilitates exploring asymptotic properties under varying conditions.

As will be shown, while this estimator is consistent over large classes of distributions of error terms, it will have a rate of convergence that discontinuously changes with tail behavior assumptions on the unobservables and the instrument. This will complicate inference in several ways. For example, as we will show, it will make the construction of pivotal statistics impossible. Furthermore, it will invalidate other approaches of constructing confidence sets such as the bootstrap.

As an alternative, we propose the idea of *locally uniform inference* that will be based on drifting parameter asymptotics. We find that with an appropriately chosen drifting sequence, the resulting estimator will have an asymptotic distribution which enables valid inference methods.

2.1 Conditionally exogenous selection

We begin our analysis with a model based on "selection on observables". In this model the mean of the error in the main equation is zero conditional on the error term in the selection equation: $E[U \mid V] = 0$. With the assumed independence of the "instrument" Z from the error terms, this also means that the mean of the error in the main equation is zero uniformly over the values of Z. We note that in this case we can directly use the system of equations of interest to show identification. In particular, we note that the mean independence condition implies that $E[U \mid V \leq z] = 0$, if the corresponding conditional density is well-defined. Then we also note that

$$E[U\,D \mid Z = z] = E[U \mid V \leq z].$$

Then we can write

$$E[Y \mid D = 1,\, Z = z] = \theta_0 + E[U \mid D = 1,\, Z = z] = \theta.$$

We note that conditioning on Z in this case is informative because even though the first moment of U conditional on V does not vary with V, the second moment may. As a result, conditioning on Z may be used, for instance, to account for heteroskedasticity. We then can re-cast the identifying conditional moment for θ_0 as

$$\theta_0 = E\left[\frac{Y}{P(Z)} \,\middle|\, Z = z\right]. \tag{2.2}$$

where $P(Z) \equiv E[D|Z]$ denotes the "propensity score". The structure of the estimator as a conditional moment of variable $Y/P(Z)$ allows us to accumulate the information over Z and the resulting estimator will not be affected by the observations where the propensity score takes values close to zero or one.

In the case where the error term in the main equation is mean independent from the error term in the selection equation, the estimator for the parameter(s) of the first equation converges at the parametric rate.

Although the estimator (2.2) provides a closed-form expression for the parameter of interest, this estimator is not robust to deviations from the "selection on observables" assumption. In case where the errors are not mean independent, the estimator will be biased and this bias cannot be estimated at a sufficiently fast rate.

Another purpose of the alternative representation below is to link the case where the error term in the main equation is mean independent from the error term in the second equation

to the case where the two error terms are correlated. In particular, we first note that

$$E\left[Y \mid Z = z\right] = \theta_0 \, P(z)$$

can be rewritten as

$$\theta_0 \, f_V(z) = \frac{\partial E\left[Y \mid Z = z\right]}{\partial z},$$

where the derivatives are well-defined under Assumption 1. Therefore

$$\theta_0 = \frac{\frac{\partial E[Y \mid Z = z]}{\partial z}}{f_V(z)}.$$

2.2 A uniformly consistent estimator for the intercept in the sample selection model

Now suppose that the only assumption that is imposed on the error terms is that $E[U] = 0$. As we previously established, this assumption is sufficient to identify the intercept in the main equation under the full support assumption. The intercept can be expressed as

$$\theta_0 = \lim_{z \to +\infty} E\left[Y \mid Z = z\right].$$

We note that by the dominated convergence theorem $\lim_{z \to -\infty} E\left[Y \mid Z = z\right] = 0$. Since working with pointwise limits of functions is often not convenient, we propose the following transformation that allows us to express the parameter of interest directly from the primitives of the model:

$$\theta_0 = \lim_{z \to \infty} \int_{-z}^{z} \frac{\partial E\left[Y \mid Z = z\right]}{\partial z} \, dz.$$

Taking the limit, we find that the parameter of interest can be represented as an improper integral

$$\theta_0 = \int_{-\infty}^{+\infty} \frac{\partial E\left[Y \mid Z = z\right]}{\partial z} \, dz$$

We can re-arrange this equation using Fubbini's theorem, and make the estimator take a form similar to that where the error term in the main equation is mean independent from the error term in the selection equation. Thus, we can obtain that under Assumption 1

$$\theta_0 = \int_{-\infty}^{+\infty} \frac{\partial E[Y \mid Z = z]}{\partial z} \frac{1}{f_Z(z)} \, f_Z(z) \, dz = E\left[\frac{\frac{\partial E[Y \mid Z]}{\partial z}}{f_Z(Z)}\right].$$

We note that this identification argument leads to a similar expression to that in Lewbel (1997), Lewbel (2007), Lewbel (1998).

Therefore, we can introduce the random variable $W = f_Z(Z)^{-1} \frac{\partial E[Y \mid Z]}{\partial z}$ and the estimator is constructed as a sample average of the draws of this random variable[2]:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i. \tag{2.3}$$

where w_i denotes realizations of W. This case clearly contrasts with the case where the error term in the main equation is mean independent from the error term in the selection equation and the estimator was written in a weighted form. We note that in both cases the variables forming the sum have a finite first moment. In particular, we note that $E[W] = \lim_{z \to \infty} E[Y \mid Z = z] < \infty$. However, the second moment of W itself may not exist. The convergence properties of the corresponding improper integral are determined by the tail behavior of the random variable $\frac{\frac{\partial E[Y \mid Z]}{\partial z}}{f_Z(Z)}$.

We note that under the i.i.d. assumption, we can apply Kolmogorov's strong law of large numbers and establish that

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i \xrightarrow{a.s.} 0,$$

where here we set the true parameter value of θ_0 to be 0.

Thus, the estimator $\widehat{\theta}$ possesses certain "stability" properties. As our results so far do not give any information regarding the characterization of the distribution of the constructed estimator, we may want to use some common method, such as the bootstrap to characterize its asymptotic distribution. However, as the results in the next section demonstrate, the traditional non-adaptive bootstrap fails in this case.

2.3 Properties of the weighted estimator

Consider a bootstrap procedure which takes the i.i.d. sample of variables $W_i = \frac{\frac{\partial E[Y_i \mid Z_i]}{\partial z}}{f_Z(Z_i)}$. Then we take an array $\{\left(I_1^{(n)}, \ldots, I_n^{(n)}\right),\ n \geq 1\}$ that is independent from W_n and such that for each n the element $I_i^{(n)}$ is uniformly distributed on $\{1, \ldots, n\}$. Then the bootstrap sample of size n is generated as $W_i^* = W_{I_i^{(n)}}$.

[2]Note that this estimator is generally infeasible as it is often the case that the density function $f_Z(Z)$ is unknown and has to be estimated. We proceed for now assuming the density function is known, but discuss later in this paper further complications that can arise when it has to be estimated.

THEOREM 1 *Suppose that identification Assumption 1 holds and one uses the bootstrap sample W_i^* to characterize the distribution of estimator (2.3). The bootstrap distribution fails to converge to true limiting distribution of the partial sum.*

The proof of the theorem is in Appendix A. This theorem is a consequence of the failure of bootstrap noted in Athreya (1987). In particular, distributions that induce finite derivative of the regression function $E[Y \mid Z = z]$ with unbounded support that lead to the non-existence of the integral $\int (1/f_Z(z))\, dz$ would exhibit this behavior.

It turns out that the failure of the bootstrap for the inverse density weighted estimator is not a particular property to that estimator. Actually, unless we impose additional assumptions, *any* uniformly consistent estimator for the intercept parameter will necessarily exhibit non-uniform behavior in terms of its convergence rate and the structure of its asymptotic distribution. We provide a general theory in Appendix A.

Later in this subsection we will be able to characterize the actual limit of the bootstrap distribution under additional structural assumptions regularizing the tail behavior of the instrument.

Given the failure of the bootstrap in this setting [3] one may consider other inference procedures employed in the literature. One such example is based on using pivotal inference. Of interest frequently is the behavior of the t-statistic corresponding to parameter $\widehat{\theta}$. In fact, this approach was proposed in Andrews and Schafagans(1998) for the selection model and Khan and Tamer (2010) as a method for analysis of parameters "identified at infinity". See Hill and Chaudhuri (2012) for another example of this approach, as well as Ma and Wang (forthcoming), Heiler and Kazak (2019) for very recent examples. In all of these papers the inference approach can be considered as "robust" in the sense that it permits valid inference across a class of bivariate distributions. However, validity for some of these examples is based on certain tail conditions which ensured a Lindeberg type condition was satisfied.

Our next result shows that without such tail conditions, the estimator (2.3), which is con-

[3] Inconsistency of the bootstrap in other settings when observations are heavy tailed is shown in Athreya (1987), Politis, Romano, and Wolf (1999) and Romano and Wolf (1999). When second moments of the observations exist, the bootstrap will be consistent. Similar results hold for the t-statistic. Interestingly, even if consistent, bootstrap draws of the t-statistic will not result in any sort of "refinement". For the bootstrap resulting in refinement, existence of third moment of W_i is both necessary and sufficient- see Bloznelis and Putter (2003). However, recent work in Müller (2017) shows an alternative method to achieve refinement without third moments existing, as long as the second moment does.

sistent uniformly over the distributions satisfying Assumption 1, is not compatible with pivotal inference.

THEOREM 2 *Suppose that Assumption 1 holds and* $E[U] = 0$. *Then the empirical distribution of*

$$\widehat{T}_\theta = \frac{\frac{1}{n}\sum_{i=1}^{n} w_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i^2}}$$

is non-pivotal. In other words, for any $\delta > 0$ *there exist two distributions of* (U, V, Z) *denoted* $F^1_{U,V,Z}$ *and* $F^2_{U,V,Z}$ *satisfying Assumption 1 such that*

$$Pr\left(\widehat{T}_\theta \leq t\right) \overset{F^k_{U,V,Z}}{\longrightarrow} F_k(t), \;\; k = 1, 2$$

and

$$\sup_{t \in R} |F_1(t) - F_2(t)| > \delta.$$

In light of these negative results for both the bootstrap and the $t-$statistic, the question remains as to what are the origins of this behavior of the estimator and whether there are ways of characterizing its actual asymptotic behavior. As it turns out, a main reason for this behavior is the non-existence of the second moments. When the second moment of the random variable does not exist, the $\sqrt{n}$-normalized centered sample average will not converge in distribution.

A natural direction to proceed in this case is to consider trimming W to obtain random variables that have a finite second moment for each n. Such a solution has been suggested in Andrews and Schafgans (2001) where it was assumed that the tail behavior of the distribution of W is given. However, in many practical settings, the tail behavior of the unobserved component of the model is unknown. Then the tail index of this unknown distribution becomes an ancillary parameter *that itself has to be estimated*. Original estimators of the tail index can be found in Hill (1975) and Pickands (1975), and for a more recent development see Müller and Wang (2017). The convergence rate of the estimator of this parameter may be extremely slow and thus its behavior will dominate the behavior of the remaining components of the trimmed estimator, see e.g. McCulloch (1986), McCulloch (1997).

This indicates that the estimators based on the oracle properties of the distribution, such as the estimator based on trimming are infeasible or they may invoke a slow adaptive rate that incorporates the fact that the tail behavior should itself be estimated.

The absence of convergence in distribution of $\sqrt{n}$-normalized centered sample averages leads to the general absence of convergence in distribution for pivotized statistics. If $\mathcal{F}$ is the class of distributions satisfying Assumption 1 then in general the distribution of the t-statistic does not converge uniformly to normal distribution. In other words even if we found a candiate $F' \in \mathcal{F}$ such that $\Pr\left(\widehat{T}_\theta \leq t\right) \to \Phi(t)$ for each $t \in \mathbb{R}$ then for any $\delta > 0$:

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \sup_{\|F-F'\|_\infty<\delta,\, F\in\mathcal{F}} \left|\Pr\left(\widehat{T}_\theta \leq t\right) - \Phi(t)\right| \not\to 0,$$

where $\Phi(\cdot)$ is the standard normal cdf.

Theorem 2 has important implications for conducting inference in this model. In previous work, Andrews and Schafgans (1998) showed that a studentized estimator in the selection model could indeed be used to conduct valid inference for a wide class of distributions of observables and unobservables, satisfying certain relative tail restrictions. Our above theorem compliments that result by demonstrating the necessity of the tail conditions for this to be a valid inference method.

As was the case in Theorem 1, it is also the case here that the conclusion in Theorem 2 for the inverse density weighted estimator is not just property to that estimator. Without additional assumptions, *any* uniformly consistent estimator for the intercept parameter will necessarily exhibit non-uniform behavior in terms of its convergence rate and the structure of its asymptotic distribution. We provide a general theory in Appendix A.

3 Locally uniform inference for the sample selection model

We noted that in case where the model is compatible with selection on observables (the error terms are mean-independent in the main and selection equations) $\widehat{\theta}_0$ is a consistent estimator for the parameter of interest in the main equation which converges at the parametric rate to the true parameter regardless of the tail behavior of the covariate density $f_Z(\cdot)$. On the other hand, for any distribution of error terms that fails to assure that the error term in the main equation is mean independent of the error term in the selection equation, we need to use estimator $\widehat{\theta}$ that uses a simple unweighted average of w_i. We recall that $\widehat{\theta}_0$ converges at a parametric $\sqrt{n}$ rate while $\widehat{\theta}$ converges at a slow rate $\ll \sqrt{n}$.

In this context it may seem attractive to use some form of pre-testing to establish whether the given model exhibits selection on unobservables. This naturally leads us to the estimator that has the structure of the Hodges estimator:

$$\widehat{\theta}^H = \begin{cases} \widehat{\theta}, & \text{if } |\widehat{\theta} - \widehat{\theta}_0| > C/\sqrt{n}, \\ \widehat{\theta}_0, & \text{if } |\widehat{\theta} - \widehat{\theta}_0| \leq C/\sqrt{n}. \end{cases}$$

This estimator however, exhibits a non-uniform behavior. In fact, for any distribution of error terms that is compatible with selection on observables we can find another distribution that will be arbitrarily close to the original distribution in the L_2 norm defined by the probability measure associated with random variable Z, but it will not be compatible with mean independence. The rate of convergence of the consistent estimator for θ under that distribution may be as slow as $\log n^{\kappa}$ for some $\kappa > 0$. Moreover, the structure of the asymptotic distribution of the consistent estimator for these two close distributions of error terms is dramatically different: while it is normal in the model with selection on observables, it is may be represented by the distribution of a stable Lévy process in the model with selection on unobservables.

It is important to note that the estimator that is based on unweighted averaging over the realizations of W is consistent in both the case of selection on observables and the selection on unobservables. The estimator that is based on the weighted average is inconsistent where the error terms in two equations are correlated. As we noticed it before, in the case where the density of the instrument Z has thin tails, the rate of convergence and the asymptotic distribution of the estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} w_i$ relies on the tail behavior of this density. An estimation procedure that is adaptive both to the convergence rate and the shape of the asymptotic distribution is hard to construct, especially of the distribution of Z has a small tail probability. On the other hand, the procedure that is based on the weighted averages of W (leading to estimator $\widehat{\theta}_0$) in general requires bias-correction. Bias correction in this case will again require the analysis of the tail behavior of the inverse density of the instrument and will lead to the same difficulties as adaptive inference for the unweighted estimator $\widehat{\theta}$.

An approach to bridge the gap between these two asymptotics is to consider a family of distributions of instruments Z that are compatible with finite (constant) second moments of random variables W. Provided that we assume that the data are i.i.d. we can apply the standard Central Limit Theorem to establish asymptotic normality. Then we can consider a distribution of instruments "local to" the distribution that has finite second moments. Formally, this means that we find a heavy tail distribution that is contiguous to the distribution that delivers the finite second moments. Specifically, what we have in mind is that the

Hellinger and L_2 distance between these two distributions converges to zero as the sample size increases. This approach may be attractive for two reasons. First, we approximate the distribution of W in the area of the support of Z that has the highest probability mass with the distribution that has finite second moments. Thus, it delivers the parametric convergence rate for the unweighted sample mean characterizing $\widehat{\theta}$. Second, given that we control the choice of contiguous heavy-tail distributions we can choose the family of continguous distributions to be sufficiently simple and thus estimation of the asymptotic distribution of $\widehat{\theta}$ will not require estimation of the tail behavior of W.

Provided that our estimator is fully characterized by the joint distribution of (Y, Z) which then determines the random variable $W = \frac{\frac{\partial E[Y \mid Z]}{\partial z}}{f_Z(Z)} Y$, we can then concentrate on analyzing this distribution.

To do so, we first introduce the class of distributions of (Y, Z) that are compatible with asymptotic normality of estimator $\widehat{\theta}$. This is class of distributions which must contain the distribution of (Y, Z) when a particular parameter that is "identified at infinity" is claimed to converge at a parametric rate to an asymptotic normal distribution.

DEFINITION 1 *Suppose that the joint distribution* (Y, Z)*, denoted* $F_{YZ}(\cdot,\cdot)$ *is defined by model (2.1). where the random elements satisfy Assumption 1. Define the class of distributions*

$$\mathcal{N} = \left\{ F_{YZ}(\cdot,\cdot) \ : \ E\left[\left(\frac{\partial E\left[Y \mid Z\right]}{\partial z} \Big/ f_Z(Z) \right)^2 \right] < \infty \right\}.$$

Also define the class

$$\mathcal{N}_2 = \left\{ F_{YZ}(\cdot,\cdot) \ : \ \textit{argsup} \left\{ \beta \in (0,+\infty), \ \ E\left[\left(\frac{Y \, \frac{\partial E[Y \mid Z]}{\partial z}}{f_Z(Z)} \right)^{\beta} \right] < \infty \right\} = 2 \right\}.$$

The defined class of distributions $\mathcal{N}$ is fundamental because it delivers the validity of the Central Limit Theorem. The class $\mathcal{N}_2$ is on the boundary of $\mathcal{N}$ in the sense that distributions in $\mathcal{N}$ can be compatible with the second and higher finite moments of W while for the distributions in $\mathcal{N}_2$, the second moment is the highest moment that exists for W.

LEMMA 1 *Suppose that Assumption 1 is satisfied and* $Pr\left(\left| \frac{\partial E[Y \mid Z]}{\partial z} \Big/ f_Z(Z) \right| > w \right)$ *is regularly varying at infinity with tail index* $-(1+\gamma)$*. Then, whenever* $\gamma \geq 1$*, the distribution* $F_{YZ}(\cdot,\cdot) \in \mathcal{N}$*. Moreover, if* $\gamma = 1$ *then* $F_{YZ}(\cdot,\cdot) \in \mathcal{N}_2$*.*

Now suppose that $F_{YZ}(\cdot,\cdot) \in \mathcal{N}_2$. We consider the distribution of W, denoted $F_W(\cdot)$ implied by such a distribution $F_{YZ}(\cdot,\cdot)$. By definition of class $\mathcal{N}_2$, we note that $\int w^2 f_W(w)\,dw < \infty$ while the integral $\int w^\beta f_W(w)\,dw$ diverges for any $\beta > 2$. One practical example where the distribution of W belongs to $\mathcal{N}_2$ is the case where $E[U|V] = 0$.

LEMMA 2 *Suppose that Assumption 1 is satisfied, Z has finite second moments and $E[U|V] = 0$ then $Pr\left(\left|\frac{\partial E[Y\,|\,Z]}{\partial z}\,/\,f_Z(Z)\right| > w\right)$ is regularly varying at infinity with tail index -2. In other words, the case where the errors are uncorrelated generates the distribution of W for which $F_{YZ}(\cdot,\cdot) \in \mathcal{N}_2$.*

The idea behind the construction of a heavy tailed distribution local to each element of $\mathcal{N}_2$ will be the following. Note that $\int_{(\cdot)} |w|^{1+c}\text{sign}(w) f_W(w)\,dw$ is a measure defined on Borel subsets of the real line for each $c \in [0,\,1)$.[4] Our further arguments will be based on the following considerations. As a "first-order approximation" we assume that distribution of W has a finite second moment. Under this approximation we can characterize the part of the asymptotic distribution around $E[W]$. Then we consider the "second-order approximation" which is taken to be an additional component that vanishes pointwise as the sample size increases, much of which characterizes the extreme tail behavior of the distribution of W.

Then for each $F_{YZ} \in \mathcal{N}_2$ the corresponding density $f_W(\cdot)$ will be used to construct the "first order" approximation to the asymptotic distribution. After an appropriate normalization, $|\cdot|^c f_w(|\cdot|^{1+c})$ is a valid density, but given that $F_{YZ} \in \mathcal{N}_2$, this density will have heavy tails and we will use the corresponding distribution to approximate the tail behavior.

The distribution $F_W(\cdot)$ has tail index 2, while the distribution with density $F_w(\text{sign}(\cdot)|\cdot|^{1/(1+c)})$ has tail index $2/(1+c)$. Then if $c = 0$, then the latter distribution has exactly two finite first moments while if $c = 1$ this distribution has only finite first moment. Now we characterize the local asymptotics for the partial sum characterizing the estimator of interest $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} w_i$. Let

$$S_W(w) = \frac{1}{2}F_W\left(\text{sign}(w)|w|^{1/(1+c)}\right) + \frac{1}{2}\left(1 - F_W\left(\text{sign}(-w)|w|^{1/(1+c)}\right)\right)$$

and $s_W(\cdot)$ be the corresponding density. For $\rho_n = n^{c/(1+c)}$ consider the local distribution

[4]We note that this constructed measure may exhibit non-regular behavior at the point $W = 0$ where function $|W|^{1+c}$ is not differentiable. We alleviate this problem by employing a technique referred to as the one-point uncompactification, which is based on re-defining the topology on $\mathbb{R}$ that avoids intersections of the elements of this topology with the origin.

for W using the density $f_W(\cdot)$ with the finite second moment up to normalization as:

$$f_W^c(w) = f_W(w) + \frac{h_c}{\rho_n}\left(s_W(w) - f_W(w)\right), \tag{3.1}$$

where $0 < h_c \leq 1$ and $h_0 = 0$ and is continuous at $c = 0$. Note that this requirement is imposed on h_c to ensure that $f_W^c(\cdot)$ is a valid density and that it converges uniformly in w and n to $f_W(\cdot)$ when c is in the neighborhood of zero.

THEOREM 3 *If the random variable W is distributed according to (3.1) then we can establish that the limiting distributions of partial sums has the following limit:*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i \xrightarrow{d} \sigma B(1) + h_c L_{2/(1+c)}(1),$$

where $B(\cdot)$ is the standard Brownian motion and $L_{2/(1+c)}(\cdot)$ is the $2/(1+c)$-stable Lévy process with $c \in [0,1]$. In other words, the asymptotic distribution is a mixture of the normal distribution and the stable distribution.

Thus, the advantage of this constructed local asymptotics is that, for one, the convergence to its asymptotic distribution will occur at parametric rate. As a result, there is no need to design an estimation procedure that will adapt both to the convergence rate and to the asymptotic distribution (as is necessary in case of standard heavy tail asymptotics). Second, our structure has a clear interpretation where the normal component characterizes the asymptotic distribution close to the expected value of W while the Lévy process component is responsible for the tail behavior of that asymptotic distribution.

The tail behavior of the asymptotic distribution as c varies from 0 to 1 changes from the case where this distribution has a finite second moment and thus asymptotically normal, to the case where this distribution only has a finite first moment and no higher moments. The object of interest will be the quality of the approximation of the asymptotic distribution uniformly over $c \in [0,\, 1)$. The following result establishes the uniform normality of the asymptotic distribution for the t-statistic constructed for $\widehat{\theta}$.

THEOREM 4 *Suppose that Assumption 1 holds and $E[W] = 0$. Let $F_T^c(w)$ be the distribution of random variable constructed as*

$$T^c = \frac{\sigma B(1) + h_c L_{2/(1+c)}(1)}{\sqrt{\sigma^2 + h_c^2 L^+_{1/(1+c)}(1)}},$$

where $L^{+}_{1/(1+c)}(\cdot)$ *is the* $1/(1+c)$*-stable Lévy process defined on* $\mathbb{R}_+ \setminus \{0\}$*. Then the distribution of random variable* T^c *uniformly approximates the distribution of the t-statistic*

$$\widehat{T}^c = \frac{\frac{1}{n}\sum_{i=1}^{n} w_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i^2}},$$

such that for some $\delta > 0$

$$\lim_{n\to\infty} \sup_{c\in[0,1-\delta]} \sup_{t\in\mathbb{R}} \left| F_{\widehat{T}^c}(t) - Pr(T^c \leq t) \right| = 0.$$

It is useful to point out the similarities and diffferences between Theorem 3 and 4 and existing results in the econometrics literature. On the one hand we note similarities between our results and on the inference in the autoregressive model with a near unit root and the models with weak instruments. The similarity of these models to ours is in the discontinuity of the distribution of the estimator for the parameter of interest with respect to the data generating process. In the near unit root case, the presence of the unit root discontinuously changes the asymptotic distribution from the normal to the non-standard Dickey-Fuller distribution. In the weak instrument case, the distribution of parameter changes from normal to Cauchy in case of the full irrelevance of an instrument.

But our model and results differ in important ways. The distribution of the estimated parameter (and its convergence rate) changes in response to any change in the parameters of the data-generating process. Therefore, it will be impossible to find a unique local parametrization of the model that makes its asymptotic distribution change continuously with respect to the model parameters. Consequently, in the choice of local parameterization we need to define, first, the "focal" data generating process. Second, given that focal data generating process we define the parametrization for local data generating processes (in the small neighborhood of the focal data generating process) that converges to that data generating process.

3.1 Approaches to inference

As we mentioned previously, one of the difficult components of inference for the parameter of interest is in the construction of its distribution theory that requires the estimation of the tail index of its domain of attraction. This index determines both the rate of convergence and the shape of the confidence set for the parameter of interest. Politis, Romano, and Wolf (1999) provide a subsampling approach that allows one to construct a valid confidence set for studentized parameter of interest.

Consider subsampling with subsample block size b and let $\widehat{\theta}_{n,b,i}$ be the parameter estimate in the i-th subsample and $\widehat{\sigma}_{n,b,i}$ be the standard deviation computed in that subsample.

THEOREM 5 *(***Politis, Romano, and Wolf (1999)***) Suppose that the tail index $1+\gamma$ is fixed. The subsampling approximation $L^*_{n,b} = \frac{1}{N_n}\sum_{i=1}^{N_n} \mathbf{1}\left\{\sqrt{b}\left(\widehat{\theta}_{n,b,i} - \widehat{\theta}\right)/\widehat{\sigma}_{n,b,i} \leq x\right\}$ converges uniformly to the distribution of variable U/V if $b \to \infty$ and $b/n \to 0$ as $n \to \infty$, where U is the domain of stable attraction of partial sums of W and V is the domain of stable attraction of partial sums of W^2.*

This is a very useful result allowing to construct approximation for the asymptotic distribution of a pivotized variable without requiring the estimation of the tail index. We note however that the quality of subsampling approximation will deteriorate when the tail index $1+\gamma$ approaches 1. The reason is that the standard deviation will be converging to the stable law with tail index $(1+\gamma)/2$ (meaning that the corresponding distribution does not have a mean) and thus the constructed statistic will be highly variable across the subsamples. This may require a more conservative inference method. The method that we propose below allows one to construct such conservative bounds under local asymptotics.

THEOREM 6 *Consider local asymptotics with a sequence of distributions (3.1). The subsampling approximation $L^*_{n,b} = \frac{1}{N_n}\sum_{i=1}^{N_n} \mathbf{1}\left\{\sqrt{b}\left(\widehat{\theta}_{n,b,i} - \widehat{\theta}\right)/\sigma \leq x\right\}$ converges uniformly to standard normal distribution if $b \to \infty$ and $n/log\, b \to \infty$ as $n \to \infty$.*

Thus, under the local asymptotics, the subsampling distribution converges to a pivotal normal distribution. The reason for that is that the component of the limiting distribution which is responsible for the "outliers" is vanishing faster than the subsample size. The distribution then converges to the non-vanishing normal limit. The subsampling is used to estimate the correct variance σ^2 of the normal component of the limiting distribution mixture.

The structure of the local distribution gives an idea for non-conservative and conservative inference based on the extracted normal distribution quantiles. The non-conservative inference will correspond to using the extracted normal quantiles for inference. The conservative inference will suggest using the "worst-case scenario" distribution for the outliers meaning that we need to take $h_c = 1$ and $L_c(\cdot)$ to be the standard stable Levy process with $c = 1$. The resulting conservative confidence set will be the sum of the normal confidence set and the confidence set constructed from adding a standard Levy process scaled by σ.

4 Simulation Results

In this section we finite sample properties of the estimation and inference procedures we propose. To do so we simulate data from the sample selection models, and we report summary statistics intended to characterize the finite sample performance of both the existing and new estimators whose asymptotic properties we established.

Simulation results are for sample sizes of 100, 200, 400 and 800 observations where we report mean bias,median bias, and RMSE and median absolute deviation (MAD) from 3000 replications. Results for the proposed inverse weight weighted (IVW) estimator of the intercept in a sample selection model are reported in tables 1-4 [5].

For our design in the sample selection model we assumed the bivariate distribution was standard bivariate normal. The selection equation has a single instrument for which we considered two designs- one where it was distributed standard normal and the other where it was distributed standard cauchy. To allow for fixed and drifting parameter sequences we adjusted the correlation between the two error terms in the selection model. For fixed parameters we simulated using 4 distinct values of this correlation- 0,0.5,0.75 and 1. For drifting parameters we divided these 4 different constants by the square root of the sample size.

As results in Tables 1-4 indicate, our finite sample results generally agree with our asymptotic theory. As we see the RMSE and MSE increase with the sample size when scaled by the square root of the sample size, indicating the estimator does not converge at the parametric rate, if at all. In one sense this is not too surprising as no trimming is used.

We also explore the sampling distribution of the estimator. We do this by creating histograms for the estimates attained from the 3000 replications. The graphs are in Figure 1 where the histograms report values of the estimator divided by the square root of the sample size. We set axis bounds as follows: for the horizontal axis the bounds where $\pm$ 5 times the standard deviation of the estimator, divided by the square root of the sample size. The vertical axis bounds were 0 and 3 times the standard deviation the estimator value, divided by the square root of the sample size. Specifically, the distribution of the estimator has a Gaussian component but also exhibits noticeably fat tails. Furthermore as the correlation between the two errors gets further away from 0, the distribution of the estimator has a noticeably skewed distribution, most notably when the instrument is Gaussian. This

[5]The tables report the RMSE and MAD multiplied by the square root of the sample size to help us indicate if the estimator converges at the parametric rate.

skewness is less pronounced when the instrument has a cauchy distribution.

To compare the finite sample procedures of other estimators we also provide histograms for different designs, in Figures 2-3. These designs include different bivariate distributions of u, v, with marginals being normal, logistic or cauchy, with varying levels of correlation. These bivariate distributions were generated using the Gaussian copula. The other estimators we report histograms for are simple OLS, the Heckman 2-step estimator, the Andrews and Schafgans estimator, and what we refer to as the Bridge estimator, which is the inverse weight estimator under local asymptotics. To implement the Andrews Schafgans estimator we used the true propensity score and only observations where it exceed 0.95. Not surprisingly, as the graphs indicate, OLS is centered away from the truth when there is correlation between the two errors as it does not account for selection bias.

We also explore the finite sample properties of our new procedure as well as others from a hypothesis testing perspective. Table 5 reports size and power by listing acceptance and rejection probabilities using the t-test for various null hypotheses when the data is generated with the true intercept being 0. These probabilities are reported for OLS, Heckman 2-step, Andrews and Schafagans (where we tried two different propensity score cutoffs, 0.95, 0.99), and our procedure. For the case where $H_0 : \alpha_0 = 0$ the probabilities reported are those of accepting the null, whereas for the case $H_0 : \alpha_0 = 0.5$ the probabilities reported are rejection probabilities.

Again, the OLS procedure does as expected having correct size and power properties only when the correlation in errors is 0. Otherwise it results in severe under rejection of the null, though it correctly rejects the null $\alpha_0 = 0.5$, for all samples sizes and all correlations most, if not all of time. The Heckman procedure tends to have low size, especially as the correlation approached one, and its power is on the low side for sample sizes of 100, but otherwise correctly rejects the null of $\alpha_0 = 0.5$ most of the time. Still, in terms of both size and power, we anticipated a better performance as in this design of bivariate normal errors the parametric Heckman model is correctly specified. The Andrews Shafgans estimator does quite well in this design both in terms of size only accepting the correct null with probabilities quite different from 0.95 when the correlation between the errors gets close to 1. However, in terms of power it was quite low for sample sizes less than 800. This might suggest the need for a sample size dependent cutoff probability in the trimming used. The inverse weight estimator appears to accept the null $\alpha_0 = 0$ too infrequently, and this problem becomes worse as the sample size increases. However it gets the right power with samples sizes of 400 or higher. Here we attribute the poor size results due to the fact that

no trimming was employed.

Tables 6 and 7 explore the properties of the bootstrap for inference. Here we report the fraction of times (from 300 bootstrapped replications and 100 simulations) that true value lies in the 95%bootstrap interval. This is done for 4 estimators (OLS, Heckman 2-Step, Andrews and Schafagans, inverse weighting) and two designs of the bivariate error distribution (bivariate normal, and marginal cauchy with Gaussian copula). For the normal case each of the four procedures resulted in overly conservative inference for all sample sizes as the probabilities are equal to 1. This illustrates our points that 1) this is a difficult parameter to conduct correct inference and 2) the invalidity of the bootstrap. Interestingly, results improve when we consider the bivariate cauchy error distribution where the probabilities are generally too small, though it appears to be the least problematic for the IVW procedure. Nonetheless, even here we are still able to illustrate the poor performance of the bootstrap.

In summary, as the graphs and tables indicate, many of the conclusions from our limiting distribution theory are reflected in finite sample outcomes. For many designs the estimators converge very slowly, and the distributions are very nongaussian for all sample sizes. Most importantly, we have shown that standard inference procedures such as the $t-$ test or the bootstrap can perform very poorly in small samples.

5 Empirical Illustration

In this section we illustrate the use of our proposed inference methods by applying them to the well known Mroz (1987) labor supply data set. This data set was also used in Ahn and Powell (1993) and Newey, Powell, and Walker (1990) to compare parametric and semiparametric methods. However, in those papers the focus was on the slope coefficients of the outcome equation, whereas here we focus on the intercept term.

In the Mroz (1987) study, the sample consists of measurements on the characteristics of 753 married women (428 employed and 325 unemployed). The dependent variable in the outcome equation, the annual hours of work, is specified to depend upon the wage rate, household income less the woman's labor income, indicators for young and older children in the household, and the woman's age and years of education. Mroz's study also used the square of experience and various interaction terms as instrumental variables for the wage rate, and were also included in his probit analysis of employment status, resulting in 18 parameters to be estimated in the first equation. Ahn and Powell (1993) use the same conditioning variables in the first equation but only the original 10 variables in their first

stage kernel regression to attain estimators of the slope coefficients in the outcome (hours worked) equation.

Our approach here will be to use their estimates of these parameters combined with our density weighted estimator to estimate the intercept term. Specifically, we will treat the 6 slope coefficients in the outcome equation as known (using the values attained in Ahn and Powell (1993)) for the coefficients on log wage, nonwife income, young children, older children, age and education), and estimate the intercept term using our density weighted expression. Recall our expression involved estimating the density of the index from the selection equation. Following Ahn and Powell (1993), we use 10 conditioning variables, but in contrast, we estimate their coefficients by estimating a Probit model. With these estimated coefficients, we can construct estimated values of the index, to which we apply kernel density estimation, using a normal kernel function and cross validation for the bandwidth, to estimate the density function of the selection equation index. Following Newey, Powell, and Walker (1990) we treat previous labor market experience, measured in total years experience, as the excluded variable that is in the employment equation but not the outcome equation.

Our estimator of α_0 is based on the moment condition:

$$\alpha_0 + E[x_i'\beta_0] = E_Z\left[\frac{\frac{d}{dz}E[y|z]}{f_Z(z)}\right] \tag{5.1}$$

where $f_Z(z)$ denotes the density function of z_i and $\frac{d}{dz}E[y|z]$ denotes the derivative of the regression function of $E[y|z]$.

To estimate α_0, note the right hand side of the above equation can be estimated by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}'(z_i)/\hat{f}(z_i) \tag{5.2}$$

where $\hat{\mu}'(z_i)$ is a local linear estimator of the derivative of the regression function and $\hat{f}(z_i)$ is a kernel estimator of the density function.

so our estimator of α_0 is

$$\hat{\alpha} = \hat{\theta} - \frac{1}{n}\sum_{i=1}^{n} x_i'\hat{\beta} \tag{5.3}$$

Using the standard bootstrap we were able create a histogram for the standardized estimator as well as provide a quantile plot.

As a comparison, we estimated α_0 from the parametric Heckman model assuming bivariate normality of the unobserved disturbances. For the parametric estimator we also bootstrap to create a histogram of the standardized estimator as well as quantile plots. Histograms and quantile plots are after the Appendix in Figure 4.

The attained results are interesting, notably that contrast between conclusions drawn from the parametric and semi parametric approaches. The parametric point estimator for α_0 is three times larger in magnitude than the semi parametric point estimator, though both point estimates are positive. Exploring the bootstrapped confidence regions, the results from the two approaches are even more strikingly different. As the quantiles plots reveal, from the parametric approach the intercept is positive at all significant levels, whereas from the semi parametric quantile plot the intercept is not significantly different from 0 at most standard significance levels (0.025, 0.05,0.1). This demonstrates how sensitive the results can be to parametric assumptions.

6 Models with behavior similar to the sample selection model

While this paper has dealt exclusively with the difficulties in conducting uniform inference for parameters of interest in the sample selection model. the same problems and difficulties arise when conducting valid inference on parameters of interest in many other widely studied (from both a theoretical and empirical perspective) nonlinear models. Examples included discrete triangular systems and non triangular systems, such as the estimation of two player games often considered in industrial organization. We illustrate the relation to our results for the sample selection model here.

6.1 Discrete triangular models

Consider the triangular model driven by the unobserved variables y_1^* and y_2^* such that the observed variables $y_1 = \mathbf{1}\{y_1^* \geq 0\}$ and $y_2 = \mathbf{1}\{y_2^* \geq 0\}$ while

$$y_1^* = z_1'\gamma_0 + \alpha\, y_2 - u,$$

and

$$y_2^* = z_2'\delta_0 - v.$$

The object of our interest will be the interaction parameter α. See Vytlacil and Yildiz (2007) for a related system without the separability conditions imposed above. Without loss of generality, we fix the coefficients of linear indices and denote $x_1 = z_1'\gamma_0$ and $x_2 = z_2'\delta_0$. We assume that the underlying data generating process is driven by the distribution of random variables (X_1, X_2, U, V). Khan and Nekipelov (2013) have considered this model for the case where the error terms (U, V) are independent from the index covariates (X_1, X_2) and demonstrated that the parameter α in this model is identified provided the large support assumption imposed on the distribution of (X_1, X_2). Khan and Nekipelov (2013) show that the large support assumption is essential meaning that without further assumptions the boundedness of the covariate support leads to a loss of point identification of parameter α.

ASSUMPTION 2 *Suppose that*

(i) X_1 *and* X_2 *have a continuous distribution with full support on* $\mathbb{R}^2$ *(which is not contained in any proper one-dimensional linear subspace);*

(ii) (U, V) *are independent of* (X_1, X_2) *and have a continuously differentiable density with the full support on* $\mathbb{R}^2$.

Under Assumption 2, the parameter α can be identified as follows. First note that

$$\lim_{x_2 \to -\infty} P(Y_1 = 1 | X_1 = x_1, X_2 = x_2) = F_U(x_1)$$

and

$$\lim_{x_2 \to +\infty} P(Y_1 = 1 | X_1 = x_1, X_2 = x_2) = F_U(x_1 + \alpha).$$

Second, we transform the limits into the conditional expectations that lead to the expresions

$$F_U(x_1) = P(Y_1 = 1 | X_1 = x_1, X_2 = 0) - E\left[\frac{\frac{\partial}{\partial x_2} P(Y_1 = 1 | X_1, X_2 = x_2)}{f_{X_2|X_1}(x_2|X_1)} \,\middle|\, X_1 = x_1, X_2 \leq 0 \right],$$

$$F_U(x_1 + \alpha) = -P(Y_1 = 1 | X_1 = x_1, X_2 = 0) + E\left[\frac{\frac{\partial}{\partial x_2} P(Y_1 = 1 | X_1, X_2 = x_2)}{f_{X_2|X_1}(x_2|X_1)} \,\middle|\, X_1 = x_1, X_2 \geq 0 \right].$$

The first equation identifies the marginal distribution of the error in the first equation and the second equation identifies the interaction parameter of interest. We note that the moment functions employed in the identification contain the conditional density of the X_2 variable in the denominator. This means that, generally speaking, the moment function does not have a finite second moment which will lead to the non-uniform behavior of the estimator for the parameter α in the underlying distributions of the errors and covariates.

6.2 Static games of complete information

Another example where the structure of the identification argument has a similar flavor to that in the selection model is a 2-player discrete game with complete information (e.g. Bjorn and Vuong (1985) and Tamer (2003)).

A simple binary game of complete information is characterized by the players' deterministic payoffs, strategic interaction coefficients, and random payoff components u and v. There are two players $i = 1, 2$ and the action space of each player consists of two points $A_i = \{0, 1\}$ with the actions denoted $y_i \in A_i$. The payoff of player 1 from choosing action $y_1 = 1$ can be characterized as a function of player 2's action:

$$y_1^* = z_1'\gamma_0 + \alpha_1 y_2 - u,$$

and the payoff of player 2 from choosing action $y_2 = 1$ is characterized as

$$y_2^* = z_2'\delta_0 + \alpha_2 y_1 - v.$$

For convenience of analysis we change notation to $x_1 = z_1'\gamma_0$ and $x_2 = z_2'\delta_0$. We normalize the payoff from action $y_i = 0$ to zero and we assume that realizations of covariates X_1 and X_2 are commonly observed by the players along with realizations of the errors U and V, which are not observed by the econometrician and thus characterize the unobserved heterogeneity in the players' payoffs. Under this information structure the pure strategy of each player is the mapping from the observable variables into actions: $(u, v, x_1, x_2) \mapsto 0, 1$. A pair of pure strategies constitute a Nash equilibrium if they reflect the best responses to the rival's equilibrium actions. The observed equilibrium actions are described by random variables (from the viewpoint of the econometrician) characterized by a pair of binary equations:

$$\begin{aligned} Y_1 &= \mathbf{1}\{X_1 + \alpha_1 Y_2 - U > 0\}, \\ Y_2 &= \mathbf{1}\{X_2 + \alpha_2 Y_1 - V > 0\}, \end{aligned} \tag{6.1}$$

where errors U and V are correlated with each other with an unknown distribution. In particular, we are interested in determining when the strategic interaction parameters α_1, α_2 can or cannot be estimated at the parametric rate.

As noted in Tamer (2003), the system of simultaneous discrete response equations (6.1) has a fundamental problem of indeterminacy as it may have the regions where it has multiple solutions or no solutions at all. If we require the signs of α_1 and α_2 to be the same, then the region where multiple solutions can occur is that where the values of $|X_1|$ and $|X_2|$ are close to those of α_1 and α_2. The way to identify the parameters of interest α_1 and α_2 as

proposed in Tamer (2003) is to use the asymptotic regions where the solution is unique, thus forming a system of asymptotic equations:

$$\begin{aligned} F_U(x_1+\alpha_1) &= \lim_{x_2\to+\infty} P(Y_1=1\,|\,X_1,X_2), \\ F_V(x_2+\alpha_2) &= \lim_{x_1\to+\infty} P(Y_2=1\,|\,X_1,X_2). \end{aligned} \tag{6.2}$$

Provided that Assumption 2 holds, we can identify the parameters of interest through the explicit expressions

$$F_U(x_1+\alpha_1) = -P(Y_1=1\,|\,X_1=x_1,X_2=0) + E\left[\left.\frac{\frac{\partial}{\partial x_2}P(Y_1=1\,|\,x_1,X_2)}{f_{X_2|X_1}(X_2|x_1)}\right| X_1=x_1, X_2\geq 0\right],$$

$$F_V(x_2+\alpha_2) = -P(Y_2=1\,|\,X_1=0,X_2=x_2) + E\left[\left.\frac{\frac{\partial}{\partial x_1}P(Y_2=1\,|\,X_1,x_2)}{f_{X_1|X_2}(X_1|x_2)}\right| X_1\geq 0, X_2=x_2\right].$$

This expression demonstrates that the parameters of interest are "identified at infinity" in the same sense as the intercept in the sample selection model and the average treatment effect parameter. As a result, we can apply our previous results to demonstrate that any uniformly consistent estimator for these parameters (i.e. the one that does not rely on an assumption regarding a particular tail structure of the distribution (U,V)) will have the properties analogous to those of the uniformly consistent estimator for the intercept. In particular, the bootsrap will not deliver a consistent approximation for the asymptotic confidence sets, and the t-statistics will not converge to the pivotal distribution. We can however, provide valid inference methods in case where the distribution of the error terms belongs to a drifting sequence which converges to the distribution with particular tail properties as the sample becomes larger. In particular, we can use the case where the error terms are independent as a focal point and construct an approximation with a drifting sequence that converges to the distribution where the joint density is equal to the product of marginal densities.

7 Conclusions

This papers considers inference for parameters of interest in nonlinear models with endogeneity. Inference becomes quite complicated for these parameters as the limiting distribution of conventional estimators is non uniform over the parameter space. To address this problem we propose a new inference procedure based on a drifting parameter sequence, loosely analogous to the "local to unity" asymptotics in the unit roots literature. We derived the limiting distribution theory which we show can be used to conduct uniformly

valid inference for the parameters of interest. This method was illustrated for the sample selection model and we informally suggest how the general method can applied to other widely studied models.

The work here suggests areas for future research. As stated many other nonlinear models will fit into this framework, so we aim to formally propose uniform inference procedures and prove their asymptotic validity.

References

Ahn, H., and J. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58(1), 3–29.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

Andrews, D., and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong and Strong Identification," *Econometrica*, 80(5), 2153–2211.

Andrews, D., and M. Schafgans (1998): "Semiparametric estimation of the intercept of a sample selection model," *The Review of Economic Studies*, 65(3), 497–517.

Athreya, K. (1987): "Bootstrap of the Mean in the Infinite Variance Case," *Annals of Statistics*, 15, 724–731.

Baker, M., D. Benjamin, A. Cegep, and M. Grant (1995): "The Distribution of Male/Female Earnings Differential, 1970-1990," *Canadian Journal of Economics*, 28, 479–501.

Bjorn, P., and Q. Vuong (1985): "Simultaneous Equations Models for Dummy Endogenous Variables: A Game Theoretic Formulation with an Application to Labor Force Participation," Caltech Working Paper 537.

Bloznelis, M., and H. Putter (2003): "Second-Order and Bootstrap Approximation to Student's t-Statistic," *Theory of Probability & Its Applications*, 47, 300–307.

Blundell, R., and J. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *ECONOMETRIC SOCIETY MONOGRAPHS*, 36, 312–357.

Chen, X., M. Ponomareva, and E. Tamer (2014): "Inference in Finite Mixture Models with an Application to Experimental Data," *Journal of Econometrics*, 182, 87–99.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

D'Hautefoeuille, X., A. Maurel, and Y. Zhang (2018): "Extremal Quantile Regressions Selection Models and the Black-white Wage Gap," *Journal of Econometrics*, 203, 129–142.

Gronau, R. (1973): "The intrafamily allocation of time: The value of the housewives' time," *The American Economic Review*, 63(4), 634–651.

Han, S., and A. McCloskey (2019): "Estimation and inference with a (nearly) singular Jacobian," *Quantitative Economics*, 10, 1019–1068.

Heckman, J. (1974): "Shadow prices, market wages, and labor supply," *Econometrica*, pp. 679–694.

——— (1990): "Varieties of selection bias," *The American Economic Review*, pp. 313–318.

Heiler, P., and E. Kazak (2019): "Valid Inference for Treatment Effect Parameters under Irregular Identification and Many Extreme Propensity Scores," University of Konstanz working paper.

Hill, B. (1975): "A Simple Approach to Inference About the Tail of a Distribution," *Annals of Statistics*, 3, 1163–1174.

Hill, J., and S. Chaudhuri (2012): "Robust Estimation of Average Treatment Effects," UNC working paper.

Honoré, B., and L. Hu (2019): "Selection Without Exclusion," FRB of Chicago Working Paper.

Khan, S., and D. Nekipelov (2018): "Information Structure and Statistical Information in Discrete Response Models," *Quantitative Economics*, 9, 995–1017.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions and Inverse Weight Estimation," *Econometrica*, 6, 2021–2042.

Lewbel, A. (1997): "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory*, 1997(1), 32–51.

Lewbel, A. (1998): "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66(1), 105–122.

——— (2007): "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 141(2), 777–806.

Ma, X., and J. Wang (forthcoming): "Robust Inference Using Inverse Probability Weighting," *Journal of the American Statistical Association*.

McCulloch, J. (1986): "Simple consistent estimators of stable distribution parameters," *Communications in Statistics-Simulation and Computation*, 15(4), 1109–1136.

——— (1997): "Measuring tail thickness to estimate the stable index α: A critique," *Journal of Business & Economic Statistics*, 15(1), 74–81.

Mroz, T. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica*, 55(1), 765–799.

Müller, U. (2017): "Refining the Central Limit Theorem Approximation via Extreme Value Theory.," Princeton University Working Paper.

Müller, U., and Y. Wang (2017): "Fixed-k Asymptotic Inference about Tail Properties," *Journal of the American Statistical Association*, 112, 1334–1343.

Newey, W., J. Powell, and J. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review Papers and Proceedings*, 80, 324–328.

Newey, W. K. (2009): "Two Step Series Estimation of Sample Selection Models," *Econometrics Journal*, 12, 217–229.

Oaxaca, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14(3), 693–709.

Peligrad, M., and H. Sang (2011): "Central limit theorem for linear processes with infinite variance," *Journal of Theoretical Probability*, pp. 1–18.

Pickands, J. (1975): "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, 3, 119–131.

Politis, D., J. Romano, and M. Wolf (1999): *Subsampling*. Springer.

Powell, J. (1986): "Censored regression quantiles," *Journal of econometrics*, 32(1), 143–155.

Resnick, S. (2006): *Heavy-tail phenomena: probabilistic and statistical modeling*, vol. 10. Springer.

Romano, J., and M. Wolf (1999): "Subsampling Inference for the Mean of a Heavy Tailed Distribution," *Metrika*, 50, 55–69.

Samorodnitsky, G., and M. Taqqu (1994): *Stable non-Gaussian random processes: stochastic models with infinite variance*. Chapman & Hall/CRC.

Smith, J., and F. Welch (1986): *Closing the Gap: Forty Years of Economic Progress for Blacks*. Rand Corporation.

Tamer, E. (2003): "Incomplete Bivariate Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70, 147–167.

Vytlacil, E., and N. Yildiz (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, 75(3), 757–779.

A General properties for consistent estimators for the intercept

Although the proposed closed form estimator delivers a convenient approach to construct a feasible consistent estimator for the intercept in the selection model, it is in general not obvious whether one can find a "better" estimator that could be used as an alternative for inference. The observable distribution of the data is fully characterized by distributions $\Pr(Y \leq y,\, D = 1 \mid Z = z, X = x)$, $F_X(\cdot)$ and $F_Z(\cdot)$. Without loss of generality for simplicity of exposition we do not analyze the case with the covariates in the selection of equation. Denote $\eta = \Pr(Y \leq y,\, D = 1,\, Z \leq z)$ the infinite-dimensional element of the model. Let $\mathcal{H} \ni \eta$ be a pseudometric space with a pseudometric $\rho(\cdot, \cdot)$. A typical choice of the pseudometric is an L_p pseudometric or a Sobolev pseudometric that also takes into considerations the derivatives.

We have established that the intercept parameter in the linear selection is identified in $\mathcal{H}$:

$$\theta = \lim_{z \to \infty} E[Y \mid D = 1, Z = z].$$

Let $\theta(\eta)$ be the intercept associated with a particular distribution structure η and let $\widehat{\theta}(\eta)$ be an estimator for $\theta(\eta)$. We call this estimator *uniformly consistent* in $\mathcal{H}$ if for any $\eta \in \mathcal{H}$: $\widehat{\theta}(\eta) \xrightarrow{p} \theta(\eta)$. Our first result shows that the process associated with a rate-normalized estimator $\widehat{\theta}(\eta)$ *cannot be stochastically equicontinuous* for any "practical" choice of $\mathcal{H}$.

THEOREM 7 *Let $\widehat{\theta}(\eta)$ be a uniformly consistent estimator for the intercept parameter and the pseudometric ρ is dominated by an L_∞ pseudometric. Suppose that for some $\eta \in \mathcal{H}$ and r_n such that $r_n/n \to 0$ for any $q \in [0,\, 1]$ there exists $C_q > 0$ such that for sufficiently large n*

$$Pr\Big(r_n|\widehat{\theta}(\eta) - \theta(\eta)| > C_q\Big) \leq q.$$

Then for any $\epsilon > 0$, $\delta \in [0, 1]$ and $\Delta > 0$ there exist $\eta' \in \mathcal{H}$ such that $\rho(\eta, \eta') < \epsilon$ and

$$Pr\Big(r_n|\widehat{\theta}(\eta') - \theta(\eta')| > \Delta\Big) > \delta.$$

Proof:
For element η consider an element η' such that the corresponding distribution $\Pr'(Y \leq y,\, D = 1,\, Z \leq z)$ has the same conditional distribution $F(y \,|\, D = 1, Z = z)$ as η whenever $z \leq \bar{z}$ while for all $z > \bar{z}$ $E_{\eta'}[Y \,|\, D = 1, Z = z] = t + E_\eta[Y \,|\, D = 1, Z = z]$. In this case $\theta(\eta') = \theta(\eta) + t$.

Let $A = \Big\{r_n|\widehat{\theta}(\eta') - \theta(\eta')| > \Delta\Big\}$. Then

$$\Pr_{\eta'}(A) \geq \Pr_{\eta'}\left(A \cap \{z_i \leq \bar{z},\, \forall i\}\right).$$

Note that by our construction

$$\Pr_{\eta'}\left(A \,|\, \{z_i \leq \bar{z},\, \forall i\}\right) = \Pr_{\eta}\left(A \,|\, \{z_i \leq \bar{z},\, \forall i\}\right)$$

At the same time, with probability exceeding $1 - C^{-1}_{\Delta/2}$, $r_n|\widehat{\theta}(\eta) - \theta(\eta)| < \Delta/2$. Therefore

$$\Pr_{\eta'}(A) \geq \Pr_{\eta'}\left(\{r_n\, t > \Delta/2\} \cap \{z_i \leq \bar{z},\, \forall i\}\right) = \mathbf{1}\{r_n\, t > \Delta/2\} F_Z(\bar{z})^n.$$

Then we can guarantee that $\rho(\eta, \eta') < \epsilon$ by choosing t and $\bar{z}$ such that $t(1 - F_Z(\bar{z})) < \epsilon$. We guarantee that the bound is exceeded whenever $r_n\, t > \Delta/2$, and $F_Z(\bar{z})^n > \delta$. That occurs for $t = n\epsilon/\log(1/\delta)$ and $\bar{z} > F_Z^{-1}(1 - \epsilon/t)$.

Q.E.D.

In other words, this theorem establishes that for each uniformly consistent estimator, in any neighborhood of a particular distribution of observable variables, we can find another distribution such that the estimator under that distribution has both a drastically different convergence rate and a drastically different asymptotic distribution.

This result implies that non-uniform behavior is not only characteristic for the closed form estimators that we consider in the paper, but for any consistent estimator for the irregularly identified parameter.

B Proof of Theorem 2

In the proof of Theorem 4 we demonstrate that if we define the process of partial sums

$$L_n(t) = \sum_{i=1}^{[nt]} \left(\frac{W_i}{a_n} - [nt] E \left[\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n \leq 1\} \right] \right),$$

then $L_n(\cdot) \Rightarrow L_{1+\gamma}(\cdot)$ where $L_{1+\gamma}(\cdot)$ is the stable Lévy process on $[0,1]$.

We note that

$$nE \left[\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n \leq 1\} \right] \to b_n.$$

Applying the continuous mapping theorem, we conclude that

$$\sum_{i=1}^{k} W_n/a_n - b_n \xrightarrow{d} \sum^{[} n\,k/n]_{i=1} \left(\frac{W_i}{a_n} - [n\,k/n] E \left[\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n \leq 1\} \right] \right)$$

Therefore, $\sum_{i=1}^{k} W_n/a_n - b_n \xrightarrow{d} L_{1+\gamma}(1)$.

C Proof of Theorem 2

Consider function $H(w) = E\left[W^2 \mathbf{1}\{|W| \leq w\}\right]$. Provided that $\psi(t) = |t|^2 \Pr\left(f_z(Z) \left| \frac{\partial E[Y \mid Z]}{\partial z} \right|^{-1} < |t|^{-1} \right)$ is slowly varying at infinity. In this case we can define function $H(w) = E\left[W^2 \mathbf{1}\{|W| \leq w\}\right]$ which is slowly varying at infinity. Next, we apply directly Theorem 2.1. in (Peligrad and Sang 2011) and establish the result of our theorem.

D Proof of Theorem 3

Imposing the normalization for $E[W]$ at zero, we conclude that the characteristic function corresponding to $f_W(\cdot)$ $\phi_W(t)$ admits the representation in the neighborhood of $t = 0$ as $\phi_W(t) = \exp(-\frac{1}{2}\sigma^2 t^2 + o(t^2)) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$. The second component corresponds to the density of the heavy tail distribution, and the re-centerining allows us to provide a simple expression for its characteristic function $\phi_c(t)$ in the neighborhood of 0 as $\phi_c(t) = \exp(-\frac{1}{2}\kappa^2 |t|^{2/(1+c)})$. The Fourier transform of the difference $(1+c)|w|^c f_W(\text{sign}(w)|w|^{1+c}) - f_W(w)$ (if $c > 0$) can be represented as

$$1 - \frac{1}{2}\kappa^2 |t|^{2/(1+c)} - 1 + o(|t|^{1/(1+c)}) = -\frac{1}{2}\kappa^2 |t|^{2/(1+c)} + o(|t|^{1/(1+c)}).$$

Now consider the random variable $\eta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i^c$, with W_i^c being the i.i.d. copies of random variable W^c with local distribution (3.1). Then

$$E\left[\exp(it\eta_n)\right] = \prod_{i=1}^{n} \int \exp\left(iw_i^c \frac{t}{\sqrt{n}}\right) f_W^c(w_i^c)\, d\, w_i^c = \left(1 - \frac{t^2\sigma^2}{2n} - \frac{1}{2\rho_n\, n^{1/(1+c)}} \kappa^2 |t|^{2/(1+c)} + o(\rho_n^{-1} n^{-2/(1+c)})\right)^n$$
$$= \exp\left(-\frac{t^2\sigma^2}{2}\right) \exp\left(-\frac{\kappa^2 |t|^{2/(1+c)}}{2}\right) + o(1).$$

Thus, as $n \to \infty$ the characteristic function of the partial sum distribution under the local distribution $f_W^c(\cdot)$ converges to the product of the characteristic function of a Gaussian random variable with variance σ^2 and a random variable with a stable distribution with tail index $2/(1+c)$. By the Leévy convergence theorem, it follows that we can characterize the asymptotic distribution as a distribution of the sum of a Gaussian random variable with variance σ^2 and an independent random variable with a stable distribution. This result is formalized in the following theorem.

E Proof of Theorem 4

Provided Theorem 37.1 in (Samorodnitsky and Taqqu 1994), if $\phi(\cdot) \in RV_{-\gamma}$, then for $\nu_{1+\gamma}(\cdot)$- $(1+\gamma)$-stable Lévy measure on $\mathbb{R}$ with tail behavior $\nu_{1+\gamma}((x, +\infty]) = x^{-(1+\gamma)}$ for some $C > 0$ and all $x > C$ and b_n selected as in Theorem 2, for all Borel subsets of $\mathbb{R}_+$ denoted B

$$n\Pr\left(\frac{W}{b_n} \in B\right) \Rightarrow \nu_{1+\gamma}(\cdot)$$

Then we consider the random measure associated with the infinite sequence of draws from the distribution of W and by Theorem 6.3. in (Resnick 2006) it follows that

$$\sum_{i=1}^{\infty} \delta_{(\frac{i}{n}, W_i/b_n)} \Rightarrow \Lambda(\text{Leb} \times \nu_{1+\gamma}),$$

where δ_x is the distribution with point mass at x, $\Lambda(\cdot,\cdot)$ is the Poisson random measure with the support on the space of Radon point measures on $\mathbb{R}_+ \times ([0, +\infty] \setminus \{0\})$ where $[0, +\infty] \setminus \{0\}$ is the set of non-negative reals that is locally uncompatified by defining a topology on its subsets that exclude the origin. Leb is the Lebesque measure of length and $\nu_{1+\gamma}$ is the $(1+\gamma)$-stable Lévy measure. Denote $\mathbb{U} = [0, +\infty] \setminus \{0\}$ and the set of Radon point measures on A by $M_r(A)$.

We consider the map $m : M_r([0,+\infty) \times \mathbb{U}) \mapsto M_r([0,+\infty) \times [\epsilon,+\infty])$, where ϵ is chosen to be the point of continuity of function $f(w) = \nu_{1+\gamma}([w,+\infty))$. This map is almost surely continuous with respect to $\Lambda(\text{Leb} \times \nu_{1+\gamma})$ by Feigin, Kratz and Resnick (1996). Also, consider functional

$$\sum_i \delta_{(\tau_i, J_i)} \mapsto \sum_{\tau_i \leq \cdot} J_i$$

mapping from $M_r([0,+\infty) \times \mathbb{U})$ into $D([0,1], \mathbb{R})$ (Skorohod space of functions defined on $[0,1]$ with values in $\mathbb{R}$) that represents summations. This function is almost surely continuous with respect to $\Lambda(\text{Leb} \times \nu_{1+\gamma})$ by Feigin, Kratz and Resnick (1996)..

As a result, we notice that

$$\sum_i \mathbf{1}\{|W_i|/a_n > \epsilon\} \delta_{(i/n, W_i/a_n)} \Rightarrow \sum_i \mathbf{1}\{j_i > \epsilon\} \delta_{(t_i, j_i)},$$

where j_i is the increment of the Poisson process defined by $\Lambda(\text{Leb} \times \nu_{1+\gamma})$ at the instant t_i. This result follows from the convergence of the empirical point measure to the Poisson random measure and the continuity of the map m (restricting the support of the Lévy measure to $[\epsilon,+\infty)$.

Also from the continuity of the summation functional, it follows that

$$\sum_{i=1}^{[nt]} \frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n > \epsilon\} \Rightarrow \sum_{t_i \leq t} j_i \mathbf{1}\{|j_i| > \epsilon\}, \;\; t \in [0,1]$$

in $D([0,1], \mathbb{R})$.

Also, by continuity of the summation functional

$$\sum_{i=1}^{[nt]} \frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n > \epsilon\} \Rightarrow \sum_{t_i \leq t} j_i \mathbf{1}\{1 \geq |j_i| > \epsilon\}, \;\; t \in [0,1]$$

in $D([0,1], \mathbb{R})$. Taking expectations, we obtain that

$$[nt] E\left[\frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n > \epsilon\}\right] \to t \int_{\epsilon < w < 1} w \nu_{1+\gamma}(dw).$$

Consider process of trimmed partial sums

$$L_n^{\epsilon}(t) = \sum_{i=1}^{[nt]} \left(\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n > \epsilon\} - [nt] E\left[\frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n > \epsilon\}\right] \right)$$

By the previous results, we conclude that

$$L_n^{\epsilon}(\cdot) \Rightarrow L_{1+\gamma}^{\epsilon}(\cdot),$$

where $L_{1+\gamma}^{\epsilon}(\cdot)$ is the "restricted" $1+\gamma$-stable Lévy process such that

$$L_{1+\gamma}^{\epsilon}(t) = \sum_{t_i \leq t} j_i \mathbf{1}\{1 \geq |j_i| > \epsilon\} - t \int_{\epsilon < w < 1} w \nu_{1+\gamma}(dw).$$

Then, using the Itô representation of the Lévy process:

$$L_{1+\gamma}^{\epsilon}(t) \to L_{1+\gamma}(t)$$

almost everywhere on w locally uniformly it $t \in [0,1]$ as $\epsilon \to 0$ If $d_s(\cdot,\cdot)$ is the Skorohod metric on $D([0,+\infty))$ then provided that local uniform convergence implies Skorohod convergence, we see that

$$d_s\left(L_{1+\gamma}^{\epsilon}(\cdot),\, L_{1+\gamma}(\cdot)\right) \to 0$$

almost surely as $\epsilon \to 0$. As a result, given that almost sure convergence implies weak convergence, then

$$L_{1+\gamma}^{\epsilon}(\cdot) \Rightarrow L_{1+\gamma}(\cdot)$$

Consider the process or regular partial sums

$$L_n(t) = \sum_{i=1}^{[nt]} \left(\frac{W_i}{a_n} - [nt] E\left[\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n \leq 1\} \right] \right)$$

Next we demonstrate the stochastic equicontinuity. Consider the following sequence of expressions:

$$\begin{aligned}
&\Pr\left(\sup_{t\in[0,1]} \|L_n^{\epsilon}(t) - L_n(t)\| > \delta \right) \\
&\quad \leq \Pr\left(\sup_{t\in[0,1]} \left| \sum_{i=1}^{[nt]} \left(\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n < \epsilon\} - E\left[\frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n < \epsilon\} \right] \right) \right| > \delta \right) \\
&\quad = \Pr\left(\max_{0\leq k\leq n} \left| \sum_{i=1}^{k} \left(\frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n < \epsilon\} - E\left[\frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n < \epsilon\} \right] \right) \right| > \delta \right)
\end{aligned}$$

Applying Doob's inequality, we conclude that

$$\begin{aligned}\Pr\left(\sup_{t\in[0,1]}\|L^{\epsilon}_{1+\gamma}(t)-L_{1+\gamma}(t)\|>\delta\right)\\ \leq\frac{\mathrm{Var}\left(\frac{W}{a_n}\mathbf{1}\{|W|/a_n<\epsilon\}\right)}{\delta^2}\end{aligned}$$

Next we note that

$$E\left[\frac{W}{a_n}\mathbf{1}\{|W|/a_n<\epsilon\}\right]\to\int_{|w|\leq\epsilon}w^2\nu_{1+\gamma}(dw)=O(\epsilon^{1-\gamma})=o(1)$$

, as $\epsilon\to 0$. Therefore, for any $\delta>0$ we show that

$$\lim_{\epsilon\to\infty}\limsup_{n\to\infty}P\left(\sup_{t\in[0,1]}\|L^{\epsilon}_{n}(t)-L_{n}(t)\|>\delta\right)=0.$$

This implies that

$$\lim_{\epsilon\to\infty}\limsup_{n\to\infty}P\left(d_s\left(L^{\epsilon}_{n}(\cdot),\,L_n(\cdot)\right)>\delta\right)=0.$$

This leads us to conclusion that $L_n(\cdot)\Rightarrow L_{1+\gamma}(\cdot)$.

TABLE 1
Design 1: normal instruments, fixed parameters

	Mean Bias				**Median Bias**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	-0.1983	-0.1374	-0.2712	0.0822	-0.0012	0.0009	0.0068	0.0086
$c = 0.50$	-0.0005	0.0074	0.0506	-0.1461	-0.2205	-0.2211	-0.2190	-0.2222
$c = 0.75$	0.0986	0.1790	0.0587	-0.2609	-0.3237	-0.3279	-0.3300	-0.3370
$c = 1.00$	0.2057	0.2835	0.1700	-0.3762	-0.4326	-0.4322	-0.4400	-0.4570
	RMSE				**MAD**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	38.3796	72.5744	157.8260	106.4073	3.1213	3.6428	4.7534	5.9159
$c = 0.50$	36.7891	72.5922	157.7288	106.4602	3.7862	4.7681	6.3180	8.6620
$c = 0.75$	36.9827	72.4332	157.7312	106.6476	4.4494	5.8431	7.8785	11.2825
$c = 1.00$	36.8312	72.5112	157.7495	106.8617	5.2608	7.0809	9.7935	14.1935

TABLE 2
Design 1: normal instruments, drifting parameters

	Mean Bias				**Median Bias**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	-0.0242	-0.1534	-0.2182	0.0517	-0.0015	0.0072	0.0043	0.0102
$c = 0.50$	-0.0436	0.1685	0.2293	-0.0438	-0.0257	0.0067	0.0070	-0.0015
$c = 0.75$	0.0535	0.1759	0.2358	-0.0397	-0.0355	-0.0134	0.0124	-0.0016
$c = 1.00$	0.0640	0.1835	0.2403	-0.0357	-0.0501	-0.0239	0.0152	-0.0063
	RMSE				**MAD**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	27.6662	134.1118	146.0589	237.0683	3.2291	3.6671	4.2015	6.0014
$c = 0.50$	27.3730	134.2254	146.0701	237.0692	3.2690	3.6758	4.2004	6.0276
$c = 0.75$	27.4384	134.2950	146.0883	237.0829	3.2707	3.6993	4.2025	6.0233
$c = 1.00$	27.3751	134.2166	146.0793	237.0718	3.3033	3.6770	4.2129	6.0072

TABLE 3
Design 3: cauchy instruments, constant parameters

	Mean Bias				**Median Bias**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	0.2714	1.4475	1.7833	-0.4549	-0.0314	0.0313	0.0245	0.0048
$c = 0.50$	0.4334	1.6107	1.6.428	-0.2749	-0.1410	0.02123	-0.1584	-0.1862
$c = 0.75$	0.5207	1.7115	1.5450	-0.1945	-0.2289	-0.3028	0.2435	-0.2713
$c = 1.00$	0.5971	1.7806	1.5421	-0.1449	-0.3123	-0.3915	0.3325	-0.3577
	RMSE ($10^3\times$)				**MAD**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	0.4116	0.5118	2.0589	0.5683	9.2291	12.6671	13.2015	14.0014
$c = 0.50$	0.43730	0.5254	2.0701	0.5692	9.2690	1 2.6758	13.2004	15.0276
$c = 0.75$	0.4384	0.5750	2.0883	0.5829	9.2707	12.6993	13.2025	15.9233
$c = 1.00$	0..4751	0.5766	2.0793	0.5718	9.3033	12.6770	13.2129	16.0072

TABLE 4
Design 3: cauchy instruments, drifting parameters

	Mean Bias				**Median Bias**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	-1.5227	-0.4932	1.8262	-1.6800	-0.0275	0.0112	0.0052	0.0005
$c = 0.50$	-1.5394	-0.5053	1.8174	-1.6863	-0.0459	-0.0029	-0.0055	- 0.0039
$c = 0.75$	-1.5472	-0.5114	1.8131	-1.6896	-0.0555	-0.0094	-0.0068	-0.0074
$c = 1.00$	-1.5554	-0.5174	1.8087	-1.6926	-0.0658	-0.0178	-0.0110	- 0.0106
	RMSE ($10^3\times$)				**MAD**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$c = 0.00$	0.2484	0.3895	0.63151	1.0910	8.8074	11.4595	13.8240	14.5414
$c = 0.50$	0.2487	0.3992	0.63142	1.0910	8.7724	11.5162	13.8324	14.5630
$c = 0.75$	0.2487	0.3957	0.6366	1.0910	8.7990	11.5741	13.8390	14.5157
$c = 1.00$	0.0.2489	0.3959	0.6325	1.0911	8.7863	11.5675	13.8244	14.4495

TABLE 5
Size and Power of t test
Design 1: Bivariate normal, correlation ρ

	OLS							
	$H_0 : \alpha_0 = 0$				$H_0 : \alpha_0 = 0.5$			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$\rho = 0.00$	0.949	0.949	0.947	0.947	0.997	1.00	1.00	1.00
$\rho = 0.5$	0.514	0.216	0.024	0.0	1.00	1.00	1.00	1.00
$\rho = 0.75$	0.153	0.024	0.00	0.00	0.991	1.00	1.00	1.00
$\rho = 0.95$	0.066	0.002	0.00	0.00	0.975	1.00	1.00	1.00
	Heckman							
$\rho = 0.00$	0.812	0.833	0.832	0.834	0.745	0..935	0.995	1.00
$\rho = 0.50$	0.803	0.827	0.820	0.807	0.746	0.935	1.00	1.00
$\rho = 0.75$	0.794	0.717	0.786	0.736	0.862	0.928	1.00	1.00
$\rho = 0.95$	0.738	0.694	0.579	0.394	0.675	0.888	0.985	1.00
	Andrews Schafagans							
	$H_0 : \alpha_0 = 0$				$H_0 : \alpha_0 = 0.5$			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$\rho = 0.00$	0.972	0.981	0.984	0.987	0.330	0.605	0.917	0.999
$\rho = 0.50$	0.979	0.966	0..950	0.929	0.235	0.440	0.781	0.986
$\rho = 0.75$	0.940	0.947	0.940	0.800	0.195	0.345	0.637	0.905
$\rho = 0.95$	0.911	0.907	0.805	0..580	0.149	0.268	0.548	0.879
	IVW							
$\rho = 0.00$	0.775	0.732	0.653	0.598	0.696	0.806	0.933	0.982
$\rho = 0.50$	0.782	0.736	0.662	0.607	0.694	0.814	0.924	0.983
$\rho = 0.75$	0.785	0.740	0.664	0.616	0.688	0.816	0.930	0.988
$\rho = 0.95$	0.791	0.754	0.694	0.629	0.687	0.825	0.935	0.998

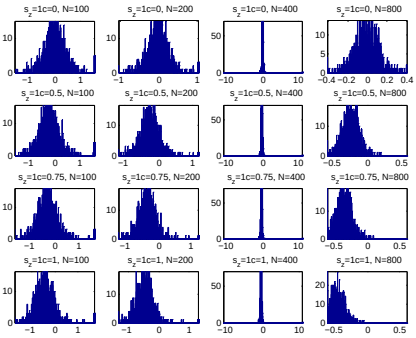
TABLE 6
Bootstrap Coverage Probabilities
Design 1: Bivariate normal, correlation ρ

	OLS				**Heckman 2 Step**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$\rho = 0.00$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\rho = 0.5$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\rho = 0.75$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\rho = 0.95$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Andrews and Schafagans				**IVW**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$\rho = 0.00$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\rho = 0.5$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\rho = 0.75$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\rho = 0.95$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99

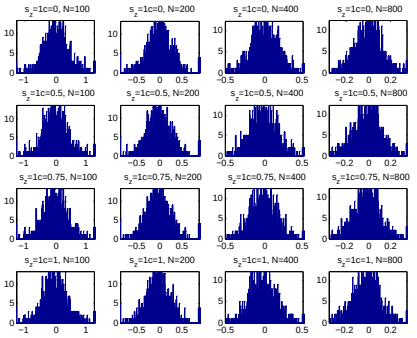
TABLE 7
Bootstrap Coverage Probabilities
Design 1: Marginal cauchy, Gaussian Copula correlation ρ

	OLS				**Heckman 2 Step**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$\rho = 0.00$	0.90	0.90	0.89	0.85	0.71	0.70	0.76	0.83
$\rho = 0.5$	0.86	0.82	0.83	0.83	0.77	0.71	0.60	0.58
$\rho = 0.75$	0.86	0.83	0.87	0.87	0.67	0.65	0.61	0.57
$\rho = 0.95$	0.85	0.84	0.88	0.89	0.57	0.57	0.53	0.50
	Andrews and Shafagans				**IVW**			
	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$\rho = 0.00$	0.66	0.83	0.72	0.77	0.91	0.90	0.89	0.85
$\rho = 0.5$	0.75	0.71	0.65	0.54	0.86	0.82	0.83	0.83
$\rho = 0.75$	0.65	0.62	0.40	0.20	0.901	0.83	0.85	0.85
$\rho = 0.95$	0.62	0.40	0.14	0.03	0.95	0.85	0.88	0.89

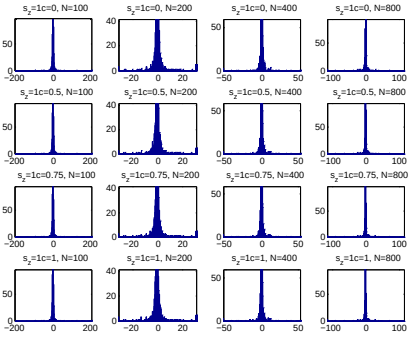
Figure 1: **Results for Inverse Weight Estimator**



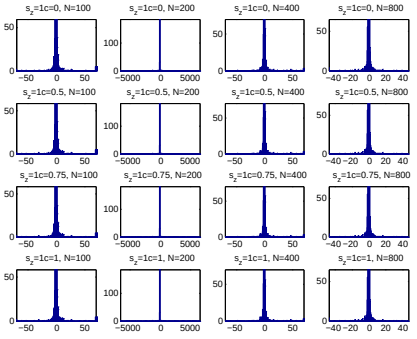
(a) Normal Instrument, Fixed Parameters



(b) Normal Instrument, Drifting Parameters

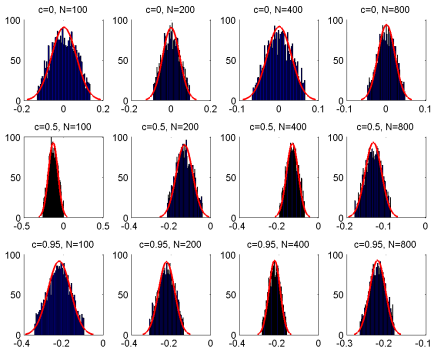


(c) Cauchy Instrument, Fixed Parameters

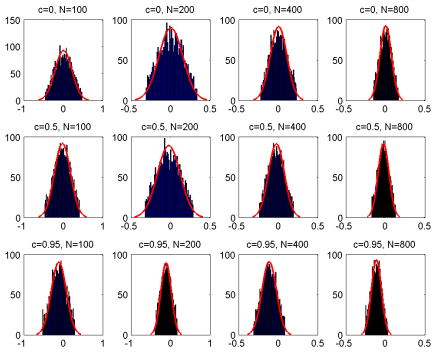


(d) Cauchy Instrument, Drifting Parameters

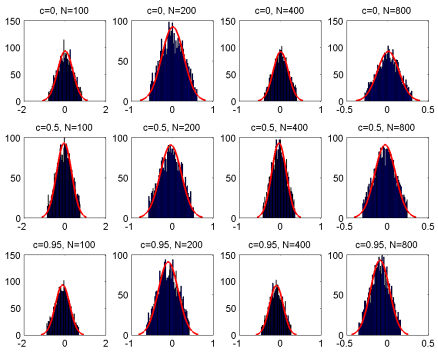
Figure 2: **Results for other Estimators, Gaussian Disturbances**



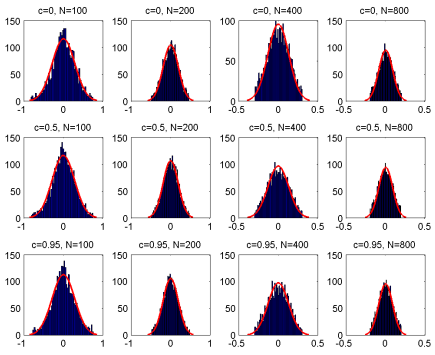
(a) OLS



(b) Heckman

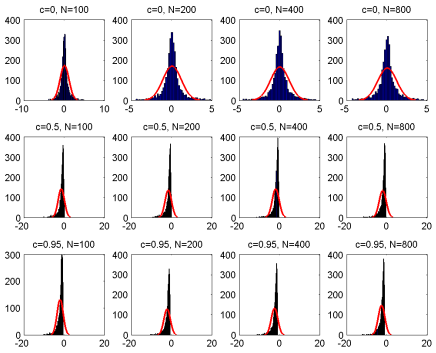


(c) Andrews and Schafgans

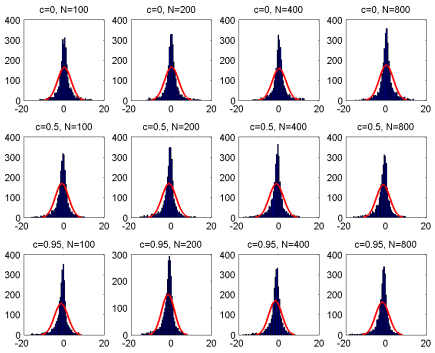


(d) Bridge estimator

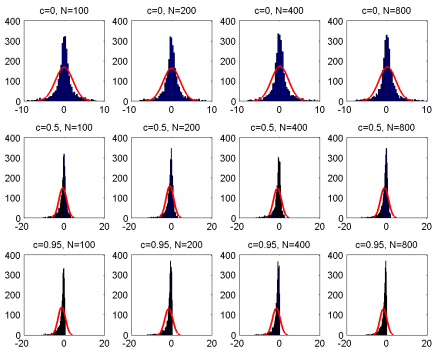
Figure 3: **Results for other Estimators, Cauchy Disturbances**



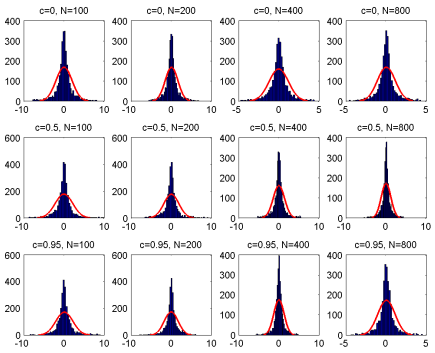
(a) OLS



(b) Heckman

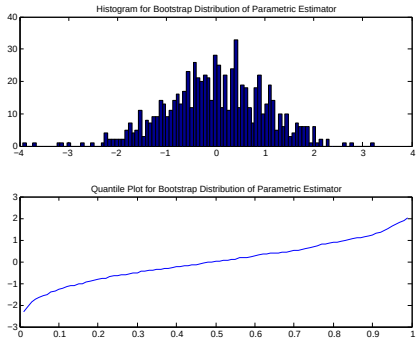


(c) Andrews and Schafgans

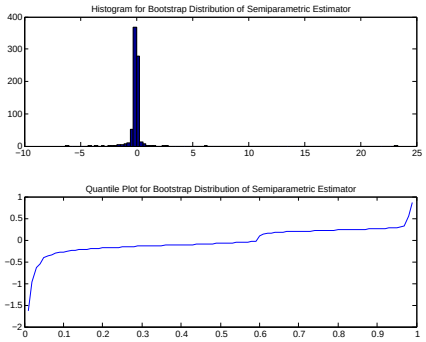


(d) Bridge estimator

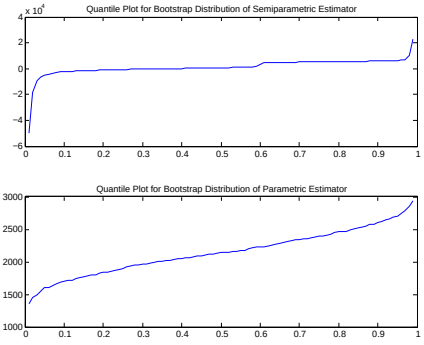
Figure 4: **Application using Mroz Data**



(a) Parametric Estimation



(b) Semiparametric



(c) Comparison