On Uniform Inference in Nonlinear Models with Endogeneity

Shakeeb Khan
Boston College

Denis Nekipelov
University of Virginia

First Version: October 2014
This Version: September 2019

Abstract

(2009). In Sections 2 and 3 we show that the large sample behavior of existing inference methods vary discontinuously with the degree of selection on unobserved variables, with one extreme case being when selection is on observed variables only. As we will show this discontinuity results in impossibility results for valid uniform inference, and motivates our new inference procedures. Sections 4 and 5 explores the nite sample properties of our new

What complicates inference on for $_0$ is that how well one can estimate $_0$ depends on the type of selection in the model, something which is unknown to the econometrician. For example, if the selection in the model is on *observables only*, which corresponds to U;V being uncorrelated with each other, than $_0$ can be consistently estimated at the standard parametric rate by, for example OLS or WLS only using the observations where D = 1. However, both OLS and WLS will be *inconsistent* if there is any amount of *selection on unobservables*. An alternative estimator would be to take into account selection on unobservables. One such estimator is proposed in Heckman (1990) and Andrews and Schafgans (1998). We propose a di erent one in this paper that will be the basis of our inference procedure.

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

To introduce an inference procedure that allows for both types of selection we consider the

that U and V

Alternatively, we can write

$$_0 = \frac{E[Y \mid Z = z]}{P(z)} \quad E[U \mid V \quad z];$$

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

$$= \lim_{z! \ +1} E[Y \mid Z = z]:$$

We note that this expresses the parameter of interest in terms of the observable conditional

2.1 Conditionally exogenous selection

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

$$E[U D \mid Z = z] = E[U \mid V \quad z]:$$

Then we can write

$$E[Y \mid D = 1; Z = z] = \ _0 + E[U \mid D = 1; Z = z] = \ :$$

We note that conditioning on Z

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

$$E[Y \mid Z = z] = \beta_0 P(z)$$

can be rewritten as

$$\beta_0 f_V(z) = \frac{\partial E[Y \mid Z = z]}{\partial z};$$

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

$$\beta_0 = \frac{\frac{\partial E[Y \mid Z = z]}{\partial z}}{f_V(z)}:$$

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

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

$$\beta_0 = \lim_{z \to +\infty} E[Y \mid Z = z]:$$

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

$$\beta_0 = \lim_{z \to \infty} \int_{\underline{z}}^{z} \frac{\partial E[Y \mid Z = z]}{\partial z} dz:$$

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

$$\beta_0 = \int_{-\infty}^{+\infty} \frac{\partial E[Y \mid Z = z]}{\partial z} dz$$

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

$$\beta_0 = \int_{-\infty}^{+\infty} \frac{\partial E[Y \mid Z = z]}{\partial z} \frac{1}{f_Z(z)} f_Z(z)\, dz = E\left[\frac{\frac{\partial E[Y \mid Z]}{\partial z}}{f_Z(Z)}\right]:$$

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

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

$$b = \frac{1}{n} \sum_{i}^{n}$$

THEOREM 1 *Suppose that identi cation Assumption 1 holds and one uses the bootstrap sample* W_i

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

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

$$\hat{p} = s\frac{\frac{1}{n}\sum_{i=1}^{n} w_i}{\frac{1}{n}\sum_{i=1}^{n} w_i^2}$$

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

$$Pr\ \hat{p} \quad t \ \overset{F^k_{U;V;Z}}{!}\ F_k(t); \quad k = 1;2$$

and

$$\sup_{t2R} jF_1(t) \quad F_2(t)j > :$$

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

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

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

The absence of convergence in distribution of $\sqrt{n}$-normalized centered sample averages leads to the general absence of convergence in distribution for pivotized statistics. If F is the class of distributions satisfying Assumption 1 then in general the distribution of the t-statistic does not converge uniformly to normal distribution. In other words even if we found a candiate $F^0 2 F$ such that $\Pr \hat{p} \quad t \quad ! \quad (t)$ for each $t 2 R$ then for any > 0:

$$\lim_{n!1} \sup_{t2R} \sup_{kF \quad F^0 k_1 < ; F \ 2F} \Pr \hat{p} \quad t \quad (t) \ 6! \ 0;$$

where $( )$ is the standard normal cdf.

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

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:

$$\hat{b}^H = \begin{cases} \hat{b}; & \text{if } |\hat{b} - \hat{b}_0| > C/\sqrt{\bar{n}}; \\ \hat{b}_0; & \text{if } |\hat{b} - \hat{b}_0| \quad C/\sqrt{\bar{n}}: \end{cases}$$

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 nd another distribution that will be arbitrarily close to the original distribution in the L_2 norm de ned 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 under that distribution may be as slow as $\log n$ for some > 0. Moreover, the structure of the asymptotic distribution of the consistent estimator for these two close distributions of error terms is dramatically di erent: while it is normal in the model with selection on observables, it is may be represented by the distribution of a stable Levy process in the model with

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 β. Second, given that we control

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

$$f_W^c(w) = f_W(w) + \frac{h_c}{n}(s_W(w) - f_W(w)); \tag{3.1}$$

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

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

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

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

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

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

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

where $L^{+}_{1=(1+c)}(\ )$ *is the* $1=(1+c)$*-stable Levy process de ned on* $R_{+}\ n\ f0g$*: Then the distribution of random variable* T^{c} *uniformly approximates the distribution of the* t*-statistic*

$$\hat{p}^{c} = q\frac{\frac{1}{n}\sum_{i=1}^{n} w_i}{\frac{1}{n}\sum_{i=1}^{n} w_i^2};$$

such that for some > 0

$$\lim_{n!1} \sup_{c2\,p;1} \sup_{\ ]\ t2R} F_{\hat{p}^{c}}(t) \quad Pr(T^{c} \quad t) = 0:$$

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

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

3.1 Approaches to inference

As we mentioned previously, one of the di cult 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 con dence set for the parameter of interest. Politis, Romano, and Wolf

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

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

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

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

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

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

4 Simulation Results

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

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

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

As results in Tables 1-4 indicate, our nite 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 rInagm(3337s4-3344vR-42Dd44

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

Our approach here will be to use their estimates of these parameters combined with our density weighted estimator to estimate the intercept term. Speci cally, we will treat the 6 slope coe cients in the outcome equation as known (using the values attained in Ahn and Powell (1993)) for the coe cients 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 coe cients by estimating a Probit model. With these estimated coe cients, 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 $_0$ is based on the moment condition:

$$_0 + E[x_i^0 \ _0] = E_Z \left[\frac{\frac{d}{dz}E[y|z]}{f_Z(z)} \right] \tag{5.1}$$

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

To estimate $_0$, note the right hand side of the above equation can be estimated by

$$\hat{} = \frac{1}{n}\sum_{i=1}^{n} \hat{}^0(z_i)/\hat{f}(z_i) \tag{5.2}$$

where $\hat{}^0(z_i)$ 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 $_0$ is

$$\hat{} = \hat{} \ \frac{1}{n}\sum_{i=1}^{n} x_i^0 \hat{} \tag{5.3}$$

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

As a comparison, we estimated $_0$ from the parametric Heckman model assuming bivariate normality of the unobserved disturbances. For the parametric estimator we also bootstrap to create a histogram of the standardized estimator as well as quantile plots. Histograms and quantile plots are after the Appendix in Figure 4.

The attained results are interesting, notably that contrast between conclusions drawn from the parametric and semi parametric approaches. The parametric point estimator for $_0$ is three times larger in magnitude than the semi parametric point estimator, though both point estimates are positive. Exploring the bootstrapped con dence regions, the results from the two approaches are even more strikingly di erent. As the quantiles plots reveal, from the parametric approach the intercept is positive at all signi cant levels, whereas from the semi parametric quantile plot the intercept is not signi cantly di erent from 0 at most standard signi cance levels (0.025, 0.05,0.1). This demonstrates how sensitive the results can be to parametric assumptions.

6 Models with behavior similar to the sample selection model

While this paper has dealt exclusively with the di culties in conducting uniform inference for parameters of interest in the sample selection model. the same problems and di culties