HSR Health Services Research

DOI: 10.1111/1475-6773.12714
METHODS ARTICLE

Two-Stage Residual Inclusion Estimation in Health Services Research and Health Economics

Joseph V. Terza

Objectives. Empirical analyses in health services research and health economics often require implementation of nonlinear models whose regressors include one or more endogenous variables—regressors that are correlated with the unobserved random component of the model. In such cases, implementation of conventional regression methods that ignore endogeneity will likely produce results that are biased and not causally interpretable. Terza et al. (2008) discuss a relatively simple estimation method that avoids endogeneity bias and is applicable in a wide variety of nonlinear regression contexts. They call this method two-stage residual inclusion (2SRI). In the present paper, I offer a 2SRI how-to guide for practitioners and a step-by-step protocol that can be implemented with any of the popular statistical or econometric software packages.

Study Design. We introduce the protocol and its Stata implementation in the context of a real data example. Implementation of 2SRI for a very broad class of nonlinear models is then discussed. Additional examples are given.

Empirical Application. We analyze cigarette smoking as a determinant of infant birthweight using data from Mullahy (1997).

Conclusion. It is hoped that the discussion will serve as a practical guide to implementation of the 2SRI protocol for applied researchers.

Key Words. Endogeneity, instrumental variables, causal interpretability, estimation protocol, computer implementation

Empirical analyses in health services research and health economics often require implementation of nonlinear models whose regressors include one or more *endogenous variables*—regressors that are correlated with the unobserved random components of the model. Failure to account for such correlation leads to biased estimation results that are not causally interpretable. Terza, Basu, and Rathouz (2008) discuss a relatively simple estimation method that avoids endogeneity bias and is applicable in a wide variety of nonlinear regression contexts. They call this method two-stage residual inclusion (2SRI). This study focuses on the practical aspects of 2SRI implementation.

The discussion begins with an example, by way of reviewing the 2SRI protocol. We revisit Mullahy's (1997) model of prenatal smoking and infant birthweight. He estimated the model using the generalized method of moments (GMM); we re-estimate the model with 2SRI implemented in Stata/Mata 14. In this example, both stages of the model are specified as exponential regressions in keeping with the non-negativity of the outcome ($Y \equiv$ birthweight) and the endogenous variable ($X_e \equiv$ cigarette smoking by the mother). We show that the 2SRI protocol can be easily implemented using packaged Stata commands. We also outline how asymptotically correct standard errors (ACSE) for the 2SRI parameter estimates can be calculated. Analytic details and requisite Stata code for the ACSE are detailed in Appendix SA1. We extend the discussion to a very general version 2SRI framework. In this context, as in the birthweight example, we note that the 2SRI protocol can be easily applied via packaged Stata commands that implement either nonlinear least squares (NLS) or maximum-likelihood (ML) methods. In particular, the discussion makes clear that NLS or ML can be used in any combination in the first and second stages of the 2SRI estimator. We also discuss the formulation and calculation of ACSE for the general 2SRI estimator. Details are given in Appendix SA2 along with a heuristic for practical implementation of the 2SRI estimation protocol in the general case. Examples of applications of the general 2SRI protocol are also provided therein. Corresponding Stata code for these examples are given in Appendices SA3–SA6. The final section summarizes and concludes.

TWO-STAGE RESIDUAL INCLUSION BY EXAMPLE

Consider the regression model of Mullahy (1997) in which the objective is to draw causal inferences regarding the effect of prenatal smoking (X_e) on infant birthweight (Y) while controlling for infant birth order (PARITY), race (WHITE), and sex (MALE). The regression model for the birthweight outcome that he proposed can be written as[1]

$$Y = \exp(X_e\beta_e + X_o\beta_o + X_u\beta_u) + \varepsilon \tag{1}$$

where X_u is a scalar representing unobservable variables that are potentially correlated with prenatal smoking (e.g., general "health mindedness" of the mother), ε is the regression error term, X_o = [PARITY WHITE MALE] is a row vector of regressors that are uncorrelated with X_u, and ε, and the βs are the

Address correspondence to Joseph V. Terza, Ph.D., Department of Economics, Indiana University Purdue University Indianapolis, Indianapolis, IN 46202; e-mail: jvterza@iupui.edu.

regression parameters.[2] At issue here is the fact that there exist unobservables (as captured by X_u) that are correlated with both $Y$and X_e. In other words, X_e is *endogenous*. Such endogeneity confounds the identification and estimation of the possible causal effect of prenatal smoking (or any of the other regressors in the model for that matter). If, for instance, the presence of X_u is ignored in applying a conventional regression method to (1), then the estimates of β_e and β_o will likely be biased because they will be picking up effects that should instead be attributed to X_u. Terza, Basu, and Rathouz (2008) discuss a method, which they call two-stage residual inclusion (2SRI), designed to correct for such endogeneity bias. They show that for a very broad class of nonlinear regression models [which subsumes (1) as a special case], 2SRI produces unbiased (consistent) parameter estimates. To apply 2SRI to (1), one must first specify an auxiliary regression model of the following form

$$X_e = \exp(W\alpha) + X_u \tag{2}$$

where α is a column vector of regression parameters, $W = [X_o\ W^+]$ and $W^+ =$ [EDFATHER EDMOTHER FAMINCOM CIGTAX] with

EDFATHER = paternal schooling in years
EDMOTHER = maternal schooling in years
FAMINCOME = family income
and
CIGTAX = cigarette tax.

Equation (2) formalizes the correlation between X_u and X_e—the essence of the endogeneity problem. The variables in W^+ are the *identifying instrumental variables* which, by definition, must satisfy the following three conditions: (1) they are correlated with neither X_u nor e; (2) they can be legitimately excluded from the outcome regression (1); and (3) they are strongly correlated with X_e. Under these assumptions, the relevant version of the 2SRI estimation protocol is as follows:

First Stage

To get a consistent estimate of α, apply NLS to (2). This can be accomplished with one line of computer code via the Stata "glm" command.[3] The residuals from this regression are

$$\hat{X}_u = X_e - \exp(W\hat{\alpha}) \tag{3}$$

where $\hat{\alpha}$ denotes the first-stage consistent estimate of α. The residuals (3) can be saved using the Stata "predict" postestimation command.[4]

Second Stage

To obtain a consistent estimate of $\beta' = [\beta_e \ \beta_o' \ \beta_u]$, apply NLS to (1) with X_u replaced by $\hat{X}_u$. This too can be accomplished with just one line of computer code via the Stata "glm" command.[5]

As is made clear by the present example, consistent estimation of the model parameters via the 2SRI method is very easy. The correct standard errors of the estimates (for use in confidence interval estimation and hypothesis testing) cannot, however, be obtained as direct regression outputs from a statistical package. Nonetheless, because the more popular computer packages offer matrix programming capability, calculating the correct standard errors typically requires only a modicum of additional coding.[6] There are three possible approaches to calculation of the corrected standard errors: (1) bootstrapping; (2) the resampling method proposed by Krinsky and Robb (1986, 1990) [KR]; and (3) ACSE derived from standard asymptotic theory. For detailed discussions, and pro/con evaluations, of the bootstrapping and KR methods, see Dowd, Greene, and Norton (2014).[7] In Appendix SA1, we show how the relatively simple general ACSE formulations suggested by Terza (2016b) can be implemented in Stata for the present example. In this illustration, the ACSE for the kth element of $\hat{\beta}$ is the square root of the kth diagonal element of the following matrix

$$B_1^{-1} \ B_2 \ V(\hat{\alpha}) \ B_2' \ B_1^{-1} + \ V(\hat{\beta}) \tag{4}$$

where $\hat{\alpha}$ and $\hat{\beta}' = [\hat{\beta}_e \ \hat{\beta}_o' \ \hat{\beta}_u]$ are the first- and second-stage 2SRI estimates; $V(\hat{\alpha})$ and $V(\hat{\beta})$ are the estimated variance–covariance matrices of the first- and second-stage 2SRI estimators of α and β, respectively, as output by Stata; and B_1 and B_2 are matrices that are functions of the observable data and the estimated parameters.[8] $V(\hat{\alpha})$ and $V(\hat{\beta})$ are routinely saved by Stata; B_1 and B_2 are not. Stata coding for the latter must be user supplied. In Appendix SA1, we detail the formulations of B_1 and B_2 for the present example and give the corresponding requisite Stata code. Confidence interval estimates and hypothesis tests for the kth element of β can be based on the following asymptotic "t-statistic"

$$\frac{\hat{\beta}(k) - \beta(k)}{\sqrt{\hat{D}(k)}} \tag{5}$$

where $\hat{\beta}(k)$ [$\beta(k)$] denotes the kth element of $\hat{\beta}$ [β] and $\hat{D}(k)$ denotes the kth diagonal element of (4).

I applied the 2SRI estimation protocol to the same dataset analyzed by Mullahy (1997) and obtained the estimates of α and β reported in Tables 1 and 2, respectively. The correct asymptotic t-statistics for the 2SRI estimate of β, reported in column 2 of Table 2, were calculated using (4) and (5). In Table 2, we also display Mullahy's GMM estimates and, as a baseline, we report the conventional NLS estimates that ignore potential endogeneity. As an indicator of the strength of the instrumental variables (i.e., the elements of W^{+}), we conducted a Wald test of their joint significance. The value of the chi-square test statistic is 49.33 so that the null hypothesis that their coefficients are jointly zero is roundly rejected at any reasonable level of significance. The second-stage 2SRI estimates shown in Table 2 (column 1) are virtually identical to Mullahy's GMM estimates (column 4), but the former, unlike the latter, provide a direct test of the endogeneity of the prenatal smoking variable via the asymptotic t-stat (5) for the coefficient of X_u [$\hat{\beta}_u = \hat{\beta}(5)$] with H_0:$\beta_u = \beta(5) = 0$. According to the results of this test, the exogeneity null hypothesis is rejected at nearly the 1% significance level. To obtain a sense of the bias from neglecting to take account of the two-stage nature of the estimator in the calculation of the asymptotic standard errors, in Table 2 (last column), we also display the "packaged" second-stage glm t-stats as reported in the Stata output. The mean absolute bias across these uncorrected asymptotic t-stats for the four control variables and X_u is nearly 9 percent.

THE GENERAL 2SRI FRAMEWORK

The framework underlying the above example generalizes to a very broad class of nonlinear models. The general forms of the outcome and

Table 1: 2SRI First-Stage Estimates

Variable	*Estimate*	*Asymp. t-stat*
PARITY	0.04	1.14
WHITE	0.28	0.86
MALE	0.15	−1.84
EDFATHER	−0.03	−3.34
EDMOTHER	−0.10	−2.65
FAMINCOM	−0.02	1.44
CIGTAX	0.02	5.60
Constant	2.04	0.56

n = 1,388.

auxiliary regressions exemplified in (1) and (2), respectively, can be defined based on minimal parametric (MP) regression structure [as in (1) and (2)] or they can be derived from more fully parametric (FP) assumptions. In the 2SRI framework, one can specify the outcome model [exemplified in (1)] as either:

$$Y = \mu(X_e, X_o, X_u; \beta) + e \text{ [minimally parametric (MP) specification]} \quad (6)$$

or

$$f(Y|X_e, X_o, X_u; \beta) \text{ [fully parametric (FP) specification]} \quad (7)$$

where $\mu(X_e, X_o, X_u; \beta)$ denotes the conditional mean of Y given X_e, X_o, and X_u; β is a vector of parameters; and $f(Y \mid X_e, X_o, X_u; \beta)$ is the conditional probability density function of Y given X_e, X_o, and X_u. Similarly, for the auxiliary regression, one can posit either:

$$X_e = r(W; \alpha) + X_u \text{ [MP specification]} \quad (8)$$

or

$$g(X_e|W; \alpha) \text{ [FP specification]} \quad (9)$$

where $r(W; \alpha)$ denotes the conditional mean of X_e given W and $g(X_e \mid W; \alpha)$ is conditional probability density function of X_e given W. Equation (8) [or (9)] formalizes the correlation between X_u and X_e which, as we saw in the above example, lies at the heart of the endogeneity problem. In the example discussed in the previous section, both the outcome and the auxiliary regression specifications were MP. Specifically, we had

$$\mu(X_e, X_o, X_u; \beta) = \exp(X_e\beta_e + X_o\beta_o + X_u\beta_u) \quad (10)$$

$$r(W; \alpha) = \exp(W\alpha). \quad (11)$$

Table 2: 2SRI Second Stage, GMM, and NLS Estimates

	2SRI			*GMM*		*NLS*	
Variable	*Estimate*	*Correct Asymp. t-stat*	*Uncorrected Asymp. t-stat*	*Estimate*	*Asymp. t-stat*	*Estimate*	*Asymp. t-stat*
CIGS	−0.01	−3.68	−4.08	−0.01	−3.46	0.00	−5.62
PARITY	0.02	3.18	3.41	0.02	3.33	0.01	2.99
WHITE	0.05	4.22	4.55	0.05	4.44	0.06	4.75
MALE	0.03	3.13	3.35	0.03	2.95	0.03	2.90
X_u	0.01	2.56	2.83	–	–	–	–
Constant	1.95	117.64	123.74	1.94	121.71	1.93	133.70

n = 1,388.

The general 2SRI protocol is as follows:

First Stage

Apply the appropriate NLS [or maximum-likelihood (ML)] estimator to (8) [or (9)] to obtain a consistent estimate of α.[9] This can usually be accomplished with one line of computer code in Stata. The residuals from this regression are

$$\hat{X}_u = X_e - r(W;\hat{\alpha}) \tag{12}$$

where $\hat{\alpha}$ denotes the first-stage consistent estimate of α. Note that the FP specification in (9) will always imply the existence of a regression specification akin to (8) from which residuals, as defined in (12), can be obtained. To complete the 2SRI first stage, save the residuals (12) using the appropriate Stata postestimation command.

Second Stage

To obtain a consistent estimate of β, apply the appropriate NLS [ML] estimator to (6) [(7)] with X_u replaced by $\hat{X}_u$.[10] This too can typically be accomplished with just one line of Stata code.[11]

Note that one can use any combination of MP/FP specifications for the first and second stages of the 2SRI estimator.

Here, as in the birthweight example, the second-stage standard errors as output by Stata are incorrect. As Terza (2016b) shows, the exact form of the ACSE depends on the estimation method used in the second stage of 2SRI—NLS vs. MLE. When NLS is used in the 2SRI second stage, the ACSE will be the square roots of the diagonal elements of an estimated variance–covariance matrix with a formulation akin to that of (4). On the other hand, if MLE is used in the second stage, the ACSE for the kth element of $\hat{\beta}$ is the square root of the kth diagonal element of a matrix of the following form

$$V(\hat{\beta})AV(\hat{\alpha})A'V(\hat{\beta}) + V(\hat{\beta}) \tag{13}$$

where A is a matrix that is formulated exclusively in terms of the observable data and the estimated parameters. As was the case for (4), $V(\hat{\alpha})$ and $V(\hat{\beta})$ are routine Stata postestimation outputs but the formulation of A, and its coding in Stata, must be user supplied. In Appendix SA2, we offer a heuristic for practical implementation of the 2SRI estimation protocol in the general case, complete with details on deriving and coding B_1 and B_2 (A) for MP (FP) outcome models for which the second-stage 2SRI estimator is NLS (ML).

SUMMARY AND DISCUSSION

We discuss practical aspects of the 2SRI method for consistent estimation of nonlinear models with endogenous regressors and illustrate its application in Stata for the case in which both X_e and Y are non-negative. The implementation of the 2SRI protocol is detailed in the context of this illustration and generalized to a very broad class of nonlinear applications. Details of the relevant mathematics and computer coding are given in the supplementary appendices. Therein, we also detail Stata/Mata applications of the protocol for four additional oft encountered model configurations involving binary and/or fractional X_e and Y. It is hoped that these additional examples will serve to demonstrate the ease with which the protocol can be extended to models involving other variable type configurations not explicitly covered here. In particular, the class of nonnegative-dependent variables encompasses important subtypes, for example, count variables, continuous variables whose support includes 0 (e.g., two-part models), and continuous variables for which 0 is excluded.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This research was supported by a grant from the Agency for Healthcare Research and Quality (R01 HS017434-01).

Disclosure: None.

Disclaimer: None.

NOTES

1. Mullahy does not explicitly specify the model in terms of the unobservable X_u. Nevertheless, (1) is substantively the same as Mullahy's model (see Terza [2006]).
2. When a is a row vector and b is a column vector, ab denotes their vector (or dot) product. For example, $X_o\beta_o$ denotes the dot product of X_o and the column vector of corresponding coefficient parameters for its elements, β_o.
3. See Appendix SA1.
4. See Appendix SA1.
5. See Appendix SA1.
6. "Mata" is the matrix programming option in Stata.
7. Dowd, Greene, and Norton (2014) also discuss the ASCE approach, but the formulation they offer (in particular, equation (18)) is based on an assumption that is seldom valid in empirical HSR. See Terza (2016a) for details.

8. C^{-1} and F′ denote the matrix inverse of the square matrix C and matrix transpose of the rectangular matrix F, respectively.
9. The first-stage ML estimator is the maximizer of $\sum_{i=1}^{n} \ln[g(X_{ei}|W_i;\alpha)]$ with respect to α where X_{ei} and W_i denote the observed values of X_e and W for the ith observation in the sample; and $i = 1, \ldots, n$.
10. The second-stage ML estimator is the maximizer of $\sum_{i=1}^{n} \ln[f(Y_i \mid X_{ei}, X_{oi}, \hat{X}_{ui};\beta)]$ with respect to β where Y_i and X_{oi} denote the observed values of Y and X_o for the ith observation in the sample; and $\hat{X}_{ui}$ is the first-stage residual for the ith observation in the sample.
11. See Appendix SA2 for generic computer code for this general 2SRI protocol. A variety of examples are also detailed therein.

REFERENCES

Dowd, B. E., W. H. Greene, and E. C. Norton. 2014. "Computation of Standard Errors." *Health Services Research* 49: 731–50.

Krinsky, I., and A. L. Robb. 1986. "On Approximating the Statistical Properties of Elasticities." *Review of Economics and Statistics* 68: 715–9.

———. 1990. "On Approximating the Statistical Properties of Elasticities: A Correction." *Review of Economics and Statistics* 72: 189–90.

Mullahy, J. 1997. "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior." *Review of Economics and Statistics* 79: 586–93.

Terza, J. V. 2006. "Estimation of Policy Effects Using Parametric Nonlinear Models: A Contextual Critique of the Generalized Method of Moments." *Health Services and Outcomes Research Methodology* 6: 177–98.

———. 2016a. "Inference Using Sample Means of Parametric Nonlinear Data Transformations." *Health Services Research* 51: 1109–13.

———. 2016b. "Simpler Standard Errors for Two-Stage Optimization Estimators." *The Stata Journal* 16: 368–85.

Terza, J. V., A. Basu, and P. Rathouz. 2008. "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics* 27: 531–43.

SUPPORTING INFORMATION

Additional supporting information may be found online in the supporting information tab for this article:

Appendix SA1: Analytic and Stata Coding Details for Mullahy Birthweight Example.

Appendix SA2: Analytic and Stata Coding Details for the General 2SRI Framework.

Appendix SA3: Stata/Mata Code for Example in Section B.1 of Appendix SA2.

Appendix SA4: Stata/Mata Code for Example in Section B.2 of Appendix SA2.

Appendix SA5: Stata/Mata Code for Example in Section B.3 of Appendix SA2.

Appendix SA6: Stata/Mata Code for Example in Section B.4 of Appendix SA2.