The Stata Journal (2016)
16, Number 2, pp. 368–385

Simpler standard errors for two-stage optimization estimators

Joseph V. Terza
Department of Economics
Indiana University Purdue University Indianapolis
Indianapolis, IN
jvterza@iupui.edu

Abstract. Aiming to lessen the analytic and computational burden faced by practitioners seeking to correct the standard errors of two-stage estimators, I offer a heretofore unexploited simplification of the conventional formulation for the most commonly encountered cases in empirical application—two-stage estimators that involve maximum likelihood or pseudomaximum likelihood estimation. With the applied researcher in mind, I focus on the two-stage residual inclusion estimator designed for nonlinear regression models involving endogeneity. I demonstrate the analytics and Stata and Mata code for implementing my simplified standard-error formula by applying the two-stage residual inclusion method to the birthweight model of Mullahy (1997, *Review of Economics and Statistics* 79: 586–593) using his original data.

Keywords: st0436, two-stage optimization estimators, standard errors, asymptotic theory, endogeneity, two-stage residual inclusion, sandwich estimator

1 Introduction

Asymptotic theory for the two-stage optimization estimator (2SOE) (in particular, correct formulation of the asymptotic standard errors) has been available to applied researchers for decades (see Murphy and Topel [1985] for cases in which each of the stages involves a maximum likelihood estimator [MLE] and Newey and McFadden [1994] and White [1994] for more general classes of 2SOE). Despite textbook treatments of the subject (Cameron and Trivedi 2005, Greene 2012, and Wooldridge 2010) and articles offering requisite computer code (in Stata) (Hardin 2002 and Hole 2006), when conducting statistical inference based on two-stage estimates, applied researchers often implement bootstrapping methods or ignore the two-stage nature of the estimator and report the uncorrected second-stage outputs from packaged statistical software. In this article, aiming toward easy software implementation (in Stata), I offer the practitioner a heretofore largely unexploited simplification of the textbook asymptotic covariance matrix formulations (and their estimators—standard errors) for the most commonly encountered versions of the 2SOE—those involving an MLE or the nonlinear least squares (NLS) method in either stage. In addition, and perhaps more importantly from a practitioner's standpoint, I focus on regression models involving endogeneity—a sampling problem whose solution often requires a 2SOE. In particular, to fix ideas, I detail my simplified covariance specifications for the two-stage residual inclusion (2SRI) estimator

 st0436

suggested by Terza, Basu, and Rathouz (2008). I illustrate the analytics and Stata code for my approach by applying 2SRI to a model of birthweight as a function of endogenous smoking during pregnancy (see Mullahy [1997]).

The remainder of the article is organized as follows. In section 2, I review the asymptotic theory of 2SOE and give the conventional textbook formulation of the corresponding correct asymptotic covariance matrix. I also show how this formulation can be simplified when the estimator implements either NLS or MLE. In section 3, I detail the 2SRI estimator and, in light of the discussion in section 2, derive its correct (and simplified) asymptotic standard errors. In section 4, I discuss the birthweight illustration in full analytic and Stata and Mata coding detail. Section 5 summarizes. Technical details are given in appendixes.

2 2SOEs and their asymptotic standard errors

The most commonly applied parametric estimators reside in the class of optimization estimators (OEs)—statistical methods that produce estimates as optimizers of well-specified objective functions (sometimes called M-estimators). The most prominent OE examples are the MLE and the NLS estimators. Model design or computational convenience often dictates that an OE be implemented in two stages. In such cases, the parameter vector of interest is partitioned as $\boldsymbol{\omega}' = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']$ and conformably estimated in two stages. First, an estimate of $\boldsymbol{\alpha}$ is obtained as the optimizer of an appropriately specified first-stage objective function,

$$\sum_{i=1}^{n} q_1(\boldsymbol{\alpha}, \mathbf{V}_{1i}) \tag{1}$$

where $\mathbf{V}_{1i}$ denotes the relevant subvector of the observable data for the ith sample individual ($i = 1, \dots, n$). Next, an estimate of $\boldsymbol{\beta}$ is obtained as the optimizer of

$$\sum_{i=1}^{n} q_2\left(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbf{V}_{2i}\right) \tag{2}$$

where $\mathbf{V}_{2i}$ denotes the second-stage analog to $\mathbf{V}_{1i}$ and $\widehat{\boldsymbol{\alpha}}$ is the first-stage estimate of $\boldsymbol{\alpha}$.

It is well established that under general conditions, this 2SOE is consistent and asymptotically normal.[1] My interest here is in simplifying the typical textbook formulation of the corresponding asymptotic covariance matrix of $\widehat{\boldsymbol{\omega}}' = [\widehat{\boldsymbol{\alpha}}' \ \widehat{\boldsymbol{\beta}}']$, where $\widehat{\boldsymbol{\beta}}$ denotes the second-stage estimator obtained from (2).[2] Before proceeding, I establish the following notational conventions:

1. See Newey and McFadden (1994) or White (1994) for details.
2. For textbook discussions of this issue, see Wooldridge (2010, sec. 12.4.2); Greene (2012, sec. 14.7); and Cameron and Trivedi (2005, sec. 6.6).

1. q_1 is shorthand notation for $q_1(\boldsymbol{\alpha}, \mathbf{V}_1)$ as defined in (1), where $\mathbf{V}_1$ is the random vector representing the population from which the observed data in $\mathbf{V}_{1i}$ was sampled.

2. q_2 is shorthand notation for $q_2(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbf{V}_2)$ as defined in (2), where $\mathbf{V}_2$ is the random vector representing the population from which the observed data in $\mathbf{V}_{2i}$ was sampled.

3. $\boldsymbol{\nabla}_s \mathbf{q}$ denotes the gradient of q with respect to parameter subvector $\mathbf{s}$—a row vector.

4. $\boldsymbol{\nabla}_{st} \mathbf{q}$ denotes the matrix whose typical element is $\partial^2 q / \partial s_j \partial t_m$—its row dimension corresponds to that of its first subscript, and the column dimension corresponds to that of its second subscript.

From Newey and McFadden (1994) or White (1994), we have that the correct asymptotic covariance matrix of $\widehat{\boldsymbol{\omega}}' = [\widehat{\boldsymbol{\alpha}}' \ \widehat{\boldsymbol{\beta}}']$ is

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{12}' & \mathbf{D}_{22} \end{bmatrix}$$

where

$$\mathbf{D}_{11} = \mathbf{AVAR}\left(\widehat{\boldsymbol{\alpha}}\right) = \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha}\mathbf{q}_1\right)^{-1} \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha}\mathbf{q}_1' \boldsymbol{\nabla}_{\alpha}\mathbf{q}_1\right) \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha}\mathbf{q}_1\right)^{-1} \tag{3}$$

$$\begin{aligned} \mathbf{D}_{12} = {} & \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha}\mathbf{q}_1\right)^{-1} \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2' \boldsymbol{\nabla}_{\alpha}\mathbf{q}_1\right)' \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \\ & - \mathbf{AVAR}\left(\widehat{\boldsymbol{\alpha}}\right) \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)' \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \end{aligned} \tag{4}$$

$$\begin{aligned} \mathbf{D}_{22} = {} & \mathbf{AVAR}\left(\widehat{\boldsymbol{\beta}}\right) \\ = {} & \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \left\{ \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) \mathbf{AVAR}\left(\widehat{\boldsymbol{\alpha}}\right) \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)' \right. \\ & - \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2' \boldsymbol{\nabla}_{\alpha}\mathbf{q}_1\right) \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha}\mathbf{q}_1\right)^{-1} \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)' \\ & \left. - \ \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha_{\mathbf{q}_1}}\right)^{-1} \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2' \boldsymbol{\nabla}_{\alpha}\mathbf{q}_1\right)' \right\} \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \\ & + \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2' \boldsymbol{\nabla}_{\beta}\mathbf{q}_2\right) \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \end{aligned} \tag{5}$$

and $\mathbf{AVAR}(\mathbf{c})$ denotes the asymptotic covariance matrix of the estimator $\mathbf{c}$. For cases in which the ultimate estimation objective is $\boldsymbol{\beta}$, only $\mathbf{D}_{22}$ is of interest. In most cases, however, an estimate of $\mathbf{D}$ will be needed. Hence, my interest is in simplifying the details of the full formulation of $\mathbf{D}$.

Hardin (2002) rediscovers the general formulation in (2) through (5) but focuses on the special case in which MLEs are applied in both stages of the 2SOE estimator (henceforth MLE-MLE-2SOE). He refers to (2) through (5) as the "sandwich" formulation of the asymptotic covariance matrix. In this case, based on the well-known asymptotic theory for the MLE combined with results given in Murphy and Topel (1985), I show in *Appendix A* that the following legitimate substitutions can be made in (2) through (5),[3]

$$\mathbf{D}_{11} = \mathbf{AVAR}\left(\widehat{\boldsymbol{\alpha}}\right) = -\mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha}\mathbf{q}_1\right)^{-1} \tag{6}$$

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right) = -\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\beta}\mathbf{q}_2\right) \tag{7}$$

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) = -\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right) \tag{8}$$

I refer to (2) through (5) with substitutions (2) through (4) as the MT (Murphy–Topel) formulation of the asymptotic covariance matrix. Clearly, (2) through (4) are simplifying in that they substantially reduce both analytic and computational demands. Nevertheless, Hardin (2002) advocates using the sandwich formulation on the grounds that it 1) is robust, 2) is easy to calculate, and 3) admits Wald tests of hypotheses across the two stages of the model. It is clear, however, that 1) the sandwich and MT formulations are equally robust given the assumptions underlying the MLE-MLE-2SOE; 2) the MT formulation is easier to calculate; and 3) the full MT formulation also facilitates tests of hypotheses involving elements of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Thus, for the MLE-MLE-2SOE, I see no apparent justification for incurring the higher analytic and computational costs of the sandwich formulation in practice. Hole (2006) apparently agrees and shows how calculation of the MT formulation of the asymptotic covariance matrix of the MLE-MLE-2SOE can be simplified using the `scores` option of the Stata `predict` command.

I extend this line of research by including a broader class of 2SOE estimators that subsumes the MLE-MLE-2SOE as a special case. I first note that in general, $\mathbf{D}_{11}$ warrants no discussion, because neither its formulation nor its estimation is affected by the two-stage nature of the estimator—$\boldsymbol{\beta}$ does not appear in (1). Therefore, the correct standard errors of $\widehat{\boldsymbol{\alpha}}$, and other inferential statistics pertaining to $\widehat{\boldsymbol{\alpha}}$, can be obtained from the "packaged" output of the software used for first-stage estimation. In other words, we can rewrite (2) as

$$\mathbf{D}_{11} = \mathbf{AVAR}^{*}\left(\widehat{\boldsymbol{\alpha}}\right) \tag{9}$$

where $\mathbf{AVAR}^{*}(\mathbf{c})$ denotes the matrix to which the packaged estimate of the asymptotic covariance matrix of the estimator $\mathbf{c}$ converges in probability (by "packaged", I refer to what would be obtained from any econometrics computer package for the estimator $\mathbf{c}$, ignoring the fact that it is a component of a 2SOE). By the same token, I need only consider how the choice of method for the second-stage determines the formulation and estimation of $\mathbf{D}_{12}$ and $\mathbf{D}_{22}$ in (4) and (5), respectively. For these reasons, I extend the discussion to the class of 2SOEs whose first stage is any OE and whose second stage is either MLE or NLS (the two most commonly implemented OEs). This class of estimators

3. In fact, Hardin (2002) imposes (4) in his specification of the sandwich formulation of the asymptotic covariance matrix of the MLE-MLE-2SOE. Strictly speaking, this is not in keeping with his formulation of the matrix A in (5)—the "bread" matrix in his sandwich formulation.

will henceforth be denoted 2SOE*. I will show that for 2SOE*, heretofore unexploited simplifications in the general asymptotic covariance formulation [(2) through (5)] are possible. First, as I show in *Appendix B*, for 2SOE*,

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_1\right) = \mathbf{0} \tag{10}$$

Until now, this condition was not recognized and therefore was not incorporated in either the textbook treatments of 2SOE or the studies by Hardin (2002) and Hole (2006). Applying (3) and (5) and noting that

$$\mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1}\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\beta}\mathbf{q}_2\right)\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1}$$

we see that (4) and (5) can be substantially simplified as, respectively,

$$\mathbf{D}_{12} = -\mathbf{AVAR}^*\left(\widehat{\boldsymbol{\alpha}}\right)\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)'\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \tag{11}$$

and

$$\begin{aligned}\mathbf{D}_{22} &= \mathbf{AVAR}\left(\widehat{\boldsymbol{\beta}}\right)\\ &= \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1}\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)\mathbf{AVAR}^*\left(\widehat{\boldsymbol{\alpha}}\right)\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)'\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1}\\ &\quad + \mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right)\end{aligned} \tag{12}$$

Equations (6) and (7) highlight the covariance matrix components that can be obtained directly from packaged econometric software versus those that require special programming. If the second stage $(\widehat{\boldsymbol{\beta}})$ is MLE, using (3) and (4) and noting that $\mathbf{AVAR}^*(\widehat{\boldsymbol{\beta}}) = -\mathbf{E}(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2)^{-1}$, we can rewrite (6) and (7) as

$$\mathbf{D}_{12} = \mathbf{AVAR}^*\left(\widehat{\boldsymbol{\alpha}}\right)\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right)'\mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right) \tag{13}$$

and

$$\begin{aligned}\mathbf{D}_{22} &= \mathbf{AVAR}\left(\widehat{\boldsymbol{\beta}}\right)\\ &= \mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right)\mathbf{AVAR}^*\left(\widehat{\boldsymbol{\alpha}}\right)\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right)'\mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right)\\ &\quad + \mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right)\end{aligned} \tag{14}$$

In this case, the only component that must be analytically derived and coded is $\mathbf{E}(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2)$. A consistent estimator of this component is

$$\widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right) = \frac{\sum_{i=1}^{n}\boldsymbol{\nabla}_{\beta}\widehat{\mathbf{q}}_{2i}'\boldsymbol{\nabla}_{\alpha}\widehat{\mathbf{q}}_{2i}}{n} \tag{15}$$

where $\widehat{q}_{2i}$ is shorthand notation for $q_2(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \mathbf{V}_{2i})$. Equation (3) can be coded in Mata. Here consistent estimators of the components of the asymptotic covariance matrix given in (2), (3), and (8) are

$$\widehat{D}_{11} = n\widehat{\mathbf{AVAR}}^* (\widehat{\boldsymbol{\alpha}}) \tag{16}$$

$$\widehat{\mathbf{D}}_{12} = \left\{n\widehat{\mathbf{AVAR}}^* (\widehat{\alpha})\right\} \widehat{\mathbf{E}} \left(\boldsymbol{\nabla}_\beta \mathbf{q}_2' \boldsymbol{\nabla}_\alpha \mathbf{q}_2\right)' \left\{n\widehat{\mathbf{AVAR}}^* \left(\widehat{\beta}\right)\right\} \tag{17}$$

$$\widehat{\mathbf{D}}_{22} = \left\{n\widehat{\mathbf{AVAR}}^* \left(\widehat{\beta}\right)\right\} \widehat{\mathbf{E}} \left(\boldsymbol{\nabla}_\beta \mathbf{q}_2' \boldsymbol{\nabla}_\alpha \mathbf{q}_2\right) \left\{n\widehat{\mathbf{AVAR}}^* (\widehat{\alpha})\right\} \widehat{\mathbf{E}} \left(\boldsymbol{\nabla}_\beta \mathbf{q}_2' \boldsymbol{\nabla}_\alpha \mathbf{q}_2\right)'$$
$$\left\{n\widehat{\mathbf{AVAR}}^* \left(\widehat{\beta}\right)\right\} + \left\{n\widehat{\mathbf{AVAR}}^* \left(\widehat{\beta}\right)\right\} \tag{18}$$

$\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})$ and $\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\beta}})$ are the estimated covariance matrices obtained from the first- and second-stage packaged regression outputs, respectively.

When the second stage is NLS, the main component of (2) can be written as

$$q_2 \left(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbf{V}_{2i}\right) = -\left\{Y_i - J\left(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbf{Z}_{2i}\right)\right\}^2$$

where $J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z}_2)$ denotes the relevant nonlinear regression function and $\mathbf{V}_{2i} = [Y_i \;\; \mathbf{Z}_{2i}]$. In this case, two components of (6) and (7) must be analytically derived—namely, $\mathbf{E}(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2)$ and $\mathbf{E}(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2)$. In *Appendix C*, I show that these components can be consistently estimated using

$$\widehat{\mathbf{E}} \left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) = \frac{\sum_{i=1}^{n} \boldsymbol{\nabla}_\beta \widehat{\mathbf{J}}_i' \boldsymbol{\nabla}_\alpha \widehat{\mathbf{J}}_i}{n} \tag{19}$$

and

$$\widehat{\mathbf{E}} \left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right) = \frac{\sum_{i=1}^{n} \boldsymbol{\nabla}_\beta \widehat{\mathbf{J}}_i' \boldsymbol{\nabla}_\beta \widehat{\mathbf{J}}_i}{n} \tag{20}$$

respectively, where $\widehat{J}_i$ is shorthand notation for $J(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \mathbf{Z}_{2i})$.[4] Expressions (19) and (20) appear neither in textbook treatments of the subject nor in the studies by Hardin (2002) and Hole (2006) and constitute a useful simplification in 2SOE with second-stage NLS in that they require the derivation and coding of only first-order partial derivatives. Using these results, we have that when the second stage is NLS, consistent estimators of the components of the asymptotic covariance matrix are

4. Strictly speaking, both (19) and (20) should be multiplied by −2, but because the 2s cancel in the formulations of (22) and (23), we suppress them here.

$$\widehat{\mathbf{D}}_{11} = n\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}}) \tag{21}$$

$$\widehat{\mathbf{D}}_{12} = \left\{n\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})\right\} \widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)' \widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \tag{22}$$

$$\widehat{\mathbf{D}}_{22} = \widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} \widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) \left\{n\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})\right\} \times \widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)' \widehat{\mathbf{E}}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right)^{-1} + n\widehat{\mathbf{AVAR}}^*\left(\widehat{\boldsymbol{\beta}}\right) \tag{23}$$

and $\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})$ and $\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\beta}})$ are defined as in (3) through (18).

So, for example, in either case (second-stage MLE or NLS), the t statistic

$$\frac{\sqrt{n}\left(\widehat{\beta}_k - \beta_k\right)}{\sqrt{\widehat{D}_{22(k)}}} \tag{24}$$

for the kth element of β is asymptotically standard normally distributed (where $\widehat{D}_{22(k)}$ denotes the kth diagonal element of $\widehat{\mathbf{D}}_{22}$) and can be used to test the hypothesis that $\beta_k = \beta_k^0$ for β_k^0, a given null value of β_k. In practice, the following version of (24) is used for this purpose,

$$\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{D}^{\dagger}_{22(k)}}} \tag{25}$$

where $\widehat{D}^{\dagger}_{22(k)}$ denotes the kth diagonal element of $\widehat{\mathbf{D}}^{\dagger}_{22}$, which is the same as $\widehat{\mathbf{D}}_{22}$ except that all the "n's" are suppressed, including those in the denominators of (3), (19), and (20). $\widehat{\mathbf{D}}^{\dagger}_{22}$ takes the following forms for the MLE and NLS cases, respectively,

$$\widehat{\mathbf{D}}^{\dagger}_{22} = \widehat{\mathbf{AVAR}}^*\left(\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\mathbf{A}}_{\beta\alpha}\right)\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})\left(\widehat{\mathbf{A}}_{\beta\alpha}\right)'\widehat{\mathbf{AVAR}}^*\left(\widehat{\boldsymbol{\beta}}\right) + \widehat{\mathbf{AVAR}}^*\left(\widehat{\boldsymbol{\beta}}\right) \tag{26}$$

and

$$\widehat{\mathbf{D}}^{\dagger}_{22} = \left(\widehat{\mathbf{B}}_{\beta\beta}\right)^{-1}\left(\widehat{\mathbf{B}}_{\beta\alpha}\right)\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})\left(\widehat{\mathbf{B}}_{\beta\alpha}\right)'\left(\widehat{\mathbf{B}}_{\beta\beta}\right)^{-1} + \mathbf{AVAR}^*\left(\widehat{\boldsymbol{\beta}}\right) \tag{27}$$

where

$$\widehat{\mathbf{A}}_{\beta\alpha} = \sum_{i=1}^{n} \boldsymbol{\nabla}_{\beta}\widehat{\mathbf{q}}'_{2i}\boldsymbol{\nabla}_{\alpha}\widehat{\mathbf{q}}_{2i}$$

$$\widehat{\mathbf{B}}_{\beta\alpha} = \sum_{i=1}^{n} \boldsymbol{\nabla}_{\beta}\widehat{\mathbf{J}}'_{i}\boldsymbol{\nabla}_{\alpha}\widehat{\mathbf{J}}_{i}$$

and

$$\widehat{\mathbf{B}}_{\beta\beta} = \sum_{i=1}^{n} \boldsymbol{\nabla}_{\beta}\widehat{\mathbf{J}}'_{i}\boldsymbol{\nabla}_{\alpha}\widehat{\mathbf{J}}_{i}$$

Expressions in (26) and (27), along with the analogous formulations corresponding to (3), (17), (21), and (22), constitute substantial simplification of a) the textbook formulations given in (2), (4), and (5) and b) the covariance formulations offered for the special case (MLE-MLE-2SOE) considered by Hardin (2002) and Hole (2006).

3 2SRI

Consider a nonlinear regression model with dependent variable Y and a particular regressor of policy interest X_p. Suppose that the data on X_p are sampled endogenously—that is, it is correlated with an unobservable variable, X_u, that is also correlated with Y. To formalize this, we follow Terza, Basu, and Rathouz (2008) and assume that

$$E\left(Y|X_p,\mathbf{X}_o,X_u\right)=\mu\left(X_p,\mathbf{X}_o,X_u\left(\mathbf{W},\boldsymbol{\alpha}\right);\boldsymbol{\beta}\right)\quad \text{[outcome regression]} \tag{28}$$

and

$$X_p=r(\mathbf{W},\boldsymbol{\alpha})+X_u\quad \text{[auxiliary regression]} \tag{29}$$

where $\mathbf{X}_o$ denotes a vector of observable confounders (observable variables that are possibly correlated with both Y and X_p), $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are parameter vectors, $\mathbf{W}=[\mathbf{X}_o\ \ \mathbf{W}^+]$, $\mathbf{W}^+$ is a vector of identifying instrumental variables (possibly comprising a single element), and $\mu()$ and $r()$ are known functions. Note that (29) implies that X_u can be written as the following function of W and α:

$$X_u(\mathbf{W},\boldsymbol{\alpha})=X_p-r(\mathbf{W},\boldsymbol{\alpha}) \tag{30}$$

Hence, its parametric representation in (28). Stated succinctly, the relevant outcome regression is

$$Y=\mu\left(X_p,\mathbf{X}_o,X_u\left(\mathbf{W},\boldsymbol{\alpha}\right);\boldsymbol{\beta}\right)+e \tag{31}$$

where e is the random-error term, tautologically defined as $e=Y-\mu(X_p,\mathbf{X}_o,X_u;\boldsymbol{\beta})$ so that $E(e|X_p,\mathbf{X}_o,X_u)=0$. The β parameters in (31) are not directly estimable (for example, via NLS) because of the presence of the unobservable confounder X_u, which embodies the endogeneity of X_p. However, Terza, Basu, and Rathouz (2008) show that the following two-stage estimator of $\boldsymbol{\omega}=[\boldsymbol{\alpha}'\ \ \boldsymbol{\beta}']$ is feasible and consistent.

First stage: Obtain a consistent estimate of α by applying NLS to (29) and compute the residual as

$$\widehat{X}_u=X_p-r(\mathbf{W},\widehat{\boldsymbol{\alpha}})$$

where $\widehat{\boldsymbol{\alpha}}$ is the first-stage estimate of $\boldsymbol{\alpha}$.

Second stage: Estimate $\boldsymbol{\beta}$ by applying NLS to

$$Y=\mu\left(X_p,\mathbf{X}_o,\widehat{X}_u;\boldsymbol{\beta}\right)+e^{\mathrm{2SRI}}$$

where e^{2SRI} denotes the regression-error term. Terza, Basu, and Rathouz (2008) call this method 2SRI.

To detail the asymptotic covariance matrix of this 2SRI estimator, I cast it in the framework of the generic 2SOE discussed above. In this case, the main components of the first- and second-stage objective functions are, respectively,

$$q_1(\boldsymbol{\alpha}, \mathbf{V}_{1i}) = -\left\{X_{pi} - r\left(\mathbf{W}_i, \boldsymbol{\alpha}\right)\right\}^2$$

and

$$q_2\left(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \mathbf{V}_{2i}\right) = -\left\{Y - \mu\left(X_p, \mathbf{X}_o, \widehat{X}_u; \boldsymbol{\beta}\right)\right\}^2$$

where $\mathbf{V}_{1i} = [X_{pi} \quad \mathbf{W}_i]$ and $\mathbf{V}_{2i} = [Y_i \quad X_{pi} \quad \mathbf{W}_i]$. This version of the 2SRI estimator implements NLS in its second stage. Therefore, expressions (19) through (23) are relevant for calculating the correct asymptotic standard errors of $\widehat{\boldsymbol{\omega}} = [\widehat{\boldsymbol{\alpha}}' \quad \widehat{\boldsymbol{\beta}}']$ (the 2SRI estimator of $\boldsymbol{\omega} = [\boldsymbol{\alpha}' \quad \boldsymbol{\beta}']$). In particular, if conventional t testing of the elements of $\boldsymbol{\beta}$ is the objective, then one may focus on the requisite analytics and coding of (25).

Note that MLE can be implemented in either of the stages of the 2SRI method. For MLE to be implemented in the first stage, the auxiliary regression in (29) must be replaced by an assumption that specifies a known form for the conditional density of $(X_p|\mathbf{W})$, say, $g(X_p|\mathbf{W};\boldsymbol{\alpha})$. Of course, such an assumption would imply a formulation for the relevant conditional mean $E(X_p|\mathbf{W})$, say, $r(\mathbf{W},\boldsymbol{\alpha})$. Therefore, in this case, the first stage of the estimator would consist of maximizing (1) with $q_1(\boldsymbol{\alpha},\mathbf{V}_{1i})$ replaced by $\ln\{g(X_{pi}|\mathbf{W}_i;\boldsymbol{\alpha})\}$ and computing the residuals as in (30). For MLE to be implemented in the second stage, the outcome regression in (28) must be replaced by an assumption that specifies a known form for the conditional density of $(Y|X_p,\mathbf{X}_o,X_u)$, say, $f(Y|X_p,\mathbf{X}_o,X_u;\boldsymbol{\beta})$. The second stage of the estimator would then consist of maximizing (2) with $q_2(\widehat{\boldsymbol{\alpha}},\boldsymbol{\beta},\mathbf{V}_{2i})$ replaced by $\ln[f(Y_i|X_{pi},\mathbf{X}_{oi},\widehat{X}_{ui};\boldsymbol{\beta})]$. To obtain the correct asymptotic covariance matrix, we would use the expressions in (3) through (18).

4 An example: Smoking and infant birthweight—Testing for endogeneity

Here we revisit the regression model of Mullahy (1997) in which

$$\begin{aligned} Y &= \text{infant birthweight in lbs.} \\ X_p &= \text{number of cigarettes smoked per day during pregnancy} \\ \mathbf{X}_o &= [\texttt{PARITY WHITE MALE}] \\ \mathbf{W}^+ &= [\texttt{EDFATHER EDMOTHER FAMINCOME CIGTAX88}] \end{aligned}$$

where

Variable	Description
PARITY	birth order
WHITE	1 if white, 0 otherwise
MALE	1 if male, 0 otherwise
EDFATHER	paternal schooling in years
EDMOTHER	maternal schooling in years
FAMINCOME	family income
CIGTAX88	cigarette tax

The relevant outcome regression in Mullahy's (1997) model can be written as[5]

$$E\left(Y|X_p, \mathbf{X}_o, X_u\right) = \exp\left(X_p\beta_p + \mathbf{X}_o\boldsymbol{\beta}_o + X_u\beta_u\right) \tag{32}$$

In the original study, the model was fit via a generalized method of moment (GMM) procedure that does not require specification of an auxiliary regression for X_p. However, this GMM method does not permit identification and estimation of β_u. Note that this precludes a direct test of endogeneity because under the assumed regression specification in (32), the null hypothesis that X_p is exogenous is tantamount to $\beta_u = 0$. However, such a test is supported in the 2SRI estimation framework. To this end, we specify the following version of (29),

$$X_p = \exp\left(\mathbf{W}\boldsymbol{\alpha}\right) + X_u$$

and, using Mullahy's data, apply the 2SRI method discussed in the previous section with (32) as the relevant version of (28). In this illustration, we applied NLS in both of the stages of 2SRI. The first- and second-stage 2SRI parameter estimates ($\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_p \quad \widehat{\boldsymbol{\beta}}_o' \quad \widehat{\beta}_u]$, respectively) were obtained via the Stata **glm** command with the **family(gaussian)**, **link(log)**, and **vce(robust)** options. We focus here on asymptotic t testing of the conventional null hypotheses for the individual elements of $\boldsymbol{\beta}$ (that is, $H_o\colon \beta_k = 0$, where β_k denotes the kth element of $\boldsymbol{\beta}$). Therefore, the expression for $\widehat{\mathbf{D}}_{22}^{\dagger}$ in (27) is relevant with

$$\begin{aligned} J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z}_2) &= \mu\left(X_p, \mathbf{X}_o, X_p - r\left(\mathbf{W}, \boldsymbol{\alpha}\right); \boldsymbol{\beta}\right) \\ &= \exp\left[X_p\beta_p + \mathbf{X}_o\boldsymbol{\beta}_o + \left\{X_p - \exp\left(\mathbf{W}\boldsymbol{\alpha}\right)\right\}\beta_u\right] \end{aligned} \tag{33}$$

where $\mathbf{Z}_2 = [X_p \quad \mathbf{X}_o \quad \mathbf{W}^+]$. After each of the 2SRI estimation stages, we saved the parameter vectors ($\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$) and their corresponding packaged covariance matrix estimators [$\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\alpha}})$ and $\widehat{\mathbf{AVAR}}^*(\widehat{\boldsymbol{\beta}})$] in Mata. Moreover, using (33), we get

$$\boldsymbol{\nabla}_\alpha \mathbf{J}\left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \mathbf{Z}_{2i}\right) = -2\widehat{\beta}_u \exp\left(\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)\exp\left(\mathbf{W}_i\widehat{\boldsymbol{\alpha}}\right)\mathbf{W}_i$$

5. Equation (32) is written in my notation, not Mullahy's. He does not explicitly specify the model in terms of the unobservable confounder X_u. Nevertheless, (32) is substantively identical to Mullahy's model (see Terza [2006]).

and

$$\nabla_{\beta}\mathbf{J}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\mathbf{Z}_{2i}\right)=2\exp\left(\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)\mathbf{X}_i$$

where $\mathbf{X}_i = [X_{pi} \quad \mathbf{X}_{oi} \quad \widehat{X}_{ui}]$. The requisite Mata code for calculating the elements of $\widehat{\mathbf{D}}^{\dagger}_{22}$ in (27) is[6]

$\exp\left(\mathbf{W}_i\widehat{\boldsymbol{\alpha}}\right)$: `expWalpha` (saved using `predict` for 1st-stage glm)

$\exp\left(\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)$: `expXbeta` (saved using `predict` for 2nd-stage glm)

$\nabla_{\alpha}\mathbf{J}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\mathbf{Z}_{2i}\right)$: `paJ=-bxu:*expXbeta:*expWalpha:*W`

$\nabla_{\beta}\mathbf{J}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\mathbf{Z}_{2i}\right)$: `pbJ=expXbeta:*X`

$\widehat{\mathbf{B}}_{\beta\alpha}$: `Bba=pbJ'*paJ`

$\widehat{\mathbf{B}}_{\beta\beta}$: `Bbb=pbJ'*pbJ`

$\widehat{\mathbf{D}}^{\dagger}_{22}$: `D22= invsym(Bbb)*Bba*covalpha*Bba'*invsym(Bbb)+covbeta`

The first- and second-stage 2SRI results are given in tables 1 and 2, respectively. Table 2 also displays Mullahy's GMM estimates and, as a baseline, reports the simple NLS estimates that ignore potential endogeneity. To indicate the strength of the instrumental variables (that is, the elements of $\mathbf{W}^{+}$), I conducted a Wald test of their joint significance. The value of the chi-squared test statistic is 49.33, and the null is roundly rejected at any reasonable level of significance. The second-stage 2SRI estimates shown in table 2 are virtually identical to the GMM estimates, but the former, unlike the latter, provide a direct test of the endogeneity of the cigarette consumption variable via the asymptotic t statistic for the coefficient of X_u. According to this test, the exogeneity null hypothesis is rejected at nearly the 1% significance level. To get a sense of the bias from neglecting to consider the two-stage nature of the estimator in calculating the asymptotic standard errors, I also display in table 2 the packaged second-stage `glm` t statistics as reported in the Stata output. The mean absolute bias across the uncorrected asymptotic t statistics given in column 3 of table 2 for the four regressors and X_u is nearly 9%.

6. See *Appendix D* for the complete do-file.

Table 1. 2SRI first-stage estimates

Variable	Estimate	Asymptotic t statistic
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

$n = 1388$

Table 2. 2SRI second-stage, GMM, and NLS estimates

	2SRI			GMM		NLS	
Variable	Estimate	Correct asymptotic t statistic	Uncorrected asymptotic t statistic	Estimate	Asymptotic t statistic	Estimate	Asymptotic t statistic
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

$n = 1388$

5 Summary

To aid applied researchers seeking to implement 2SOE, I offer simplified versions of the textbook formulations of the correct asymptotic standard errors for cases in which the second-stage method is MLE or NLS. My results apply to, and thus simplify, the standard-error formulations offered by Hardin (2002) and Hole (2006) for the special case in which both stages of the 2SOE are MLEs (MLE-MLE-2SOE). As an illustration, I detail 2SRI estimation of a nonlinear model with an endogenous regressor in which the second-stage estimator is NLS.

6 Acknowledgments

This research was supported by a grant from the Agency for Healthcare Research and Quality (R01 HS017434-01). This article was presented at the 2012 Stata Conference in San Diego, CA, July 26–27, 2012, and as part of the keynote presentation at the Mexican Stata Users Group Meeting in Mexico City, November 14, 2014.

7 References

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Hardin, J. W. 2002. The robust variance estimator for two-stage models. *Stata Journal* 2: 253–266.

Hole, A. R. 2006. Calculating Murphy–Topel variance estimates in Stata: A simplified procedure. *Stata Journal* 6: 521–529.

Mullahy, J. 1997. Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics* 79: 586–593.

Murphy, K. M., and R. H. Topel. 1985. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics* 3: 370–379.

Newey, W. K., and D. McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, ed. R. F. Engle and D. L. McFadden, vol. 4, 2111–2245. Amsterdam: Elsevier.

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: 177–198.

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: 531–543.

White, H. 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the author

Joseph Terza is a health economist and econometrician in the Department of Economics at Indiana University–Purdue University Indianapolis. His research focuses on the development

and application of methods for estimating qualitative and limited dependent variable models with endogeneity. Two of his methods have been implemented as Stata commands. He was keynote speaker at the Stata Users Group meeting in Mexico City in November 2014.

Appendix A: Establishment of (3) and (4) in the text

When the second stage is MLE so that $q_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_2)$ is a log likelihood, the true value of the parameter vector $\boldsymbol{\omega}_0' = [\boldsymbol{\alpha}_0' \ \ \boldsymbol{\beta}_0']$, the unconditional information matrix equality (see expression 13.27 on page 479 of Wooldridge [2010]), yields

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\omega\omega}\mathbf{q}_2\right) = -\mathbf{E}\left(\boldsymbol{\nabla}_{\omega}\mathbf{q}_2'\boldsymbol{\nabla}_{\omega}\mathbf{q}_2\right)$$

or

$$\begin{bmatrix} \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha\alpha}\mathbf{q}_2\right) & \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right)' \\ \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) & \mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right) \end{bmatrix} = -\begin{bmatrix} \mathbf{E}\left(\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right) & \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right)' \\ \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right) & \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\beta}\mathbf{q}_2\right) \end{bmatrix}$$

Therefore,

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right) = -\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\beta}\mathbf{q}_2\right)$$

and

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) = -\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_2\right)$$

so (3) and (4) in the text hold.

Appendix B: Establishment of (5) in the text

When the second stage is NLS,

$$q_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_2) = -\{Y - J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z}_2)\}^2$$

where $J(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z}_2)$ denotes the relevant nonlinear regression function, $\mathbf{V}_{2i} = [Y_i \ \ \mathbf{Z}_{2i}]$,

$$J(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \mathbf{Z}_2) = \mathbf{E}\left(Y|\mathbf{Z}_2\right) \tag{34}$$

and $\boldsymbol{\omega}_0' = [\boldsymbol{\alpha}_0' \ \ \boldsymbol{\beta}_0']$ denotes the true value of the parameter vector. Moreover, we can write

$$\begin{aligned} \mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_1\right) &= \mathbf{E}\left(\mathbf{E}\left(\boldsymbol{\nabla}_{\beta}\mathbf{q}_2\left(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \mathbf{V}_2\right)|\mathbf{Z}_2\right)'\boldsymbol{\nabla}_{\alpha}\mathbf{q}_1(\boldsymbol{\delta}_0, \mathbf{V}_1)\right) \\ &= 2\mathbf{E}\left\{\left(\mathbf{E}\left[\left\{Y - J\left(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \mathbf{Z}_2\right)\right\}|\mathbf{Z}_2\right]\boldsymbol{\nabla}_{\beta}\mathbf{J}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \mathbf{Z}_2)\right)' \right. \\ &\qquad \left. \boldsymbol{\nabla}_{\alpha}\mathbf{q}_1(\boldsymbol{\alpha}_0, \mathbf{V}_1)\right\} \end{aligned}$$

because $q_1(\boldsymbol{\alpha}_0, \mathbf{V}_1)$ is not a function of Y and $\mathbf{Z}_2$ is a subvector of $\mathbf{V}_1$. But using (34), we have that

$$\mathbf{E}\left[\left\{Y - J\left(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \mathbf{Z}_2\right)\right\}|\mathbf{Z}_2\right] = 0$$

Therefore, (5) in the text holds.

When the second stage is MLE, we have that

$$q_2(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_{2i}) = \ln f(Y|\mathbf{Z}_2;\boldsymbol{\alpha},\boldsymbol{\beta})$$

where $f(Y|\mathbf{Z}_2;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$ denotes the true conditional density of Y given $\mathbf{Z}_2$. Accordingly, we can write

$$\mathbf{E}\left(\boldsymbol{\nabla}_\beta \mathbf{q}_2' \boldsymbol{\nabla}_\alpha \mathbf{q}_1\right) = \mathbf{E}\left[\mathbf{E}\left\{\boldsymbol{\nabla}_\beta \mathbf{f}(Y|\mathbf{Z}_2;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)|\mathbf{Z}_2\right\}' \boldsymbol{\nabla}_\alpha \mathbf{q}_1(\boldsymbol{\alpha}_0,\mathbf{V}_1)\right]$$

Now, using (13.20) on page 477 of Wooldridge (2010), we have that

$$\mathbf{E}\left\{\boldsymbol{\nabla}_\beta \mathbf{f}\left(Y|\mathbf{Z}_2;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0\right)|\mathbf{Z}_2\right\} = \mathbf{0}$$

because it is the score of the second-stage log-likelihood function. Therefore, (5) in the text holds.

Appendix C: Derivation of consistent estimators (19) and (20) in the text

When the second stage is NLS

$$q_2(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_2) = -\{Y - J(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{Z}_2)\}^2 \tag{35}$$

where $J(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{Z}_2)$ denotes the relevant nonlinear regression function and $\mathbf{V}_{2i} = [Y_i \;\; \mathbf{Z}_{2i}]$. From (35), we get

$$\boldsymbol{\nabla}_\beta \mathbf{q}_2 = 2e\boldsymbol{\nabla}_\beta \mathbf{J}$$

so

$$\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2 = 2(\boldsymbol{\nabla}_\alpha e \boldsymbol{\nabla}_\beta \mathbf{J} + e\boldsymbol{\nabla}_{\beta\alpha}\mathbf{J})$$

and

$$\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2 = 2(\boldsymbol{\nabla}_\beta e \boldsymbol{\nabla}_\beta \mathbf{J} + e\boldsymbol{\nabla}_{\beta\beta}\mathbf{J})$$

where $e = (Y - J(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{Z}_2))$ and J is shorthand notation for $J(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{Z}_2)$. Now,

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) = 2\mathbf{E}\left\{\boldsymbol{\nabla}_\beta \mathbf{J}'\boldsymbol{\nabla}_\alpha e + \mathbf{E}\left(e|\mathbf{Z}_2\right)\boldsymbol{\nabla}_{\beta\alpha}\mathbf{J}\right\}$$

where

$$\boldsymbol{\nabla}_\alpha e = -\boldsymbol{\nabla}_\alpha \mathbf{J}$$

and at the true value of $\boldsymbol{\omega}$ (say, $\boldsymbol{\omega}_0' = [\boldsymbol{\alpha}_0' \;\; \boldsymbol{\beta}_0']$),

$$E\left(e|\mathbf{Z}_2\right) = 0$$

Therefore,

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\alpha}\mathbf{q}_2\right) = -2\mathbf{E}\left(\boldsymbol{\nabla}_\beta \mathbf{J}'\boldsymbol{\nabla}_\alpha \mathbf{J}\right) \tag{36}$$

It can similarly be shown that

$$\mathbf{E}\left(\boldsymbol{\nabla}_{\beta\beta}\mathbf{q}_2\right) = -2\mathbf{E}\left(\boldsymbol{\nabla}_\beta \mathbf{J}'\boldsymbol{\nabla}_\beta \mathbf{J}\right) \tag{37}$$

The consistent estimators in (17) and (18) are the sample analogs to (36) and (37).

Appendix D: The do-file for 2SRI estimation of the birthweight model

```
/*************************************************
** Purpose: Mullahy (1997) birthweight model.
** Estimation of the model using the 2SRI and
** GMM approaches to account for endogeneity
** of smoking.
*************************************************/

/*************************************************
** Read in the data.
*************************************************/
use mullahy-birthweight-data.dta

/*************************************************
** Transform birthweight ounces to pounds.
*************************************************/
generate BIRTHWTLB=BIRTHWT/16

*************************************************
** Compute and display descriptive statistics.
*************************************************/
summarize

/*************************************************
** Simple NLS estimation.
*************************************************/
glm BIRTHWTLB CIGSPREG PARITY WHITE MALE, ///
   family(gaussian) link(log) vce(robust)

/*************************************************
** 2SRI estimation begins here.
*************************************************/
/*************************************************
** First-stage NLS estimation of the auxiliary
** exponential regression (via GLM). The
** exponential regression for the 2SRI first
** stage is   xp = exp(w*a)+ xu.
** Conduct Wald test of joint significance of
** the instruments.
** Save xuhat and the predicted values from the
** regression.
*************************************************/
glm CIGSPREG PARITY WHITE MALE EDFATHER EDMOTHER FAMINCOM CIGTAX88, ///
   family(gaussian) link(log) vce(robust)
test (EDFATHER = 0) (EDMOTHER = 0) (FAMINCOM = 0) (CIGTAX88 = 0)
predict xuhat, response
predict expWalpha, mu

/*************************************************
** Load the coefficient vector and covariance
** matrix from first-stage GLM into Mata
** matrices.
*************************************************/
mata: alpha=st_matrix("e(b)")´
mata: covalpha=st_matrix("e(V)")

/*************************************************
```

```
** Apply GLM for the 2SRI second stage.
*************************************************/
glm BIRTHWTLB CIGSPREG PARITY WHITE MALE xuhat, ///
   family(gaussian) link(log) vce(robust)
predict expXbeta, mu

/*************************************************
** Load the coefficient vector and covariance
** matrix from second-stage GLM into Mata
** matrices.
*************************************************/
mata: beta=st_matrix("e(b)")´
mata: covbeta=st_matrix("e(V)")

/*************************************************
** Do GMM estimation.
*************************************************/
gmm (BIRTHWTLB/exp(xb:CIGSPREG PARITY WHITE MALE + b0)-1), ///
   instruments(PARITY WHITE MALE EDFATHER EDMOTHER FAMINCOM CIGTAX88)

/*************************************************
** Use the Stata "putmata" command to send
** Stata data variables into Mata vectors.
*************************************************/
putmata CIGSPREG BIRTHWTLB PARITY WHITE MALE EDFATHER ///
   EDMOTHER FAMINCOM CIGTAX88 xuhat expWalpha expXbeta

/*************************************************
** Mata start-up.
*************************************************/
mata:

/*************************************************
** Make the estimate of betau explicit.
*************************************************/
bxu=beta[5]

/*************************************************
** Load the W-variables for the RHS of the
** first-stage GLM equation into a Mata matrix.
** -- Don´t include the policy variable or xuhat.
** -- Do include the IVs.
** -- Do include a constant term.
*************************************************/
W=PARITY, WHITE, MALE, EDFATHER, EDMOTHER, FAMINCOM, ///
   CIGTAX88, J(rows(PARITY),1,1)

/*************************************************
** Load the X-variables for the RHS of the
** second-stage GLM equation into a Mata matrix.
*************************************************/
X=CIGSPREG,PARITY, WHITE, MALE, xuhat,J(rows(CIGSPREG),1,1)

/*************************************************
** Compute the Xbeta index by multiplying the
** matrix of exogenous variables (X) by the
** coefficient vectors.
*************************************************/
Xbeta=X*beta
```

```
/************************************************
** Compute the asymptotic covariance matrix of
** the 2SRI estimate of beta.
************************************************/
paJ=-bxu:*expXbeta:*expWalpha:*W
pbJ=expXbeta:*X
Bba=pbJ´*paJ
Bbb=pbJ´*pbJ
d22=invsym(Bbb)*Bba*covalpha*Bba´*invsym(Bbb)+covbeta

/************************************************
**  2SRI estimate of beta with correct
**  asymptotic t statistic
************************************************/
/************************************************
**  2SRI second-stage results for beta
************************************************/
/************************************************
**  First, the uncorrected t statistics
************************************************/
tstatwrong=beta:/sqrt(diagonal(covbeta))

/************************************************
**  Now, the corrected t statistics
************************************************/
tstatbeta=beta:/sqrt(diagonal(d22))

/************************************************
**  Display the results.
************************************************/
"Second-Stage Estimates, True Asy t-stats and p-values"
pvalues=2:*(1:-normal(abs(tstatbeta)))
header="variable","estimate","t-stat","wrong-t-stat","p-value" \
"","","","",""
varnames="CIGSPREG", "PARITY", "WHITE", "MALE", "xuhat","constant"
results=beta,tstatbeta,tstatwrong,pvalues
resview=strofreal(results)
header \ (varnames´,resview)

/************************************************
** Close Mata.
************************************************/
end
```