Treatment Effect Estimation Using Nonlinear Two-Stage Instrumental Variable Estimators: Another Cautionary Note

Cole G. Chapman and John M. Brooks

Objective. To examine the settings of simulation evidence supporting use of nonlinear two-stage residual inclusion (2SRI) instrumental variable (IV) methods for estimating average treatment effects (ATE) using observational data and investigate potential bias of 2SRI across alternative scenarios of essential heterogeneity and uniqueness of marginal patients.

Study Design. Potential bias of linear and nonlinear IV methods for ATE and local average treatment effects (LATE) is assessed using simulation models with a binary outcome and binary endogenous treatment across settings varying by the relationship between treatment effectiveness and treatment choice.

Principal Findings. Results show that nonlinear 2SRI models produce estimates of ATE and LATE that are substantially biased when the relationships between treatment and outcome for marginal patients are unique from relationships for the full population. Bias of linear IV estimates for LATE was low across all scenarios.

Conclusions. Researchers are increasingly opting for nonlinear 2SRI to estimate treatment effects in models with binary and otherwise inherently nonlinear dependent variables, believing that it produces generally unbiased and consistent estimates. This research shows that positive properties of nonlinear 2SRI rely on assumptions about the relationships between treatment effect heterogeneity and choice.

Key Words. Instrumental variables, econometrics, applied methods, residual inclusion

---

The need to estimate treatment effect parameters using observational data is a core component of comparative effectiveness research. However, research using these data is often complicated by unmeasured variables related to selection into treatment, treatment effectiveness, and outcomes that bias direct estimation of treatment effects. Given a valid instrument, instrumental variable (IV) estimators offer a pathway to generating unbiased treatment effect estimates despite unmeasured confounders. Under a modest set of assumptions,

IV estimators such as linear two-stage least squares (2SLS) have been shown to yield consistent estimates of the local average treatment effect (LATE)—the average absolute treatment effect for patients whose treatment choices were influenced by their instrument value (Imbens and Angrist 1994; McClellan, McNeil, and Newhouse 1994; Angrist, Imbens, and Rubin 1996; Harris and Remler 1998; Newhouse and McClellan 1998; Brooks and Fang 2009). The patients in this subset are referred to as compliers, or marginal patients in the health IV literature (Imbens and Angrist 1994; McClellan, McNeil, and Newhouse 1994; Angrist, Imbens, and Rubin 1996; Harris and Remler 1998; Newhouse and McClellan 1998). While many have argued that LATE estimates have policy relevance in and of themselves (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996; Harris and Remler 1998; Angrist 2004; Brooks and Chrischilles 2007), others have bemoaned the limited ability to generalize IV estimates across a broader patient population (Heckman and Vytlacil 2005; Heckman, Urzua, and Vytlacil 2006; Basu et al. 2007).

It is well known that IV estimates can be generalized more broadly if treatment effectiveness is either homogeneous across patients or is heterogeneous across patients but unrelated to treatment choice (Heckman, Urzua, and Vytlacil 2006; Basu et al. 2007; Brooks and Fang 2009; Angrist and Fernandez-Val 2013). However, recent simulation research by Terza and colleagues has suggested that IV-based nonlinear two-stage residual inclusion (2SRI) estimators—a special case of control function methods (Hausman 1978; Garrido et al. 2012; Lee and Kim 2012)—can directly estimate the population average treatment effect (ATE) in models with binary, or otherwise “inherently nonlinear,” dependent variables without relying on these assumptions (Terza, Bradford, and Dismuke 2007; Terza, Basu, and Rathouz 2008). As a result, the authors of this research suggest that nonlinear 2SRI should always be employed for analysis of empirical models with inherently nonlinear dependent variables and unmeasured confounding risk (Terza, Basu, and Rathouz 2008). Their suggestion has taken hold; this work has been cited more than 200 times as the basis for choosing variants of 2SRI in research (Stuart, Doshi, and Terza 2009; Black, Spetz, and Harrington 2010; Gibson et al. 2010; Gore et al. 2010; Hadley et al. 2010; Trogon, Nurmagametov, and Thompson 2010; Fang et al. 2011; Baughman and Smith 2012; Hadley

and Reschovsky 2012; Li and Jensen 2012; Prada, Salkever, and MacKenzie 2012; Tan et al. 2012; Grabowski et al. 2013; Li et al. 2013).

Simulation evidence in Terza, Bradford, and Dismuke (2007) and Terza, Basu, and Rathouz (2008) demonstrated that nonlinear 2SRI *can* yield unbiased estimates of the ATE. However, 2SRI is analogous to 2SLS when estimated using linear models and the paper does not explain *why* the nonlinearity of 2SRI estimation, in and of itself, enables researchers to ignore assumptions previously required to identify concepts beyond the LATE using IV methods (Hausman 1978; Terza, Basu, and Rathouz 2008). It is possible that the simulation approaches used to demonstrate positive properties of nonlinear 2SRI created a unique set of simulated patients having characteristics aligned with unstated assumptions required to obtain ATE. If so, the simulation results may misguide researchers estimating effects in scenarios without this alignment.

In many real-world scenarios, the relationships between treatment and outcomes for marginal patients can be expected to differ from the remaining patients and the LATE will be distinct from the ATE. This is particularly relevant to settings of *essential heterogeneity* where treatment effects are heterogeneous across patients and treatment choices are related to this heterogeneity (Heckman, Urzua, and Vytlačil 2006; Basu et al. 2007). For example, Basu et al. (2007) reported strong evidence of essential heterogeneity when examining the choice of mastectomy or breast-conserving surgery plus radiation therapy (BCS+R) for breast cancer treatment (Basu et al. 2007). There likely exists patients for whom this treatment decision is very clear—patients who will always receive either mastectomy or BCS+R, regardless of their instrument value (e.g., region of the country, reimbursement levels). There are also likely to be patients for whom the relative benefits or harms are less clear and whose treatment choices will be more likely influenced by instrument values. These are the marginal patients. It seems implausible that average treatment effectiveness across the marginal patients in this scenario—for whom the best treatment is uncertain—will provide meaningful information on the effects of mastectomy versus BCS+R for patients considered well suited for these treatments. Rather, it could be expected that relationships between treatment and outcomes for the marginal patients would differ substantially from these other patients. Whether nonlinear 2SRI can yield consistent estimates of ATE in clinical scenarios in which the relationships between treatment and outcomes for marginal patients are unique from the general patient population remains to be shown.

In this article, we first describe the population characteristics underlying the binary treatment and outcome simulation model used by Terza, Bradford, and Dismuke (2007) with respect to essential heterogeneity and the *uniqueness* of marginal patients (Terza, Bradford, and Dismuke 2007). Uniqueness is defined as the extent to which the distribution of treatment effectiveness for the marginal patients differs from that of the general patient population. We then modify the Terza simulation approach to manipulate the extent of uniqueness. Finally, we examine the effects of gradually increasing the intensity of essential heterogeneity (and therefore uniqueness) through varying the importance of treatment effect heterogeneity relative to instruments in the treatment decision. The ability to generate unbiased estimates of ATE and LATE using nonlinear 2SRI estimators is evaluated in each scenario. We also generate LATE estimates from 2SLS for comparison.

Treatment Effect Heterogeneity and Choice: Essential Heterogeneity

Treatment effects can be heterogeneous with variables that are either related or unrelated to treatment choices, and these variables may be either measured or unmeasured by an analyst. Scenarios in which treatment effect heterogeneity is unrelated to treatment choice have been termed scenarios of *nonessential* heterogeneity (Basu et al. 2007). Under nonessential heterogeneity, treatment effectiveness varies across patients, but average treatment effects will not vary across treated and untreated subpopulations. Nonessential heterogeneity can occur, for example, if treatment is randomized across patients. If it can be argued that treatment effects are nonessential in a specific empirical scenario, then 2SLS will generate estimates of LATE that will be generalizable to the ATE (i.e., $LATE = ATE$) regardless of whether the variables driving heterogeneity are measured by the analyst.

Conversely, essential heterogeneity is said to exist in scenarios in which patients or providers are cognizant of factors associated with treatment effect heterogeneity and use this information when making treatment decisions (Basu et al. 2007). Under essential heterogeneity, if treatment effect heterogeneity across patients is fully explained by variables measured by the analyst, then this heterogeneity can be modeled directly and inferences can be made on population average treatment effects as well as average treatment effects across subgroups. However, if variables contributing to essential heterogeneity are unobserved by the analyst, then treatment choice will vary in unmeasured ways associated with treatment effectiveness and the analyst will be limited in the average treatment effect concepts that can be identified.

Instrumental variable estimators will generate consistent average treatment effect estimates specifically representing the subgroup of marginal patients whose treatment decisions were influenced by the instrument(s) in the model. The literature refers to this estimated effect as the LATE, or the average complier effect (ACE).

If factors driving essential heterogeneity are unmeasured by the analyst, then we theorize two general scenarios in which treatment effect estimates derived by IV estimators may be generalizable beyond marginal patients. We expect IV estimates will generalize to ATE if the analyst has an exceptionally strong instrument, such that all subjects in the sample are potential marginal patients. Such strong instruments will overwhelm any variation in treatment choice due to factors related to heterogeneity, making the scenario practically nonessential. This is, in essence, the premise on which randomized trials yield ATE estimates. Alternatively, if the instruments are not exceptionally strong but treatment effectiveness distributions for the marginal and the full populations happen to be very similar, then IV estimates can again be generalized to ATE.

METHODOLOGY

The simulation approach used by Terza, Bradford, and Dismuke (2007) uses two distinct equations to simulate treatment choice and outcome for individual patients (Terza, Bradford, and Dismuke 2007). Treatment choice (T_i), is a binary variable equaling 1 if a simulated patient's treatment value (T_i^*) is positive and 0 otherwise. Outcome (Y_i) is a binary variable representing whether an individual has been cured and is simulated conditional on each patient's treatment choice and other covariates. The treatment choice and outcome equations are linked through a set of factors (X_i) that are specified in both equations. Treatment effectiveness in this model is a nonlinear function of the X_i factors specified in the outcome equation and is heterogeneous across patients by their X_i characteristics. Treatment value, and in turn treatment choice, is determined by X_i factors and other factors (Z_i) that are unrelated to outcome and serve as the instruments in the model. This setting is characterized by essential heterogeneity because both treatment effectiveness and treatment choice are functions of X_i variables.

Our base simulation model (Scenario 1) recreates the binary model of Terza, Bradford, and Dismuke (2007), the details of which were provided by the authors in an appendix (Terza, Bradford, and Dismuke 2007). Our first

objective was to characterize this scenario with respect to essential heterogeneity and the uniqueness of the marginal patients relative to the full patient sample. In Scenarios 2 and 3, we made minor adjustments to coefficient values of the Terza model to change the uniqueness of marginal patients and assess robustness of IV estimators to these changes. The marginal population was made less unique in Scenario 2 and more unique in Scenario 3. We then created a final scenario (Scenario 4) that is based on Scenario 3 but varies the intensity of essential heterogeneity across subscenarios—by varying the instrument coefficients—to more clearly illustrate the relationship between uniqueness of marginal patients and bias of alternative estimators while holding all else constant.

For all scenarios, observed cure (Y_i) is a binary variable generated from an index function on an underlying continuous latent variable, Y_i^* . The model generating outcome is:

$$Y_i^* = \beta_T T_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (1)$$

$$Y_i = 1(Y_i^* > 0) = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \quad (2)$$

T_i is the observed binary treatment status of individual i . β_T , β_1 , and β_2 are parameters relating T_i , X_{1i} , and X_{2i} to latent outcome Y_i^* , respectively. ε_i is the value of a random disturbance term drawn from a standard normal distribution (i.e., $\varepsilon \sim N(0, 1)$).

The absolute effect of treatment on outcome (TE) is a nonlinear function of all factors affecting outcome. The error term in the outcome model is drawn from the standard normal distribution ($\varepsilon \sim N(0, 1)$) and the absolute effect of treatment on outcome for individual i is:

$$TE_i = \Phi(\beta_T + \beta_1 X_{1i} + \beta_2 X_{2i}) - \phi(\beta_1 X_{1i} + \beta_2 X_{2i}) \quad (3)$$

Φ denotes the standard normal distribution function. The absolute effect of treatment (TE_i) varies with X_{1i} and X_{2i} because ϕ is a nonlinear function. Full details regarding coefficient values and variable distributions used in the simulation models are available in the Appendix.

Treatment Choice Model

Treatment choice (T_i) is generated as the result of a cost–benefit decision by patients and providers where individuals are treated ($T_i = 1$) if the value of treatment, net costs, is positive. The underlying treatment value is represented

by T_i^* . The treatment choice model for all scenarios is illustrated by:

$$T_i^* = \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 Z_{1i} + \alpha_4 Z_{2i} \quad (4)$$

$$T_i = 1(T_i^* > 0) \equiv \begin{cases} 1 & \text{if } T_i^* > 0 \\ 0 & \text{if } T_i^* \leq 0 \end{cases} \quad (5)$$

Z_{1i} and Z_{2i} are instruments related to treatment choice but unrelated to outcome. α_1 , α_2 , α_3 , and α_4 are parameters relating X_{1i} , X_{2i} , Z_{1i} and Z_{2i} to treatment value T_i^* , respectively. TE_i does not affect treatment choice directly but varies across patients in a nonlinear manner related to T_i because both T_i and TE_i are functions of X_{1i} and X_{2i} . Essential heterogeneity exists because X_{1i} and X_{2i} directly affect both treatment choice and treatment effectiveness. The relationship between treatment effectiveness and treatment choice—and therefore the uniqueness of marginal patients—will depend upon the coefficient values in the vectors α and β .

It is not straightforward to assess and compare the intensity of essential heterogeneity across Scenarios 1–3 *a priori* because treatment effect heterogeneity does not affect treatment choice directly in the nonlinear model. Scenario 4 is intended to more clearly demonstrate relationships between intensity of essential heterogeneity, the uniqueness of marginal patients, and bias of treatment effect estimates. All else equal, increasing the influence of instruments on treatment choice will decrease the impact of factors related to treatment effectiveness on treatment choice and the intensity of essential heterogeneity will diminish. As essential heterogeneity is reduced, marginal patients will become less unique relative to the full population. Scenario 4 takes advantage of this and varies the strength of instruments across subsenarios where all other parameters are held constant (i.e., by varying α_3 and α_4 on Z_{1i} and Z_{2i} respectively). At larger values of α_3 and α_4 , instruments will become more significant drivers of treatment choice relative to X_i factors, heterogeneity will be less essential, and the marginal patients will be less unique from the full sample. F -statistics from testing exclusion of Z_1 and Z_2 from the treatment choice model will be calculated in each setting of Scenario 4 to demonstrate the instrument strength required for 2SRI estimates to yield consistent estimates of ATE in this model.

Treatment Effect Estimation

The mean effect of treatment on outcome is estimated in each scenario using nonlinear 2SRI models where Y_i , T_i , X_{1i} , Z_{1i} and Z_{2i} are measured and X_{2i} is

unmeasured. These 2SRI estimates are compared to estimates of true ATE and true LATE in each scenario. True values of ATE are calculated as the average causal effect of treatment across all observations, or

$$ATE = \frac{1}{N} \sum_{i=1}^N \Phi(\beta_T + \beta_1 X_{1i} + \beta_2 X_{2i}) - \Phi(\beta_1 X_{1i} + \beta_2 X_{2i}) = \frac{1}{N} \sum_{i=1}^N TE_i \quad (6)$$

In models with a single binary instrument, true LATE is the mean treatment effect across the subpopulation of compliers, or marginal patients, who are directly identifiable as those whose treatment choice changes with a discrete change in the instrument. Compliers are not identifiable in our models because Z_{1i} and Z_{2i} are continuous. As such, this traditional definition of true LATE is not directly calculable in our models. True LATE is estimated instead as the ratio of conditional covariances between the instrument and outcome, and the instrument and treatment choice. More specifically, true LATE for each simulated population is:

$$LATE = \frac{\text{Cov}(\mathbf{Z}, Y|\mathbf{X})}{\text{Cov}(\mathbf{Z}, T|\mathbf{X})} \quad (7)$$

where $\mathbf{Z}$ is the vector containing all $\mathbf{Z}_i = \alpha_3 Z_{1i} + \alpha_4 Z_{2i}$ and $\mathbf{X}$ is the matrix containing X_{1i} and X_{2i} (Angrist, Imbens, and Rubin 1996; Greene 2003; Angrist and Pischke 2008). Equations to calculate the covariances in equation (7) are available in the Appendix. The mean of equation (7) across all iterations of the simulation represents a consistent method of moments estimate of the true LATE in each scenario (Angrist, Imbens, and Rubin 1996; Greene 2003; Angrist and Pischke 2008).

In addition to the primary ATE and LATE measures, we also calculate the probability that each simulated patient will be a complier (marginal) to investigate the uniqueness of marginal patients through comparing treatment effect distributions for the full population and marginal patients in each scenario. The probability of being marginal ($P[Marginal]$) is approximated as the probability that treatment choice, given X_{1i} and X_{2i} , will change with random values of instruments Z_{1i} and Z_{2i} . The validity of this approach is assessed through a comparison of true LATE from equation (7) against an alternative estimate calculated as the $P[Marginal]$ -weighted mean of TE_i across observations. Further detail and discussion regarding approaches for calculating $P[Marginal]$ and LATE are in the Appendix.

The focus of this research is on assessing bias in average causal treatment effect (ATE) estimates generated by nonlinear 2SRI models across scenarios

varying the uniqueness of marginal patients. 2SRI is carried out by first estimating the treatment choice model as a function of measured covariates (X_{1i}) and instrumental variables (Z_{1i} and Z_{2i}). The residual term from this model ($u_i = T_i - \hat{T}_i$) is then included as a covariate in a regression of outcome on observed treatment choice (T_i) and X_{1i} . Absolute treatment effect estimates are generated by calculating the average change in predicted probability of outcome associated with a discrete change in observed treatment choice. The first and second stages of nonlinear 2SRI models are estimated using a probit model, the appropriate method for models of this form with error term $\varepsilon \sim N(0,1)$ and the functional form commonly applied by the developers of 2SRI (Terza 1999, 2002; Wooldridge 2002; Bhattacharya, Goldman, and McCaffrey 2006; Terza, Bradford, and Dismuke 2007; Stuart, Doshi, and Terza 2009). We also estimate 2SRI models using a first-stage linear probability model, as it has been suggested that treatment effect estimates generated using a two-stage approach with first-stage OLS model and a correctly specified nonlinear second-stage outcome model will be consistent regardless of whether the first-stage model is linear in truth (Angrist 2001; Angrist and Krueger 2001; Kaplan and Zhang 2012). ATE estimates are calculated as the average estimated effect of a discrete change in treatment status on outcome across all observations.

For comparison purposes, we also provide estimates of LATE generated by 2SLS models. 2SLS is the linear form of the general class of two-stage predictor substitution (2SPS) methods. For 2SLS, we first estimate the treatment choice model as described above for 2SRI using linear OLS. Predicted treatment ($\hat{T}_i$) from this first-stage regression is then substituted for the observed treatment choice variable in the second-stage linear OLS regression equation predicting outcome. The raw coefficient estimate on $\hat{T}$ ($\hat{\beta}_T$) is the absolute LATE estimate. 2SLS and 2SRI are analogs when 2SRI is estimated by OLS in both first- and second-stage regressions.

Final true and estimated values of ATE and LATE are generated for each scenario using Monte Carlo simulation methods—a common strategy in simulation studies examining properties of IV estimators (Bhattacharya, Goldman, and McCaffrey 2006; Terza, Bradford, and Dismuke 2007; Terza, Basu, and Rathouz 2008). Each simulation is completed with 1,000 iterations of 20,000 observations per iteration. Results from simulations run with 1,000 iterations of 500,000 observations are also reported to show sensitivity of results at larger sample sizes. *STATA 13* is used for simulating all data and for all analyses (StataCorp 2013). The primary statistic of interest for each scenario is the extent to which estimates generated by 2SRI and 2SLS are biased for esti-

mates of true ATE and true LATE. Bias is reported as the percentage difference between mean estimated and true values, calculated across the 1,000 iterations as:

$$\frac{\overline{Truth} - \overline{Estimate}}{\overline{Truth}} \times 100 \quad (8)$$

We also generate samples of 1,000,000 simulated observations to illustrate differences across scenarios with respect to summary statistics and the distributions of true treatment effectiveness for the marginal and full populations.

RESULTS

Table 1 shows mean approximated probability of being marginal (i.e., a complier), the true mean treatment effect, and the mean treatment value across true treatment effectiveness deciles for Scenarios 1–3 and two settings of Scenario 4 with low and high intensity of essential heterogeneity. Table 1 shows that the percentages of marginal patients in Scenario 1, using the parameters found in the Terza simulations, were fairly consistent across treatment effectiveness deciles. This consistency increased in Scenario 2, reflecting that marginal patients were less unique relative to the full sample in Scenario 1 than in Scenario 2. In contrast, the marginal patients in Scenario 3 were concentrated near the middle of the treatment effectiveness distribution relative to the full population, illustrating that marginal patients were more unique in Scenario 3 than in Scenarios 1 or 2 (Table 1).

Simulation results for Scenario 1 are provided in Table 2. True ATE was 0.425 and true LATE was 0.463. Nonlinear 2SRI estimates were slightly biased for true ATE (6.4 and 6.6 percent absolute bias for the linear-probit and probit-probit specification of 2SRI, respectively) but were substantially more biased for LATE (14.1–14.3 percent). 2SRI bias for ATE and LATE did not attenuate at larger sample size. 2SLS estimates were nearly unbiased for LATE (0.2 percent).

Simulation results for Scenario 2 are provided in Table 2. True ATE equaled 0.425 and true LATE equaled 0.402. Because marginal patients were less unique in Scenario 2 compared to Scenario 1, true ATE and true LATE were closer in value. Nonlinear 2SRI estimates of ATE were practically unbiased (0.6–2.1 percent) and, because ATE and LATE were fairly close in value, 2SRI bias for LATE was also low (3.4–5.0 percent). The linear-probit specification of 2SRI produced less biased estimates of both ATE and LATE, relative

Table 1: Distribution of Marginal Patients across Treatment Effectiveness Decile, All Scenarios ($N = 1,000,000$)

	Treatment Effectiveness Decile										Total
	1 (Low)	2	3	4	5	6	7	8	9	10 (High)	
Scenario 1											
$P[Marginal]^*$	.02	.14	.18	.16	.14	.13	.13	.14	.16	.18	.14
TE (mean)	.01	.07	.16	.27	.39	.51	.61	.70	.75	.78	.43
T* (mean)	1.40	.69	.33	.071	-.14	-.30	-.42	-.50	-.56	-.58	-.002
Scenario 2											
$P[Marginal]^*$	.15	.10	.10	.10	.10	.10	.10	.10	.10	.10	.11
TE (mean)	.01	.07	.15	.27	.39	.51	.61	.70	.75	.78	.42
T* (mean)	-.19	.19	.26	.21	.13	.02	-.08	-.14	-.2	-.22	-.002
Scenario 3											
$P[Marginal]^*$	.00	.00	.01	.09	.25	.34	.18	.07	.02	.0004	.10
TE (mean)	.01	.07	.16	.27	.39	.51	.61	.70	.75	.78	.43
T* (mean)	2.40	1.40	.81	.30	-.16	-.53	-.82	-1.00	-1.20	-1.20	-.002
Scenario 4—Low essential [†]											
$P[Marginal]^*$	.28	.35	.39	.42	.44	.45	.42	.40	.39	.38	.39
TE (mean)	.01	.07	.16	.27	.39	.51	.61	.7	.75	.78	.43
T* (mean)	2.40	1.40	.79	.32	-.17	-.52	-.81	-1.00	-1.20	-1.20	.003
Scenario 4—High essential [‡]											
$P[Marginal]^*$	.00	.00	.00	.00	.01	.17	.00	.00	.00	.00	0.02
TE (mean)	.01	.07	.16	.27	.39	.51	.61	.70	.75	.78	.43
T* (mean)	2.40	1.40	.81	.30	-.16	-.53	-.83	-1.00	-1.20	-1.20	-.002

*Defined as the mean approximated probability of being marginal of observations in each decile. Definition of $P[Marginal]$ described in more detail in the Appendix.

[†]Absolute value of IV coefficients (α_{IV}) = 2. F -Statistic for instrument exclusions = 16,189.

[‡]Absolute value of IV coefficients (α_{IV}) = 0.04. F -Statistic for instrument exclusions = 29.8.

Table 2: Average Percent Bias of Estimate, by Estimator and Sample Size, across 1,000 Simulated Samples

<i>Estimator</i>	<i>% Bias for ATE</i>		<i>% Bias for LATE</i>	
	<i>N = 20,000</i>	<i>N = 500,000</i>	<i>N = 20,000</i>	<i>N = 500,000</i>
Scenario 1				
2SLS	9.1	9.1	−0.2	−0.01
2SRI (Linear-Probit)	−6.6	−6.8	−14.3	−14.6
2SRI (Probit-Probit)	−6.4	−6.7	−14.1	−14.5
Scenario 2				
2SLS	−5.1	−5.1	0.1	−0.04
2SRI (Linear-Probit)	−0.6	−0.5	5.0	4.8
2SRI (Probit-Probit)	−2.1	−2.0	3.4	3.2
Scenario 3				
2SLS	15.6	15.4	0.2	0.03
2SRI (Linear-Probit)	−16.7	−17.2	−27.8	−28.2
2SRI (Probit-Probit)	−9.0	−9.3	−21.1	−21.4
Scenario 4—Low essential*				
2SLS	1.6	1.5	0.1	0.01
2SRI (Linear-Probit)	−1.1	−1.2	2.5	−2.6
2SRI (Probit-Probit)	−1.7	−1.8	3.1	−3.3
Scenario 4—High essential†				
2SLS	13.3	13.4	0.1	0.02
2SRI (Linear-Probit)	−13.7	−13.6	−23.7	−23.8
2SRI (Probit-Probit)	−9.0	−8.9	−19.5	−19.7

*Absolute value of IV coefficients (α_{IV}) = 2. *F*-Statistic for instrument exclusions = 16,189.

†Absolute value of IV coefficients (α_{IV}) = 0.04. *F*-Statistic for instrument exclusions = 29.8.

to the probit-probit specification. Estimates generated by 2SLS were nearly unbiased for LATE (0.1 percent).

Simulation results for Scenario 3 are provided in Table 2. True ATE equaled 0.425 and true LATE equaled 0.490. Both specifications of the nonlinear 2SRI estimator generated estimates of ATE that were substantially biased for true ATE (9.0–16.7 percent). The magnitude of bias did not decrease with larger sample size. Rather, bias increased slightly with the larger sample. 2SLS estimates were nearly unbiased for LATE (0.2 percent), while nonlinear 2SRI estimates were 21.1–27.8 percent biased for LATE. Bias of nonlinear 2SRI for LATE remained large regardless of sample size.

Effects of Directly Varying the Intensity of Essential Heterogeneity on Treatment Choice

Intensity of essential heterogeneity in Scenario 4 was modified by varying the coefficients on instruments across subscenarios. For convenience, let α_{IV}

denote the absolute value of instrument coefficients across subscenarios of Scenario 4 (i.e., $\alpha_{IV} = -\alpha_3 = \alpha_4$). As α_{IV} increased, the instruments explained a larger portion of the variation in treatment choice relative to other factors and marginal patients became less unique relative to the full population. Uniqueness of marginal patients across alternative subscenarios is demonstrated in Figure 1, which shows the distributions of treatment effects for the full population and marginal subpopulation at different levels of essentialness in Scenario 4.

Figure 1: Comparison of Treatment Effectiveness (TE) Distributions of Marginal and Full Populations Across Subscenarios of Scenario 4 Varying by Instrument Strength ($\alpha_{IV} = -\alpha_3 = \alpha_4$)

Simulation results for two subscenarios of Scenario 4 are provided in Table 2. True ATE equaled 0.425 across all subscenarios. True LATE equaled 0.480 when heterogeneity was very essential and marginal patients were very unique ($\alpha_{IV} = 0.04$; F-statistic = 29.8) and true LATE equaled 0.431 when heterogeneity was less essential and marginal patients were less unique ($\alpha_{IV} = 2$; F-statistic = 16,189). For the high essential subscenario, $\alpha_{IV} = 0.04$ was chosen because the F-statistic for the instruments was at a level considered not “weak” in the IV literature (Staiger and Stock 1997). For the low essential subscenario, $\alpha_{IV} = 2$ was chosen because it was the first scenario in which 2SRI estimates were practically unbiased for true ATE (i.e., percent bias <2 percent) and remained so for larger values of α_{IV} . Estimates generated by 2SRI were largely biased for both ATE (9.0–13.7 percent) and LATE (19.5–23.7 percent) when heterogeneity was very essential (i.e., $\alpha_{IV} = 0.04$). Bias of estimates generated by 2SRI estimators was low for both ATE (1.1–1.7 percent) and LATE (2.5–3.1 percent) when heterogeneity was less essential (i.e., $\alpha_{IV} = 2$).

Overall, 2SRI estimates became less biased as heterogeneity was made less essential in Scenario 4. 2SLS generated practically unbiased estimates of LATE across all subscenarios and unbiased estimates of ATE when heterogeneity was less essential. Figure 2A shows the bias of 2SRI and 2SLS estimates for ATE and LATE, and Figure 2B shows the F-Statistic associated with testing exclusion of instruments across all subscenarios varying α_{IV} .

Summary of Simulation Results

Nonlinear 2SRI generated practically unbiased estimates of the ATE only when uniqueness of marginal patients relative to the total population was very low with regard to their treatment effectiveness distributions. Our results also showed that nonlinear 2SRI estimates were similarly biased for LATE. On the other hand, 2SLS generated practically unbiased estimates of LATE in all scenarios. The magnitude of bias of linear-probit 2SRI estimates for ATE appeared positively correlated with the uniqueness of marginal patients and the related difference between estimates of true ATE and LATE across our simulation models. Figure 2A shows that bias of probit-probit 2SRI estimates for ATE was at first large and negative, then bias decreased toward 0 on the way to increasing again in the positive direction. We were not able to establish any certain expectations for direction of bias of nonlinear 2SRI estimates for ATE, only that the magnitude of bias generally decreased as marginal patients were less unique from the full population. Results were consistent when true LATE was calculated as the $P[Marginal]$ weighted mean treatment effect.

Figure 2: (A) Percent Bias of IV Estimates Generated by Nonlinear 2SRI and 2SLS for True ATE and True LATE across Subscenarios of Scenario 4 Varying by Instrument Strength ($\alpha_{IV} = -\alpha_3 = \alpha_4$). (B) F -Statistic from Testing Exclusion of Instruments (Z_1 and Z_2) from the Treatment Choice Model across Subscenarios of Scenario 4 Varying by Instrument Strength ($\alpha_{IV} = -\alpha_3 = \alpha_4$)

DISCUSSION AND CONCLUDING REMARKS

Taken together, our results support ideas that IV-based estimators only identify treatment effects for marginal patients. IV methods cannot be used to make more general inferences across the population without additional distributional or theoretical assumptions regarding the nature of treatment effect heterogeneity and treatment choice. Our results suggest that when unmeasured factors are contributing to essential heterogeneity, nonlinear 2SRI only yields unbiased estimates of the ATE (or LATE) if the distribution of treatment effects for the marginal population is not unique from that of the full population. This occurred in our simulated scenarios when instruments were extremely strong (IV exclusion F -statistic $>16,189$). This level of instrument strength is vastly beyond any current threshold for assessing a strong instrument in the literature. If, in a given clinical scenario, physicians routinely use their expertise to personalize care to patients in consideration of treatment effectiveness, then these assumptions likely would not apply (Harris and Remler 1998). For example, the frailest patients may not be recommended to receive invasive treatments no matter how close they live to a treatment facility because the expected benefit from treatment is too low when accounting for the risks related to frailty (McClellan, McNeil, and Newhouse 1994; Harris and Remler 1998). Consequently, when factors such as frailty are unmeasured in these clinical scenarios, IV estimators should not be expected to yield estimates that can be generalized to frail patients.

The results of this study contradict general statements that 2SLS produces inconsistent estimates when applied to models with inherently nonlinear dependent variables (Terza, Bradford, and Dismuke 2007; Bonsang 2009). Traditional 2SLS, but not nonlinear 2SRI, generated consistent estimates of LATE in all scenarios. This is not a surprising or new finding; it is well established that 2SLS generates consistent estimates of LATE with minimal assumptions. It may be surprising, however, that nonlinear 2SRI estimates were not unbiased or consistent for LATE. The value of LATE for inference, particularly under settings of essential heterogeneity, should not be underappreciated. Ideas of sorting-on-the-gain or passive personalization represent the artistry of medicine. In cases of essential heterogeneity, the patients for whom treatment decisions are least clear may also be those for whom naturally random factors unrelated to outcomes may affect treatment decisions. These marginal patients may be those whose treatment choices are sensitive to

policies seeking to influence treatment rates and therefore those for whom research efforts may be most pertinent and beneficial.

While specifying our simulation model to include only a single binary instrument would have allowed for direct identification of marginal patients (i.e., compliers) and a more straightforward calculation of LATE, we chose to remain consistent with the simulation models used by Terza. Estimates of true LATE generated by our primary definition (equation 7) were consistent with estimates generated as the probability weighted mean treatment effect. The finding that 2SLS estimates were approximately equal to our estimates of true LATE provides us with added confidence in our approach.

We show that the ability to generalize IV estimates beyond marginal patients requires acceptance of seemingly strong assumptions about the nature of treatment effect heterogeneity and choice. It is our hope that this research highlights the importance of theory related to treatment choice when estimating and interpreting average treatment effect concepts from observational data using IV estimators. Specifically, researchers should consider treatment effect heterogeneity and its relationship with treatment choice. Researchers seeking to generalize IV estimates should state and defend these assumptions based upon theory and evidence specific to their clinical setting. If these assumptions cannot be reasonably accepted, then it may not be possible to estimate ATE or other concepts beyond the LATE using IV-based estimators. The pursuit of innovative statistical methodologies furthering our ability to identify patient-centered treatment effect concepts from observational data is laudable and certainly a goal that we should strive for. However, researchers must resist using these methods for the *sake* of using them. Without full understanding and careful thought of the added strong assumptions underlying positive properties of these methods, it is all too easy to make potentially inappropriate inferences that can misguide patients, clinicians, or policy makers.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This research is based upon the graduate thesis of Cole G. Chapman, the results of which were presented at the 2014 ASHEcon meeting. This research was supported by funding received from the National Institutes of Health, National Institute on Aging (grant number RC4AG038635; PI John Brooks) for empirical research on effectiveness of antihypertensives after AMI in the Medicare population.

Disclosures: None.

Disclaimers: None.

REFERENCES

- Angrist, J. D. 2001. "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of Business & Economic Statistics* 19 (1): 173–175.
- . 2004. "Treatment Effect Heterogeneity in Theory and Practice*." *Economic Journal* 114 (494): C52–83.
- Angrist, J. D., and I. Fernandez-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics*, pp. 401–434. Cambridge, UK: Cambridge University Press.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.
- Angrist, J., and A. B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15 (4): 69–85.
- Angrist, J. D., and J. S. Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Basu, A., J. J. Heckman, S. Navarro-Lozano, and S. Urzua. 2007. "Use of Instrumental Variables in the Presence of Heterogeneity and Self-Selection: An Application to Treatments of Breast Cancer Patients." *Health Economics* 16 (11): 1133–57.
- Baughman, R. A., and K. E. Smith. 2012. "Labor Mobility of the Direct Care Workforce: Implications for the Provision of Long-Term Care." *Health Economics* 21 (12): 1402–15.
- Bhattacharya, J., D. Goldman, and D. McCaffrey. 2006. "Estimating Probit Models with Self-Selected Treatments." *Statistics in Medicine* 25 (3): 389–413.
- Black, L., J. Spetz, and C. Harrington. 2010. "Nurses Who Do Not Nurse: Factors That Predict Non-Nursing Work in the US Registered Nursing Labor Market." *Nursing Economics* 28 (4): 245–54.
- Bonsang, E. 2009. "Does Informal Care from Children to Their Elderly Parents Substitute for Formal Care in Europe?" *Journal of Health Economics* 28 (1): 143–54.
- Brooks, J. M., and E. A. Chrischilles. 2007. "Heterogeneity and the Interpretation of Treatment Effect Estimates from Risk Adjustment and Instrumental Variable Methods." *Medical Care* 45 (10): S123–30.
- Brooks, J. M., and G. Fang. 2009. "Interpreting Treatment-Effect Estimates with Heterogeneity and Choice: Simulation Model Results." *Clinical Therapeutics* 31 (4): 902–19.
- Fang, H., N. H. Miller, J. Rizzo, and R. Zeckhauser. 2011. "Demanding Customers: Consumerist Patients and Quality of Care." *B E Journal of Economic Analysis & Policy* 11 (1): 59.

- Garrido, M. M., P. Deb, J. F. Burgess Jr., and J. D. Penrod. 2012. "Choosing Models for Health Care Cost Analyses: Issues of Nonlinearity and Endogeneity." *Health Services Research* 47 (6): 2377–97.
- Gibson, T. B., X. Song, B. Alemayehu, S. S. Wang, J. L. Waddell, J. R. Bouchard, and F. Forma. 2010. "Cost Sharing, Adherence, and Health Outcomes in Patients with Diabetes." *American Journal of Managed Care* 16 (8): 589–600.
- Gore, J. L., M. S. Litwin, J. Lai, E. M. Yano, R. Madison, C. Setodji, J. L. Adams, C. S. Saigal, and Urologic Dis Amer Project. 2010. "Use of Radical Cystectomy for Patients with Invasive Bladder Cancer." *Journal of the National Cancer Institute* 102 (11): 802–11.
- Grabowski, D. C., Z. Feng, R. Hirth, M. Rahman, and V. Mor. 2013. "Effect of Nursing Home Ownership on the Quality of Post-Acute Care: An Instrumental Variables Approach." *Journal of Health Economics* 32 (1): 12–21.
- Greene, W. H. 2003. *Econometric Analysis*. Upper Saddle River, NJ: Pearson Education Inc.
- Hadley, J., and J. D. Reschovsky. 2012. "Medicare Spending, Mortality Rates, and Quality of Care." *International Journal of Health Care Finance & Economics* 12 (1): 87–105.
- Hadley, J., K. R. Yabroff, M. J. Barrett, D. F. Penson, C. S. Saigal, and A. L. Potosky. 2010. "Comparative Effectiveness of Prostate Cancer Treatments: Evaluating Statistical Adjustments for Confounding in Observational Data." *Journal of the National Cancer Institute* 102 (23): 1780–93.
- Harris, K. M., and D. K. Remler. 1998. "Who Is the Marginal Patient? Understanding Instrumental Variables Estimates of Treatment Effects." *Health Services Research* 33 (5 Pt 1): 1337.
- Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251–71.
- Heckman, J. J., S. Urzua, and E. Vytlačil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88 (3): 389–432.
- Heckman, J. J., and E. Vytlačil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation 1." *Econometrica* 73 (3): 669–738.
- Imbens, G. W., and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.
- Kaplan, C., and Y. Zhang. 2012. "Assessing the Comparative-Effectiveness of Antidepressants Commonly Prescribed for Depression in the US Medicare Population." *Journal of Mental Health Policy and Economics* 15 (4): 171–8.
- Lee, M., and Y. Kim. 2012. "Zero-Inflated Endogenous Count in Censored Model: Effects of Informal Family Care on Formal Health Care." *Health Economics* 21 (9): 1119–33.
- Li, Y., and G. A. Jensen. 2012. "Effects of Drinking on Hospital Stays and Emergency Room Visits among Older Adults." *Journal of Aging and Health* 24 (1): 67–91.
- Li, Y., X. Cai, D. B. Mukamel, and P. Cram. 2013. "Impact of Length of Stay after Coronary Bypass Surgery on Short-Term Readmission Rate an Instrumental Variable Analysis." *Medical Care* 51 (1): 45–51.

- McClellan, M., B. J. McNeil, and J. P. Newhouse. 1994. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association* 272 (11): 859–66.
- Newhouse, J. P., and M. McClellan. 1998. "Econometrics in Outcomes Research: The Use of Instrumental Variables." *Annual Review of Public Health* 19: 17–34.
- Prada, S. I., D. Salkever, and E. J. MacKenzie. 2012. "Level-I Trauma Center Effects on Return-to-Work Outcomes." *Evaluation Review* 36 (2): 133–64.
- Staiger, D., and J. H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557.
- StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.
- Stuart, B. C., J. A. Doshi, and J. V. Terza. 2009. "Assessing the Impact of Drug Use on Hospital Costs." *Health Services Research* 44 (1): 128–44.
- Tan, H., E. C. Norton, Z. Ye, K. S. Hafez, J. L. Gore, and D. C. Miller. 2012. "Long-Term Survival Following Partial vs Radical Nephrectomy among Older Patients with Early-Stage Kidney Cancer." *Journal of the American Medical Association* 307 (15): 1629–35.
- Terza, J. 1999. "Estimating Endogenous Treatment Effects in Retrospective Data Analysis." *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research* 2 (6): 429–34.
- Terza, J. V. 2002. "Alcohol Abuse and Employment: A Second Look." *Journal of Applied Econometrics* 17 (4): 393–404.
- Terza, J. V., A. Basu, and P. J. Rathouz. 2008. "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics* 27 (3): 531–43.
- Terza, J. V., W. D. Bradford, and C. E. Dismuke. 2007. "The Use of Linear Instrumental Variables Methods in Health Services Research and Health Economics: A Cautionary Note." *Health Services Research* 43 (3): 1102–20.
- Trogon, J. G., T. A. Nurmagambetov, and H. F. Thompson. 2010. "The Economic Implications of Influenza Vaccination for Adults with Asthma." *American Journal of Preventive Medicine* 39 (5): 403–10.
- Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Data S1. Simulation Methodology Details.

Table A1. Parameter and Variable Definitions for Simulation Scenarios.