On Uniform Inference in Nonlinear Models with Endogeneity

Shakeeb Khan*  Denis Nekipelov*
Duke University  University of Virginia

September 20, 2014

Abstract

This paper explores the uniformity of inference for parameters of interest in nonlinear models with endogeneity. The notion of uniformity is fundamental in these models because due to potential endogeneity, the behavior of standard estimators of these parameters is shown to vary with where they lie in the parameter space. Consequently, uniform 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 2012a), (Andrews and Cheng 2012b) and (Chen, Ponomareva, and Tamer 2011).

We illustrate this point with models widely used in empirical work. Our main example is the standard sample selection model, where the parameter is the intercept term. ((Heckman 1990), (Andrews and Schafgans 1998) and (Lewbel 1997a)). We show that with selection on unobservables, asymptotic theory for this parameter is not standard in terms of there being nonparametric rates and non-gaussian limiting distributions. In contrast if the selection is on observables only, rates and asymptotic distribution are standard, and consequently, there is a discontinuity in the limiting distribution theory for an estimator despite it being uniformly consistent. We show that this discontinuity prevents standard inference procedures from being globally uniformly valid, and this motivates the development of new inference procedures that are proven to be (locally) uniformly valid under drifting parameter asymptotics. Finite sample properties of the new inference procedure is illustrated through a simulation study as well an empirical illustration using the (Mroz 1987) data set. It is also illustrated how the new inference procedure can be useful in other nonlinear models with endogenous variables.

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, HKUST, SHUFE, SMU, NUS 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. Unfortunately, it is not theoretically appropriate to apply 2SLS to non-linear models, as the consistency of 2SLS depends critically upon the orthogonality conditions that arise in the linear-regression context.

Until recently, the standard approach for handling endogeneity in many non-linear models such as discrete choice, censored or bivariate selection has required parametric specification of the error disturbances (see, e.g. (Heckman 1990)). 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. In one sense, this semiparametric, or “distribution-free” approach can be roughly divided into two groups, depending on the source of endogeneity that arises in the nonlinear model.

In one group the data available to the econometrician is selected nonrandomly, resulting in what is now well known as sample selection bias ((Gronau 1973), (Heckman 1974)). Original inference methods were based on standard procedures such as MLE or NLLS but consistency of these approaches were based on the parametric specification of the unobserved components making such methods undesirable, and motivating the distribution free methods proposed in (Powell 1986), (Ahn and Powell 1993), (Lewbel 1997a).

In the other group, the source of endogeneity is the explanatory variables themselves. As mentioned, correct inference on the coefficients of these endogenous regressors at first required parametric specification on the unobserved components of the model, but fortunately, more recent work such as (Blundell and Powell 2003), (Vytlacil and Yildiz 2007), (Khan and Nekipelov 2011) proposed semi-parametric, distribution free methods which are robust to misspecification of the distribution of the unobserved components of the model.

In both groups, the proposed semiparametric inferences are a welcome addition to the literature when compared to parametric methods. However, 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 uniform in the parameters of the model. By this we mean that the large sample properties of these proposed semiparametric models to address endogeneity will vary depending on the values of the unknown parameters of the model. This will make inference of the parameters of interest more complicated than suggested by existing semiparametric inference methods. This paper aims to fill this gap in the literature by proposing inference methods for the parameters of interest for nonlinear models with endogeneity that are uniform in the unknown parameter values. In the examples we consider the notion of uniformity is directly linked to the endogeneity in the model. This is because, as we will show, the large sample properties of the
existing semiparametric inference procedures will vary substantially with the degree of endogeneity as well as the parameter values themselves. Specifically, as we will show, limiting distribution theory will be discontinuous in the values of parameters in the model. In that sense that makes valid inference in these models complicated, and general, standard inference procedures will be shown to not be globally uniformly valid. This motivates new inference procedures and we propose ones that are locally uniformly valid, and are based on drifting parameter value asymptotics, analogous to those used in the unit root and weak instruments literature.

The rest of the paper is organized as follows. The next section illustrates our main nonuniformity problem by reconsidering the semiparametric sample selection model (Ahn and Powell 1993), (Heckman 1990), (Lewbel 1997a), (Andrews and Schafgans 1998)). In Sections 3 and 4 we show that the large sample behavior of proposed estimation and inference methods varies 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 several impossibility results for valid uniform inference, and motivates our new inference procedures based on drifting parameter asymptotic theory, under which we prove the asymptotic validity.

Sections 5 and 6 explores the finite sample properties of our new inference methods in two ways. First (Section 4) we consider a simulation study, simulating the models considered in Section 2, and reporting finite sample properties of estimation and testing procedures. Second (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 2012).

Section 7 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. Here we show that properties of standard estimation procedures for the coefficient of a dummy endogenous variable, a fundamental parameter of interest, varies discontinuously in the degree of endogeneity, with the extreme case being that the the dummy variable is exogenous.

Section 8 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 in conducting uniform inference in the sample selection model, which has been a model of widespread interest in both theoretical and applied econometrics. This interest 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 consider inference in the following model:

\[ D = 1\{Z - V \geq 0\} \]
\[ Y = DY^* = D \cdot (\theta_0 + U) \]  

(2.1)

Where \( \theta_0 \in \mathbb{R} \) 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 \( 1\{\cdot\} \) 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 \( \beta_0 \) would also be a parameter to conduct inference on- see, e.g. (Ahn and Powell 1993)

Our focus is on the condensed model and \( \theta_0 \) only, for the following reasons. First, \( \theta_0 \) 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 estimated 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, $\theta_0$ is the parameter for which inference on can vary with the degree of selection, measured by the correlation between $U$ and $V$, and thus where the problem of *uniformity* arises. This is generally not the case for inference on $\beta_0$ as we will explain further below. Thus inference on $\beta_0$ can be handled by existing methods since estimators for it will behave similarly across varying degrees of selection in the model.

What makes inference complicated for $\theta_0$ is that how well one can estimate $\theta_0$ 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 completely on *observables only*, which corresponds to $U,V$ being uncorrelated with each other, than $\theta_0$ 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 (Andrews and Schafgans 1998). We propose a different one in this paper that will be the basis of our inference procedure.

Interestingly, 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. In other words these estimators are not *adaptive* to the type of selection. 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 discuss an inference procedure that allows for both types of selection we consider the behavior of the KN estimator under locally drifting parameter sequences. For the problem at hand we effectively consider sequences where the correlation between $U$ and $V$ converges to 0, so that in the limit, the selection is on observables only.

The following subsection formalizes these statements with the statement and proofs of various theorems concerning the large sample properties of the various estimators under various types of selection. To facilitate this discussion 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 have a temptation to pre-test for the correlation between the error terms in the two equations, and in case it becomes clear that the error terms are uncorrelated, one may use the mean of the linear outcome whenever the dummy
$D$ is not equal to zero, as an estimate for $\theta$. 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 $\theta$ may need to employ an estimated unknown function leading to a slow rate of its convergence. This means that the properties of the estimator for $\theta$ are non-uniform in the distribution of the error terms. As we will show, the main cause of the non-uniform 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 $\theta$ may have drastically different performance both in the rate of convergence and in the structure of 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 structure 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 $E[|U|^2 | V = v] < \infty$ 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 $v$.

In the subsequent discussion we set the true value of the parameter of interest $\theta$ to zero which can be done without loss of generality.

First, we establish the general identification result for the parameter of interest. We note that our only selected 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 and, although no information is available regarding the structure of correlation between $U$ and $V$, the marginal distribution of $V$ which 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[Y | D = 1, Z = z] = \theta + E[U | V \leq z]\]$$

Alternatively, we can write

$$\theta = \frac{E[Y | Z = z]}{P(z)} - E[U | V \leq z],$$

6
where \( P(z) = E[D | Z = z] \). Then given the assumption that support of \( z \) is large, we can see that
\[
\lim_{z \to +\infty} P(z) = 1 \quad \text{and} \quad \lim_{z \to \infty} E[U | V \leq z] = E[U] = 0,
\]
therefore
\[
\theta = \lim_{z \to +\infty} E[Y | Z = z].
\]

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. However, these inference procedures rely on a high-level assumption regarding the tail behavior of the joint density of error terms in the main and the selection equation. In contrast, here we demonstrate that there exists a simple analog estimator that does not rely on any tail assumptions and thus remains consistent uniformly in the class of selection models induced by the joint distribution of error terms and the instrument that have absolutely continuous densities with full support. This estimator 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 find, however, that his estimator has undesirable properties with \( t \)-ratios not converging to a pivotal distribution. We also show under some regularity conditions imposed on the density of the instrument, the estimator will converge to a non-standard distribution at a nonparametric rate.

We also show that under stronger assumptions regarding the distribution of the error terms there exist estimators with a faster convergence rate than our analog estimator. In particular, we consider the case where the error terms in the main and the selection equation are conditionally independent. In that case the regular sample mean computed from the non-censored sample is a \( \sqrt{n} \) - consistent estimator. As with the estimator in (Andrews and Schafgans 1998), the corresponding estimator will be inconsistent if the assumption regarding the joint distribution of \((U, V)\) is not satisfied.

Then we investigate whether imposing an assumption on the distribution structure of the error terms is the only option to generate estimators that converge to the pivotal distribution at the known rate. We find that any estimator that is consistent over large classes of distributions of error terms will have a rate of convergence that discontinuously changes with the tail assumptions making the construction of pivotal statistics impossible. Also, traditional approaches of constructing confidence sets such as bootstrap are invalid.

As an alternative, we propose the idea of **locally uniform inference**. We start with some restriction on the joint density \((U, V)\) that ensures that the uniformly consistent estimator for the intercept exhibits regular behavior: we choose the assumption of conditional independence between \( U \) and \( V \). Then we approximate the joint distribution of the error terms with a drifting family of distributions which in
the limit converges to the distribution that satisfies the imposed tail condition. We find that with an appropriately chosen drifting sequence, the resulting estimator will have an asymptotic distribution that can be characterized as a sum of the normal distribution component and a component that is characterized by the finite-dimensional distribution of the Levy process. This structure ensures that inference becomes robust to deviations from the imposed distributional assumption.

2.1 Conditionally exogenous selection

We start our analysis with the familiar model with “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: \( E[U|V] = 0 \). Provided the 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 \( E[U|V \leq z] = 0 \), if the corresponding conditional density is well-defined. Then we also note that

\[
E[UD|Z = z] = E[U|V \leq z].
\]

For convenience of the further discussion we introduce a “propensity score” function

\[
P(z) = E[D|Z = z].
\]

Then we can write

\[
E[Y|D = 1, Z = z] = \theta + E[U|D = 1, Z = z] = \theta.
\]

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 \( \theta \) as

\[
\theta = E\left[\frac{Y}{P(Z)}\right]_{Z = z}.
\] (2.2)

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 data-driven, 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. As a result, we consider the estimator that is not based on the efficient weighting.

Another purpose of this alternative representation 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

$$E[Y | Z = z] = \theta P(z)$$

can be rewritten as

$$\theta f_V(z) = \frac{\partial E[Y | Z = z]}{\partial z},$$

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

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

We note that once we characterized the parameter of interest in this form, the form of the corresponding semiparametric efficient estimator is implied by (Newey 1994) and (Brown and Newey 1998).

### 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

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

We note that by the dominated convergence theorem $$\lim_{z \to -\infty} E[Y | Z = z] = 0$$. 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 elements of the model:

$$\theta = \lim_{z \to +\infty} \int_{-\infty}^{z} \frac{\partial E[Y | Z = z]}{\partial z} dz.$$

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

$$\theta = \int_{-\infty}^{+\infty} \frac{\partial E[Y | Z = z]}{\partial z} dz.$$

We 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

$$\theta = \int_{-\infty}^{+\infty} \frac{\partial E[Y | Z = z]}{\partial z} \frac{1}{f_Z(z)} f_Z(z) dz = E \left[ \frac{\partial E[Y | Z]}{\partial z} \frac{1}{f_Z(Z)} \right].$$
We note that this identification argument leads to a similar expression to that in (Lewbel 1997b), (Lewbel 1997a), (Lewbel 1998). Therefore, we can introduce the random variable \( W = f_Z(Z)^{-1} \frac{\partial E[Y|Z]}{\partial z} \) and the estimator is constructed as a sample average of the draws of this random variable\(^1\):

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i.
\]  

(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 \( E[W] = \lim_{z \to \infty} E[Y|Z = z] < \infty \). 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 \( \frac{\partial E[Y|Z]}{\partial z} f_Z(Z) \).

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

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i \xrightarrow{a.s.} 0,
\]

because recall we set w.l.o.g. the true parameter value to be 0. Thus, the estimator \( \hat{\theta} \) possesses certain "stability" properties. Provided that our results so far do not give any information regarding the characterization of the distribution of the constructed estimator, we would want to use some common method, such as bootstrap to characterize its asymptotic distribution. However, as the following result demonstrates, the traditional non-adaptive bootstrap fails in this case.

### 2.3 Distributional properties of the weighted estimator

Consider a bootstrap procedure which takes the i.i.d. sample of variables \( W_i = \frac{\partial E[Y_i|z_i]}{\partial z_i} f_Z(Z_i) \). Then we take an array \( \{I_1^{(n)}, \ldots, I_n^{(n)}\}, n \geq 1 \} \) 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 \( W_i^* = W_{I_i^{(n)}} \).

**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.

Further 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.

\(^1\)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.
Another approach to inference with estimators that have non-standard properties is based on using pivotal inference. Of interest frequently is the behavior of the \( t \)-statistic corresponding to parameter \( \hat{\theta} \). 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. 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 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 consistent 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

\[
\hat{T}_\theta = \frac{\frac{1}{n} \sum_{i=1}^{n} w_i}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} w_i^2}}
\]

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

\[
Pr(\hat{T}_\theta \leq t) \xrightarrow{F_{U,V,Z}^k} F_k(t), \ k = 1, 2
\]

and

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

Finally, one can try to use an estimator that is consistent under a particular distributional assumption regarding \( (U, V) \) and then just correct the bias using some additional information that can be observed from the actual tail behavior. The following theorem establishes the fact that bias correction is infeasible: bias cannot be estimated at the rate that is faster than the required \( n^{-1/4} \).

**THEOREM 3** Suppose for \( \theta = E[W] \)

\[
Pr(\{\omega \mid \theta - E[W \mid Z(\omega)] \neq 0\}) > 0.
\]

Then the estimator (2.3) is biased. Then, unless the tail behavior of \( (U, V) \) is known a priori, the bias of estimator (2.3) cannot be estimated at the rate faster than \( n^{-1/4} \). Therefore, bias correction is infeasible.

In light of all these negative results, the main 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.
Our asymptotic theory is based on the theory of stable distributions considered in (Zolotarev 1986), (Samorodnitsky and Taqqu 1994), (Nolan 2003), (Resnick 2006). The idea of the theory of stable distributions is the following. Suppose that $X_1, X_2, \ldots, X_n$ is the sequence of random variables and $S_n = \sum_{i=1}^{n} X_i$ is the partial sum of this sequence. Then we consider some numerical sequences $a_n$ and $b_n$ with $a_n \to \infty$ as $n \to \infty$ and consider a normalized and re-centered partial sum $T_n = S_n/a_n - b_n$. It is then proposed to consider a family of distributions whose characteristic functions have the following structure

$$
\psi(t) = \begin{cases}
\exp\left(-\sigma|t|^{\alpha} \left(1 - i\beta \text{sign}(t) \tan \frac{\pi \alpha}{2}\right) + i\mu t\right), & \text{if } \alpha \neq 1,
\exp\left(-\sigma|t| \left(1 + \frac{2i\beta \pi}{\alpha} \text{sign}(t) \log |t|\right) + i\mu t\right), & \text{if } \alpha = 1.
\end{cases}
$$

(2.4)

with $\alpha \in (0, 2]$, $\sigma \geq 0$, $|\beta| \leq 1$ and $\mu \in \mathbb{R}$. Parameter $\alpha$ is called an index of $\alpha$-stable distribution.

The following theorem, which is a result by Khinchine and Gnedenko, establishes why the class of distributions with characteristic functions (2.4) is of interest.

**THEOREM 4** ((Zolotarev 1986)) Let $\mathcal{G}$ be the class of all distributions that can be weak limits of distributions of random variables $T_n$ as $n \to \infty$. Then distribution $G \in \mathcal{G}$ if and only if its characteristic function can be written in the form (2.4)

Theorem 4 states that the limits of partial sums of any partial sums of random variables that converge to a distribution can be characterized by a distribution from family (2.4), therefore, called stable distributions.

Then the class of distributions $F$ of variables $X_i$ for which random variables $T_n \xrightarrow{d} T$ with distribution $F_T(\cdot)$ whose characteristic function is described by (2.4) is called the $\alpha$-stable domain of attraction (denoted as $S_{\rho}(\sigma, \beta, \mu)$). The properties of the characteristic functions of limiting distributions in $S_{\alpha}$ turn out to be closely related to the properties of functions of regular variation.

**DEFINITION 1** A measurable function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is regularly varying at $\infty$ with index $\alpha \in \mathbb{R}$ (written as $f \in RV_\alpha$) if for $x > 0$,

$$
\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^\alpha,
$$

where $\alpha$ is called the exponent of variation.

In case where $\alpha = 0$, function $f(\cdot)$ is called slowly varying. E.g. function $f(t) = \log t$ is slowly varying. It turns out that the cumulants of distributions in $S_\alpha(\sigma, \beta, \mu)$ of a transformed argument $1/t$ are regularly varying with index $-\alpha$.

The rates for the normalizing sequence $a_n$ and the bias-correction terms $b_n$ turn out to depend on the tail structure of the distribution of $X_i$. The following theorem establishes the rate.
THEOREM 5 (Samorodnitsky and Taqqu 1994) Sequence $T_n = \sum_{i=1}^{n} X_i/a_n - b_n$ with i.i.d. copies of random variable $X \sim F(\cdot)$, converges to stable law $S_\alpha(1, \beta, 0)$ if $x^\alpha(1 - F(x) + F(-x))$ is slowly varying at infinity and

$$
\lim_{x \to \infty} \frac{F(-x)}{1 + F(-x) - F(x)} = \frac{1 - \beta}{2}.
$$

Sequence $a_n$ stabilizes the tails of $F(\cdot)$ such that:

$$
\lim_{n \to \infty} n (1 - F(a_n) + F(-a_n)) = \begin{cases} 
\Gamma(1 - \alpha) \cos \frac{\pi \alpha}{2}, & \text{if } \alpha \in (0, 1), \\
2/\pi, & \text{if } \alpha = 1, \\
\frac{\Gamma(2-\alpha)}{\alpha - 1} |\cos \frac{\pi \alpha}{2}|, & \text{if } \alpha \in (1, 2),
\end{cases}
$$

and

$$
b_n = \begin{cases} 
0, & \text{if } \alpha \in (0, 1), \\
na_n \int_{-\infty}^{+\infty} \sin \left(\frac{x}{a_n}\right) dF(x), & \text{if } \alpha = 1, \\
n \int_{-\infty}^{+\infty} x dF(x), & \text{if } \alpha \in (1, 2).
\end{cases}
$$

This theorem suggests that the tail index of the distribution of random variable allows us to evaluate quickly the corresponding rate of convergence for its partial sums. In particular, we always obtain $a_n = n^{1/\alpha}l(n)$ where $l(\cdot)$ is a function slowly varying at infinity. This leads to familiar $\propto \sqrt{n}$ normalizations for distributions with stable Gaussian limits where the characteristic function $\psi(t) \propto \exp(-|t|^2)$.

Having described the general distribution theory for partial sums of i.i.d. random variables that may not have finite moments, we will apply this theory to analyze the distribution of the estimator of the intercept in the main equation of the selection model. We start by making the following structural assumption.

ASSUMPTION 2 Suppose that function $\phi(t) = |t| Pr \left(f_z(Z) \left| \frac{\partial E[Y | Z]}{\partial z} \right|^{-1} < |t|^{-1} \right)$ is non-increasing in $t$ for all $|t| > C$ for some large constant $C$ and $\phi \in RV_{-\gamma}$ for some $\gamma > 0$.

We note that this assumption is satisfied for all parametric distributions of $(U, V, Z)$ that are commonly used in applications ranging from Normal to Cauchy. This is the assumption which will allow us to develop the exact asymptotic theory for the estimator of the intercept and understand the source of the negative results that we established earlier.

It may be convenient to express the estimation procedure in terms of expectations with respect to the distribution of the propensity score. In fact, the propensity score is a sufficient statistic in the identification argument. In fact, conditioning on the propensity score leads to

$$
E[Y | P(Z) = \pi] = \theta \pi + E[U 1\{P(V) \leq \pi\}] 
$$
Then in case where the error term in the main equation is mean independent from the error term in the selection equation, we use the law of iterated expectation to establish that
\[ \theta = \frac{E[Y \mid P(Z) = \pi]}{\pi}, \quad \forall \, \pi \in [0, 1]. \]

Thus if we consider random variable \( P = P(Z) \), then the mean-independent case can be represented as
\[ \theta = E \left[ \frac{Y}{P} \mid P = \pi \right] \]

In case where error terms \( U \) and \( V \) are correlated, we can identify parameter of interest asymptotically:
\[ \theta = \lim_{\pi \to 1} E[Y \mid P = \pi]. \]

Then we can apply the same technique that we applied before to replace the limit with an expectation, which leads to the expression
\[ \theta = E \left[ \frac{\partial E[Y \mid P]}{\partial P} f_P(P) \right] \quad (2.7) \]

This provides an alternative representation for the parameter of interest.

The behavior of \( \hat{\theta} \) in case where the errors in the two equations are uncorrelated is determined by the tail behavior of random variable \( W \). We consider the tail probability \( \Pr(\mid W \mid > w) \). Given that the expectation of \( W \) is finite by assumption, the tail index of the distribution of \( W \) is at least one (otherwise, the first moment would not exist).

Here we can provide a basic intuition for establishing the tail behavior of random variable \( W \). In particular, under Assumption 1 the “score” \( \frac{E[Y \mid Z]}{\partial z} \) has a finite second moment. Then for \( w \to +\infty \)
\[ P \left( \frac{\partial E[Y \mid Z]}{\partial z} f_Z(Z) > w \right) \leq P \left( \left| \frac{\partial E[Y \mid Z]}{\partial z} \right| > w \right) \]
\[ + P \left( \frac{1}{f_Z(Z)} > w \right) \leq E \left[ \left| \frac{\partial E[Y \mid Z]}{\partial z} \right|^2 \right] \frac{1}{w^2} + P \left( \frac{1}{f_Z(Z)} > w \right) \]

We can then evaluate
\[ P \left( \frac{1}{f_Z(z)} > w \right) = \int f_Z(z) 1 \left\{ f_Z(z) < \frac{1}{w} \right\} \, dz = 1 - F_Z(f_Z^{-1}(w^{-1})) + F_Z(-f_Z^{-1}(w^{-1})). \]

In particular, if \( Z \) is a univariate standard normal random variable, then \( f_Z^{-1}(w^{-1}) \propto \sqrt{\log w} \). Also asymptotically, \( F_Z(z) \propto \frac{z^{-1/2}}{2}. \) As a result
\[ P \left( \frac{1}{f_Z(z)} > w \right) \propto (w \sqrt{\log w})^{-1}. \]
Provided that function $\log w$ is slowly varying at infinity, the tail distribution of $W$ is stable with the tail index 1. This means that the selection models where the “instrument” $Z$ has normal distribution, the convergence rate for the parameter in the main equation can be as slow as $\sqrt{\log n}$.

It is also important to note that (2.7) allows us to provide a quick evaluation of the tail index of the distribution of $W$ using the density of the distribution of the propensity score at 1. In fact, assuming that the variance of the “score” $\frac{\partial \log f_Y}{\partial Y} |_{P(Y|P)}$ is finite, the tail probability is determined by the probability of large deviations of $1/f_P(P)$. We can then evaluate this probability as

$$\Pr (|W| > w) = \int_{f_P(p)<1/w} f_P(p) \, dp + O(w^{-2}) = O\left( w^{-1} \left(1 - f_P^{-1}(w^{-1}) \right) \right).$$

We note that, unlike the case of the estimation of treatment effects that we consider further, the distribution of $W$ will be “worse”, the less the distribution of the propensity score is concentrated around one. Thus if the propensity score concentrates such that $f_P(1-z) \propto (1-z)^{-\delta}$ with some $\delta$ as $z \to 0$, then $\Pr (|W| > w) = O\left( \frac{1}{w^{1+1/\delta}} \right)$, leading to the tails compatible with $1 + \frac{1}{\delta}$-stable distribution. We note that, the thinner is the density of the propensity score at 1 (corresponding to a larger positive value of parameter $\delta$), the lower is the tail index of $W$. In particular, if we the distribution of the propensity score is exponentially thin around 1, then the tail index of $W$ will be 1.

We note that the exact convergence rate and the bias-correcting sequence needs to be computed on a case-by-case level. We can provide the following result.

**THEOREM 6** Suppose that Assumptions 1 and 2 hold. Consider the sequence of i.i.d. random variables $W_i = \frac{a_i \mathbb{E}[Y_i | Z_i]}{f_Z(Z_i)}$ and let sequences $a_n$ and $b_n$ defined by (2.5) and (2.6) correspond to the distribution of $W_1$. Then the estimator $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n w_i$ is consistent and:

$$\frac{n}{a_n} \hat{\theta} - b_n \xrightarrow{d} L_{1+\gamma}(1),$$

where $L_{1+\gamma}(\cdot)$ is a $(1+\gamma)$-stable Lévy process on $[0,1]$.

We recall that the $\alpha$-stable Lévy process is the stochastic process driven by the measure with the tail property of the Lévy measure $\nu(x) = c_\pm ||x||^{-\alpha}$ as $x \to \pm \infty$.

Thus the coefficients in the main equation of the selection model (unless a special structure is imposed where $U$ is mean independent of $V$) converges at the rate $n/a_n$ to the distribution determined by the limiting stable Lévy process. Stable Lévy process has jumps, unlike the standard Browninan motion. As a result, characterization of the corresponding distribution becomes complicated.

The natural temptation 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 offered 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.
The convergence rate of this estimator 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). To give a concrete example, we can consider the so-called Hill estimator for
the tail index of the distribution of $W$. This estimator is based on $k$ highest-order statistics of $W$:
$W_{(1)} \geq W_{(2)} \geq \ldots W_{(n)}$. It is constructed as

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{W_{(i)}}{W_{(k+1)}}.$$ 

This estimator has poor small sample properties and its convergence, in general, requires random
centering.

**THEOREM 7** Suppose that the distribution of $W$ is regularly varying with the tail index $\alpha$, then

$$\sqrt{k} \left( H_{k,n} - \int_{W_{(k)}}^{\infty} \frac{n}{k} F_W(s) \frac{ds}{s} \right) \Rightarrow \frac{1}{\alpha} \int_{0}^{1} W(s) \frac{ds}{s}$$ 

Thus, this estimator, first of all, requires random re-centering that is based on the sample order
statistic $k$. Second, its convergence rate is determined by the number of the order statistic selected
and would require conditions beyond $k/n \to \infty$.

This demonstrates 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. One particular result that is relevant in
this context is the performance of the estimator (2.3) in case where the model exhibits the selection
on unobservables. In that case estimator (2.3) is biased and one may attempt to still use it but then
correct the bias. The following theorem establishes the fact that bias correction is infeasible: bias
cannot be estimated at the rate that is faster than the required $n^{-1/4}$.

**THEOREM 8** Suppose for $\theta = E[W]$

$$Pr\left( \{ \omega : \theta - E[W | Z(\omega)] \neq 0 \} \right) > 0.$$ 

Then the estimator (2.3) is biased. Then, unless $\phi(t) = |t| Pr \left( f_z(Z) \left| \frac{\partial E[Y | Z]}{\partial z} \right|^{-1} < |t|^{-1} \right)$ is non-
increasing in $t$ for all $|t| > C$ for some large constant $C$ and $\phi \in RV_{-\gamma}$ and $\gamma$ is known a priori, the
bias of estimator (2.3) cannot be estimated at the rate faster than $n^{-1/4}$. Therefore, bias correction
is infeasible.

We have previously demonstrated a general failure of bootstrap for approximation of the distribution
of the estimator (2.3). We can further study the limit of the bootstrap procedure. Consider a
bootstrap procedure which takes the i.i.d. sample of variables $W_i = \frac{\partial E[Y_i|Z_i]}{f_Z(Z_i)} Y_i$. Then we take an array $\{(I_1^{(n)}, \ldots, I_n^{(n)}) , n \geq 1\}$ 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 $W_i^* = W_{I_i^{(n)}}$.

**THEOREM 9** Suppose that assumptions of Theorem 10 hold. Take $\eta_1, \ldots, \eta_n$ be i.i.d. Poisson random variables with parameter $\lambda = 1$ and $J_i$ be homogeneous unit rate Poisson points. Then defining the probability by the class of Poisson random measures over the set of all probability measures on the space of Radon measures on $[0, +\infty) \backslash \{0\}$, for each measurable subset $M$ in the set of Radon measures

$$P\left(\sum_{i=1}^{n} \delta W_i^*/a_n \in M\right) \Rightarrow P\left(\sum_{i=1}^{n} \eta_i \delta J_i^{1/(1+\gamma)} \in M\right),$$

as $n \to \infty$.

We have previously established the failure of the distribution of t-statistic to converge to a pivotal distribution for a general class of distributions satisfying Assumption 1. In the following result we demonstrate the case where the t-statistic will be asymptotically normal whenever the tail behavior of $W$ is sufficiently close to the case where $W$ has finite second moments.

**THEOREM 10** Suppose that Assumption 1 holds, function $\psi(t) = |t|^2 Pr\left(f_Z(Z) \left| \frac{\partial E[Y|Z]}{\partial z}\right|^{-1} < |t|^{-1}\right)$ is non-decreasing in $t$ for all $|t| > C$ for some large constant $C$ and $\psi(\cdot)$ is slowly varying at infinity. Then if the expectation $E[W] = 0$ then

$$\hat{T}_\theta = \frac{1}{n} \sum_{i=1}^{n} w_i \sqrt{\frac{1}{n} \sum_{i=1}^{n} w_i^2} \Rightarrow N(0, 1).$$

We note that this result is a pointwise result that is only valid when tails of distribution of $W$ are “sufficiently” close to the tails that guarantee the finite second moment. In other words, the trimmed variance $E\left[W^2 1\{|W| < w\}\right]$ can only be a function slowly varying at infinity (thus, it cannot diverge to infinity at the geometric rate or faster). Thus, we cannot consider $\gamma$ belonging to some small neighborhood of 1 to be able to apply Theorem 10. In other words, the distribution of the t-statistic does not converge uniformly to normal distribution for any $\delta > 0$:

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}} \sup_{\gamma \in [1-\delta, 1]} \left| Pr\left(\hat{T}_\theta \leq t\right) - \Phi(t)\right| \rightarrow 0,$$

where $\Phi(\cdot)$ is the standard normal cdf.
2.4 General properties for consistent estimators for the intercept

Although the importance transformation delivers a convenient approach to deliver a feasible consistent estimator for the intercept in the selection model, it is in general not obvious whether one can find a “better” estimator. The observable distribution of the data is fully characterized by distributions $Pr(Y \leq y, D = 1 | Z = z, X = x)$, $F_X(\cdot)$ 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 $\eta = (Pr(Y \leq y_1 | \quad Z = z, X = x), F_X(\cdot))$ 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 $\mathcal{H}$:

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

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

**THEOREM 11** Let $\hat{\theta}(\eta)$ be a uniformly consistent estimator for the intercept parameter and the pseudometric $\rho$ is dominated by an $L_\infty$ pseudometric. Then for any $\eta \in \mathcal{H}$, any $r_n$ such that $r_n/n \to 0$ and $r_n(\hat{\theta}(\eta) - \theta(\eta)) = O_p(1)$ (where the limit is allowed to be degenerate at zero), and any $\epsilon > 0$ and $\Delta > 0$ there exist $\eta' \in \mathcal{H}$ such that $\rho(\eta, \eta') < \epsilon$ and

$$Pr\left( r_n(\hat{\theta}(\eta') - \theta(\eta')) > \Delta \right) \to 1.$$

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.

In the next two corollaries we show that essentially there is no remedy from the observed non-uniform behavior.

**COROLLARY 1** Suppose that $S_n(\eta)$ is a statistic such that $\hat{\theta}(\eta)/S_n(\eta)$ converges in distribution to a non-degenerate limit $F_\eta(\cdot)$ for each $\eta \in \mathcal{H}$. Then for any $\eta \in \mathcal{H}$ and any $\epsilon > 0$ there exist $\eta' \in \mathcal{H}$ such that $\rho(\eta, \eta') < \epsilon$, set $A(\eta, \epsilon)$ and a constant $\Delta(\eta, \epsilon)$ so that

$$|F_\eta(A) - F_{\eta'}(A)| > \Delta.$$

This corollary states that any self-normalization will not lead to a construction of a pivotal statistic. In other words, the pivotization of any uniformly consistent estimator is impossible.
3 Locally uniform inference for the sample selection model

We note that in case where the model is compatible with selection on observables (the error terms are mean-independent in the main and selection equations) \( \hat{\theta}_0 \) 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 \( \hat{\theta} \) that uses a simple unweighted average of \( w_i \). We recall that \( \hat{\theta}_0 \) converges at a parametric \( \sqrt{n} \) rate while \( \hat{\theta} \) converges at a slow rate \( n^{\gamma/(1+\gamma)} / l(n) \) (where \( l(\cdot) \) is a function slowly varying at infinity).

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:

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

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 \( \theta \) 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 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 case where the density of the instrument \( Z \) has thin tails, the rate of convergence and the asymptotic distribution of the estimator \( \hat{\theta} = 1/n \sum_{i=1}^n w_i \) 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 \( \hat{\theta}_0 \)) 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 \( \hat{\theta} \).

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
Provided that we assume that the data are i.i.d. we can apply the standard Central Limit Theorem to establish the asymptotic normality. Then we 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. 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 $\hat{\theta}$. Second, given that we control the choice of contiguous heavy-tail distributions we can choose the family of contiguous distributions to be sufficiently simple and thus estimation of the asymptotic distribution of $\hat{\theta}$ 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 random variable $W = \frac{\partial E[Y | Z]}{f_Z(Z)}$, we concentrate on analyzing this distribution.

First of all, we introduce the class of distributions of $(Y, Z)$ that are compatible with asymptotic normality of estimator $\hat{\theta}$. 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 asymptotic normal distribution.

**DEFINITION 2** Suppose that the joint distribution $(Y, Z)$, denoted $F_{YZ}(\cdot, \cdot)$ is defined by model (2.1), where random elements satisfy Assumption 1. Define the class of distributions

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

Also define the class

$$N_2 = \{ F_{YZ}(\cdot, \cdot) : \sup_{\beta \in (0, +\infty)} E \left[ \left( \frac{Y \cdot \frac{\partial E[Y | Z]}{\partial z}}{f_Z(Z)} \right)^{\beta} \right] < \infty \} = 2.$$ 

The defined class of distributions $N$ is fundamental because it delivers the validity of the Central Limit Theorem. The class $N_2$ is on the boundary of $N$ in the sense that distributions in $N$ can be compatible with the second and higher finite moments of $W$ while for the distributions in $N_2$, the second moment is the highest moment that exists for $W$.

**LEMMA 1** Suppose that Assumption 1 is satisfied and $Pr \left( \left| \frac{\partial E[Y | Z]}{\partial z} / f_Z(Z) \right| > w \right)$ is regularly varying at infinity with tail index $-(1 + \gamma)$. Then, whenever $\gamma \geq 1$, distribution $F_{YZ}(\cdot, \cdot) \in N$. Moreover, if $\gamma = 1$ then $F_{YZ}(\cdot, \cdot) \in N_2$.

Now suppose that $F_{YZ}(\cdot, \cdot) \in N_2$. We consider the distribution of $W$, denoted $F_W(\cdot)$ implied by such a distribution $F_{YZ}(\cdot, \cdot)$. By definition of class $N_2$, 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, distribution $Z$ 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 $\int (|w|^{1+c} \text{sign}(w) f_W(w) \, dw)$ is a measure defined on Borel subsets of the real line for each $c \in [0, 1)$.\(^2\) Our further logic 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 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.

Distribution $F_W(\cdot)$ has tail index 2, while 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 $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} w_i$. 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 for $W$ using the density $f_W(\cdot)$ with the finite second moment up to normalization as:

$$f^n_W(w) = f_W(w) + \frac{h_c}{\rho_n} \left( s_W(w) - f_W(w) \right), \tag{3.8}$$

where $0 < h_c \leq 1$ and $h_c \to 0$ as $c \to 0$. Note that this requirement is imposed on $h_c$ to ensure that $f^n_W(\cdot)$ is a valid density and that it converges uniformly in $w$ and $n$ to $f_W(\cdot)$ as $c \to 0$.

\(^2\)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.
THEOREM 12 If random variable $W$ is distributed according to (3.8) 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, first of all, the convergence to 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 $\hat{\theta}$.

THEOREM 13 Suppose that Assumption 1 holds and we choose normalization $E [W] = 0$. Let $F_{\hat{T}}(w)$ 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

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

such that for some $\delta > 0$

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

We note the difference between this result and the distribution results usually obtained for the local asymptotics in the autoregressive time series models. While in the autoregressive models
statistic $\frac{1}{n} \sum_{i=1}^{n} w_i^2$ has asymptotic distribution which is determined by the integrated square of the process that drives statistic $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i$, in our case it is characterized by the sum of independent stable processes defined on the positive part of the real line. The denominator in this expression is represented by the process that has heavier tails than the numerator. Thus, we can establish that as $c \to 0$, then $T^c \Rightarrow N(0, 1)$

3.1 Approaches to inference

As we mentioned this 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. The studentized parameter will not be pivotal as we mentioned in the previous section.

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

THEOREM 14 (Politis, Romano, and Wolf 1999) Suppose that the tail index $1 + \gamma$ is fixed. The subsampling approximation $L_{n,b}^* = \frac{1}{N} \sum_{i=1}^{N} 1 \left\{ \sqrt{b} \left( \hat{\theta}_{n,b,i} - \bar{\theta} \right) / \hat{\sigma}_{n,b,i} \leq x \right\}$ 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 pivotized variable without requiring the estimation of the tail index. We can note though that the quality of subsampling approximation will deteriorate when the tail index $1 + \gamma$ approaches 1. The reason for that 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 allows one to construct such conservative bounds under local asymptotics.

THEOREM 15 Consider local asymptotics with a sequence of distributions (3.8). The subsampling approximation $L_{n,b}^* = \frac{1}{N} \sum_{i=1}^{N} 1 \left\{ \sqrt{b} \left( \hat{\theta}_{n,b,i} - \bar{\theta} \right) / \sigma \leq x \right\}$ 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 $\sigma^2$ of the normal component of the limiting distribution mixture.

The structure of the local distribution gives the idea for the 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 $\sigma$.

4 Simulation Results

In this section we explore some of the finite sample implications of the main theorems in this paper. 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.

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, $\pm$ 5 times the standard deviation of the estimator,

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.
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 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 $H_0: \alpha = 0.5$ 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 s 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 die sign 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 this is a difficult parameter to correct inference on as well as the invalidity of the bootstrap. Interestingly, things change in the opposite direction 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 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 Application

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 used 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 $\alpha_0$ is based on the moment condition:

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

(5.9)

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

To estimate $\alpha_0$, note the right hand side of the above equation can be estimated by

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

(5.10)

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 $\alpha_0$ is

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

(5.11)

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 $\alpha_0$ from the parametric Heckman model assuming bivariate normality of the unobserved disturbances. For the parametric estimator we also to create 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 $\alpha_0$ is three time slaggier 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.

To conduct inference, we use the limiting distribution theory established in this paper attained with our “bridging” asymptotics to allow for both selection on observables and selection on unobservables. We construct confidence intervals for the intercept based on three approximations of the limiting mixture distribution. First, we attain a confidence interval based only on the Gaussian component of the mixture distribution. This can be considered as the “least conservative” approach. Second, we consider a conservative approach where set the tail index parameter equal to 0 in our mixture distribution. Finally, we report the “in between” case where we use an estimator of the tail index parameter in the mixture distribution. We compare these results with those attained from the parametric specification in ((Mroz 1987)).

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 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: \((u, v, x_1, x_2) \rightarrow 0, 1\). 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{align*}
Y_1 &= 1\{X_1 + \alpha_1 Y_2 - U > 0\}, \\
Y_2 &= 1\{X_2 + \alpha_2 Y_1 - V > 0\},
\end{align*}

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 \(\alpha_1, \alpha_2\) can or cannot be
estimated at the parametric rate.

**ASSUMPTION 3** 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 \(\mathbb{R}^2\).

As noted in (Tamer 2003), the system of simultaneous discrete response equations (6.12) 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 \(\alpha_1\) and \(\alpha_2\) 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 \(\alpha_1\) and
\(\alpha_2\). The way to identify the parameters of interest \(\alpha_1\) and \(\alpha_2\) as proposed in (Tamer 2003) is to use
the asymptotic regions where the solution is unique, thus forming a system of asymptotic equations:

\begin{align*}
F_U(x_1 + \alpha_1) &= \lim_{x_2 \rightarrow +\infty} P(Y_1 = 1 \mid X_1, X_2), \\
F_V(x_2 + \alpha_2) &= \lim_{x_1 \rightarrow +\infty} P(Y_2 = 1 \mid X_1, X_2).
\end{align*}

Thus, under the mean normalization \(E[U] = E[V] = 0\) provided that Assumption 3 holds, we can
identify the parameters of interest through the explicit expressions

\begin{align*}
\alpha_1 &= \lim_{x_2 \rightarrow +\infty} \int_{-\infty}^{+\infty} x_1 \frac{\partial}{\partial x_1} P(Y_1 = 1 \mid x_1, x_2) \, dx_1, \\
\alpha_2 &= \lim_{x_1 \rightarrow +\infty} \int_{-\infty}^{+\infty} x_2 \frac{\partial}{\partial x_2} P(Y_2 = 1 \mid x_1, x_2) \, dx_2.
\end{align*}

This expression clearly 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 bootstrap 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 via the drifting sequence that converges to the distribution where the joint density is equal to the product of marginal densities.

The use of the importance sampling transformation allows us to replace the limits with integrals:

\[
\alpha_1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1 \frac{\partial^2}{\partial x_1 \partial x_2} P(Y_1 = 1 \mid x_1, x_2) \, dx_1 \, dx_2,
\]

\[
\alpha_2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_2 \frac{\partial^2}{\partial x_1 \partial x_2} P(Y_2 = 1 \mid x_1, x_2) \, dx_1 \, dx_2.
\]

Then the final expression takes the form

\[
\alpha_1 = E \left[ \frac{X_1 \frac{\partial^2}{\partial x_1 \partial x_2} P(Y_1 = 1 \mid X_1, X_2)}{f_{X_1, X_2}(X_1, X_2)} \right],
\]

\[
\alpha_2 = E \left[ \frac{X_2 \frac{\partial^2}{\partial x_1 \partial x_2} P(Y_2 = 1 \mid X_1, X_2)}{f_{X_1, X_2}(X_1, X_2)} \right].
\]

We note that this estimator has the same inverse density structure as the estimator (2.3) for the intercept in the sample selection model.

7 Conclusions

This paper 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 so that the data generating process can “bridge” different models, loosely analogous to the “local to unity” asymptotics in the unit roots literature. We deriving the limiting distribution theory which we show can be used to conduct uniformly valid inference for the parameters of interest. This method is illustrate in two very relevant models in microeconometrics- the sample selection model and the two player static game theory model.

The work here suggests areas for future research. Many other nonlinear models will fit into this framework, so we aim to suggest uniform inference procedures for other parameters of interest. For example the triangular and non triangular models studied in (Khan and Nekipelov 2011) were shown
to suffer from difficult identification and nonstandard asymptotics for the coefficients of discrete endogenous variables and these are the parameters of interest in both the treatment effects, and peer effects literatures. We aim to extend our uniform inference procedures in this paper to those models as well.

References


A Proof of Theorem 10

In the proof of Theorem 13 we demonstrate that if we define the process of partial sums

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

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

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

Applying the continuous mapping theorem, we conclude that

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

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

B Proof of Theorem 10

Consider function $H(w) = E \left[ W^2 1\{|W| \leq w\} \right]$. Provided that $\psi(t) = |t|^2 \Pr \left( f_Z(Z) \left| \frac{\partial E[Y|Z]}{\partial z} \right|^{-1} < |t|^{-1} \right)$
is slowly varying at infinity. In this case we can define function $H(w) = E \left[ W^2 1\{|W| \leq w\} \right]$ 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.
C Proof of Theorem 12

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-centering 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 $\eta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n W^c_i$, with $W^c_i$ being the i.i.d. copies of random variable $W^c$ with local distribution $E[\exp(it\eta_n)] = \prod_{i=1}^n \int \exp \left( iw_i t \sqrt{n} \right) f_W(w_i^c) \, dw_i^c = \left( 1 - \frac{t^2 \sigma^2}{2n} - \frac{1}{2} \frac{1}{n^{1/(1+c)}} \kappa^2 |t|^{2/(1+c)} + o(\rho^{-1}_{n} n^{-2/(1+c)}) \right)^n \exp \left( -\frac{t^2 \sigma^2}{2} \right) + o(1).

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

D Proof of Theorem 13

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 5, 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

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

where $\delta_x$ 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 \( U = [0, +\infty) \setminus \{0\} \) and the set of Radon point measures on \( A \) by \( M_r(A) \).

We consider the map \( m : M_r((0, +\infty) \times U) \to M_r([0, +\infty) \times [\epsilon, +\infty]) \), where \( \epsilon \) 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

\[
\sum_i \delta_{(\tau_i, t_i)} \mapsto \sum_{\tau_i \leq \cdot} J_i
\]

mapping from \( M_r([0, +\infty) \times U) \) 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

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

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

\[
\sum nt_{i=1} W_{i} a_n 1\{|W_i|/a_n > \epsilon\} \Rightarrow \sum_{t_i \leq t} j_i 1\{|j_i| > \epsilon\}, \quad t \in [0, 1]
\]

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

Also, by continuity of the summation functional

\[
\sum nt_{i=1} W_{i} a_n 1\{1 \geq |W_i|/a_n > \epsilon\} \Rightarrow \sum_{t_i \leq t} j_i 1\{1 \geq |j_i| > \epsilon\}, \quad t \in [0, 1]
\]

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

Taking expectations, we obtain that

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

Consider process of trimmed partial sums

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

By the previous results, we conclude that

\[
L_n^\epsilon(\cdot) \Rightarrow L_{1+\gamma}^\epsilon(\cdot),
\]
where $L'_{1+\gamma}(\cdot)$ is the “restricted” $1+\gamma$-stable Lévy process such that

$$L'_{1+\gamma}(t) = \sum_{j_i \leq t} j_i \mathbf{1}\{1 \geq |j_i| > \epsilon\} - t \int_{\epsilon < w < 1} w \nu_{1+\gamma}(dw).$$

Then, using the Itô representation of the Lévy process:

$$L'_{1+\gamma}(t) \to L_{1+\gamma}(t)$$

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

$$d_s\left(L'_{1+\gamma}(\cdot), L_{1+\gamma}(\cdot)\right) \to 0$$

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

$$L'_{1+\gamma}(\cdot) \Rightarrow L_{1+\gamma}(\cdot)$$

Consider the process or regular partial sums

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

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

$$\Pr\left( \sup_{t \in [0,1]} \|L'_n(t) - L_n(t)\| > \delta \right)$$

$$\leq \Pr\left( \sup_{t \in [0,1]} \left| \sum_{i=1}^n \left( \frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n < \epsilon\} - \mathbb{E}\left[ \frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n < \epsilon\} \right] \right) > \delta \right)$$

$$= \Pr\left( \max_{0 \leq k \leq n} \left| \sum_{i=1}^k \left( \frac{W_i}{a_n} \mathbf{1}\{|W_i|/a_n < \epsilon\} - \mathbb{E}\left[ \frac{W_i}{a_n} \mathbf{1}\{1 \geq |W_i|/a_n < \epsilon\} \right] \right) > \delta \right)$$

Applying Doob’s inequality, we conclude that

$$\Pr\left( \sup_{t \in [0,1]} \|L'_{1+\gamma}(t) - L_{1+\gamma}(t)\| > \delta \right)$$

$$\leq \frac{\mathbb{E}\left[ \mathbf{1}\{|W|/a_n < \epsilon\} \right]}{\delta^2}$$

Next we note that

$$\mathbb{E}\left[ \frac{W}{a_n} \mathbf{1}\{|W|/a_n < \epsilon\} \right] \to \int_{|w| \leq \epsilon} w^2 \nu_{1+\gamma}(dw) = O(\epsilon^{1-\gamma}) = o(1)$$
, 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
\[
\lim_{\epsilon \to \infty} \limsup_{n \to \infty} P (d_\epsilon (L_\epsilon^n(\cdot), L_n(\cdot)) > \delta) = 0.
\]
This leads us to conclusion that $L_n(\cdot) \Rightarrow L_{1+\gamma}(\cdot)$.

E Semiparametric efficiency bound for the model with selection on observables

We can construct a semiparametrically efficient estimator for $\theta$ from (2.2). Consider a projection of the conditional moment equation representing $\theta$ on $Z$ using function $M(\cdot)$. In this case $E[M(Z)] \theta = E \left[ M(Z) \frac{Y}{P(Z)} \right]$. Without loss of generality, we re-center the true $\theta$ at zero. Then we consider the likelihood
\[
\ell(y, d, z) = f_{Y|D,Z}(y|d, x)P(z)^d (1 - P(z))^{1-d} f_Z(z).
\]
We consider a smooth parameterization of the likelihood using a single-dimensional parameter $t$. Then the score of the parametric submodel can be expressed as
\[
S^t(y, d, z) = s^t_{Y|D,Z}(y|d, z) + \dot{P}_t \left( \frac{d}{P(z)} - \frac{1 - d}{1 - P(z)} \right) + s^t_Z(z).
\]
We now evaluate the derivative of the parameter of interest along the parameterization path denoting it $\dot{\theta}^t$:
\[
E[M(Z)] \dot{\theta}^t = E \left[ M(Z) \frac{Y}{P(Z)} s_{Y|D,Z}(Y|D, Z) \right] - E \left[ M(Z) \frac{Y}{P(Z)^2} \dot{P}_t \right].
\]
The efficient influence function for $\theta$, denoted $\Psi^\theta(\cdot, \cdot, \cdot)$, can be found as a solution of
\[
\dot{\theta}^t = E \left[ \Psi^\theta(\cdot, \cdot, \cdot) S^\theta Y, D, Z \right].
\]
Provided that $E[s^t_{Y|D,Z}(Y|D, Z)|D, Z] = 0$ and $E[D - P(Z)| Z] = 0$, we notice that the solution of interest is
\[
\Psi^\theta(Y, D, Z) = E[M(Z)]^{-1} M(Z) \frac{Y}{P(Z)}.
\]
The variance of the efficient influence function is
\[
\text{Var}(\Psi^\theta(Y, D, Z)) = E[M(Z)]^{-2} E \left[ M^2(Z) \frac{E[Y^2|D = 1, Z]}{P(Z)} \right].
\]
Now we recall that we chose function $M(\cdot)$ to be arbitrary. Now we can minimize the variance by appropriately choosing $M(\cdot)$. This provides the information for parameter of interest expressed as

$$I_\theta = \inf_{M \in L^2(X)} \text{Var} \left( \Psi_\theta (Y, D, Z) \right) = E \left[ \frac{P(Z)}{E[Y^2|D = 1, Z]} \right]^{-1},$$

with the “efficient instrument” $M(z) = \frac{P(z)}{E[Y^2|D = 1, Z = z]}$. Thus, the semiparametrically efficient estimator for $\theta$ can be written as

$$\hat{\theta} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{P}(z_i)}{E[Y^2|D = 1, Z = z_i]} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{E[Y^2|D = 1, Z = z_i]}.$$

This means that

$$\sqrt{n} \hat{I}_\theta \hat{\theta} \xrightarrow{d} N(0, 1).$$

Therefore, in 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.

**F Efficiency bound for the uniform ATE estimator**

$$\ell(y, d, x) = f_{Y|D,X}(y|d, x)P(x)^d (1 - P(x))^{1-d} f_X(x).$$

We consider a smooth parameterization of the likelihood using a single-dimensional parameter $\theta$. Then the score of the parametric submodel can be expressed as

$$S^\theta(y, d, x) = s_{Y|D,X}^\theta(y|d, x) + \hat{P}^\theta \left( \frac{d}{P(x)} - \frac{1-d}{1-P(x)} \right) + s_X^\theta(x).$$

As it is more convenient to work with an unconditional moment model, we transform the model for the original parameter by projecting it on an arbitrary fixed measurable function $M(\cdot)$ to obtain

$$E \left[ M(X) \left( \frac{YD}{P(X)} - \frac{Y(1-D)}{1-P(X)} - \alpha_0 \right) \right] = 0.$$

We now evaluate the derivative of the parameter of interest along the parameterization path denoting it $\dot{\alpha}_0$:

$$E \left[ M(X) \left( \frac{YD}{P(X)} - \frac{Y(1-D)}{1-P(X)} - \alpha_0 \right) \right] s_{Y|D,X}(Y|D, X) = 0.$$

The efficient influence function for $\alpha_0$, denoted $\Psi_0(\cdot, \cdot, \cdot)$, can be found as a solution of

$$\dot{\alpha}_0 = E \left[ \Psi_0(Y, D, X) S^\theta(Y, D, X) \right].$$
Provided that $E[s_{Y|D,X}^0(Y|D,X)|D,X] = 0$ and $E[D - \mathcal{P}(X) | X] = 0$, we notice that the solution of interest is
\[
\Psi_0^\theta(Y,D,X) = E [M(X)]^{-1} M(X) \left( \frac{YD}{\mathcal{P}(X)} - \frac{Y(1-D)}{1-\mathcal{P}(X)} - \alpha_0 \right).
\]

The variance of the efficient influence function is
\[
\text{Var} \left( \Psi_0^\theta(Y,D,X) \right) = E [M(X)]^{-2} E \left[ M^2(X) \left( \frac{E[Y^2|D = 1,X]}{\mathcal{P}(X)} + \frac{E[Y^2|D = 0,X]}{1-\mathcal{P}(X)} - \alpha_0^2 \right) \right].
\]

Now we recall that we chose function $M(\cdot)$ to be arbitrary. Now we can minimize the variance by appropriately choosing $M(\cdot)$.

\[
P^{\alpha_0} = \left( \inf_{M \in L_2(X)} \text{Var} \left( \Psi_0^\theta(Y,D,X) \right) \right)^{-1} = E \left[ \left( \frac{E[Y^2|D = 1,X]}{\mathcal{P}(X)} + \frac{E[Y^2|D = 0,X]}{1-\mathcal{P}(X)} - \alpha_0^2 \right)^{-1} \right].
\]
### TABLE 1
Design 1: normal instruments, fixed parameters

<table>
<thead>
<tr>
<th>n</th>
<th>Mean Bias</th>
<th>Median Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 200</td>
</tr>
<tr>
<td>c = 0.00</td>
<td>-0.1983</td>
<td>-0.1374</td>
</tr>
<tr>
<td>c = 0.50</td>
<td>-0.0805</td>
<td>0.0074</td>
</tr>
<tr>
<td>c = 0.75</td>
<td>0.0986</td>
<td>0.1790</td>
</tr>
<tr>
<td>c = 1.00</td>
<td>0.2057</td>
<td>0.2835</td>
</tr>
</tbody>
</table>

### TABLE 2
Design 1: normal instruments, drifting parameters

<table>
<thead>
<tr>
<th>n</th>
<th>Mean Bias</th>
<th>Median Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 200</td>
</tr>
<tr>
<td>c = 0.00</td>
<td>38.3796</td>
<td>27.5744</td>
</tr>
<tr>
<td>c = 0.50</td>
<td>38.7891</td>
<td>27.5222</td>
</tr>
<tr>
<td>c = 0.75</td>
<td>36.9827</td>
<td>27.4332</td>
</tr>
<tr>
<td>c = 1.00</td>
<td>38.8312</td>
<td>27.5112</td>
</tr>
</tbody>
</table>

### TABLE 3
Design 3: cauchy instruments, constant parameters

<table>
<thead>
<tr>
<th>n</th>
<th>Mean Bias</th>
<th>Median Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 200</td>
</tr>
<tr>
<td>c = 0.00</td>
<td>0.2714</td>
<td>1.4475</td>
</tr>
<tr>
<td>c = 0.50</td>
<td>0.4334</td>
<td>1.6107</td>
</tr>
<tr>
<td>c = 0.75</td>
<td>0.5207</td>
<td>1.7115</td>
</tr>
<tr>
<td>c = 1.00</td>
<td>0.5971</td>
<td>1.8006</td>
</tr>
</tbody>
</table>

### TABLE 4
Design 3: cauchy instruments, drifting parameters

<table>
<thead>
<tr>
<th>n</th>
<th>Mean Bias</th>
<th>Median Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 200</td>
</tr>
<tr>
<td>c = 0.00</td>
<td>0.4116</td>
<td>0.5118</td>
</tr>
<tr>
<td>c = 0.50</td>
<td>0.4730</td>
<td>0.5254</td>
</tr>
<tr>
<td>c = 0.75</td>
<td>0.4384</td>
<td>0.5750</td>
</tr>
<tr>
<td>c = 1.00</td>
<td>0.4751</td>
<td>0.5766</td>
</tr>
</tbody>
</table>

### TABLE 5
Design 3: cauchy instruments, drifting parameters

<table>
<thead>
<tr>
<th>n</th>
<th>Mean Bias</th>
<th>Median Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 100</td>
<td>n = 200</td>
</tr>
<tr>
<td>c = 0.00</td>
<td>0.2484</td>
<td>0.3095</td>
</tr>
<tr>
<td>c = 0.50</td>
<td>0.2487</td>
<td>0.3092</td>
</tr>
<tr>
<td>c = 0.75</td>
<td>0.2487</td>
<td>0.3057</td>
</tr>
<tr>
<td>c = 1.00</td>
<td>0.2489</td>
<td>0.3059</td>
</tr>
</tbody>
</table>
### Table 5

Size and Power of the Bootstrap Coverage Probabilities

<table>
<thead>
<tr>
<th>Design 1: Bivariate normal, correlation $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: \alpha = 0$</td>
</tr>
<tr>
<td>OLS</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>$\rho = 0.00$</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
</tr>
<tr>
<td>$\rho = 0.75$</td>
</tr>
<tr>
<td>$\rho = 0.95$</td>
</tr>
</tbody>
</table>

### Table 6

Bootstrap Coverage Probabilities

Design 1: Bivariate normal, correlation $\rho$

<table>
<thead>
<tr>
<th>OLS</th>
<th>Heckman 2 Step</th>
<th>IVW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.00$</td>
<td>$1.00$</td>
<td>$1.00$</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td>$1.00$</td>
<td>$1.00$</td>
</tr>
<tr>
<td>$\rho = 0.75$</td>
<td>$1.00$</td>
<td>$1.00$</td>
</tr>
<tr>
<td>$\rho = 0.95$</td>
<td>$1.00$</td>
<td>$1.00$</td>
</tr>
</tbody>
</table>

Andrews and Schafagans

IVW

| $\rho = 0.00$ | $1.00$ | $1.00$ | $1.00$ |
| $\rho = 0.5$ | $1.00$ | $1.00$ | $1.00$ |
| $\rho = 0.75$ | $1.00$ | $1.00$ | $1.00$ |
| $\rho = 0.95$ | $1.00$ | $1.00$ | $1.00$ |

### Table 7

Bootstrap Coverage Probabilities

Design 1: Marginal Cauchy, Gaussian Copula correlation $\rho$

<table>
<thead>
<tr>
<th>OLS</th>
<th>Heckman 2 Step</th>
<th>IVW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.00$</td>
<td>$0.50$</td>
<td>$0.85$</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td>$0.86$</td>
<td>$0.83$</td>
</tr>
<tr>
<td>$\rho = 0.75$</td>
<td>$0.86$</td>
<td>$0.83$</td>
</tr>
<tr>
<td>$\rho = 0.95$</td>
<td>$0.85$</td>
<td>$0.84$</td>
</tr>
</tbody>
</table>

Andrews and Schafagans

IVW

| $\rho = 0.00$ | $0.50$ | $0.85$ | $0.85$ |
| $\rho = 0.5$ | $0.86$ | $0.83$ | $0.83$ |
| $\rho = 0.75$ | $0.86$ | $0.83$ | $0.87$ |
| $\rho = 0.95$ | $0.85$ | $0.84$ | $0.89$ |
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