The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
14, Number 1, pp. 60–75

miivfind: A command for identifying model-implied instrumental variables for structural equation models in Stata

Shawn Bauldry
University of Alabama at Birmingham
Birmingham, AL
sbauldry@uab.edu

Abstract. This article presents a new Stata command, `miivfind`, that implements an algorithm developed by Bollen and Bauer (2004, *Sociological Methods and Research* 32: 425–452) to find the model-implied instrumental variables (MIIVs) from an identified structural equation model. MIIVs allow researchers to draw on instrumental-variable estimators, such as two-stage least-squares estimators, to obtain estimates for the parameters of a hypothesized structural equation model. It can be difficult to identify MIIVs by inspection of either a diagram of the model or the model equations. Two examples are provided that illustrate the use of `miivfind` to identify MIIVs and some of the advantages of a MIIV estimator as compared with a maximum likelihood estimator. By assisting in the process of finding MIIVs, `miivfind` facilitates the use of an alternative class of estimators, instrumental-variable estimators, to the standard maximum-likelihood and asymptotic-distribution free estimators available for structural equation models.

Keywords: st0324, miivfind, structural equation models, instrumental-variable estimators, model-implied instrumental variables

1 Introduction

Interest in the use of Stata for structural equation models (SEMs) has increased with the introduction of a suite of commands for fitting SEMs. The `sem` command comes with the option of using one of three popular estimators: a maximum likelihood estimator (ML), a direct maximum-likelihood mean- and variance-adjusted estimator for missing data, and an asymptotic distribution free estimator. Though not with the `sem` command, one can also obtain parameter estimates for SEMs by using instrumental-variable (IV) estimators and other procedures available in Stata. Madansky (1964), Hägglund (1982), and Jöreskog (1983) provided the initial development of IV estimators for factor analytic models, and Jöreskog and Sörbom (1993) discussed a procedure involving an IV estimator for obtaining estimates for the parameters in the latent-variable model. However, these were restricted in a variety of ways ranging from assuming uncorrelated errors to not providing significance tests for statistical inferences. In a series of articles, Bollen and colleagues derived alternative IV estimators for the general SEM, for higher-order factor analysis, and for SEMs involving endogenous categorical or censored observed variables (Bollen 1996b, 2001; Bollen and Biesanz 2002; Bollen and Maydeu-Olivares 2007).

Instrumental-variable estimators hold a number of advantages over traditional ML estimators for SEMs. Most IV estimators are noniterative and thus not subject to computational issues that can lead to nonconvergence with ML estimators (or other iterative estimators). In addition, IV estimators for SEMs can be used equation by equation (that is, treated as limited-information estimators), which helps isolate bias resulting from structural misspecifications in other parts of the model (Bollen et al. 2007). In contrast, ML estimators are full-information estimators, which have the potential to spread bias because of misspecification throughout the model parameters.[1] Overidentification tests have been derived for IV estimators that in the context of SEMs can help identify structural misspecifications in any given equation (Kirby and Bollen 2009). IV estimators for SEMs are asymptotically "distribution free" and allow for heteroskedastic-consistent standard errors (Bollen 1996a). Furthermore, IV estimators are readily available in Stata with the `ivregress` command.

The key challenge in using IV estimators for SEMs with latent variables lies in rewriting the model equations in terms of observed variables and identifying which of the observed variables, if any, are suitable instruments for each equation. According to Bollen's (1996b) terminology, such instruments are referred to as model-implied instrumental variables (MIIVs) because they arise from the specification of the model (discussed in more detail in the following section). Bollen (1996b, 2001) demonstrated how to rewrite the general SEM equations involving latent variables in terms of observed variables only, and Bollen and Bauer (2004) developed an algorithm for identifying the MIIVs for each equation.

This article presents a new Stata command, `miivfind`, that uses the algorithm developed by Bollen and Bauer (2004) to output the MIIVs for a given SEM supplied by the user. Though the algorithm was developed for SEMs involving latent variables, it is also applicable to other SEMs in which there are no latent variables. For instance, `miivfind` can also be used to identify instruments in classic simultaneous equation models without latent variables. The article is organized as follows: The next two sections provide a brief overview of MIIVs and the logic of the algorithm for finding them. The fourth section provides a detailed discussion of the inputs and outputs of `miivfind`. Finally, the fifth section provides two examples of using `miivfind` to identify the MIIVs for two different SEMs. The first example illustrates the use of `miivfind` for a simple model. The second example is designed to illustrate both the use of `miivfind` and some of the potential benefits of an IV estimator. This example compares the estimates from an IV estimator (based on using the output from `miivfind`) and an ML estimator.

1. Full-information estimators are typically more efficient than limited-information estimators, but simulation results suggest that the efficiency gains of ML over two-stage least-squares (2SLS) estimators with respect to SEMs are minimal (Bollen et al. 2007).

2 What are MIIVs?

Readers may be unfamiliar with instrumental variables in relation to latent-variable SEMs. In econometrics texts, IVs are usually introduced to address endogenous covariates (for example, Cameron and Trivedi [2005] and Wooldridge [2010]). When a model includes a potentially endogenous covariate, then the analyst searches for additional variables that could serve as instruments. In some cases, the additional variables are wholly outside the model (for example, using settler mortality rates as an instrument for institutions in models of economic growth, as in Acemoglu, Johnson, and Robinson [2001]); in other cases, time-lagged or spatially lagged variables can serve as instruments (for example, Anselin [1988] and Arellano and Bond [1991]). More recently, econometricians have focused on instrumental variables that arise in randomized or quasi-randomized experiments (for example, Angrist and Pischke [2009]; Heckman, Urzua, and Vytlacil [2006]; and Imbens and Angrist [1994]). By definition, an instrumental variable, z, is a variable that satisfies two conditions: 1) z must be correlated with an endogenous regressor x and 2) z must be uncorrelated with the disturbance term ϵ (Cameron and Trivedi 2005). With a little algebra, it is possible to show how such variables arise naturally in nearly all identified SEMs.

Using a slightly modified version of Jöreskog's (1977) notation, we write the latent-variable model as

$$\boldsymbol{\eta} = \boldsymbol{\alpha}_{\boldsymbol{\eta}} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1}$$

where $\boldsymbol{\eta}$ is a vector of endogenous latent variables, $\boldsymbol{\xi}$ is a vector of exogenous latent variables, and $\boldsymbol{\zeta}$ is a vector of disturbances. The matrix of coefficients $\mathbf{B}$ gives the effects of the latent endogenous variables on each other, and the matrix of coefficients $\boldsymbol{\Gamma}$ gives the effects of the latent exogenous variables on the latent endogenous variables. We assume that $E(\boldsymbol{\zeta}) = \mathbf{0}$, $E(\boldsymbol{\xi}'\boldsymbol{\zeta}) = \mathbf{0}$, and $(\mathbf{I} - \mathbf{B})$ is nonsingular. The measurement model includes two equations,

$$\mathbf{y} = \boldsymbol{\alpha}_{\mathbf{y}} + \boldsymbol{\Lambda}_{\mathbf{y}}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{2}$$

$$\mathbf{x} = \boldsymbol{\alpha}_{\mathbf{x}} + \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{3}$$

where $\mathbf{y}$ and $\mathbf{x}$ are, respectively, vectors of indicators of the latent endogenous and exogenous variables, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are vectors of disturbances. The matrices of factor loadings $\boldsymbol{\Lambda}_{\mathbf{y}}$ and $\boldsymbol{\Lambda}_{\mathbf{x}}$ give the effects of the latent variables on the observed indicators. We assume that $E(\boldsymbol{\epsilon}) = E(\boldsymbol{\delta}) = \mathbf{0}$ and that $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are uncorrelated with each other and with $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$.

Three additional matrices are used to identify the variances and covariances among the disturbances in (1), (2), and (3). The matrix for the variances and covariances among the disturbances for the latent-variable model, $\boldsymbol{\zeta}$, is the $\boldsymbol{\Psi}$ matrix. The matrices for the variances and covariances among the disturbances for the measurement models, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$, are, respectively, $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$.

Equations (1), (2), and (3) are quite general and include many well-known models as special cases. If one assumes no measurement error (that is, $\mathbf{y} = \mathbf{I}\boldsymbol{\eta}$ and $\mathbf{x} = \mathbf{I}\boldsymbol{\xi}$), then (1) reduces to the classic simultaneous equations model from econometrics. If

in addition, there is only a single outcome, y, then (1) further reduces to a standard multiple regression model. Alternatively, if one does not specify structural effects among the latent variables (that is, $\mathbf{B} = \mathbf{\Gamma} = \mathbf{0}$) and does not specify $\mathbf{y}$, then only (3) remains, a confirmatory factor analysis model. Because the general SEM includes many well-known models as special cases, MIIV estimators are widely applicable to models not typically thought of as SEMs.

To use an IV estimator, the equations in (1), (2), and (3) need to be rewritten in terms of only observed variables. This is accomplished by choosing an indicator to serve as a scaling indicator for each latent variable and solving the measurement equations for these indicators for the latent variables. If we denote the vectors of scaling indicators for the endogenous and exogenous latent variables as $\mathbf{y}_1$ and $\mathbf{x}_1$, then we have

$$\boldsymbol{\eta} = \mathbf{y}_1 - \boldsymbol{\epsilon}_1 \tag{4}$$

$$\boldsymbol{\xi} = \mathbf{x}_1 - \boldsymbol{\delta}_1 \tag{5}$$

Following Bollen (2001), (4) and (5) can be substituted into (1), (2), and (3), and with the remaining nonscaling indicators denoted by $\mathbf{y}_2$ and $\mathbf{x}_2$, we have

$$\mathbf{y}_1 = \boldsymbol{\alpha}_{\boldsymbol{\eta}} + \mathbf{B}\mathbf{y}_1 + \mathbf{\Gamma}\mathbf{x}_1 + \boldsymbol{\epsilon}_1 - \mathbf{B}\boldsymbol{\epsilon}_1 - \mathbf{\Gamma}\boldsymbol{\delta}_1 + \boldsymbol{\zeta} \tag{6}$$

$$\mathbf{y}_2 = \boldsymbol{\alpha}_{\mathbf{y}_2} + \mathbf{\Lambda}_{\mathbf{y}_2}\mathbf{y}_1 - \mathbf{\Lambda}_{\mathbf{y}_2}\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2 \tag{7}$$

$$\mathbf{x}_2 = \boldsymbol{\alpha}_{\mathbf{x}_2} + \mathbf{\Lambda}_{\mathbf{x}_2}\mathbf{x}_1 - \mathbf{\Lambda}_{\mathbf{x}_2}\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2 \tag{8}$$

With the equations for the general SEM expressed in terms of observed variables, it is now possible to identify potential MIIVs. As noted above, a vector of instruments, $\mathbf{z}$, must in general satisfy two conditions: 1) $\mathbf{z}$ must be correlated with the covariates in the equation and 2) $\mathbf{z}$ must be uncorrelated with the disturbance term in the equation. The first condition is not difficult to check. After identifying candidate MIIVs, one can regress the covariates in the equation on the candidate MIIVs and use standard procedures for assessing the strength of the instruments (Cameron and Trivedi 2005; Stock and Yogo 2005). The second condition is typically more challenging to assess.

The model equations (6), (7), and (8), written in terms of observed variables, introduce composite disturbance terms. For instance, the disturbance term in (6) is a function of $\boldsymbol{\epsilon}, \boldsymbol{\delta}$, and $\boldsymbol{\zeta}$. All the observed variables are candidate instruments for any given equation; however, those that are associated with any of the error terms in the composite disturbance fail to meet the second condition and thus are not suitable instruments for the given equation. The specification of the model determines which observed variables are correlated with each disturbance, hence the term MIIVs. Of course, the proposed model may be misspecified, in which case the MIIVs based on the proposed model may in fact be correlated with the composite disturbance and therefore fail to meet the second condition. As long as there are more MIIVs than right-hand-side variables for a given equation, standard overidentification tests for instrumental-variable estimators can be used to help assess whether the instruments are uncorrelated with the error term (Sargan 1958; Basmann 1960). The overidentification tests serve a similar function of detecting model misspecification as the χ^2 test for model fit available with ML estimators (Bollen 1989).

3 Algorithm for finding MIIVs

In practice, it can be difficult to determine the MIIVs for a given equation by examining a set of equations representing a SEM. Bollen and Bauer (2004) developed an algorithm for identifying MIIVs in any SEM that can be expressed using equations (1), (2), and (3). The algorithm involves four steps, which are outlined below. For more details about the algorithm, readers should consult Bollen and Bauer (2004).

In the first step of the algorithm, the composite disturbances are worked out for each equation [see equations (6)–(8)]. As described below, the inputs for the algorithm include vectors and matrices with indices for the observed variables and error terms and indicators for the parameters associated with a given SEM. The first step simply involves systematically searching through these inputs to identify each equation in the model and the associated error terms. If we let J be the number of equations with K disturbances, then in the `miivfind` command, the composite disturbances are stored in a matrix $\mathbf{C}$ with J rows and $K+1$ columns. The entries in the first column of the matrix are indices for the dependent variable for a given equation. The entries in the remaining columns are indices for the disturbances associated with a given equation or a 0 if necessary to ensure conformability (that is, when a given equation has fewer disturbances than the maximum number across all equations).

The second step of the algorithm is to identify the total effects of each disturbance. Although it is unusual to consider the total effects of an error term, it is possible to treat error terms, because they are variables, in the same way as other variables to calculate their total effects on all the observed variables. The total effects of the disturbances can be obtained from the coefficients from the following reduced form equations (see Bollen [1987, 1989] for details regarding derivations and stability conditions):

$$\mathbf{y} = \boldsymbol{\alpha}_{\mathbf{y}} + \boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\alpha}_{\boldsymbol{\eta}} + \boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\zeta} + \boldsymbol{\epsilon}$$
$$\mathbf{x} = \boldsymbol{\alpha}_{\mathbf{x}} + \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\xi} + \boldsymbol{\delta}$$

These equations indicate that the total effect of $\boldsymbol{\epsilon}$ on $\mathbf{y}$ and $\boldsymbol{\delta}$ on $\mathbf{x}$ is simply $\mathbf{I}$ and the total effect of $\boldsymbol{\zeta}$ on $\mathbf{y}$ is $\boldsymbol{\Lambda}_{\mathbf{y}}(\mathbf{I}-\mathbf{B})^{-1}$. The algorithm uses these coefficients to construct a matrix $\mathbf{T}$ that has rows equal to the number of observed variables (O) and $K+1$ columns. The first column is reserved for the indices of the observed variables, and the remaining columns contain indices for the disturbances that have an effect on a given variable or a 0 if necessary to ensure conformability.

The third step of the algorithm involves comparing the matrices $\mathbf{C}$ and $\mathbf{T}$ to identify the initial set of potential instruments for each equation. For each equation, the initial set of potential instruments consists of the observed variables that are unaffected by the disturbances for the given equation. This is determined by working through each row of $\mathbf{C}$, which corresponds to each equation, and checking whether the observed variables indexed in $\mathbf{T}$ are affected by the disturbances in a given equation identified in $\mathbf{C}$. In other words, if j indexes equations, k indexes disturbances, and o indexes observed variables, then if $C_{jk} = T_{jo}$, the observed variable o cannot serve as an instrument in equation j, because the observed variable is associated with at least one component of the composite disturbance in equation j. The initial set of instruments, L for each

equation, is stored in a matrix $\mathbf{P}$ with the J rows and the $L+1$ columns. The first column is reserved to index the equation, and the remaining columns contain indices for the observed variables that are potential instruments or zeros to ensure conformability.

The fourth step of the algorithm involves checking whether the potential instruments identified in the third step are ineligible because of the potential presence of correlations among disturbances. This is determined by checking whether the disturbances that have an effect on a given potential instrument are associated with disturbances that are present in the composite error term for a given equation. Information about the associations among disturbances are contained in the error term matrices defined by users as inputs for the algorithm (see discussion in next section). Any potential instrument affected by a disturbance that is correlated with a disturbance in the composite error term for a given equation is removed from further consideration. The remaining instruments, if any, for each equation are stored in a matrix $\mathbf{IV}$ with the J rows and columns equal to the maximum number of instruments across all equations plus one. The matrix $\mathbf{IV}$ thus contains the MIIVs, if any, for each equation of a given SEM.

4 The miivfind command

The new Stata command `miivfind` implements Bollen and Bauer's (2004) algorithm and outputs the MIIVs for each equation from a user-supplied SEM. A few preliminary steps are required to prepare the inputs (vectors and matrices) for `miivfind`. The first step is to specify a SEM and assign each of the observed variables a unique index number (for example, $y_1 = 1, y_2 = 2, \ldots, x_1 = 5, x_2 = 6$). The second step is to identify the scaling indicators for each of the latent variables and construct row vectors for $\mathbf{y}_1, \mathbf{x}_1, \mathbf{y}_2$, and $\mathbf{x}_2$ using the index numbers for the variables. If there are no variables that correspond with one of the given vectors, the user should specify it as a zero vector. For instance, if there are no observed variables in $\mathbf{x}_2$, then it should be set to [0].

The third step is to construct coefficient matrices that correspond with the parameters in the specified model. The parameter matrices from equations (6), (7), and (8) are $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Lambda}_{\mathbf{y}_2}$, and $\mathbf{\Lambda}_{\mathbf{x}_2}$. These matrices should be defined using 1s to indicate free parameters and 0s to indicate fixed parameters. As with the row vectors, if the matrix is not an element of the model, then the matrix should be set to a zero matrix. It is important to note that $\mathbf{\Lambda}_{\mathbf{y}_2}$ and $\mathbf{\Lambda}_{\mathbf{x}_2}$ refer to the matrices of factor loadings for the variables contained in $\mathbf{y}_2$ and $\mathbf{x}_2$, respectively.

Finally, the fourth step is to specify the covariance matrices for the disturbances. There are three matrices that need to be specified: $\mathbf{\Theta}_{\boldsymbol{\epsilon}}$, $\mathbf{\Theta}_{\boldsymbol{\delta}}$, and $\mathbf{\Psi}$. These matrices also require unique index numbers. It is convenient to use the same index numbers assigned to the observed variables to reference the error variances associated with the given variable. For instance, if the variable y_1 is assigned the index 1, then it is convenient to also assign the error variance for y_1, the $[1,1]$ cell in $\mathbf{\Theta}_{\boldsymbol{\epsilon}}$, the index 1. Any remaining covariances among disturbances need to be assigned their own index numbers; however, as symmetric matrices, the index number assigned to the i,jth cell should be the same as the number assigned to the j,ith cell. Any variance or covariance among the distur-

bances that is not estimated should be set to 0. As with all the previous vectors and matrices, if one of these matrices is not an element of the model, then the matrix should be set to the zero matrix.

The command `miivfind` takes the 11 matrices outlined above in the following order as arguments.

`miivfind` *y1 y2 x1 x2 Beta Gamma Lam_y2 Lam_x2 Theta_ep Theta_del Psi*

The vectors and matrices need to be defined by the user before invoking the `miivfind` command. They can have any name, but the order must be maintained. The command is designed to check the dimensions of each matrix for accuracy and conformability (for example, the vector $\mathbf{y}_1$ should always be a row vector with the number of columns equal to the number of columns in $\mathbf{B}$). If the command returns an error, it is an indication that either the noted vector or matrix was incorrectly specified or the vectors and matrices were entered out of order. The command outputs a table that lists each equation based on the index assigned to the dependent variable and the indices of any observed variables that are MIIVs for each equation.

5 Two examples

This section illustrates the use of the `miivfind` command with two examples. The first example is a simple measurement model with one latent variable, four indicators of the latent variable, and a correlation between two of the disturbances (see figure 1). The equations for this model are

$$\begin{aligned} x_1 &= \xi_1 + \delta_1 \\ x_2 &= \alpha_2 + \lambda_2\xi_1 + \delta_2 \\ x_3 &= \alpha_3 + \lambda_3\xi_1 + \delta_3 \\ x_4 &= \alpha_4 + \lambda_4\xi_1 + \delta_4 \end{aligned}$$

with $\text{Cov}(\delta_2, \delta_3) \neq 0$. As depicted in figure 1, the xs are assigned the index numbers 1 through 4. The latent variable, ξ_1, is scaled to x_1; therefore, $\mathbf{x}_1 = [1]$ and $\mathbf{x}_2 = [2\ 3\ 4]$. The only free parameters in the model are the factor loadings for x_2, x_3, and x_4, so we have

$$\mathbf{\Lambda}_{\mathbf{x}_2} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Finally, the same index numbers used for the xs are used for the covariance matrix among the disturbances with the addition of 5 for $\text{Cov}(\delta_2, \delta_3)$. This results in

$$\mathbf{\Theta}_{\boldsymbol{\delta}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 5 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

All the remaining vectors and matrices are set to the zero vector or matrix.

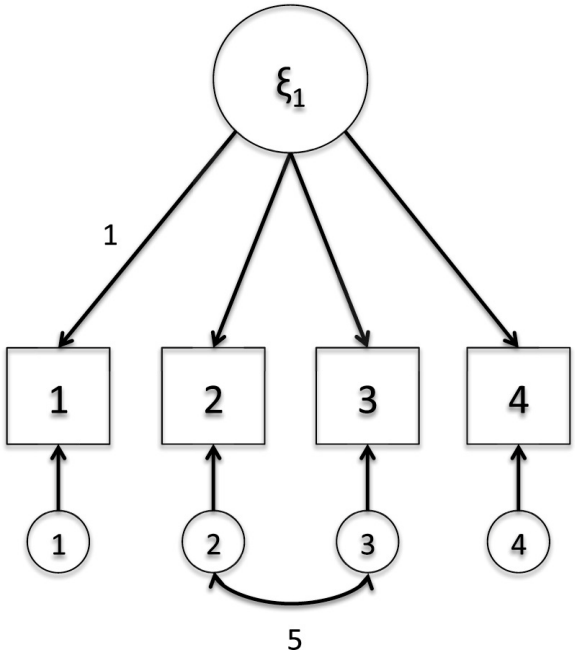

Figure 1. Diagram of model for example 1

The following log illustrates the Stata commands to enter the matrices, the use of `miivfind`, and the resulting output for this model.

```
. *** Input vectors and matrices
. matrix y1 = (0)
. matrix y2 = (0)
. matrix x1 = (1)
. matrix x2 = (2, 3, 4)
. matrix Beta = (0)
. matrix Gamma = (0)
. matrix Ly2 = (0)
. matrix Lx2 = (1 \ 1 \ 1)
. matrix ThetaE = (0)
. matrix ThetaD = (1, 0, 0, 0 \ 0, 2, 5, 0 \ 0, 5, 3, 0 \ 0, 0, 0, 4)
. matrix Psi = (0)
```

```
. *** Invoking miivfind
. miivfind y1 y2 x1 x2 Beta Gamma Ly2 Lx2 ThetaE ThetaD Psi

List of MIIVs (if any) by Equation DV
------------------------------------------------------------------------------
DV    MIIVs
------------------------------------------------------------------------------
2     4
3     4
4     2, 3
------------------------------------------------------------------------------
Note: numbers in table are indices assigned to variables.
```

The output indicates that x_4 is the only MIIV available for the x_2 and x_3 equations. Both x_2 and x_3 are MIIVs for the x_4 equation. With information about the MIIVs for each equation, the MIIV-2SLS estimator (Bollen 1996b) can be used with Stata's `ivregress` command in combination with various postestimation options to assess the potential presence of weak instruments and overidentification tests when available (see [R] **ivregress** for options).

The second example is designed to illustrate the use of `miivfind` to identify MIIVs in a more complicated model and to compare a MIIV-2SLS estimator with an ML estimator in a context in which there are advantages to using the MIIV-2SLS estimator. Figure 2 depicts a model in which the latent variable ξ_1 has a direct effect on the latent variable η_1. Each of the latent variables is measured by three indicators. The exogenous latent variable, ξ_1, also affects two of the indicators for η_1. In the figure, these cross loadings are drawn with dashed lines to indicate these effects are present in the population but unknown to the analyst and are thus a source of misspecification.

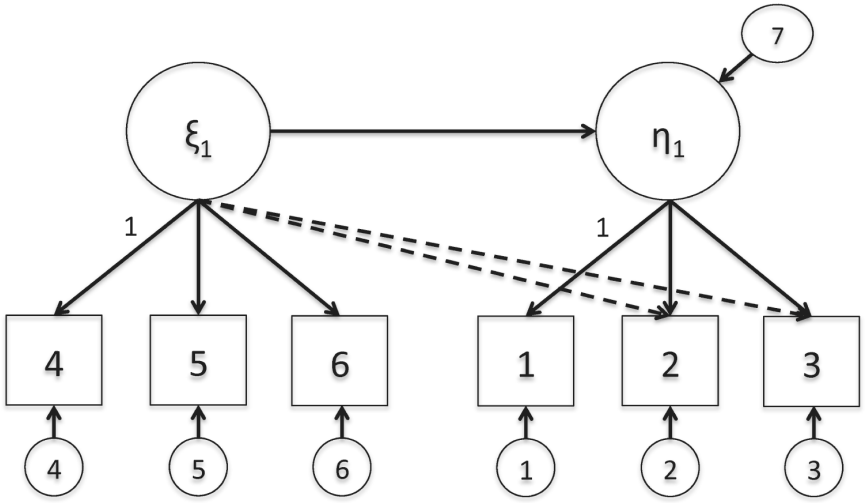

Figure 2. Diagram of model for example 2. Dashed lines indicate effects in the population that are unknown to the analyst.

As limited-information estimators, MIIV estimators are more robust to model misspecifications than full-information estimators, such as the standard ML estimator for SEMs. In this example, as we will demonstrate, the MIIV-2SLS estimator is able to recover the structural parameter (the effect of ξ_1 on η_1), while the standard ML estimator returns a biased estimate. This illustrates one potential advantage of MIIV estimators and underscores their usefulness as an alternative option for fitting SEMs. This is not to suggest, however, that MIIV estimators are always to be preferred to ML estimators.

The equations for this model (from the perspective of the analyst who is unaware of the cross loadings) are

$$\begin{aligned} \eta_1 &= \alpha_{\eta_1} + \gamma_{11}\xi_1 + \zeta_1 \\ x_j &= \alpha_{x_j} + \lambda_{x_j}\xi_1 + \delta_j \\ y_j &= \alpha_{y_j} + \lambda_{y_j}\eta_1 + \epsilon_j \end{aligned}$$

for $j = 1, 2, 3$. The latent variables are both scaled to the first indicator; that is, $\alpha_{x_1} = \alpha_{y_1} = 0$ and $\lambda_{x_1} = \lambda_{y_1} = 1$.

As shown in figure 2, the ys are indexed 1, 2, and 3, and the xs are indexed 4, 5, and 6. The two scaling indicators are x_1 and y_1; thus $\mathbf{y}_1 = [1]$ and $\mathbf{x}_1 = [4]$. This leaves $\mathbf{y}_2 = [2\ 3]$ and $\mathbf{x}_2 = [5\ 6]$. The matrices for the free structural parameters are

$$\mathbf{B} = \begin{bmatrix} 0 \end{bmatrix} \text{ and } \mathbf{\Gamma} = \begin{bmatrix} 1 \end{bmatrix}$$

The matrices for the free parameters in the measurement model are

$$\mathbf{\Lambda}_{\mathbf{y}_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \mathbf{\Lambda}_{\mathbf{x}_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Note that the failure to consider the cross loadings from ξ_1 to the indicators of η_1 means that the analyst's model is misspecified. The covariance matrices for the disturbances are

$$\mathbf{\Theta}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \mathbf{\Theta}_{\boldsymbol{\delta}} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 6 \end{bmatrix}, \text{ and } \mathbf{\Psi} = \begin{bmatrix} 7 \end{bmatrix}$$

The following log illustrates the Stata commands used to enter the matrices for the second example and the use of `miivfind`.

```
. *** Input vectors and matrices
. matrix y1 = (1)
. matrix y2 = (2,3)
. matrix x1 = (4)
. matrix x2 = (5,6)
. matrix Beta = (0)
. matrix Gamma = (1)
. matrix Ly2 = (1 \ 1)
. matrix Lx2 = (1 \ 1)
```

```
. matrix ThetaE = (1, 0, 0 \ 0, 2, 0 \ 0, 0, 3)
. matrix ThetaD = (4, 0, 0 \ 0, 5, 0 \ 0, 0, 6)
. matrix Psi = (7)
. *** Invoking miivfind
. miivfind y1 y2 x1 x2 Beta Gamma Ly2 Lx2 ThetaE ThetaD Psi

List of MIIVs (if any) by Equation DV
------------------------------------------------------------------------------
DV    MIIVs
------------------------------------------------------------------------------
1     5, 6
2     3, 4, 5, 6
3     2, 4, 5, 6
5     1, 2, 3, 6
6     1, 2, 3, 5
------------------------------------------------------------------------------
Note: numbers in table are indices assigned to variables.
```

The output indicates that x_2 and x_3 (that is, the variables labeled "5" and "6") are MIIVs for the y_1 equation, which is used to estimate the structural parameter γ_{11}. The MIIVs for the remaining equations, which are used to estimate the measurement model parameters (λs), include all the observed variables except for the appropriate scaling variable (that is, y_1 for the indicators of η_1 and x_1 for the indicators of ξ_1) and the dependent variable for the given equation.

To illustrate how one can use the information from `miivfind` and the potential value of a MIIV estimator, I constructed a simple simulation treating the model illustrated in figure 2 as the population model. I parameterized the model as follows: I set $\gamma_{11} = 0.5$; I set the free-factor loadings to 0.7 and the cross loadings to 0.4; I set all intercepts to 0; and I set the distributions for all the disturbances and the latent exogenous variable to $N(0, 1)$. I drew 1,000 samples of size $N = 1{,}000$. For each sample, I estimated the analyst's model (that is, the misspecified model that omits the cross loadings) by using Stata's `sem` command, and I estimated γ_{11} and λ_{y_2} by using MIIV-2SLS estimators via Stata's `ivregress` command with the information provided by `miivfind` concerning which observed variables to use as instruments.

The following log provides the Stata code for the simulation program.

```
. *** Defining simulation program
. capture program drop miivsim
. program miivsim, rclass
  1.   clear
  2.   *** Generating data
.   set obs 1000
  3.   gen xi1 = rnormal(0,1)
  4.   gen eta1 = 0.5*xi1 + rnormal(0,1)
  5.   gen x1 = xi1 + rnormal(0,1)
  6.   gen x2 = 0.7*xi1 + rnormal(0,1)
  7.   gen x3 = 0.7*xi1 + rnormal(0,1)
  8.   gen y1 = eta1 + rnormal(0,1)
  9.   gen y2 = 0.7*eta1 + 0.4*xi1 + rnormal(0,1)
 10.   gen y3 = 0.7*eta1 + 0.4*xi1 + rnormal(0,1)
 11.   *** ML estimator
.   qui sem (Xi -> x1 x2 x3) (Eta -> y1 y2 y3) (Xi -> Eta)
 12.   matrix b1 = e(b)
 13.   return scalar chisq_pval = e(p_ms)
 14.   return scalar ml_gamma11 = b1[1,1]
 15.   return scalar ml_lambda21 = b1[1,10]
 16.   *** MIIV-2SLS estimators
.   qui ivregress 2sls y1 (x1 = x2 x3)
 17.   qui estat overid
 18.   matrix b2 = e(b)
 19.   return scalar tsls_gamma11 = b2[1,1]
 20.   return scalar y1_sargan_pval = r(p_sargan)
 21.   qui ivregress 2sls y2 (y1 = x1 x2 x3 y3)
 22.   qui estat overid
 23.   matrix b3 = e(b)
 24.   return scalar tsls_lambda21 = b3[1,1]
 25.   return scalar y2_sargan_pval = r(p_sargan)
 26. end
```

After simulating data from the population model, the program first estimates the analyst's model by using Stata's `sem` command and stores estimates for γ_{11}, λ_{y_2}, and the p-value for the χ^2 test of overall model fit. The program then uses Stata's `ivregress 2sls` command to invoke MIIV-2SLS estimators to obtain estimates for γ_{11} by using the y_1 equation with the instruments identified by `miivfind` and for λ_{y_2} by using the y_2 equation with instruments identified by `miivfind`. The program stores the estimates for γ_{11}, λ_{y_2}, and, because these equations are overidentified, the p-values from Sargan's (1958) χ^2 test of the overidentifying restrictions. Because the y_1 equation is correctly specified (that is, there are no misspecifications in the structural component of the model), one would expect roughly 5% of the simulation samples to return a significant p-value (less than 0.05). In contrast, because the y_2 equation is part of the measurement model, which is misspecified, we would expect a substantial proportion of the samples to return a significant p-value. I set the seed to 69,185,391 with the simulation command.

Table 1 presents the mean parameter estimates and root mean square errors (RMSEs) for γ_{11} and λ_{y_2} from the two estimators. As expected, the MIIV-2SLS estimator recovers the structural parameter γ_{11}, while the ML estimator is biased. Also as expected, both estimators are biased for λ_{y_2} because of the analyst's misspecification of the measurement model. In this case, the RMSEs for the MIIV-2SLS estimator are lower than

the RMSEs for the ML estimator for both parameters, but this is due to the greater bias in the ML estimator. In general, ML estimators are more efficient than MIIV-2SLS estimators, though the gains in efficiency can be quite small in practice (Bollen et al. 2007).

Table 1. Parameter estimates from simulation study

Parameter	Population	MIIV-2SLS		ML	
		Mean	RMSE	Mean	RMSE
γ_{11}	0.5	0.502	0.069	0.653	0.162
λ_{y_2}	0.7	1.017	0.323	1.077	0.382

The χ^2 test for the overall model fit is significant in 99% of the samples, which is an indication that the model is misspecified. The test, however, is for the model as a whole and does not provide any guidance about the nature of the misspecification. Sargan's (1958) test for the y_1 equation, which is used to estimate γ_{11}, is significant in 4.8% of the samples, which is within sampling fluctuation of the nominal 5%. Sargan's (1958) test for the y_2 equation, which is used to estimate λ_{y_2}, is significant in 99% of the samples. This test points to a problem with the measurement model for η_1, which might help analysts narrow down the source of misspecification.

This example and simulation illustrate how the MIIVs identified by `miivfind` can be used in conjunction with the `ivregress` command to obtain parameter estimates for a given SEM in which the MIIV-2SLS estimator has some advantages over the standard ML estimator. This is not to suggest that the MIIV-2SLS estimator will always or even often be preferred over an ML estimator. For additional information about MIIV estimators, readers should consult Bollen and his colleagues' work (Bollen 1996b, 2001; Bollen and Biesanz 2002; Bollen and Maydeu-Olivares 2007). Cameron and Trivedi (2010) provide a nice general discussion of IV estimators as implemented in Stata. Finally, readers should consult Stata's documentation for the various options available with the `ivregress` command (see [R] **ivregress**).

6 Conclusion

This article presents a new Stata command, `miivfind`, that assists users in finding MIIVs from an identified SEM. MIIV estimators have a number of benefits relative to the more commonly used ML estimators for SEMs. MIIV estimators have the potential to better isolate misspecification errors (as demonstrated in the second simulation); can make use of overidentification tests, when available; are asymptotically "distribution-free"; and can allow for heteroskedastic errors. However, it can be difficult to identify MIIVs by inspection of the model equations. By assisting in the process of finding MIIVs,

`miivfind` facilitates the use of an alternative class of estimators to the standard ML estimators available for SEMs.

7 References

Acemoglu, D., S. Johnson, and J. A. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–1401.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Arellano, M., and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–297.

Basmann, R. L. 1960. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 55: 650–659.

Bollen, K. A. 1987. Total, direct, and indirect effects in structural equation models. *Sociological Methodology* 17: 37–69.

———. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

———. 1996a. A limited-information estimator for LISREL models with and without heteroscedastic errors. In *Advanced Structural Equation Modeling Techniques*, ed. G. A. Marcoulides and R. E. Schumacker, 227–241. Mahwah, NJ: Lawrence Erlbaum.

———. 1996b. An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika* 61: 109–121.

———. 2001. Two-stage least squares and latent variable models: Simultaneous estimation and robustness to misspecifications. In *Structural Equation Modeling: Present and Future: A Festschrift in Honor of Karl Jöreskog*, ed. R. Cudeck, S. D. Toit, and D. Sörbom, 119–138. Lincolnwood, IL: Scientific Software.

Bollen, K. A., and D. J. Bauer. 2004. Automating the selection of model-implied instrumental variables. *Sociological Methods and Research* 32: 425–452.

Bollen, K. A., and J. C. Biesanz. 2002. A note on a two-stage least squares estimator for higher-order factor analyses. *Sociological Methods and Research* 30: 568–579.

Bollen, K. A., J. B. Kirby, P. J. Curran, P. M. Paxton, and F. Chen. 2007. Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods and Research* 36: 48–86.

Bollen, K. A., and A. Maydeu-Olivares. 2007. A polychoric instrumental variable (PIV) estimator for structural equation models with categorical variables. *Psychometrika* 72: 309–326.

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

———. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.

Hägglund, G. 1982. Factor analysis by instrumental variables methods. *Psychometrika* 47: 209–222.

Heckman, J. J., S. Urzua, and E. J. Vytlacil. 2006. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88: 389–432.

Imbens, G. W., and J. D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62: 467–475.

Jöreskog, K. G. 1977. Structural equation models in the social sciences: Specification, estimation, and testing. In *Applications of Statistics*, ed. P. R. Krishnaiah, 265–287. Amsterdam: North-Holland.

———. 1983. Factor analysis as an error-in-variables models. In *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, ed. H. Wainer and S. Messick, 185–196. Hillsdale, NJ: Lawrence Erlbaum.

Jöreskog, K. G., and D. Sörbom. 1993. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Lincolnwood, IL: Scientific Software International.

Kirby, J. B., and K. A. Bollen. 2009. Using instrumental variable (IV) tests to evaluate model specification in latent variable structural equation models. *Sociological Methodology* 39: 327–355.

Madansky, A. 1964. Instrumental variables in factor analysis. *Psychometrika* 29: 105–113.

Sargan, J. D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.

Stock, J. H., and M. Yogo. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock, 80–108. 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

Shawn Bauldry is an assistant professor in the Department of Sociology at the University of Alabama at Birmingham. Previously, he was a postdoctoral research associate in the Department of Sociology at the University of North Carolina at Chapel Hill. His methodological research interests focus on the development of SEMs and approaches to handling missing data.