Even Simpler Standard Errors for Two-Stage Optimization Estimators:
Mata Implementation via the DERIV Command

by

Joseph V. Terza

Department of Economics
Indiana University Purdue University Indianapolis
Indianapolis, IN 46202

(July, 2018)
Two-Stage Estimation: Example -- Smoking and Infant Birth Weight

-- Consider the regression model of Mullahy (1997) in which

\[ Y = \text{infant birth weight in lbs.} \]

\[ X_p = \text{number of cigarettes smoked per day during pregnancy.} \]

-- Objective to regress Y on \( X_p \) with a view toward the estimation of (and drawing inferences regarding) the causal effect of the latter on the former.

Smoking and Infant Birth Weight (cont’d)

-- Two complicating factors:

-- the regression specification is nonlinear because Y is non-negative.

-- \( X_p \) is likely to be *endogenous* – correlated with unobservable variates that are also correlated with Y.

-- For example, unobserved unhealthy behaviors may be correlated with both smoking and infant birth weight.

-- If the endogeneity of \( X_p \) is not explicitly accounted for in estimation, effects on Y due to the unobservables will be attributed to \( X_p \) and the regression results will not be causally interpretable (CI).
Remedy: Two-Stage Residual Inclusion (2SRI) Estimation

-- Can use a 2SRI estimator (Terza et al., 2008, Terza 2017a and 2018) to account for endogeneity and avoid bias.

-- The two stage are:

  -- Estimate “auxiliary” regression of $X_p$ on some controls [including instrumental variables (IV)].

  -- Estimate “outcome” regression of $Y$ on $X_p$, controls (not including IV), and the residuals from the auxiliary regression.


Two-Stage Estimation: Example – Education and Family Size

-- As another example, we revisit the regression model of Wang and Famoye (1997).

-- We diverge a bit from the authors and begin the analysis by specifying the potential outcome (PO) version of the model in which

\[ X_p^* \equiv \text{exogenously imposed (EI) version of relevant causal variable} \]

\[ \equiv \text{EI wife’s years of education} \]

\[ Y_{X_p^*} \equiv \text{relevant PO for EI version of relevant causal variable} \]

\[ \equiv \text{potential number of children in the family if EI wife’s education is } X_p^*. \]

Education and Family Size (cont’d)

-- For the sake of argument we assume the following PO specification

\[
\text{pdf}(Y^*_{X_p} \mid X_o) = f(Y^*_{X_p} \mid X_o)(X^*_{X_p}, X_o; \pi) \equiv \text{POI}(Y^*_{X_p}, \lambda^*)
\] (1)

where \( Y^*_{X_p} = 0, 1, \ldots, \infty \)

\[
\text{POI}(A, \ b) \equiv \text{the pdf of the Poisson random variable } A \text{ with parameter } b
\]

\[
\equiv \frac{b^A \exp(-b)}{A!}.
\]

\[
\lambda^* = E[Y^*_{X_p} \mid X_o] = \exp(X^*_{X_p} \beta_p + X_o \beta_o).
\] (2)

and \( X_o \) is a vector of regression controls (no endogeneity here).

-- Here \( \pi' = \beta' = [\beta_p \ \beta_o] \).
Two-Stage Marginal Effect (2SME) Estimation: Education and Family Size

Suppose that our estimation objective is the average incremental effect (AIE) of an additional year of education on the number of children in the family, i.e.,

\[
\text{AIE}(1) = E[Y_{X_p^{pre}+1}] - E[Y_{X_p^{pre}+1}]
\]  
(3)

where \(X_p^{pre}\) is the pre-increment EI wife’s education.

Given (2) we can rewrite (3) as

\[
\text{AIE}(1) = E\left[\exp([X_p^{pre} + 1]\beta_p + X_o\beta_o)\right] - E\left[\exp(X_p^{pre}\beta_p + X_o\beta_o)\right]
\]  
(4)
2SME Estimation: Education and Family Size (cont’d)

-- Assuming we have consistent estimates of $\beta_p$ and $\beta_o$ (say $\hat{\beta}_p$ and $\hat{\beta}_o$) and taking $X_{p}^{\text{pre}}$ to be the EI version of observable wife’s education ($X_{pi}$), (4) can be consistently estimated using:

$$\overline{\text{AIE}}(1) = \sum_{i=1}^{n} \frac{1}{n} \left\{ \exp([X_{pi} + 1]\hat{\beta}_p + X_{oi}\hat{\beta}_o) - \exp(X_{pi}\hat{\beta}_p + X_{oi}\hat{\beta}_o) \right\}$$

where $X_{oi}$ represents the observed vector of controls.

*Note that substituting the observed values ($Y_i, X_{pi}$, and $X_{oi}$) for $Y_{X_p^*}, X_p^*$ and $X_o$ in (1) will not necessarily yield consistent maximum likelihood estimates (MLE) of $\beta_p$ and $\beta_o$. The specific conditions under which such MLE are consistent are detailed in Terza (2018).

2SME Estimation: Education and Family Size (cont’d)

-- The two stages are:

-- Estimate $\beta' = [\beta_p \quad \beta'_o]$ by Poisson regressing Y on $X_p$ and $X_o$.

-- Estimate AIE of an additional year of wife’s education using (5).
Asymptotically Correct Standard Errors (ACSE) for Two-Stage Estimators:

Using the Mata DERIV Command

-- The objective here is to show how the Mata DERIV command can be used to simplify otherwise daunting coding and calculation of ACSE for the class of two-stage estimators of which 2SRI and 2SME are members.

-- For brevity and ease of exposition, I focus here on 2SME estimators.
A Somewhat General Form of the 2SME Estimator

-- Let’s first consider a more general form of the 2SME estimator

\[
\hat{\text{ME}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\text{me}}_i
\]  

(6)

where \( \hat{\text{me}}_i \) is shorthand notation for \( \text{me}(X_{pi}, \Delta_i, X_{oi}, \hat{\pi}) \), \( \hat{\pi} \) is the first-stage estimator of \( \pi \) and

\[
m(1, X_o; \pi) - m(0, X_o; \pi)
\]  

(6-a)

\[
\text{me}(X_{p}^{\text{pre}}, \Delta, X_o, \pi) = m(X_{p}^{\text{pre}} + \Delta, X_o, \pi) - m(X_{p}^{\text{pre}}, X_o, \pi)
\]  

(6-b)

\[
\left. \frac{\partial m(a, b; \pi)}{\partial a} \right|_{a=X_{p}^{\text{pre}}, b=X_o}
\]

(6-c)
The 2SME Estimator (cont’d)

-- (14-a) defines the general form of the *average treatment effect* (ATE)

-- (14-b) defines the general form of the *average incremental effect* (AIE)

-- (14-c) defines the general form of the *average marginal effect* (AME)
ACSE for 2SME Estimators

-- In this case, we seek the estimated asymptotically correct variance of $\hat{\text{ME}}$ [i.e. $\text{EACV}(\hat{\text{ME}})$] the square root of which is the correct asymptotic standard error.

-- Based on general results for two-stage optimization estimators (2SOE) and the fact that 2SME estimators are 2SOE, Terza (2016a and b) shows that the formulation of the $\text{EACV}(\hat{\text{ME}})$ is


ACSE for 2SME Estimators (cont’d)

\[
\left( \frac{1}{n} \sum_{i=1}^{n} \nabla_\pi \hat{m}_{ei} \right) \left( \text{AVAR}(\hat{\pi}) \right) \left( \frac{1}{n} \sum_{i=1}^{n} \nabla_\pi \hat{m}_{ei} \right) + \frac{1}{n} \sum_{i=1}^{n} \left( \hat{m}_{ei} - \text{ME} \right)^2
\]

(7)

where

\text{AVAR}(\hat{\beta}) \) is the estimated asymptotic covariance matrix of \( \hat{\pi} \)

\( \nabla_\pi \hat{m}_{ei} \) denotes the gradient of \( \hat{m}_{ei} \) with respect to \( \pi \)

and

\( \nabla_\pi \hat{m}_{ei} \) represents \( \nabla_\pi \hat{m}_{ei} \) with \( X_{pi}^{\text{pre}}, X_{oi} \) and \( \hat{\pi} \) substituted for \( X_p^{\text{pre}}, X_o \) and \( \pi \); respectively.
ACSE for 2SME Estimators (cont’d)

-- $\text{AVAR}(\hat{\pi})$ can be obtained directly from the Stata output for the relevant Stata regression command.

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{\text{me}}_i - \hat{\text{ME}})^2$$

is easily calculated using Mata, given that $\hat{\text{ME}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\text{me}}_i$ has already been calculated (i.e., $\hat{\text{me}}_i$ and $\hat{\text{ME}}$ are already in hand).

-- Direct calculation of the remaining component of (7), viz. $\frac{\sum_{i=1}^{n} \nabla_{\pi} \hat{\text{me}}_i}{n}$, requires analytic derivation of $\nabla_{\pi} \text{me}$ and Mata coding of $\nabla_{\pi} \hat{\text{me}}_i$. 
To the above education and family size model we add:

$$X_o = [\text{employed } \text{eduwe } \text{agewife } \text{faminc } \text{race } \text{city } 1]$$

where

- employed = 1 if employed, 0 if not
- agewife = wife’s age in years
- faminc = family income
- race = 1 if wife is white, 0 if not
- city = if the family is situated in a county whose largest city has more than 50K people.
ACSE for 2SME Estimators: Education and Family Size (cont’d)

-- Recall that in this case we seek to estimate the AIE of an additional year of wife’s education using

$$
\overline{\text{AIE}(1)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \exp(\beta_p X_i + \beta_o X_{oi}) - \exp(\hat{\beta}_p X_i + \hat{\beta}_o X_{oi}) \right\}
$$

where \( \hat{\beta}' = [\hat{\beta}_p \ \hat{\beta}_o] \) is the vector of Poisson parameter estimates.

-- Following Terza (2016b, 2017b), in this example we have

$$
\nabla_{\beta} \widehat{\text{me}_i} = \exp(\beta_p X_i + \beta_o X_{oi}) \left[ \begin{array}{c} X_i + 1 \\ X_o \end{array} \right] - \exp(\hat{\beta}_p X_i + \hat{\beta}_o X_{oi}) \left[ \begin{array}{cc} X_p & X_{oi} \end{array} \right]
$$

ACSE for 2SME Estimators: Education and Family Size

-- I estimated $\beta$ using the Stata POISSON command and obtained \[ \frac{\sum_{i=1}^{n} \nabla_{\beta} \hat{m}e_i}{n} \] using (9) and direct Mata coding. Following are the results

+-----------------------------------------------------+
  1 |         AIE       asy-se   asy-t-stat      p-value  |
  2 |                                                     |
  3 |  -.0458791     .0140945    -3.255099     .0011335  |
+-----------------------------------------------------+

-- Alternatively, we can use the Mata DERIV command to calculate the ACSE and corresponding t-stat without having: a) the exact formulation of $\nabla_{\beta} \hat{m}e_i$; and b) to directly Mata code of $\nabla_{\beta} \hat{m}e_i$. 
The Mata DERIV Command: Basic Elements of Implementation

-- Requisite matrix and vector initializations.

-- User-supplied Mata evaluator function subroutine for calculation of the relevant function

-- e.g., me(X_{p}^{pre}, \Delta, X_{o}, \beta) with vector argument \beta.

-- DERIV also accommodates vector-valued functions, say F(b), of a vector argument b. In this case DERIV calculates the Jacobian matrix of F(b) with respect to b. Such Jacobian matrices are required, for example, in the 2SRI context).

-- Name the project using:

\[ <\text{user-supplied project name}>=\text{deriv\_init()} \]
-- Identify the relevant evaluator function using:

    deriv_init_evaluator(<project name>, &<evaluator function name>())

-- Identify the evaluator type using:

    deriv_init_evaluatortype((<project name>, "v")

    ONLY NEEDED IF RELEVANT FUNCTION IS VECTOR-VALUED.

-- Give the value of the argument vector at which the gradient (Jacobian) is to be evaluated using:

    deriv_init_params(<project name>, <name of vector of argument values>)
-- Invoke DERIV using:

```
deriv(<project name>,1)
```

-- Load the Jacobian into a specified matrix using:

```
<specified Jacobian matrix name>=deriv_result_scores((<project name>))
```

**ONLY NEEDED IF RELEVANT FUNCTION IS VECTOR-VALUED.**
Education and Family Size: ACSE via the Mata DERIV Command

-- Recall that to get the correct standard error of our AIE estimate we needed to calculate the following vector

$$\sum_{i=1}^{n} \nabla_{\beta} \hat{m}_{ei} \overline{n}$$

--- (10)---

-- Use of the Mata DERIV command allows you to avoid having to derive the explicit form of (10) because it affords a way to numerically approximate the components of this gradient vector.
Education and Family Size: ACSE via the DERIV Command (cont’d)

\[ \sum_{i=1}^{n} \nabla_\beta \hat{m}_{ei} \bigg/ n = \nabla_\beta \left( \sum_{i=1}^{n} \hat{m}_{ei} \bigg/ n \right). \]

-- Note that we can write \( \frac{\sum \nabla_\beta \hat{m}_{ei}}{n} = \nabla_\beta \left( \frac{\sum \hat{m}_{ei}}{n} \right) \).

-- Note also that the entity inside the parentheses is a scalar-valued function of a vector... one of the function types for which the DERIV command is designed.

-- We assume that the Stata POISSON command has been used to estimate \( \beta \).

-- We also assume that relevant Mata commands have been used to save the vector of parameter estimates in the Mata vector “\( \text{betahat} \)” along with \( \text{AVAR}^*(\hat{\beta}) \) in the Mata matrix “\( \text{Vbetahat} \)”. See Terza (2017b).
Education and Family Size: ACSE via the DERIV Command (cont’d)

-- Mata coding for the DERIV command:

/********************************************************* 
** User-supplied Evaluator function for deriv( ). 
**********************************************************/ 
function MEfunct(bbeta,MME) 
{ 
  external me 
  external XD 
  external X 
  me=exp(XD*bbeta'):-exp(X*bbeta') 
  MME=mean(me) 
}

/********************************************************* 
** Name the project. 
**********************************************************/
MECALC=deriv_init()

/********************************************************* 
** Identify the relevant evaluator function. 
**********************************************************/ 
deriv_init_evaluator(MECALC,&MEfunct())
Education and Family Size: The Mata DERIV Command (cont’d)

/*************************************************
** Give the parameter vector value at which the ** gradient is to be evaluated.
*************************************************/
deriv_init_params(MECALC,betahat)

/*************************************************
** Invoke DERIV and load gradient into specified ** vector.
*************************************************/
gradbetape=deriv(MECALC,1)

/*************************************************
** Invoke DERIV and load function value into ** specified scalar.
*************************************************/
ME=deriv(MECALC,0)

/*************************************************
** Compute the estimated asymptotically ** correct variance of the 2SME estimator.
*************************************************/
varME=gradbetape*(n:*'(betaVhat))' * gradbetape'/*
*:+=mean((me:-ME):^2)
--- Results using DERIV:

<table>
<thead>
<tr>
<th></th>
<th>AIE</th>
<th>asy-se</th>
<th>asy-t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-.0458791</td>
<td>.0140945</td>
<td>-3.255099</td>
<td>.0011335</td>
</tr>
</tbody>
</table>

--- Results using analytic gradient and direct Mata coding:

<table>
<thead>
<tr>
<th></th>
<th>AIE</th>
<th>asy-se</th>
<th>asy-t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-.0458791</td>
<td>.0140945</td>
<td>-3.255099</td>
<td>.0011335</td>
</tr>
</tbody>
</table>
So What??? One Can Apply the Stata “margins” Command

-- Yes this is true but…

-- The above example is merely intended to illustrate the simplicity of using DERIV in cases for which:

   a) “margins” is not available

and

   b) the formulation of \( \text{me}(X_p, \Delta, X_i, \pi) \) is analytically daunting.

-- For example, in the education and family size example, suppose that we want to accommodate potential under-dispersion, as is typical of fertility data, by replacing the Poisson assumption for the distribution of the PO (family size) with the Conway-Maxwell Poisson (CMP).
So What??? One Can Apply the Stata “margins” Command (cont’d)

-- The CMP accommodates equi-, over- and under-dispersed data and in this context has the following conditional mean function

\[
E[Y_{X^*_p} | X_o] = \lambda^* \left( \sum_{j=0}^{\infty} \frac{(\lambda^*)^j}{(j!)^\sigma} \right) \frac{\sum_{j=1}^{\infty} j(\lambda^*)^{j-1}}{\sum_{j=0}^{\infty} (\lambda^*)^j} \quad (28)
\]

with

\[
\lambda^* = \exp(X^*_p \beta_p + X_o \beta_o)
\]

and \(\sigma > 0\) being the dispersion parameter
So What??? One Can Apply the Stata “margins” Command (cont’d)

-- The CMP nests the standard Poisson distribution when $\sigma = 1$. The over- (under-) dispersion case corresponds to if $\sigma < 1$ ($\sigma_2 > 1$).

-- In this case, the “margins” command is not available and the formulation of $\text{me}(X_p, \Delta, X_i, \pi)$ for the targeted AIE ($\Delta = 1$) is relatively daunting.
By the Way…

-- Under the Poisson PO assumption, I calculated the AIE and its asymptotic standard error (asymptotic t-stat) using the “margins” command and got:

<table>
<thead>
<tr>
<th></th>
<th>Terza (2016a, 2016b)</th>
<th>margins command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic Standard Error</td>
<td>.0140945</td>
<td>.0124141</td>
</tr>
<tr>
<td>Asymptotic t-statistic</td>
<td>-3.255099</td>
<td>-3.695725</td>
</tr>
</tbody>
</table>

-- Note the difference in the asymptotic t-stats.

-- For a detailed discussion see Terza (2017b).