xtewreg: Estimating the errors-in-variables model using high-order cumulants and moments

Timothy Erickson
Bureau of Labor Statistics
Washington, DC
Erickson.Timothy@bls.gov

Robert Parham
University of Rochester
Rochester, NY
robert.parham@simon.rochester.edu

Toni M. Whited
University of Michigan
Ann Arbor, MI
twhited@umich.edu

Abstract. We consider a multiple mismeasured regressor errors-in-variables (EIV) model. We present `xtewreg`, a command for using two-step generalized method of moments (GMM) and minimum distance estimators that exploit overidentifying information contained in high-order cumulants or moments of the data. The command supports either cumulant or moment estimation, internal support for the bootstrap with moment condition recentering, an arbitrary number of mismeasured regressors and perfectly measured regressors, and cumulants or moments up to an arbitrary degree. We also demonstrate how to use the estimators in the context of a corporate leverage regression.

Keywords: xtewreg, errors-in-variables, high-order moments, high-order cumulants

1 Introduction

We present the Stata command `xtewreg` for implementing the estimators in Erickson and Whited (2000, 2002, 2012), and Erickson et al. (2014) for the classical errors-in-variables (EIV) model. The model is of a multiple linear regression in which any number of the explanatory variables can be measured with an additive error. The estimators produce consistent regression slope estimates by exploiting information contained in the third- and higher-order cumulants or moments of the data. Such estimators are of interest because ordinary least squares (OLS) is inconsistent when the independent

variables of a linear regression are replaced with error-laden measurements or proxy variables. Typically, researchers address this issue by finding additional observable variables that can serve as instruments, but in many situations no such variables are available. Consistent estimators based on the original, unaugmented set of observable variables are therefore potentially quite valuable.

The article proceeds as follows: We start with a description of the EIV model and the moment and cumulant estimators in Section 2. In section 3, we describe the `xtewreg` command. Section 4 contains a demonstration of the use of the `xtewreg` command.

2 Background

This section draws from Erickson and Whited (2002) and Erickson et al. (2014) to sketch the errors-in-variables model and the high-order moment and cumulant estimators. For details, see Erickson and Whited (2002) and Erickson et al. (2014).

2.1 Notation and Model

Let (y_i, x_i, z_i), $i = 1, \ldots, n$, be a sequence of observable vectors, where $x_i \equiv (x_{i1}, \ldots, x_{iJ})$ and $z_i \equiv (1, z_{i1}, \ldots, z_{iM})$. Let $(u_i, \varepsilon_i, \chi_i)$ be a sequence of *un*observable vectors, where $\chi_i \equiv (\chi_{i1}, \ldots, \chi_{iJ})$ and $\varepsilon_i \equiv (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$. We consider a multiple-regressor version of the classical errors-in-variables model, where (y_i, x_i, z_i) is related to $(u_i, \varepsilon_i, \chi_i)$ and

unknown parameters $\alpha \equiv (\alpha_0, \alpha_1, \ldots, \alpha_M)'$ and $\beta \equiv (\beta_1, \ldots, \beta_J)'$ according to:

$$y_i = z_i\alpha + \chi_i\beta + u_i \tag{1}$$

$$x_i = \chi_i + \varepsilon_i. \tag{2}$$

Equation (1) is a linear regression model containing J regressors χ_i that are imperfectly measured by x_i according to (2), and M perfectly measured regressors, z_i. The assumption of unit slopes and zero-valued intercepts in (2) is required to identify the parameters in (1). We assume the variables in (1) and (2) satisfy the following assumptions:

Assumption 1. (i) $(u_i, \varepsilon_i, \chi_i, z_i)$, $i = 1, \ldots, n$, *is an i.i.d. sequence;* (ii) u_i *and the elements of* ε_i, χ_i, *and* z_i *have finite moments of every order;* (iii) (u_i, ε_i) *is independent of* (χ_i, z_i), *and the individual elements in* (u_i, ε_i) *are independent of each other;* (iv) $E(u_i) = 0$ *and* $E(\varepsilon_i) = 0$*;* (vi) $E\left[(\chi_i, z_i)'(\chi_i, z_i)\right]$ *is positive definite.*

Before sketching the estimators, we partial out the perfectly measured variables, and rewrite the model in terms of population residuals. The $1 \times J$ residual from the population linear regression of x_i on z_i is $x_i - z_i\mu_x$, where:

$$\mu_x \equiv \left[E(z_i'z_i)\right]^{-1} E(z_i'x_i)\,. \tag{3}$$

The corresponding $1 \times J$ residual from the population linear regression of χ_i on z_i is:

$$\eta_i \equiv \chi_i - z_i\mu_x, \tag{4}$$

where μ_x appears because (2) and the independence of ε_i and z_i imply

$$\mu_x = \left[E(z_i'z_i)\right]^{-1} E[z_i'\left(\chi_i + \varepsilon_i\right)] = \left[E(z_i'z_i)\right]^{-1} E(z_i'\chi_i)\,.$$

Note that subtracting $z_i\mu_x$ from both sides of (2) gives:

$$x_i - z_i\mu_x = \eta_i + \varepsilon_i. \tag{5}$$

Similarly, the residual from the population linear regression of y_i on z_i is $y_i - z_i\mu_y$, where $\mu_y \equiv [E(z_i'z_i)]^{-1} E(z_i'y_i)$. Equation (1) and the independence of u_i and z_i imply:

$$\begin{aligned} \mu_y &= [E(z_i'z_i)]^{-1} E[z_i' (z_i\alpha + \chi_i\beta + u_i)] \\ &= \alpha + \mu_x\beta. \end{aligned} \tag{6}$$

Therefore, subtracting $z_i\mu_y$ from both sides of (1) gives:

$$y_i - z_i\mu_y = \eta_i\beta + u_i. \tag{7}$$

2.2 Estimators

Both the cumulant and moment estimators are based on a two-step approach to estimation, where the first step is to substitute the least squares estimates

$$\begin{aligned} \hat{\mu}_x &\equiv \left[\sum_{i=1}^{n} z_i'z_i\right]^{-1} \sum_{i=1}^{n} z_i'x_i \\ \hat{\mu}_y &\equiv \left[\sum_{i=1}^{n} z_i'z_i\right]^{-1} \sum_{i=1}^{n} z_i'y_i \end{aligned}$$

into (5) and (7), and the second step is to estimate β using sample cumulants or moments of $y_i - z_i\hat{\mu}_y$ and $x_i - z_i\hat{\mu}_x$.

Regarding the practical implementation of this step, it is important that the researcher classify all possibly mismeasured variables as belonging to the vector χ_i, and

not to the vector z_i. Correct classification is important even if one or more of the mismeasured variables is not of primary economic interest. If any mismeasured regressor is classified as perfectly measured, then the ordinary least squares estimates, $\hat{\mu}_x$ and $\hat{\mu}_y$, will be biased. In this case, equations (5) and (7) will be misspecified.

Moments

The high-order moment estimators are based on moment conditions derived from (5) and (7) by taking powers of these two equations, multiplying the results together and then taking expectations of both sides. The resulting equations express observable higher order moments and cross-moments of the data as nonlinear functions of β and moments of unobservable variables, where these latter moments are treated as parameters. The general form for these moment equations is given by:

$$E\left[(y_i - z_i\mu_y)^{r_0}\prod_{j=1}^{J}(x_{ij} - z_i\mu_{xj})^{r_j}\right] = \sum_{v\in V}\sum_{k\in K} a_{v,k}\left(\prod_{j=1}^{J}\beta_j^{v_j}\right)E\left(\prod_{j=1}^{J}\eta_{ij}^{(v_j+k_j)}\right)\left(\prod_{j=1}^{J}E\left(\varepsilon_{ij}^{(r_j-k_j)}\right)\right)E(u_i^{v_0})\,, \tag{8}$$

where $v \equiv (v_0, v_1, \ldots, v_J)$ and $k \equiv (k_1, ..., k_J)$ are vectors of nonnegative integers, $V \equiv \left\{v : \sum_{j=0}^{J} v_j = r_0\right\}$, $K \equiv \left\{k : \sum_{j=1}^{J} k_j \leq \sum_{j=0}^{J} r_j,\ k_j \leq r_j,\ j = 1, ..., J\right\}$, and

$$a_{v,k} \equiv \frac{r_0!}{v_0!v_1!\cdots v_J!}\prod_{j=1}^{J}\frac{r_j!}{k_j!(r_j-k_j)!}.$$

A GMM estimator can then be constructed by using subsets of these moment conditions, where the weight matrix is simply the covariance matrix of the observable

moments on the left-hand side of (8), adjusted to account for the sampling variation in the estimates of μ_x and μ_y. As explained in more detail in Erickson and Whited (2002), it is natural to consider sets of equations based on moment up to a certain order $N = r_0 + r_1 + \cdots + r_J$, so the `xtewreg` command considers sets of moment equations based on moments of orders $3, 4, 5, \ldots$.

We now describe a simple example of (8) that can be used to construct an estimator. We consider the case of a single mismeasured regressor, so $J = 1$. First, we square (7), multiply the result by (5), and take expectations of both sides, obtaining:

$$E\left((y_i - z_i\mu_y)^2(x_i - z_i\mu_x)\right) = \beta^2 E\left(\eta_i^3\right). \tag{9}$$

Similarly if we square (5), multiply the result by (7), and take expectations, we obtain:

$$E\left((y_i - z_i\mu_y)(x_i - z_i\mu_x)^2\right) = \beta E\left(\eta_i^3\right). \tag{10}$$

If $\beta \neq 0$ and $E\left(\eta_i^3\right) \neq 0$, dividing (9) by (10) produces a consistent estimator for β:

$$\begin{aligned} \beta &= \frac{\beta^2 E\left(\eta_i^3\right)}{\beta E\left(\eta_i^3\right)} \\ &= \frac{E\left((y_i - z_i\mu_y)^2(x_i - z_i\mu_x)\right)}{E\left((y_i - z_i\mu_y)(x_i - z_i\mu_x)^2\right)}. \end{aligned} \tag{11}$$

An estimator can be derived from (11) by replacing the population moments by sample moments.

Cumulants

As shown in Erickson et al. (2014), the cumulant estimators are asymptotically equivalent to the moment estimators, but they have a convenient closed form. The following

outline of the estimators draws from Erickson et al. (2014). Let $K(s_0, s_1, \ldots, s_J)$ be the cumulant of order s_0 in $y_i - z_i\mu_y$ and s_j in $x_{ij} - z_i\mu_{xj}$. The cumulant estimators are based on the result from Geary (1942) that for any $(s_0, s_1, \ldots, s_J)$ containing two or more positive elements, the following relationship between cumulants holds:

$$K(s_0 + 1, s_1, \ldots, s_J) = \beta_1 K(s_0, s_1 + 1, \ldots, s_J) + \cdots + \beta_J K(s_0, s_1, \ldots, s_J + 1). \quad (12)$$

There are an infinity of equations given by (12), one for each admissible vector $(s_0, s_1, \ldots, s_J)$. Let

$$K_y = K_x\beta \quad (13)$$

denote a system of M equations of the form (12). If $M = J$ and $\det K_x \neq 0$, then it is possible to solve for β.

We consider possibly overidentified estimators for β, so $M \geq J$. Let $\hat{K_y}$ and $\hat{K_x}$ be consistent estimates of K_y and K_x, and let $\hat{W}$ be a symmetric positive definite matrix. The estimator $\hat{\beta}$ solves:

$$\hat{\beta} \equiv \text{argmin}_{b \in \Re^J} \left(\hat{K_y} - \hat{K_x}b\right)' \hat{W} \left(\hat{K_y} - \hat{K_x}b\right). \quad (14)$$

Because $\hat{K_y} - \hat{K_x}b$ is linear in b, (14) has the solution

$$\hat{\beta} = \left(\hat{K_x}'\hat{W}\hat{K_x}\right)^{-1} \hat{K_x}'\hat{W}\hat{K_y}, \quad (15)$$

whenever $\hat{K_x}$ has full column rank. As in the case of the moment estimators, we consider estimators based on sets of cumulant equations up to a certain integer order, $N = s_0 + s_1 + \ldots + s_J$.

2.3 Identifying assumptions

Both the cumulant and moment estimators obtain identification from the third and higher order moments or cumulants of the regression variables. In particular, as shown in Erickson and Whited (2002), identification requires that the distribution of η satisfies $E\left[(\eta_i c)^3\right] \neq 0$ for every vector of constants $c = (c_1, ..., c_J)$ having at least one nonzero element. For practical problems, this requirement boils down to having nonnormally distributed mismeasured regressors. An example of this requirement can be seen intuitively in equation (11), which contains the third moment of η_i in the denominator. Without a skewed distribution, this particular third-order moment estimator is undefined. The assumption of nonnormality clearly limits the applicability of these estimators. For instance, asset returns are often approximately normally distributed, and many aggregate variables are often approximately lognormally distributed, and typically expressed as natural logarithms. In both of theses cases, the cumulant or moment estimators are unlikely to be of use. However, in many microeconometric settings, especially those in corporate finance and accounting, many regression variables are plausibly nonnormally distributed.

2.4 Other estimates and test statistics

Both the moment and cumulant estimators can produce estimates of the coefficients on the perfectly measured regressors, α, which can be recovered from the identity (6). The estimators can also produce estimates of the population coefficient of determination for

(1), which can be written as:

$$\rho^2 = \frac{\mu_y' \text{var}(z_i)\,\mu_y + \beta' \text{var}(\eta_i)\,\beta}{\mu_y' \text{var}(z_i)\,\mu_y + \beta' \text{var}(\eta_i)\,\beta + E(u_i^2)}. \tag{16}$$

Similarly, the estimators can produce an estimate of the population coefficients of determination for (2):

$$\tau_j^2 = \frac{\mu_{xj}' \text{var}(z_i)\,\mu_{xj} + \text{var}(\eta_{ij})}{\mu_{xj}' \text{var}(z_i)\,\mu_{xj} + \text{var}(\eta_{ij}) + E(\varepsilon_{ij}^2)}. \tag{17}$$

In (17), the j subscript refers to the j^{th} mismeasured regressor. The standard errors for α, ρ^2, and τ^2 are calculated by stacking the influence functions for their various components to obtain the covariance matrix of these components and then using the delta method.

Finally, except for the case of the third-order moment estimator with one mismeasured regressor, all of the estimators included in `xtewreg` are overidentified. Both the cumulant and moment estimators are accompanied by the standard Hansen-Sargan test statistic for the overidentifying restrictions.

3 The `xtewreg` command

3.1 Syntax

```
xtewreg depvar misindepvars [ indepvars ] [ if ] [ in ] , maxdeg(#) [
   mismeasured(#) method(string) panmethod(string) bxint(numlist)
   centmom(string) hascons nocons noprn ]
```

Here, *misindepvars* are independent variables assumed to be mismeasured, and *indepvars* are independent variables assumed to be perfectly measured. For more than one

mismeasured variable, use the `mismeasured()` option to specify the number of mismeasured independent variables.

3.2 Options

`maxdeg(#)` sets the highest order of cumulants/moments to use. The minimum value is 3. Very high values (above 8) are not advised, as the computational time for these models increases sharply with maxdeg. `xtewreg` does not provide a default value for `maxdeg()`. This choice is left to the researcher and is an empirical choice. Generally speaking, the more data one has, the higher the order moment or cumulant one can use. A reasonable starting value for applied work is `maxdeg(5)`, but the sensitivity of the estimates to different values of `maxdeg()` should be explored on a case by case basis.

`mismeasured(#)` declares the number of mismeasured independent variables in the model. The default value is 1. `xtewreg` uses this value to distinguish between *misindepvars* and *indepvars*. The first `mismeasured()` independent variables are taken to be *misindepvars*, and the rest are taken to be *indepvars*.

`method(`*string*`)` chooses whether to use high-order cumulants (`cml`, the default) or high-order moments (`mom`). While `xtewreg` supports both high-order cumulant and moment estimators, using high-order moment estimators is not advised because the high-order moment estimators require a numerical minimization procedure when computing the GMM objective function, whereas the cumulant-based estimators are linear and naturally have a closed-form solution.

`panmethod(`*string*`)` chooses whether to perform panel estimation by using a clustered weight matrix used for the cumulant or moment estimators (`cls`, the default) or by combining cross-sections using a minimum distance estimator (`cmd`). While `xtewreg` supports panel data using both clustered weighting matrices and classical minimum distance, classical minimum distance can entail long computation time for panels with a large time dimension.

`bxint(`*numlist*`)` is a list of starting values for the coefficients on *misindepvars*. This option requires setting `method(mom)`. The high-order moment estimators require numerical minimization of a nonlinear objective function and thus require starting values. The default (if `bxint()` is omitted) is to use both the OLS coefficients and the coefficients from `maxdeg(3)` as possible starting values. If there are J *misindepvars* and one wishes to provide K sets of possible starting values, `numlist` should be of order $J \times K$.

`centmom(`*string*`)` is a directive supporting the centering of the moment conditions for bootstrap computation of t-test and overidentification test critical values. The option takes one of the values [`set, use, reset`]. `centmom(set)` saves the value of the moment conditions for the entire sample, and should be used *before* using the `bootstrap` command. `centmom(use)` should be specified *when* using `bootstrap` with `xtewreg`. `centmom(reset)` resets the value of saved moment conditions, and is rarely used.

`hascons` indicates that *indepvar* already contains a constant variable, and so a constant should not be added by the estimation procedure.

`nocons` requires that a constant will not be added by the estimation procedure. Note that when this option is used, the researcher should verify all variables included in the estimation have mean zero, or regression results will be inconsistent.

`noprn` suppresses the printing of the results table.

3.3 Saved Results

`xtewreg` saves the following in `e()`:

Scalars

`e(N)`	number of observations
`e(rho)`	estimate of ρ^2
`e(SErho)`	standard error for ρ^2
`e(Jstat)`	Sargan-Hansen J statistic
`e(Jval)`	p-value for Jstat
`e(dfree)`	degrees of freedom for Jstat
`e(obj)`	minimized value of the GMM objective function

Macros

`e(bxint)`	initial guesses for β
`e(method)`	method used for estimation
`e(panmethod)`	panel method used for estimation

Matrices

`e(b)`	regression coefficients
`e(V)`	covariance matrix for e(b)
`e(serr)`	standard errors for e(b)
`e(tau)`	estimates of proxy accuracy, τ^2
`e(SEtau)`	standard errors for τ^2
`e(vcrhotau)`	covariance matrix for ρ^2 and τ_j^2
`e(w)`	weighting matrix used for estimation

Additionally, `xtewreg` sets two global `MATA` variables:

`EWSAVEDprb` holds the problem structure (i.e. the `Symbolic` estimation equations) for a given number of mismeasured independent variables J and a given maximum cumulant/moment degree M. Generating these equations is computationally intensive, and `xtewreg` saves the last estimated problem structure to optimize repeated estimations of the same problem structure, such as when using the `bootstrap`. When `xtewreg` is

called with a problem structure different from the one last used, it will notify by printing the message "Problem structure different from last executed. Rebuilding problem." `xtewreg` will rebuild and save the new `Symbolic` estimating equations.

`EWSAVEDfCent` holds the centered moment conditions generated by specifying `centmom(set)` and used when specifying `centmom(use)`.

3.4 The `Symbolic` class

To implement a moment system of arbitrary degree M and with arbitrarily many mismeasured variables J, `xtewreg` needs to be able to construct a large set of equations of the type described by the general form in 8. These equations then need to be evaluated for the data provided to calculate the moments and cumulants. To construct these equations, we implement a symbolic algebra class in MATA, `Symbolic`, which supports the complete algebra over the polynomial ring with arbitrarily many indeterminates and with coefficients from the real field. The class is similar in capabilities to Stata's `polyeval` command, with two important differences: it is a stateful MATA class, which allows superior encapsulation, and it supports arbitrarily many indeterminates (e.g. polynomials of the form $\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2^2$), whereas `polyeval` only supports a single indeterminate (i.e. x is a scalar, not a vector). Further discussion of the `Symbolic` class is outside the scope of this paper, but a stand-alone version of the `Symbolic` class is available from the authors upon request.

The `EWSAVEDprb` problem structure contains the set of `Symbolic` equations corresponding to the current degree M and mismeasured variable count J, and these equa-

tions can then be re-evaluated given a set of data. This way, the problem structure is only constructed once, and can then be evaluated multiple times given different data.

4 Investment and leverage example

A firm leverage dataset from Compustat is included with this distribution. It contains over 121,000 firm-year observations for approximately 11,000 firms. We include the following variables, defined in terms of Compustat mnemonic variable names:

- *gvkey* - The Compustat unique firm identifier.
- *fyear* - The firm fiscal year.
- *lever* - Firm leverage, defined as (DLTT+DLC)/AT.
- *mtb* - Firm market-to-book ratio, where the numerator is AT+PRCC_F times CSHO minus CEQ minus TXDB, and the denominator is AT.
- *tangib* - Fixed assets, defined as PPENT/AT.
- *logsales* - The natural log of firm sales (SALE).
- *oi* - Firm operating income, defined as OIBDP/AT.

with all cash items provided in term of deviations from firm means and year means.

First, we provide a summary of the `EPW.dta` dataset:

```
. use "EPW.dta", clear
. xtset gvkey
       panel variable:  gvkey (unbalanced)
. summarize fyear gvkey lever mtb tangib logsales oi
```

Variable	Obs	Mean	Std. Dev.	Min	Max
fyear	121733	1991.063	11.70219	1970	2011
gvkey	121733	21855.43	35329.08	1000	287462
lever	121733	-1.89e-10	.1482152	-.9990212	.9677935
mtb	121733	3.92e-10	.6669923	-9.285594	12.85808
tangib	121733	-2.04e-11	.1167422	-.8033313	.7093889
logsales	121733	8.19e-12	.5162762	-6.318236	4.589146
oi	121733	1.78e-10	.0930696	-1.030859	.6828895

We begin by estimating the relation between leverage and market-to-book and tangibility using an OLS regression. The market-to-book ratio is a proxy for firm growth opportunities, and the ratio of fixed to total assets is a proxy for asset tangibility. We cluster standard errors at the firm level, and use the nocons option in the regression as the dependent variable was de-meaned:

```
. regress lever mtb tangib , vce(cluster gvkey) nocons
Linear regression                               Number of obs =  121733
                                                F(  2, 10795) =  497.72
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0390
                                                Root MSE      =  .14529
                                (Std. Err. adjusted for 10796 clusters in gvkey)
------------------------------------------------------------------------------
             |               Robust
       lever |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mtb |  -.0242871   .0011228   -21.63   0.000    -.0264881   -.0220862
      tangib |   .2049681   .0098809    20.74   0.000     .1855996    .2243365
------------------------------------------------------------------------------
```

Compare these results with those of `xtewreg`, assuming both regressors are measured with error, as indicated by the `mismeasured(2)` option:

```
. xtewreg lever mtb tangib , maxdeg(5) mismeasured(2) nocons
5(2) EIV results                                          N =    121733
                                                          Rho^2 = 0.171
                                                                  (0.009)
------------------------------------------------------------------------------
       lever |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mtb |  -.0339095   .0037329    -9.08   0.000    -.0412259   -.0265932
      tangib |   1.185099   .0373822    31.70   0.000     1.111831    1.258367
------------------------------------------------------------------------------
Tau1^2: 0.570 (0.083)
Tau2^2: 0.172 (0.010)
```

```
Sargan-Hansen J statistic: 210.285   (p=0.000, d=20)
```

Note that the coefficient on tangibility rises by a factor of six, and the coefficient of determination (ρ^2) for the model rises considerably. These are explained by the estimates of the errors in market-to-book and tangibility, measured by the τ_1^2 and τ_2^2 coefficient. These errors, when ignored in OLS, leads to attenuation bias. Further note that the estimation uses cumulants and a clustered weighting matrix (the defaults), and we set `maxdeg(5)` for an estimator based on cumulants up to fifth order.

Next, we add several perfectly measured controls and estimate the model using an OLS regression. We again cluster standard errors at the firm level:

```
. regress lever mtb tangib logsales oi , vce(cluster gvkey) nocons
Linear regression                               Number of obs =  121733
                                                F(  4, 10795) =  530.73
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0728
                                                Root MSE      =  .14272
                                (Std. Err. adjusted for 10796 clusters in gvkey)
------------------------------------------------------------------------------
             |               Robust
       lever |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mtb |  -.0149454   .0011126   -13.43   0.000    -.0171263   -.0127645
      tangib |   .1991992   .0099499    20.02   0.000     .1796956    .2187028
    logsales |   .0394179   .0025019    15.76   0.000     .0345138    .0443221
          oi |  -.2411662   .0092385   -26.10   0.000    -.2592753   -.2230571
------------------------------------------------------------------------------
```

Compare these results with those of `xtewreg`, assuming again that *mtb* and *tangib* are measured with error:

```
. xtewreg lever mtb tangib logsales oi , maxdeg(5) mismeasured(2) nocons
5(2) EIV results                                          N =    121733
                                                          Rho^2 = 0.199
                                                                  (0.009)
------------------------------------------------------------------------------
       lever |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mtb |  -.0318794   .0044092    -7.23   0.000    -.0405212   -.0232376
      tangib |   1.207097   .0378421    31.90   0.000     1.132928    1.281266
    logsales |   .0579218   .0036609    15.82   0.000     .0507467     .065097
```

```
          oi |  -.0566342    .015969    -3.55   0.000    -.0879328   -.0253356
------------------------------------------------------------------------------
Tau1^2: 0.478 (0.080)
Tau2^2: 0.186 (0.010)
Sargan-Hansen J statistic: 245.977   (p=0.000, d=20)
```

Note the message printed by `xtewreg` regarding rebuilding the problem, as the parameters of the problem are different from those used during the latest call to `xtewreg`.

Note that the *J*-statistic for the test of overidentifying restrictions is quite large. This result indicates a violation of one of the conditions in Assumption 1, with the likely culprit being a regression error, u_i, that is independent of the regressors, χ_i and z_i. The leverage regression we have chosen as an example, although widely used, likely suffers from problems of omitted variables.

Repeating the estimation with `maxdeg(8)` so as to use all moments condition up to degree 8 yields:

```
. xtewreg lever mtb tangib logsales oi , maxdeg(8) mismeasured(2) nocons
Problem structure different from last executed. Rebuilding problem.
8(2) EIV results                                              N =    121733
                                                              Rho^2 = 0.204
                                                                     (0.008)
------------------------------------------------------------------------------
       lever |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mtb |  -.0241135   .0008971   -26.88   0.000    -.0258717   -.0223552
      tangib |   1.264688   .0079716   158.65   0.000     1.249064    1.280312
    logsales |   .0599253   .0036729    16.32   0.000     .0527266    .0671239
          oi |   -.061634   .0119372    -5.16   0.000    -.0850305   -.0382375
------------------------------------------------------------------------------
Tau1^2: 0.611 (0.066)
Tau2^2: 0.179 (0.008)
Sargan-Hansen J statistic: 1289.990   (p=0.000, d=96)
```

4.1 Using `bootstrap` with `xtewreg`

To calculate the bootstrapped critical values for the test statistics, we need to recenter the moment conditions for every bootstrap iteration (see Hall and Horowitz 1996, for details). To do so, we first execute `xtewreg` on the entire dataset while setting the `centmom(set)` option. Next, we prefix xtewreg with the bootstrap command, while setting the `centmom(use)` option:

```
. xtewreg lever mtb tangib logsales oi , maxdeg(5) mismeasured(2) centmom(set) nocons
(output omitted)
. bootstrap t_mtb=(_b[mtb]/el(e(serr),1,1)) t_tangib=(_b[tangib]/el(e(serr),2,1)) ///
>           t_logsales=(_b[logsales]/el(e(serr),3,1)) t_oi=(_b[oi]/el(e(serr),4,1)) , ///
>           rep(100) seed(1337) cluster(gvkey) notable: ///
>           xtewreg lever mtb tangib logsales oi , maxdeg(5) mismeasured(2) centmom(use) nocons
(running xtewreg on estimation sample)
Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
Bootstrap results                                 Number of obs      =    121733
                                                  Replications       =       100

      command:  xtewreg lever mtb tangib logsales oi, maxdeg(5) mismeasured(2) centmom(use) nocons
        t_mtb:  _b[mtb]/el(e(serr),1,1)
     t_tangib:  _b[tangib]/el(e(serr),2,1)
   t_logsales:  _b[logsales]/el(e(serr),3,1)
         t_oi:  _b[oi]/el(e(serr),4,1)

. estat bootstrap, p
Bootstrap results                                 Number of obs      =    121733
                                                  Replications       =       100

      command:  xtewreg lever mtb tangib logsales oi, maxdeg(5) mismeasured(2) centmom(use) nocons
        t_mtb:  _b[mtb]/el(e(serr),1,1)
     t_tangib:  _b[tangib]/el(e(serr),2,1)
   t_logsales:  _b[logsales]/el(e(serr),3,1)
         t_oi:  _b[oi]/el(e(serr),4,1)

                              (Replications based on 10796 clusters in gvkey)
------------------------------------------------------------------------------
             |   Observed               Bootstrap
             |      Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
       t_mtb |  -7.2302641   1.893839    1.009294  -7.611491  -3.683594   (P)
    t_tangib |   31.898225  -8.269736   1.3878318   20.88363   25.95456   (P)
  t_logsales |   15.821917  -4.489249   .97056268   9.338326   13.13195   (P)
        t_oi |  -3.5465207   .7876053   .89791996  -4.257066  -1.154614   (P)
------------------------------------------------------------------------------
(P)    percentile confidence interval
```

Note that we use the bootstrap to calculate the critical value for the t-statistic, as it is an asymptotically pivotal statistic (see Horowitz 2001, for details). Furthermore, we use the percentile method to derive confidence intervals and p-values (by `estat bootstrap, p` after executing `bootstrap`).

5 References

Erickson, T., C. Jiang, and T. M. Whited. 2014. Minimum Distance Estimation of the Errors-in-Variables Model Using Linear Cumulant Equations. *Journal of Econometrics* 183: 211–221.

Erickson, T., and T. M. Whited. 2000. Measurement Error and the Relationship Between Investment and q. *Journal of Political Economy* 108: 1027–1057.

———. 2002. Two-step GMM estimation of the errors-in-variables model using high-order moments. *Econometric Theory* 18(3): 776–799.

———. 2012. Treating measurement error in Tobin's q. *Review of Financial Studies* 25: 1286–1329.

Geary, R. C. 1942. Inherent Relations between Random Variables. *Proceedings of the Royal Irish Academy A* 47: 63–76.

Hall, P., and J. L. Horowitz. 1996. Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators. *Econometrica* 64(4): 891–916.

Horowitz, J. L. 2001. The Bootstrap. In *Handbook of Econometrics*, vol. 5, 3159 – 3228. Elsevier.

About the authors

Timothy Erickson is a research economist at the Bureau of Labor Statistics in the Division of Price and Index Number Research.

Robert Parham is a PhD student at the University of Rochester.

Toni M. Whited is the Dale L. Dykema Professor of Business Administration at the University of Michigan and a Research Associate at the National Bureau of Economic Research.