AMERICAN SOCIOLOGICAL REVIEW

OFFICIAL JOURNAL OF THE AMERICAN SOCIOLOGICAL ASSOCIATION

ONLINE SUPPLEMENT
to article in
AMERICAN SOCIOLOGICAL REVIEW, 2017, VOL. 82

Culture Remains Elusive: On the Identification of Cultural Effects with Instrumental Variables

Winston Chou
Princeton University

Simulation Study

This simulation study illustrates the effect of collider bias on instrumental variables (IV) methods that analyze immigrants separately from non-immigrants. I focus on the SISTER method introduced by Polavieja (2015) in the *American Sociological Review*. Several scenarios are analyzed; however, across all scenarios, I assume the following model for an outcome *Y*

$$Y_i = -T_i + \varepsilon_i \quad (1)$$

with

$$\varepsilon_i \sim N(0,1) \quad (2)$$

and

$$T_i \sim \text{Bern}(\pi_i), \quad (3)$$

where the value of π_i varies depending on the scenario.

To illustrate the effect of collider bias, I further assume that migration is correlated with *Y* and *T*. In particular,

$$M_i \sim Bern(\mu_i), \quad (4)$$

where *M* indicates the decision to migrate,

$$\mu_i = \Phi(\varepsilon_i - \pi_i), \quad (5)$$

and Φ is the standard normal distribution function.

The behavioral interpretations for these equations are as follows. Equation 1 states that "traditional" individuals $(T = 1)$ will tend to have lower outcomes than non-traditional individuals. Equations 1, 3, 4, and 5 state that, *ceteris paribus*, individuals are more likely to migrate when they have higher values of the outcome (e.g., post-migration labor

force participation) and are not traditional. In other words, individuals select into migration on the basis of post-migration outcomes and their pre-migration values.

I examine the following scenarios. In Scenario 1, I show that collider bias can result in inconsistent estimates even when the researcher has an instrument that is exogenous in the population. This is because the instrument and the omitted variable jointly influence the propensity to migrate. In Scenario 2, I show that collider bias can result in inconsistent estimates even when the omitted variable affects the treatment for migrants only. In Scenario 3, I show that collider bias remains an issue when the instrument is imputed, as in the SISTER method.

Scenario 1

In the baseline scenario, I assume that π_i is a function of ε_i, so that *T* is endogenous. However, the researcher has a valid instrument *Z* (i.e., is exogenous, satisfies the exclusion restriction, and affects *T*). In particular,

$$Z_i \sim N(0,1), \tag{6}$$

and

$$\pi_i = \Phi(Z_i - \varepsilon_i). \tag{7}$$

I estimate four models in all scenarios. First, Equation 1 is estimated without ε_i by ordinary least squares (OLS):

$$\hat{Y} = \hat{\tau} T_i \tag{8}$$

Because of the omitted variable, the OLS estimate is inconsistent for $\tau = -1$.

Second, I estimate τ with two-stage least squares using *Z* as the instrument. The whole sample is used. This is the standard IV approach, and the IV estimate is consistent for τ.

Third, *and for migrants only*, I estimate τ with two-stage least squares using *Z* as the instrument. Due to collider bias, this IV estimate is inconsistent for τ.

Fourth, to illustrate that this is due to collider bias, I reestimate the IV model but allow *M* to be exogenously determined by letting $\mu_i = 0.5$ for all *i*. In other words, I resimulate the data but allow individuals to migrate completely at random. The IV estimate for τ from this model is also consistent.

The results, based on 4 × 5,000 Monte Carlo simulations, are shown in Figure S1, which plots the root-mean-squared error (RMSE) of the estimates of τ as the sample size grows

from 500 to 10,000. The RMSE of an unbiased and consistent estimator should converge to 0 as the sample size goes to infinity. Figure S1 also shows the coverage of the 95 percent confidence interval (CI) at these sample sizes. The true value of the parameter should be covered by the 95 percent CI in .95 of simulations.

As the red lines in Figure S1 show, the IV with migrants only under non-random migration is both inconsistent and has poor coverage even as the sample size grows to infinity. Note that these issues are not specific to SISTER, as I do not use the imputation step of SISTER until Scenario 3. Rather, these results illustrate some general implications of collider bias when analyzing migrants and non-migrants separately. Also note that, consistent with expectations, OLS with an omitted variable performs poorly, having high RMSE and low coverage even as $n \to \infty$.

On the other hand, IV with the full sample and under random migration is consistent, and the estimated confidence intervals cover the true value of $\tau = -1$ in the correct proportion of trials. The latter has higher variance at each sample size because half of the sample (the non-migrants) is discarded.

Scenario 2

In the second scenario, traditionalism is endogenous only for migrants. For precision, I introduce some additional notation. Let T_i^0 denote pre-migration traditionalism. It is distributed according to the Bernoulli distribution with parameter π_i^0. Then, as before, *Z* is exogenous:

$$Z_i \sim N(0,1), \tag{9}$$

while π_i^0 is determined by

$$\pi_i^0 = \Phi(Z_i). \tag{10}$$

However, after *M* is realized, the value of traditionalism, again denoted by *T*, is endogenous for migrants only. Specifically,

$$T_i \sim \text{Bern}(\pi_i), \tag{11}$$

and

$$\pi_i = \Phi(Z_i - M_i\varepsilon_i). \tag{12}$$

However, migrants' pre-migration values (T_i^0) are observed and can be used as an instrument for *T*.

For this scenario, I reestimate all the models but replace Z with T_i^0. As Figure S2 shows, with the exception of the IV regression that uses migrants only, the other estimators have lower RMSE relative to Scenario 1. This is because the treatment variable is endogenous for migrants only. Indeed, the OLS estimator has slightly better coverage when the sample size is small, which is intuitive because the confidence intervals are larger.

However, the problematic IV estimator does not improve; in fact, it has a slightly higher RMSE relative to Scenario 1. This is because the instrument is now a mapping from Z to $\{0,1\}$, and not Z itself.

Scenario 3

Finally, I examine the scenario in which T is endogenous only for migrants, but their pre-migration values must be estimated using observationally equivalent non-migrants. This scenario corresponds most closely to SISTER. The only observed covariate is Z, which satisfies the instrumental variables assumptions. As recommended by Polavieja (2015), migrants' pre-migration values are estimated using linear regression, that is,

$$\hat{T}_i^0 = \hat{\gamma} Z_i \qquad (13)$$

where $\hat{\gamma}_i^0$ is estimated from a linear regression of traditionalism on Z for migrants.

As with Scenarios 1 and 2, I estimate the following four models: a linear regression model, an IV regression with $\hat{T}_i^0$ as the instrument for the full sample, an IV regresion with $\hat{T}_i^0$ as the instrument for migrants only when migration is endogenous, and an IV regression with $\hat{T}_i^0$ as the instrument for migrants when migration is exogenous.

As Figure S3 shows, the SISTER estimator does not perform well. It has a slightly better RMSE relative to Scenario 2, as the instrument is now a bijective function of Z rather than a mapping to $\{0,1\}$. However, its RMSE is even higher than OLS and its coverage remains poor.

Concluding Remarks

This simulation study examines the statistical properties of SISTER and other IV methods that condition on migration outcomes. It shows that SISTER can produce biased and inconsistent estimates when individuals do not migrate at random.

Reference

Polavieja, Javier G. 2015. “Capturing Culture: A New Method to Estimate Exogenous Causal Effects using Migration Populations.” *American Sociological Review* 80(1):166–91.

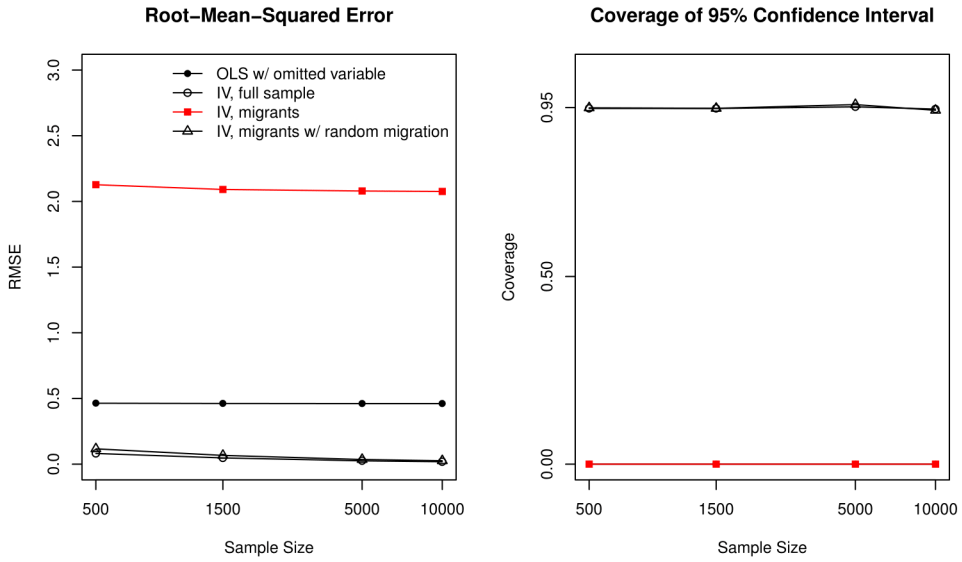

Figure S1. Conditioning on Migration Leads to Inconsistent Estimates and Misleading Confidence Intervals when Migration Is Nonrandom

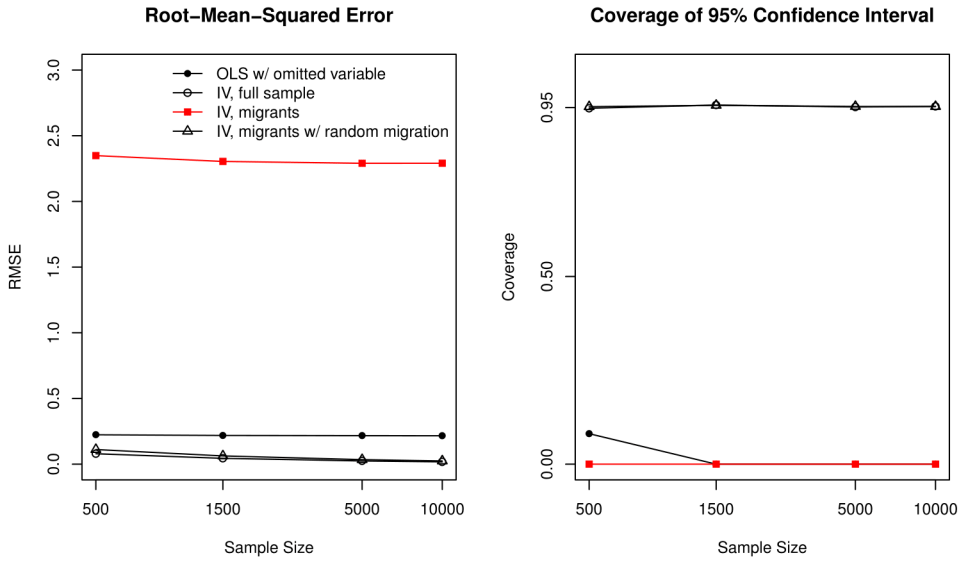

Figure S2. Estimates Remain Inconsistent when Traditionalism Is Endogenous for Migrants Only

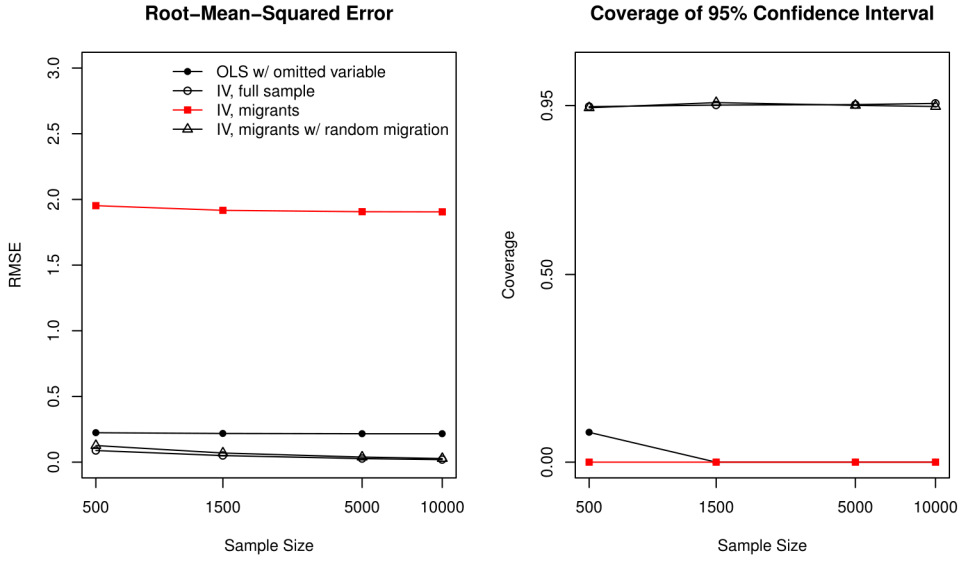

Figure S3. Estimates Remain Inconsistent when Traditionalism Is Imputed