Review in Advance first posted online on May 8, 2012. (Changes may still occur before final publication online and in print.)

Instrumental Variables in Sociology and the Social Sciences

Kenneth A. Bollen

Department of Sociology, University of North Carolina, Chapel Hill, North Carolina 27599-3210; email: bollen@unc.edu

Annu. Rev. Sociol. 2012. 38:22.1–22.36

The *Annual Review of Sociology* is online at soc.annualreviews.org

This article's doi: 10.1146/annurev-soc-081309-150141

0360-0572/12/0811-0001$20.00

Keywords

structural equation models, potential outcomes, latent variables, identification, limited dependent variables, endogeneity problem

Abstract

Instrumental variable (IV) methods provide a powerful but underutilized tool to address many common problems with observational sociological data. Key to their successful use is having IVs that are uncorrelated with an equation's disturbance and that are sufficiently strongly related to the problematic endogenous covariates. This review briefly defines IVs, summarizes their origins, and describes their use in multiple regression, simultaneous equation models, factor analysis, latent variable structural equation models, and limited dependent variable models. It defines and contrasts three methods of selecting IVs: auxiliary instrumental variable, model implied instrumental variable, and randomized instrumental variable. It provides overidentification tests and weak IV diagnostics as methods to evaluate the quality of IVs. I review the use of IVs in models that assume heterogeneous causal effects. Another section summarizes the use of IVs in contemporary sociological publications. The conclusion suggests ways to improve the use of IVs and suggests that there are many areas in which IVs could be profitably used in sociological research.

INTRODUCTION

Sociology and the social sciences face a host of problems when analyzing observational data. Few of our variables are free of measurement error. Feedback relations between variables are sometimes suspected. Omitted variables might influence both explanatory and dependent variables. Or sample selectivity can influence who is in or out of a sample or who is missing values on one or more variables.

Although these issues are diverse, they have in common that they can create a correlation between one or more explanatory variables (covariates) and the error of an equation. This correlation renders commonly used estimators such as ordinary least squares (OLS) or probit and logistic regressions biased and inconsistent, and this in turn undermines our confidence in our estimates of the effects of one variable on another. In a diverse literature over a long period, instrumental variable (IV) estimators have been proposed as solutions to each of these problems. Under assumptions that I describe more fully below, IV estimators of coefficients can restore asymptotic unbiasedness and consistency *if* the conditions for their use are satisfied.

Researchers from a broad spectrum of disciplines have applied IV methods (see Didelez et al. 2010 for a recent review from an epidemiological perspective, Sovey & Green 2011 for a recent review of IVs' use in political science, and Angrist & Krueger 2001 for a review of economic applications).The multifarious applications across fields have resulted in researchers in one area who are unaware of IV applications in other areas. Sociologists are no exception. Sociologists have familiarity with IVs in the context of an endogenous explanatory variable with a continuous dependent variable, and they use IV approaches in models that include sample selectivity and spatial effects. They are becoming more familiar with the use of IVs in the potential outcome (or counterfactual) causality literature (Morgan & Winship 2007), but they are less familiar with IVs in latent variable models or factor analysis estimation. Perhaps of greater concern is that sociologists too frequently ignore the problems that lead to endogenous covariates.

The purpose of this article is to give a broad overview of IVs in diverse areas and of their current use in sociology. By necessity I draw on literature from several disciplines, so that in this sense the review might be useful to nonsociologists as well. In addition, references are provided for the technical details of the estimators and their properties, but the focus is on key results and formulas. I describe what IVs are, major areas of applying IV methods, the different approaches to finding IVs, how they apply with homogeneous and heterogeneous causal effects, diagnostics for determining the quality of IVs, and the use of IVs in sociological practice.

The next section addresses what IVs are. A brief section on the origins of IV methods follows. I then present several major application areas for IVs, followed by three approaches to finding IVs and a section on evaluating the quality of IVs. A section on heterogeneous causal effects comes next. After that is a section that summarizes the applications of IVs in contemporary sociology, followed by a conclusion.

WHAT ARE INSTRUMENTAL VARIABLES?

The Problem: $COV(x, \epsilon) \neq 0$

To describe the nature of IVs, it is useful to begin with a simple regression model,

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad 1.$$

where y_i is a continuous[1] dependent variable for the ith case influenced by the explanatory variable or covariate x_i, α is the regression intercept, β is the regression coefficient, ϵ_i is the error or disturbance for the ith case with a mean of zero $[E(\epsilon_i) = 0]$ and $i = 1, 2, \ldots, N$, with

[1]In practice, y_i at least approximates a continuous variable. Dichotomous and ordinal dependent variables are discussed in a later section.

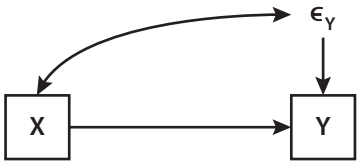


Figure 1

Path diagram of simple regression model with error-covariate (X) correlation.

N the total sample size. For now I assume that β is the same value for all individuals. A common assumption in a regression model is that the error (ϵ_i) is uncorrelated with the covariate (x_i) or $COV(x_i, \epsilon_i) = 0$ for all i. When these assumptions hold, then the OLS estimator applied to Equation 1 has excellent properties, including unbiasedness and consistency. However, the assumption that $COV(x_i, \epsilon_i) = 0$ is crucial to these desirable OLS properties [some presentations use the stronger assumption of mean independence, $E(\epsilon_i|x_i) = 0$, in place of $COV(\epsilon_i, x_i) = 0$ (e.g., Cameron & Trivedi 2005, p. 78; Angrist & Pischke 2009); I use the weaker assumption here to simplify the presentation].

Unfortunately, there are a variety of conditions that lead to $COV(x_i, \epsilon_i) \neq 0$. **Figure 1** illustrates a simple regression with $COV(x_i, \epsilon_i) \neq 0$ in a path diagram. In path diagrams, boxes represent observed variables, ovals stand for latent variables, and errors or disturbance are not enclosed. To simplify the notation, I have suppressed the subscript i. The single-headed straight arrows signify the impact of the variable at the base of the arrow on the variable to which the arrow points. **Figure 2** presents the path diagram for a series of models that, if true, would lead to a correlation of the covariate (x) and error (ϵ) if we were to estimate the simple regression model of **Figure 1**. **Figure 2*a*** portrays the true model as a feedback relation, where y and x are locked in a feedback loop with their error terms correlated (e.g., Paxton et al. 2011). An example is y being the respondent's amount of smoking and x being a best friend's smoking. If a best friend's smoking (x) influences the smoking of the respondent (y), then it makes sense to think that the respondent's smoking will affect the best friend's smoking. These reciprocal linkages and correlated errors create a correlation between the error (ϵ) and the covariate (x) if treated as Equation 1 in **Figure 1**.

Or assume that there is no feedback relation, but there is random measurement error in the covariate (x), with ξ being the true covariate without error. See **Figure 2*b***. Here ξ has a direct effect on y, the dependent variable, and on ξ's measure, x. If, as is commonly done, we ignore the measurement error in x and estimate the simple regression in Equation 1, then this will create a correlation between the error (ϵ) and the covariate (x) shown in **Figure 1**. This is true even if the measurement error in x is random and uncorrelated with ξ. A hypothetical example is if x is a measure of student motivation and y is grade point average. The measurement error in the motivation measure would create a correlation between the equation error (ϵ) and x.

Omitted common causes of y and x as illustrated in **Figure 2*c*** are yet another source of correlation between x and ϵ [$COV(x, \epsilon) \neq 0$]. The omitted latent variable L could be a collection of many variables that influence both y and x. The omission of these variables from the model means that L is part of ϵ, and this in turn leads ϵ to correlate with x as in **Figure 1**. For instance, x could be the industrialization and y the level of liberal democracy of a country. The variable L might consist of a variety of factors such as the timing of industrialization, cultural traditions, colonial history, or peripheral position in the world system, all variables that affect both democracy and industrialization. Ignoring these omitted factors means they are part of ϵ, and using Equation 1 would lead to $COV(x, \epsilon) \neq 0$.

Figure 2*d* is a model where the error of y (ϵ_y) is influenced by the error term for x (ϵ_x). A common situation where this might occur is when x is a lagged value of y and the errors of the variable tend to persist, creating the autoregressive relation. Ignoring this autoregression of the errors and estimating the simple

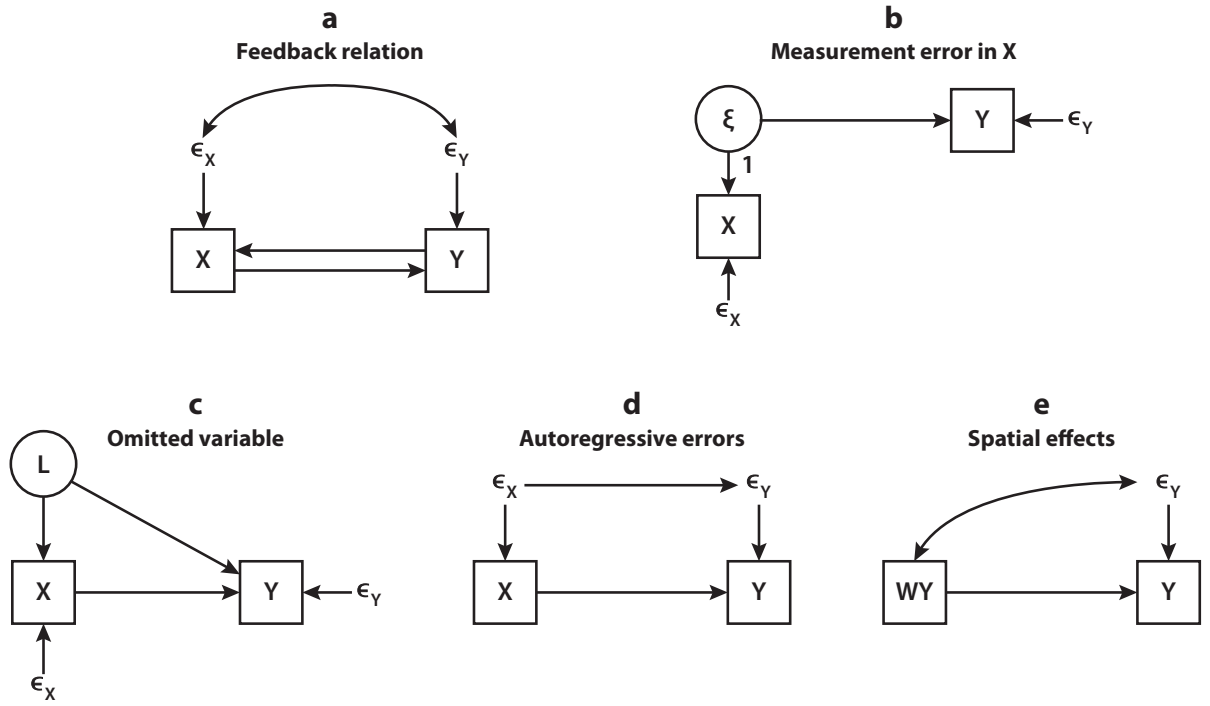


Figure 2

Possible causes of error-covariate (X) correlation.

regression leads to the correlation of error and covariate shown in **Figure 1**.

Finally, **Figure 2*e*** represents a spatial effects model where the value of the dependent variable of each case is influenced by a weighted sum of the dependent variable values of spatially nearby cases. The **Wy** term has **W** as the $N \times N$ weight matrix, **y** as the $N \times 1$ vector of values of the dependent variable, and ρ as the coefficient that gives the spatial effects. OLS estimation of the equation $y_i = \alpha + \rho(\mathbf{Wy})_i + \epsilon_i$ would result in biased and inconsistent estimates because ϵ_i correlates with $(\mathbf{Wy})_i$ due to the spatial effects. Intuitively, this correlation occurs because under the spatial effects model y_j affects y_i through its presence in **Wy**, but in turn y_i affects y_j, and this induces a correlation between ϵ_i and $(\mathbf{Wy})_i$.

Each situation illustrated in **Figure 2** occurs in practice, and each would create the problem of $COV(x, \epsilon) \neq 0$ [or $COV(\mathbf{Wy}, \epsilon) \neq 0$ for the spatial effect example]. This covariance would bias the OLS estimator applied to Equation 1 and hence bias our estimate of x's effect on y. Of course, the same is true with multiple regression when the error correlates with one or more of the explanatory variables. IVs can sometimes help, as I explain in the next section.

The Instrumental Variable Solution

IV estimators were devised for equations such as in the previous subsection where the equation error correlates with one or more of the covariates in the model. IVs have two primary conditions to satisfy. Suppose I call the variable that is a candidate for an IV z. The z must (*a*) correlate with x and (*b*) be uncorrelated with ϵ. Implicit in *b* is that z has no direct effect on y and is not included in the equation to explain y. Under these conditions, the covariance of y and z is

$$COV(y, z) = COV(\alpha + \beta x + \epsilon, z) \quad \text{[By Equation (1)]} \quad 2.$$

$$= \beta COV(x, z), \quad 3.$$

so we can get β by

$$\beta = \frac{COV(y, z)}{COV(x, z)}. \quad 4.$$

This ratio of the covariance of y and z to the covariance of x and z is the IV estimator of β when the sample values are substituted for the population covariances. This provides an asymptotically unbiased and consistent estimator of β, the impact of x on y (the IV estimator is asymptotically unbiased, even though in finite samples it is not unbiased).

This easily generalizes to multiple regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad 5.$$

where $\mathbf{y}$ is an $N \times 1$ vector for the dependent variable, $\mathbf{X}$ is the $N \times K$ matrix of K covariates, $\boldsymbol{\beta}$ is the $K \times 1$ vector of regression coefficients, and ϵ is the $N \times 1$ vector of errors with mean of zero. Suppose that I partition the $\mathbf{X}$ variables into two groups so that $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$ with $\boldsymbol{\beta}' = [\boldsymbol{\beta}'_1\ \boldsymbol{\beta}'_2]$ their corresponding coefficients. There are K_1 variables in $\mathbf{X}_1$ and K_2 variables in $\mathbf{X}_2$. The multiple regression in Equation 5 is then equivalent to

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon \quad 6.$$

Assume that ϵ is uncorrelated with $\mathbf{X}_2$, but correlates with the variables in $\mathbf{X}_1$ for one or more of the reasons illustrated in **Figure 2**. If I used the OLS estimator of

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad 7.$$

then I would have an asymptotically biased and inconsistent estimator of $\boldsymbol{\beta}$, and this is generally true even for $\boldsymbol{\beta}_2$.[2] However, suppose I have a matrix of variables $\mathbf{Z} = [\mathbf{X}_2\ \mathbf{X}_3]$, where $\mathbf{X}_2$ are the exogenous variables from the original $\mathbf{X}$ that are uncorrelated with ϵ and $\mathbf{X}_3$ are additional K_3 exogenous variables that are not part of the original covariates but are correlated with $\mathbf{X}$ and uncorrelated with ϵ.[3] Then I can form an IV estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_{IV} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y}, \quad 8.$$

where $\mathbf{P}_Z$ is a "projection matrix" equal to $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ [to form $(\mathbf{Z}'\mathbf{Z})^{-1}$ requires that $(\mathbf{Z}'\mathbf{Z})$ is nonsingular which means that perfect collinearity does not exist among IVs]. The use of the IV estimator assumes that (*a*) $\mathbf{Z}$ is uncorrelated with ϵ, (*b*) the covariance matrix of $\mathbf{Z}$ is nonsingular, and (*c*) the rank of the covariance matrix of $\mathbf{Z}$ and $\mathbf{X}$ equals $K_1 + K_2$. Assumption *b* eliminates the situation of perfect multicollinearity among any IVs. Assumption *c* concerns whether there is sufficient association between the IVs and $\mathbf{X}$. Marginal satisfaction of this assumption can create the problem of weak IVs, which I discuss below. A necessary but not sufficient identification condition for the number of IVs is that there be at least as many additional IVs in $\mathbf{X}_3$ as there are variables in $\mathbf{X}_1$ (i.e., $K_3 \geq K_1$). So if $\mathbf{X}_1$ has five variables that correlate with ϵ, then $\mathbf{X}_3$ must have at least five variables. If assumption *c* is satisfied and a researcher has the bare minimum number of IVs ($K_3 = K_1$), then the equation is exactly identified, and Equation 8 simplifies to $\hat{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$. Equation 8 is general enough also to include overidentified equations where assumption *c* holds and there is more than the necessary minimum number of IVs. In the case of overidentified equations with an excess number of IVs, the $\hat{\boldsymbol{\beta}}_{IV}$ is called the two-stage least squares (2SLS) estimator, and it is quite popular in multiple regression and simultaneous equation models. I refer to the exactly identified and overidentified 2SLS estimator as $\hat{\boldsymbol{\beta}}_{IV}$ to highlight its connection to IVs. The IV estimator $\hat{\boldsymbol{\beta}}_{IV}$ is asymptotically unbiased (Theil 1958, Richardson 1970), consistent (Basmann 1957, Theil 1958), asymptotically efficient among single-equation limited

[2] If the correlation between ϵ and $\mathbf{X}_1$ is weak, the bias and inconsistency might not be too large and the OLS estimator might be a reasonable approximation given the added variability introduced with an IV estimator. However, it is rare to know the magnitude of the correlation prior to estimation.

[3] This "correlation" should be interpreted as partial correlations that are nonzero. In other words, $\mathbf{X}_3$ should have partial associations with $\mathbf{X}_1$ that persist after holding constant the variables in $\mathbf{X}_2$. We return to this point when discussing weak IVs.

information estimators (Bowden & Turkington 1984, pp. 110–11), and asymptotically normally distributed (Basmann 1960) with an asymptotic covariance matrix of

$$\widehat{VAR}(\hat{\boldsymbol{\beta}}_{IV}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}, \qquad 9.$$

where $\hat{\sigma}^2$ is the residual variance from Equation 5 when $\hat{\boldsymbol{\beta}}_{IV}$ is substituted for $\boldsymbol{\beta}$ (e.g., Johnston 1984). In addition, the endogenous covariates ($\mathbf{X}_1$) can include dummy variables (Kelejian 1971; Angrist & Krueger 2001, p. 80) as well as continuous variables, and the same formulas apply. A heteroscedastic-robust covariance matrix also is available (e.g., Wooldridge 2002, pp. 100–1).

Although it might not be immediately evident, the multiple regression model includes spatial effects models as well.[4] I can use $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon$ with the understanding that one of the variables in $\mathbf{X}_1$ is the spatial effect term of $\mathbf{Wy}$ with its coefficient ρ placed in the corresponding row of $\boldsymbol{\beta}_1$. This endogenous covariate would be treated the same as the other endogenous covariates in $\mathbf{X}_1$, and the same requirements for IVs in $\mathbf{Z}$ hold.

Thus, the IV method provides a solution to the problem of a disturbance that correlates with one or more covariates (or explanatory variables) in an equation in many situations in which that correlation is present. It relies on having variables that satisfy the conditions necessary for IVs. In a later section, I discuss several ways of finding IVs. But in the next section I briefly describe the origins of the IV method.

ORIGINS OF INSTRUMENTAL VARIABLE METHODS

There is some controversy about the first development of IV methods. Goldberger (1972) argues that the first statistical use of IVs was in Sewall Wright's (1925) path analysis study of the relationship between corn and hogs in a recursive model with latent variables [Stock & Trebbi (2003, p. 179) argue that IVs were unnecessary in this model and that OLS would have sufficed but do not explain how OLS could have been used given the key role played by latent variables in Wright's (1925) model]. Others have emphasized Appendix B of Philip Wright's (1928) *The Tariff on Animal and Vegetable Oils*, where the first IV solution to estimating the classic supply-demand curve relationship is provided. Philip Wright was the father of Sewall Wright, and this has contributed to disagreements over which Wright was responsible for the IV solution described in Appendix B (see, e.g., Crow 1978, Manski 1988, Morgan 1990, Angrist & Krueger 2001, Stock & Trebbi 2003). The first use of IVs appears to be S. Wright (1925), whereas the first use of IVs in a supply and demand simultaneous equation is in P. Wright (1928), with credit for the IV idea of Appendix B under dispute (Grootendorst 2007 suggests that Snow's study of origins of cholera was based on IV ideas, although the connection to contemporary IV methods is less evident than is true for the Wrights' work). Years after both Wrights' work, the name "instrumental variable" was developed and the estimator applied to error of measurement in variables in Reiersøl (1941, 1945) and Geary (1949). Madansky (1964) suggested an IV estimator for factor analysis, and Imbens & Angrist (1994) and Angrist et al. (1996) took an IV approach to measure heterogeneous responses to interventions or treatments. Further details on the history of IVs are in Angrist & Krueger (2001), Bowden & Turkington (1984), Goldberger (1972), Morgan (1990), and Stock & Trebbi (2003).

APPLICATION AREAS

IV methods are used in a broad range of areas. I have already introduced their use in multiple regression analysis in a previous section. In this section, I describe their roles in simultaneous equation models, factor analysis, latent variable

[4]Another application area is panel models with lagged dependent variables and first differences. The Arellano-Bond IV estimator that uses lagged values of endogenous and exogenous variables as IVs is a common approach (Holtz-Eakin et al. 1988, Arellano & Bond 1991).

structural equation models (SEMs), and equations with limited dependent variables.

Simultaneous Equation Models

After multiple regression, simultaneous equation models are probably the major application area for IVs. These multiequation models consist of endogenous and exogenous variables as below in

$$\mathbf{y} = \boldsymbol{\alpha}_{\mathbf{y}} + \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \epsilon_{\mathbf{y}}, \qquad 10.$$

where $\mathbf{y}$ is a vector of endogenous variables, $\boldsymbol{\alpha}_{\mathbf{y}}$ is the vector of intercepts with one intercept per variable in $\mathbf{y}$, $\mathbf{B}$ contains the regression coefficients of any effects of endogenous variables on each other, $\mathbf{x}$ is the vector of exogenous variables with $\boldsymbol{\Gamma}$ their coefficients for their effects on $\mathbf{y}$, and $\epsilon_{\mathbf{y}}$ is the vector of equation disturbances. The i subscript is omitted from this equation and the variables to simplify the notation.

This model contrasts with the preceding multiple regression model in several ways. For one thing, it is a multiequation system with each individual equation similar to a multiple regression. The $\mathbf{y}$ vector has as many ys as there are dependent variables in the model. The $\boldsymbol{\alpha}_{\mathbf{y}}$ collects together the intercepts for each of these equations, while the regression coefficients that give the impact of one y on another y are in $\mathbf{B}$ and the regression coefficients of the exogenous x s are in the $\boldsymbol{\Gamma}$ matrix. The endogenous variables ($\mathbf{y}$) are explained within the model where such variables might be influenced by other variables in $\mathbf{y}$ or in $\mathbf{x}$. The exogenous variables ($\mathbf{x}$) are taken as given and are not explained by any other variables in the model, though they typically are allowed to correlate with each other. The model assumes that the disturbances (errors) are uncorrelated with $\mathbf{x}$ [$COV(\mathbf{x}, \epsilon_{\mathbf{y}}) = \mathbf{0}$] and that they have means of zero [$E(\epsilon_{\mathbf{y}}) = \mathbf{0}$].

There are special cases of the simultaneous Equation 10 when the error or disturbance of the equation does not correlate with the y covariates on the right-hand side. An example is a fully recursive model in which there are no feedback relationships ($\mathbf{B}$ is lower triangular) and equation errors are uncorrelated across equations. The classic Blau & Duncan (1967) status attainment models were fully recursive models. In the 1960s and 1970s, when path analysis versions of simultaneous equations were first introduced into sociology, assuming recursive models was quite common despite the lack of support for such a strong assumption in many applications. A more realistic assumption for an equation from a simultaneous equation model is to have its disturbance or error term correlated with one or more of the ys that serve as explanatory variables in the equation.

An IV estimator like 2SLS as is given in Equation 8 can handle such problems, provided there are sufficient variables to serve as IVs. A researcher can estimate each equation separately. All exogenous variables in the model can serve as IVs in that they commonly correlate with the endogenous covariate and by assumption they are uncorrelated with all equation errors or disturbances. Those exogenous variables that do not directly appear in the selected equation serve as the extra IVs and help to identify the equation. When all equation errors are assumed to correlate, then the exogenous variables form the only pool of IVs in a model. But if some errors are uncorrelated, certain endogenous variables (ys) may be uncorrelated with the error of the estimated equation and may qualify as an IV for an equation. With the exception of fully recursive models, it is rare to see a sociological application of this latter assumption.

A sociology paper by Tomaskovic-Devey & Skaggs (2002) provides an example. The authors are interested in whether the sex composition of jobs affects wages. The simultaneous equations of their model are below:

$$\begin{aligned}
y_1 &= \alpha_{y1} + B_{12}y_2 + B_{13}y_3 + B_{14}y_4 + \boldsymbol{\Gamma}_{11}\mathbf{x}_1 + \epsilon_{y1} \\
y_2 &= \alpha_{y2} + B_{21}y_1 + B_{23}y_3 + B_{24}y_4 + \boldsymbol{\Gamma}_{22}\mathbf{x}_2 + \epsilon_{y2} \\
y_3 &= \alpha_{y3} + B_{31}y_1 + B_{32}y_2 + B_{34}y_4 + \boldsymbol{\Gamma}_{33}\mathbf{x}_3 + \epsilon_{y3} \\
y_4 &= \alpha_{y4} + B_{41}y_1 + B_{42}y_2 + B_{43}y_3 + \boldsymbol{\Gamma}_{44}\mathbf{x}_4 + \epsilon_{y4} \\
y_5 &= \alpha_{y5} + B_{51}y_1 + B_{52}y_2 + B_{53}y_3 + B_{54}y_4 \\
&\quad + \boldsymbol{\Gamma}_{55}\mathbf{x}_5 + \epsilon_{y5},
\end{aligned} \qquad 11.$$

where again I omit the case subscript i from these variables. The primary endogenous variables are job training (y_1), percent female in the job (y_2), task complexity (y_3), supervisory

control (y_4), and earnings (y_5). Their starting model permits feedback relations among y_1 to y_4, but not earnings (y_5). Earnings (y_5) is regressed on y_1 to y_4 without earnings influencing any of the other variables. The exogenous variables for each equation are represented by the vectors $\mathbf{x}_1$ to $\mathbf{x}_5$. The errors ϵ_{y1} to ϵ_{y4} are all correlated. The feedback relations among y_1 to y_4 combined with the correlation of their errors makes the ys on the right-hand side of the y_1 to y_4 equations endogenous covariates. For the y_1 equation, there are three endogenous covariates (y_2 to y_4). To use an IV estimator such as 2SLS, there must be at least three exogenous variables that are not part of $\mathbf{x}_1$, but are among the other exogenous variables in $\mathbf{x}_2$ to $\mathbf{x}_5$. The y_2 equation also has three endogenous covariates (y_1, y_3, y_4) and needs at least three exogenous variables that are omitted from $\mathbf{x}_2$ but included in $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$, or $\mathbf{x}_5$. An analogous condition must hold for the y_3 and y_4 equations.

Tomaskovic-Devey & Skaggs (2002) assume that ϵ_{y5} is uncorrelated with all other errors, and this means that IV methods are not needed and that OLS is a consistent estimator of the coefficients of the y_5 equation. As I discuss in the section on diagnostics for IVs, there are tests of whether all IVs of an overidentified equation are uncorrelated with the equation disturbance.

Factor Analysis

The section on historical origins discussed IV estimators of exploratory and confirmatory factor analysis (CFA). Readers not exposed to this idea might be puzzled by the use of IVs in factor analysis because it appears so different from the other applications of IVs. As I demonstrate shortly, the need for IVs is evident once I transform the factor analysis model to eliminate the latent factors and substitute scaling indicators in their places. The IV estimators are advantageous compared to alternatives such as maximum likelihood (ML) in that they are more robust to structural misspecifications, have overidentification tests available for individual indicators, are asymptotically distribution-free, are noniterative estimators that do not suffer from nonconvergence, and are estimable for any subset of equations instead of all equations.

There are several IV approaches (Madansky 1964, Hägglund 1982, Jöreskog & Sörbom 1993), but nearly all assume uncorrelated "errors" (i.e., unique factors), ignore the means and intercepts of the model, and fail to consider higher-order factor analysis. Bollen's (1996, 2001; Bollen & Biesanz 2002) model implied instrumental variable (MIIV) estimator for factor analysis addresses each of these issues. The MIIV factor analysis is a special case of the general latent variable SEM developed in a series of papers. I rely on this MIIV approach given that it covers a broader range of conditions than do other IV estimators for factor analysis.

The factor analysis model is

$$\mathbf{x} = \boldsymbol{\alpha}_{\mathbf{x}} + \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\xi} + \boldsymbol{\epsilon}_{\mathbf{x}}, \qquad 12.$$

where $\mathbf{x}$ is a vector of indicators of the factors or latent variables, $\boldsymbol{\alpha}_{\mathbf{x}}$ is the vector of intercepts for each indicator, $\boldsymbol{\Lambda}_{\mathbf{x}}$ is the factor loading matrix, $\boldsymbol{\xi}$ is the vector of factors, and $\boldsymbol{\epsilon}_{\mathbf{x}}$ is the vector of unique factors (errors). The notation is simplified by omitting the i subscript. Each factor must be assigned a scale and origin to interpret them and to help identify the model. A simple way to do this is to choose one scaling indicator for each factor and set the scaling indicator's intercept to zero and its factor loading to one (Bollen 1989, pp. 306–8). For instance, suppose the equation for the first indicator is $x_1 = \alpha_{x1} + \Lambda_{11}\xi_1 + \epsilon_{y1}$. If I use x_1 as the scaling indicator for ξ_1, then the first equation becomes

$$x_1 = \xi_1 + \epsilon_{y1}, \qquad 13.$$

where $\alpha_{x1} = 0$ and $\Lambda_{11} = 1$. What is interesting about this choice is that I can manipulate this equation to solve for the latent factor leading to

$$\xi_1 = x_1 - \epsilon_{y1}. \qquad 14.$$

The same process follows for all latent variables or factors in the model. I choose a scaling indicator and write the latent variable (factor) as equal to the scaling indicator minus its unique factor.

If I reorder the variables in $\mathbf{x}$ so that the scaling indicators come first and the nonscaling indicators second, then I can partition the indicator vector as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_n \end{bmatrix}, \qquad 15.$$

where $\mathbf{x}_s$ is the vector of scaling indicators and $\mathbf{x}_n$ is the vector of nonscaling indicators, and I can write

$$\boldsymbol{\xi} = \mathbf{x}_s - \epsilon_{\mathbf{x}_s}, \qquad 16.$$

with latent factors equal to the scaling indicators minus their individual respective unique factors.

By substituting Equations 16 into 12 and focusing on the nonscaling indicators ($\mathbf{x}_n$), I get

$$\mathbf{x}_n = \boldsymbol{\alpha}_{\mathbf{x}_n} + \boldsymbol{\Lambda}_{\mathbf{x}_n}\mathbf{x}_s - \boldsymbol{\Lambda}_{\mathbf{x}_n}\epsilon_{\mathbf{x}_s} + \epsilon_{\mathbf{x}_n}, \qquad 17.$$

where $\boldsymbol{\Lambda}_{\mathbf{x}_n}$ contains the factor loadings for the nonscaling indicators. The result is that the latent variable factor analysis model is transformed into a model that replaces the latent factors with their scaling indicators but keeps the same intercept ($\boldsymbol{\alpha}_{\mathbf{x}_n}$) and factor loading coefficients ($\boldsymbol{\Lambda}_{\mathbf{x}_n}$). This appears to be a multivariate regression, and the temptation is to estimate the intercepts and loadings by regressing the nonscaling indicators ($\mathbf{x}_n$) on the scaling indicators ($\mathbf{x}_s$). However, the composite disturbance undermines this strategy because $\epsilon_{\mathbf{x}_s}$ is correlated with $\mathbf{x}_s$, contrary to a key assumption of multivariate regression. The correlation between error and covariate is the reason that I turn to an IV estimator.

This application of IVs differs in a few ways from the more typical econometric approach that I described for simultaneous equation models. I mention these differences here for those who are familiar with IV use in simultaneous equations. First, in simultaneous equations the primary and often exclusive source of IVs is the exogenous observed variables in a model. In factor analysis, there are no exogenous observed variables, so they cannot be the source of IVs. In factor analysis, all observed variables are endogenous variables, a fact that might lead one from the econometric tradition to question whether any IVs are possible. However, the key requirement of IVs is that they are variables that are uncorrelated with the disturbance of the equation of interest,[5] and as I explain below, other indicators that are not part of the estimated equation often meet this condition.

Mention of the disturbance brings me to a second contrast with IVs in simultaneous equation models. The disturbance or error term in this MIIV factor analysis approach is a composite. Equation 17 shows it as $-\boldsymbol{\Lambda}_{\mathbf{x}_n}\epsilon_{\mathbf{x}_s} + \epsilon_{\mathbf{x}_n}$. For an individual equation, this is the factor loading times the unique factor of the scaling indicator for each factor that directly affects an indicator plus the unique factor for the indicator equation being estimated. For instance, suppose that the x_2 indicator measures only the first factor ξ_1 and the scaling indicator for ξ_1 is x_1 so that $\xi_1 = x_1 - \epsilon_{x1}$. The x_2 indicator equation from Equation 17 would be

$$x_2 = \alpha_{x2} + \Lambda_{21}x_1 - \Lambda_{21}\epsilon_{x1} + \epsilon_{x2}, \qquad 18.$$

and the composite error includes ϵ_{x1} and ϵ_{x2}. All the other indicators in the factor analysis model would be suitable MIIVs as long as the unique factors of the indicators do not correlate with either ϵ_{x1} or ϵ_{x2}. Thus, a number of endogenous observed variables (indicators) can serve as MIIVs in factor analysis (Bollen 1996).

Another difference in the use of IVs in factor analysis versus simultaneous equation models is that in simultaneous equation models the IV methods estimate the regression coefficients that give the impact of observed variables on other observed variables. In contrast, in factor analysis I estimate the impact of the latent variables on observed variables even though the model is transformed to an observed variable form in order to get at these factor loadings. Another way of saying this is that for simultaneous equations the estimated form is the original form of the structural model. In factor analysis estimation with MIIVs, the original model with latent variables is transformed to an observed variable form that eliminates the latent

[5] Implicit is that the MIIVs of an equation do not directly affect the dependent indicator of the equation.

variables while preserving their original coefficients and intercepts. This is done for estimation purposes, but the interpretation of the coefficients and intercept are with respect to the original factor analysis structural form of the model.

What the MIIV factor analysis shares with the IV approach to simultaneous equations is that the estimation procedures developed for simultaneous equations are easily adaptable to the MIIV approach to factor analysis. For instance, the IV estimator in Equation 8 applies to the MIIV factor analysis approach, where the variables in $\mathbf{Z}$ are the MIIV indicators for the equation, the $\mathbf{X}$ consists of the scaling indicators for the factor(s) that influence a given indicator, and the $\mathbf{y}$ is the indicator of the factor analysis equation being estimated (see Bollen 1996, 2001 for more details).

As an illustration, sociologist Regoeczi (2002) analyzes the impact of population density on withdrawn behavior. She has seven measures of withdrawn behavior at two points in time. The measures indicate a respondent's agreement with statements such as "It is hard for me to feel close to other people" (x_1); "I keep other people at a distance too much" (x_2); "It is hard for me to experience a feeling of love for another person" (x_3); "It is hard for me to show affection to other people" (x_4); "It is hard for me to socialize with other people" (x_5); "It is hard for me to introduce myself to new people" (x_6); and "It is hard for me to join in on groups" (x_7). A CFA model for these measures for the first wave is

$$\begin{aligned} x_1 &= \xi_1 + \delta_1 \\ x_2 &= \alpha_2 + \Lambda_{21}\xi_1 + \delta_2 \\ x_3 &= \alpha_3 + \Lambda_{31}\xi_1 + \delta_3 \\ &\vdots \\ x_7 &= \alpha_7 + \Lambda_{71}\xi_1 + \delta_7, \end{aligned} \qquad 19.$$

where x_1 is the scaling indicator with its intercept set to zero and its factor loading set to one (following the author, I treat these indicators as adequately approximating continuous variables). The errors among these seven variables are uncorrelated, and each has a mean of zero. With $x_1 = \xi_1 + \delta_1$, I can solve for $\xi_1 = x_1 - \delta_1$ and substitute $(x_1 - \delta_1)$ for ξ_1 in the remaining equations for the other indicators. This results in

$$\begin{aligned} \xi_1 &= x_1 - \delta_1 \\ x_2 &= \alpha_2 + \Lambda_{21}x_1 - \Lambda_{21}\delta_1 + \delta_2 \\ x_3 &= \alpha_3 + \Lambda_{31}x_1 - \Lambda_{31}\delta_1 + \delta_3 \\ &\vdots \\ x_7 &= \alpha_7 + \Lambda_{71}x_1 - \Lambda_{71}\delta_1 + \delta_7. \end{aligned} \qquad 20.$$

Bollen's (1996) MIIV-2SLS could be used to estimate the x_2 to x_7 equations. For the x_2 equation, the MIIVs are x_3 to x_7; for the x_3 equation, the MIIVs are x_2 and x_4 to x_7; and so on. More generally, the MIIVs for each of these equations are the other indicators that are not part of the equation.[6]

Structural Equation Models

Another application area for IVs is in general SEMs that include latent variables. Sometimes called "LISREL" or "covariance structure" models but more commonly just called SEMs, these widely used models are similar to simultaneous equations in that they are multiequation systems that consist of endogenous and exogenous variables, but these relationships include latent variables. Like factor analysis, SEMs include measurement models of the ties between indicators and latent variables as part of the same model structure.

The general SEM is

$$\boldsymbol{\eta} = \boldsymbol{\alpha}_\eta + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \epsilon_\eta \qquad 21.$$

$$\mathbf{y} = \boldsymbol{\alpha}_\mathbf{y} + \boldsymbol{\Lambda}_\mathbf{y}\boldsymbol{\eta} + \epsilon_\mathbf{y} \qquad 22.$$

$$\boldsymbol{x} = \boldsymbol{\alpha}_\mathbf{x} + \boldsymbol{\Lambda}_\mathbf{x}\boldsymbol{\xi} + \epsilon_\mathbf{x}, \qquad 23.$$

where $\boldsymbol{\eta}$ is the vector of latent endogenous variables, $\boldsymbol{\alpha}_\eta$ is the vector of intercepts for each latent endogenous variable, $\mathbf{B}$ contains the regression coefficients given the expected impact of latent endogenous variables on each other,

[6] The selection rules are modified if correlated unique factors (errors) are present. Then those variables with errors correlated with the errors of the equation of interest are no longer suitable IVs.

$\boldsymbol{\xi}$ is a vector containing the latent exogenous variables, $\boldsymbol{\Gamma}$ contains their coefficients, and ϵ_{η} is the vector of the errors in the equation. The model assumes that all disturbances have means of zero, that is, $E(\epsilon_{\eta}) = \mathbf{0}$, $E(\epsilon_{\mathbf{y}}) = \mathbf{0}$, and $E(\epsilon_{\mathbf{x}}) = \mathbf{0}$. All disturbances are uncorrelated with $\boldsymbol{\xi}$. Typically, it is assumed that each disturbance vector is uncorrelated with other disturbance vectors.[7]

The advantages of an IV estimation approach to the full latent variable SEM are the same as I mentioned for factor analysis: (*a*) greater robustness to structural specification errors, (*b*) overidentification tests by equation, (*c*) lack of nonconvergence as an issue, and (*d*) ability to estimate and test any subset of equations rather than the whole model. The MIIV approach in Bollen (1996, 2001) to IV estimation of this model starts by transforming the latent variable model into an observed variable model with composite disturbances. Collect all the scaling indicators for η and $\boldsymbol{\xi}$ into vectors $\mathbf{y}_s$ and $\mathbf{x}_s$, respectively. The nonscaling indicators are $\mathbf{y}_n$ and $\mathbf{x}_n$. I can write each latent variable as equal to its scaling indicator minus the unique factor as I did with factor analysis. Following the same process for all latent variables and substituting into Equations 21 to 23 leads to

$$\mathbf{y}_s = \boldsymbol{\alpha}_{\eta} + \mathbf{B}\mathbf{y}_s + \boldsymbol{\Gamma}\mathbf{x}_s + (\mathbf{I} - \mathbf{B})\epsilon_{\mathbf{y}_s} - \boldsymbol{\Gamma}\epsilon_{\mathbf{x}_s} + \epsilon_{\eta} \quad 24.$$

$$\mathbf{y}_n = \boldsymbol{\alpha}_{\mathbf{y}_n} + \boldsymbol{\Lambda}_{\mathbf{y}_n}\mathbf{x}_s - \boldsymbol{\Lambda}_{\mathbf{y}_n}\epsilon_{\mathbf{y}_s} + \epsilon_{\mathbf{y}_n} \quad 25.$$

$$\mathbf{x}_n = \boldsymbol{\alpha}_{\mathbf{x}_n} + \boldsymbol{\Lambda}_{\mathbf{x}_n}\mathbf{x}_s - \boldsymbol{\Lambda}_{\mathbf{x}_n}\epsilon_{\mathbf{x}_s} + \epsilon_{\mathbf{x}_n}. \quad 26.$$

Though this expression looks formidable, it is similar to the equations for the factor analysis model that I have already discussed; indeed, Equation 26 is identical to the factor analysis one in Equation 17 from the last subsection. The first, in Equation 24, has the parameters of the latent variable model in Equation 21 and the second, in Equation 25, contains the parameters for the factor analysis model relating the $\mathbf{y}_n$ to $\boldsymbol{\eta}$. An IV estimator is possible for each equation. The observed variables that will satisfy the conditions required for IVs in each equation are implicitly determined by the structure of the model. I say more about these MIIVs in the section on selecting IVs.

Limited Dependent Variable Models

The multiple regression example I gave in an earlier section is an example of a situation in which a single dependent variable (y) and hence single equation is the focus. There I mentioned that the endogenous covariate could be dichotomous or continuous for the usual IV methods to apply as long as the dependent variable approximates a continuous variable. Another single equation area where IVs apply is for limited dependent variable models (Maddala 1983, Long 1997). Though limited dependent variable models are widely employed in sociology, IV estimators for limited dependent variables are rare. Because endogenous covariates are no less common with categorical rather than continuous dependent variables, I briefly review IV methods in this context. In these categorical dependent variable models, the continuous dependent variable is not fully observed, but it might appear as a dichotomous, ordinal, or censored variable because of the way the data were collected or because of the difficulty of collecting the continuous version of the variable. For instance, attitudes toward abortions might be a continuous variable, but a survey question typically collects it as a dichotomy ("support," "oppose") or an ordinal variable indicating the degree of support or opposition. Or the propensity to divorce might be tapped only by responding to the question of whether a person has ever been divorced. In these and many other cases, the continuous propensity dependent variable is representable in a multiple regression–like setup such as

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \epsilon \quad 27.$$

where the model and assumptions are defined the same way that they were in multiple regression except that $\mathbf{y}^*$ is the underlying or latent

[7] Other common assumptions are that the disturbances are not autocorrelated over cases and that they are homoscedastic. Corrective procedures are available when these are violated.

propensity for $\mathbf{y}$. If y_i is the ith value of $\mathbf{y}$ for a dichotomous variable, then

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0, \end{cases} \qquad 28.$$

where the 0 is the threshold value that determines whether y_i^* falls in one category or the other. Similarly, if y_i is an ordinal variable, then

$$y_i = c, \text{ if } \tau_c \leq y_i^* < \tau_{c+1}, \qquad 29.$$

where τ are the thresholds that determine categories into which y_i^* falls, y_i has C categories with $c = 0, 1, 2, \ldots, C-1$, $\tau_0 = -\infty$, and $\tau_C = \infty$. I could represent censored variables in an analogous fashion (e.g., Long 1997).

If I assume that ϵ comes from a standardized normal distribution with a mean of zero and variance of one, then the model is the usual dichotomous probit or ordinal probit regression model. Alternatively, if I assume that ϵ comes from a standardized logistic distribution with a mean of zero and variance of $\frac{\pi^2}{3}$, then it is either the dichotomous logistic or ordinal logistic regression model. Regardless of which of these models we have, they all assume that ϵ is uncorrelated with $\mathbf{X}$.

Suppose I partition the $\mathbf{X}$ variables into two groups as I did for multiple regression where $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$ with $\boldsymbol{\beta}' = [\boldsymbol{\beta}'_1 \;\; \boldsymbol{\beta}'_2]$ and rewrite the y^* Equation 27 as

$$\mathbf{y}^* = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon, \qquad 30.$$

where ϵ is uncorrelated with $\mathbf{X}_2$ but correlates with the variables in $\mathbf{X}_1$. Violating this assumption undermines the consistency and asymptotic unbiasedness of the probit or logistic ML estimator in the same way that the endogenous covariates undermined these properties for OLS with a continuous dependent variable (Yatchew & Griliches 1985).[8]

IV methods provide several ways to address this problem.

Treat y_i as if continuous. One approximation is to treat the dichotomous or ordinal y_i as if it were continuous. Then apply the usual IV estimator $\hat{\boldsymbol{\beta}}_{IV}$, as was done for the models in the earlier sections. In the case of dichotomous dependent variables, heteroscedasticity will be present, so the asymptotic covariance matrix needs to be heteroscedastic-consistent (Goldberger 1964). Cameron & Trivedi (2005, p. 473) and Wooldridge (2002, p. 472) mention this approach, particularly for exploratory stages of research.

Instrumental variable probit method. The preceding IV method takes account of the endogeneity of $\mathbf{X}_1$ but does not incorporate the auxiliary measurement model that gives the nonlinear relation between the dichotomous or ordinal y_i and y_i^*. Other IV methods incorporate the auxiliary measurement model. One method that does so is the IV probit (IVP) method (Lee 1981, Rivers & Vuong 1988). As above, the matrix $\mathbf{X}_1$ contains the endogenous covariates from $\mathbf{X}$. The IVP method has two steps. The first step uses OLS multivariate regression of $\mathbf{X}_1$ on all exogenous variables ($\mathbf{Z} = [\mathbf{X}_2 \;\; \mathbf{X}_3]$):

$$\mathbf{X}_1 = \mathbf{Z}\boldsymbol{\Pi} + \epsilon_{\mathbf{x}_1} \qquad 31.$$

and forms the predicted value of $\mathbf{X}_1$ as $\hat{\mathbf{X}}_1 = \mathbf{Z}\hat{\boldsymbol{\Pi}}$ and the predicted residuals as $\hat{\epsilon}_{\mathbf{x}_1} = \mathbf{X}_1 - \mathbf{Z}\hat{\boldsymbol{\Pi}}$.

The second step of the IVP method regresses the dichotomous y on $\hat{\mathbf{X}}_1$ and $\mathbf{X}_2$ using the probit ML estimator (Maddala 1983, Bollen et al. 1995), where now

$$\mathbf{y}^* = \hat{\mathbf{X}}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon^* \qquad 32.$$

and $\epsilon^* = \hat{\epsilon}_{\mathbf{x}_1}\boldsymbol{\beta}_1 + \epsilon$.

This method has the advantage that the correlation between the covariate and error is asymptotically removed through the use of IVs, and this should remove the bias in the probit ML estimator of the coefficients in large samples. Remember, though, that the scaling in probit is somewhat arbitrary in that the variance of the error of a probit regression is set to one, no matter what composes the error. In the original Equation 30, this meant that the

[8] If the correlation of the error (ϵ) with $\mathbf{X}_1$ is weak, then the usual probit or logistic ML estimator could have only slight bias. But until this correlation is estimated, this information is usually not known.

$VAR(\epsilon) = 1$ and the coefficients were scaled accordingly. With IVP, the $VAR(\epsilon^*) = 1$ and the coefficients are scaled under this constraint. This means that the coefficients from the original equation and the equation using the $\hat{\mathbf{X}}_1$ are not directly comparable.

Another complication with the IVP method is that the usual ML probit asymptotic covariance matrix for the coefficients from the second stage does not take account of the two-step nature of this estimator, where $\hat{\mathbf{X}}_1$ is estimated separately from the probit equation. Maddala (1983, p. 245) and Rivers & Vuong (1988) provide the correct asymptotic standard errors for the coefficients from the IVP estimator. Guilkey et al.'s (1992) Monte Carlo simulations raise questions as to whether these corrected standard errors perform that much better than the uncorrected second-stage standard errors. Bootstrap estimates of standard errors or confidence intervals is another possibility.

Two-stage conditional probit. A second IV method is two-stage conditional probit (2SCP) as developed by Vuong (1984), Rivers & Vuong (1988), and Smith & Blundell (1986). The 2SCP is similar to the just discussed IVP method in that the reduced-form Equation 31 is the first step. But 2SCP differs in that the second step estimates the ML probit equation of

$$\mathbf{y}^* = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 - \hat{\epsilon}_{\mathbf{x}_1}\boldsymbol{\beta}_{\epsilon_{x1}} + \epsilon^*. \quad 33.$$

Comparing Equations 33 and 32 reveals the differences of the IVP and 2SCP estimation methods. One is that the 2SCP uses the original $\mathbf{X}_1$ as explanatory variables whereas IVP uses $\hat{\mathbf{X}}_1$. A second difference is that 2SCP includes the reduced-form error matrix, $\hat{\epsilon}_{\mathbf{x}_1}$, as additional explanatory variables in the second stage estimation whereas IVP does not. Rivers & Vuong (1988) provide the asymptotic covariance matrix for the coefficient estimators of 2SCP for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ because the usual probit covariance matrix from the second-stage probit does not take account of the two-stage nature of the procedure.

Polychoric instrumental variables. The polychoric instrumental variable (PIV) estimator proposed in Bollen & Maydeu-Olivares (2007) is devised for general latent variable SEM when some of the endogenous observed variables are dichotomous, ordinal, or censored variables. However, the PIV estimator also applies to a single probit equation with endogenous covariates. The first step is to form the polychoric covariance (correlation) matrix of $\mathbf{y}^*$, $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$. The polychoric covariance matrix assumes that $\mathbf{y}^*$ comes from a standardized normal distribution, and based on this assumption and the distributions of the categorical variable $\mathbf{y}$ and $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, the covariance (correlation) of $\mathbf{y}$ with each of these variables is estimated using ML methods (Olsson 1979, Olsson et al. 1982). The usual IV estimator in Equation 8 applies to this covariance matrix. One limitation of this approach is the assumption that each pair of variables among $\mathbf{y}$, $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ is bivariate normal. Another possibility is to estimate the conditional polychoric covariance matrix of $(\mathbf{y}^*, \mathbf{X}_1 | \mathbf{X}_2, \mathbf{X}_3)$ under the assumption that this conditional distribution is multivariate normal. This is a less restrictive distributional assumption because it permits any distribution for $\mathbf{X}_2$, $\mathbf{X}_3$ and requires only that the conditional distribution, not the original distributions of $\mathbf{y}^*$, $\mathbf{X}_1$, be normal. Bollen & Maydeu-Olivares (2007) provide the asymptotic covariance matrix of the PIV regression coefficients.

Comparisons of instrumental variable estimators. The preceding gives an overview of some of the possible IV estimators for limited dependent variable models (this is not an exhaustive list; for example, Iwata 2001 describes a rescaling and recentered GMM method that differs from the methods I have reviewed). Analytic and simulation information that compares these different estimators for limited dependent variable equations is sparse. Rivers & Vuong (1988) look at the IVP and 2SCP estimators and find that analytically a constant ordering of efficiency under all conditions cannot be found. However, in finite samples

they discovered that the efficiency of 2SCP had some advantage over IVP, though Adkins (2012) observed that Rivers & Vuong's simulations had strong IVs. Among the IV estimators reviewed above, Adkins (2012) compares IVP and 2SCP. He finds that with weak IVs (very low correlations of IVs with endogenous covariate) none of these alternatives works well. His simulation shows that, as sample size gets larger and the correlations of the IVs with endogenous variable grow, the IV estimators are essentially unbiased. The PIV estimator is too recent to be considered in any of these studies. At this point, the evidence is not sufficient to unambiguously point toward any one of these estimators as being superior to the others.

I should also note that my discussion is limited to IV estimators of limited dependent variable models and that there are other estimators such as full information ML, limited information ML, and Amenta's GLS estimators for limited dependent variables with endogenous covariates (see, e.g., Maddala 1983, Adkins 2012).

GMM-IV Estimators

Generalized method of moment (GMM) estimators provide a general estimation framework from which to view existing estimators (e.g., OLS, ML) as well as to develop new estimators. Hansen (1982) provided the term and impetus for much of the contemporary research on GMM, though its history goes back much further (see Hall 2005, ch. 1). The GMM estimator has been particularly useful when combined with IV methods. The IV method for multiple regression has been given a GMM interpretation, as has the IV estimation of simultaneous equation models (Hansen 1982; Hall 2005; Mátyás 1999; Wooldridge 2002, ch. 8). Bollen et al. (2011) provide GMM-IV estimators to factor analysis and latent variable SEMs. Other work has extended GMM to limited dependent variable models (e.g., Rassen et al. 2009, Wilde 2008).

In addition to providing a unified estimation approach, the GMM-IV estimators have several other desirable characteristics. One is that the GMM-IV estimators for linear models have valuable properties such as consistency, asymptotic unbiasedness, and asymptotic normality under a broad range of conditions. In addition, GMM-IV estimators readily permit heteroscedastic-consistent significance tests. This includes an overidentification test when the number of IVs exceeds the minimum needed. It also is possible to test whether subsets of IVs are uncorrelated with equation disturbances rather than testing just all IVs. Space does not permit the presentation of the GMM-IV estimator in more detail, but this continues to be a promising area of development that is receiving a great deal of attention in economics, but is less considered in sociology. Furthermore, the finite sample behavior of GMM deserves further investigation because in modest sample sizes its asymptotic properties can be misleading (see, e.g., Hall 2005, ch. 6).

FINDING INSTRUMENTAL VARIABLES

Finding suitable IVs is key to any of the IV methods I have discussed. Indeed, a common caution in the presentation of IVs is to warn researchers about the difficulties of finding good IVs. The dominant IV strategy asks researchers to find IVs once they encounter an endogenous covariate. A second strategy draws IVs from among the observed variables in the model based on the structure of the model. I refer to the former type of IV as auxiliary instrumental variables (AIVs) and to the latter as MIIVs. Finally, a third class of IVs comes from randomization or "natural experiments" of an intervention or treatment, and I refer to these as randomization instrumental variables (RIVs). The next three subsections discuss these in more detail.

Auxiliary Instrumental Variables

AIVs are IVs brought into a model to supplement the other model variables so as to satisfy the necessary condition for IVs to identify the model parameters. For instance, a researcher's

original specification might contain a feedback relationship and be underidentified. This leads the researcher to search for variables that correlate with the endogenous variables of the equation but that do not correlate with the equation disturbance. The relationships between occupational prestige and mental health provides an illustration. It seems likely that there is a feedback relation between occupation and mental health such that low prestige jobs can adversely affect mental health and mental health can influence a person's occupation. A preliminary model might have several determinants of occupational prestige and mental health, but there might not be unique exogenous variables that influence each of these two variables to permit identification of the feedback relationship. The researcher then seeks to supplement the original set of variables with variables that will identify the feedback relationship. These additional variables are AIVs.

A common type of AIV is lagged values of variables. With panel data, a researcher might use the lagged values as IVs for the current values of variables. Or in a spatial effect model in which neighboring values of the dependent variable are influential, the spatially weighted lagged values of the exogenous variables might serve as IVs for the spatial effects variables (Anselin 1988).[9]

These variables are AIVs in that they were initially not conceived as an integral part of the model, but are largely added to help identify coefficients. AIVs are the most common type of IVs. An example from sociology is Axinn & Barber's (2001) analysis of fertility behavior determined by, among other things, the proportion of their children parents will send to school, an endogenous covariate. Axinn & Barber use childhood characteristics of the respondent and her husband and characteristics of their parents as IVs for the propensity to send children to school. Burris (2004) seeks the effect of departmental social capital on the prestige of sociology departments and uses faculty size as an IV for departmental social capital. Angrist & Krueger (1991) want to estimate the causal effect of education on wages and use quarter of birth as an IV for education. Numerous other examples of AIV selection are available, and their plausibility in meeting the conditions of IVs varies.

Advantages and disadvantages of the auxiliary instrumental variable method. AIVs have several advantages and disadvantages. The most obvious advantage is that an AIV has the potential to identify an otherwise underidentified parameter or equation. Another aspect of AIVs that researchers find attractive is that the exact relation of the AIV to the endogenous variables need not be specified (Bowden & Turkington 1984, p. 9). In other words, a structural relation is not needed to use the AIV in the model. The structural relation could be built, but that is not always needed or done. However, at least some justification for the selection of the IVs is needed to convince readers of their validity.

When more than the minimum number of AIVs needed for identification is available, then one can construct overidentification tests such as I discuss below. These are tests of the validity of the assumption that *all* AIVs in an equation are uncorrelated with the equation error. Another advantage of AIVs is that researchers commonly use these in probability samples, and as such the results of the analysis should be representative of a population. This contrasts with some intervention or experimental research in which the representativeness of the sample is open to question.

The disadvantages of the AIV selection method largely run in parallel to its advantages. Although AIVs might permit identification of underidentified equations, this offers little comfort if the AIVs correlate with the equation error contrary to assumptions. Sometimes the ad hoc selection of AIV raises doubts about the

[9]One could argue that the lagged variables are not AIVs in that they are implicitly part of the model. The line separating the AIV method from the MIIV method that I describe shortly is not always clean. However, in the case of lagged variables I treat them as AIVs in that they do not initially appear as part of the original structural form of the model.

validity of this condition. This is the other side of the lack of specifying how the AIV structurally relates to the other variables that are in the model. Though this lack of specification is sometimes thought to be an advantage, it often means that less thought has gone into how the AIV might enter the model structure. More thought could raise doubts about its suitability. For instance, I might want to know the effect of a respondent's education on the prestige of first occupation but suspect that education might be an endogenous covariate due to omitted variables affecting both respondent's education and prestige. Father's education might be selected as an AIV for respondent's education. After all, the temporal priority condition is met, and it might seem reasonable to assume that the error of respondent's occupational prestige correlates with respondent's education, but not with father's education. But suppose that there are personality characteristics shared among the respondent and the father that affect both of their educations and occupational achievements. Such omitted variables would enter the error for prestige and render father's education unsuitable as an IV.

A related disadvantage of AIV is the tendency to seek just enough AIVs to identify the equation. For instance, just enough lagged values of the endogenous variables might be chosen as IVs to identify the model. This leads to exactly identified models and the inability to perform overidentification tests that could reveal problematic IVs. This cautionary note suggests that if AIVs are used, then a sufficient number should be collected to permit IV overidentification tests.

Model Implied Instrumental Variables

MIIV selection is less common than AIV selection. The term is taken from Bollen (1996; Bollen & Bauer 2004). It refers to a model-dependent method of selecting IVs. With MIIVs, the structure of a model comes first. The researcher incorporates the best existing knowledge to build a model. The model structure implies that particular variables correlate with the endogenous covariates of an equation and are uncorrelated with the error of the same equation. In other words, the structure of the model tells the researcher which observed variables can serve as IVs and which cannot. The observed variables are included as part of the original structure, and their presence in the model is not primarily determined by identification needs.

To illustrate this, suppose I have a CFA model with four indicators and a single factor:

$$\begin{aligned} x_{1i} &= \xi + \epsilon_{1i} \\ x_{2i} &= \alpha_2 + \Lambda_2 \xi + \epsilon_{2i} \\ x_{3i} &= \alpha_3 + \Lambda_3 \xi + \epsilon_{3i} \\ x_{4i} &= \alpha_4 + \Lambda_4 \xi + \epsilon_{4i} \end{aligned} \qquad 34.$$

with the usual assumption of CFA (see previous section on Factor Analysis). Because x_{1i} is the scaling indicator, I write $\xi = x_{1i} - \epsilon_{1i}$. Substituting this into the x_{2i} equation, I get

$$x_{2i} = \alpha_2 + \Lambda_2 x_{1i} - \Lambda_2 \epsilon_{1i} + \epsilon_{2i}. \qquad 35.$$

Because of ϵ_{1i}, the composite error correlates with x_{1i}. To use an IV estimator, I need to find IVs that correlate with x_{1i}, but are uncorrelated with ϵ_{1i}. Referring to the model in Equation 34, the MIIVs that meet these conditions are x_{3i} and x_{4i}. Note that these variables are part of the original model and were not introduced as auxiliary variables. Their qualifications as IVs are based on model structure. In a case like this in which there are more MIIVs than the bare minimum needed, I can test whether all MIIVs are uncorrelated with the disturbance and hence test the model. I discuss these tests shortly. But the important point is that the MIIVs come from the model structure and refer to observed variables that are part of the original model.

Observed exogenous variables are another source of MIIVs. By assumption, all exogenous observed variables are uncorrelated with the disturbances of all equations. Because of this, each exogenous observed variable is a MIIV that a researcher can use for each equation of the model. These MIIVs are typical of simultaneous equation models (see prior section) in which measurement error is ignored and all equation

disturbances are correlated, but exogenous observed variables are MIIVs in any SEM in which they appear. Of course, a researcher declaring variables as exogenous does not make them so. In overidentified equations, a researcher can test whether all such variables are uncorrelated with the equation disturbance.

In latent variable SEMs with many equations, it can be hard to determine the MIIVs. Bollen & Bauer (2004) have proposed a procedure to automate MIIV selection based on the structure of the model, and they provide a SAS macro to implement it. The MIIV approach covers latent variable SEMs, factor analysis, simultaneous equation models, and other special cases of SEMs.

Directed acyclical graph–selected instrumental variables. Closely related to the MIIV method is the directed acyclical graph (DAG) selection methods. Pearl (2000, 2010), Brito (2010), and Brito & Pearl (2002) have proposed the use of DAGs to help in the selection of IVs for a given model. In the standard language of SEMs, a DAG is a recursive (no feedback) multiequation model of observed variables. All observed variables are treated as if they are free of measurement error. The only latent variables are the errors or disturbances. Correlation of at least some errors is permitted.

Brito & Pearl (2002) give rules to establish the identification of coefficients in DAG models and the variables that can serve as IVs. They also discuss the idea of conditional IVs. These are variables that meet the criteria for IVs once conditioned on other variables in the model. Their theorem extends beyond conditional IVs to cover additional situations in which less obvious IVs are available.

Their methods are a type of MIIVs in that the selection is based on a prespecified model structure and the IVs are chosen from among the observed variables (or among some function of them) in the model. The DAG approach was devised for particular types of SEMs. More specifically, DAG methods apply to recursive models of observed variables usually with correlated errors. The MIIV method applies to these models, but also to latent variable SEMs.

Furthermore, some applications of the DAG methods that appear as novel are covered by the MIIV or more traditional IV selection methods. For instance, **Figure 3*a*** is a model from Brito & Pearl (2002, figure 5*a*) that I have rewritten in

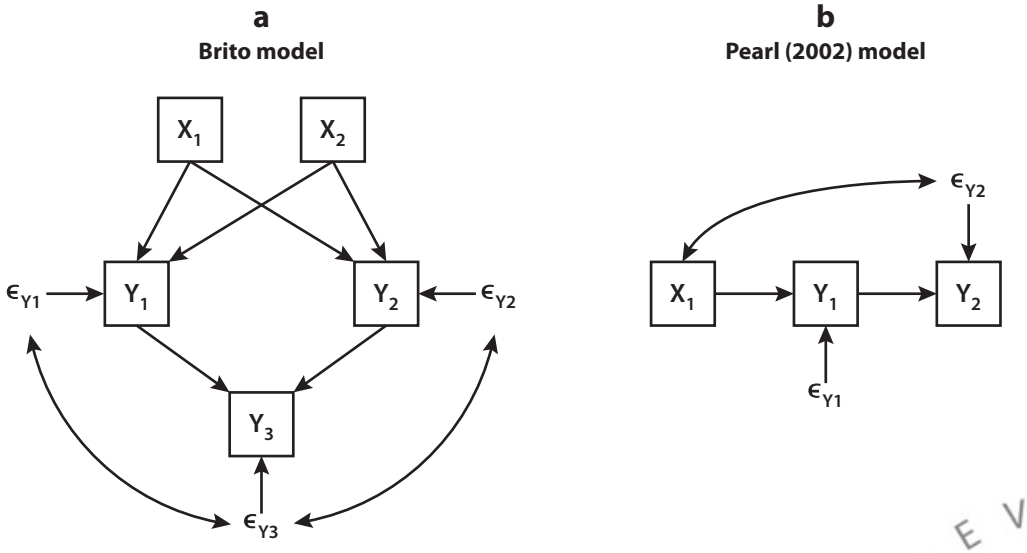


Figure 3

Brito & Pearl (2002) models.

conventional SEM notation. Most researchers familiar with simultaneous equations would not be surprised to see that x_1 and x_2 are IVs for the y_3 equation. The MIIV method enables the selection of the IVs for many of the models that they present. However, it would be valuable to focus on models in which the DAG and MIIV selection methods diverged. **Figure 3*b*** is also a model based on Brito & Pearl (2002, figure 1b). The usual MIIV method does not handle the y_2 equation well unless both y_1 and x_1 are included as covariates. The coefficient for y_1 would be a structural coefficient, whereas the coefficient for x_1 would not, but would control for the correlation of the error (ϵ_{y2}) of y_2 with x_1. If so done, no IV would be needed for either the y_1 or y_2 equations. More contrasts between the MIIV and DAG selection methods would be helpful.

Advantages and disadvantages of the MIIV method. The MIIV method of IV selection has several advantages and disadvantages. On the plus side, a more sustained effort goes into the construction of the SEM rather than directed to searching for AIVs. Ideally, this would force more attention to the model and the assumptions being made as part of the model. MIIVs flow from the structure of the model.

A related advantage is that the model specification clearly shows the assumptions being made about the latent (when present) and observed variables of the model. A researcher need not struggle to see which errors are assumed uncorrelated with which other errors or variables. Assumptions about exogenous variables are part of the specification.

When the model structure leads to an excess of MIIVs, then overidentification tests are possible. These are tests not only of whether all the MIIVs meet the condition of being uncorrelated with the equation error, but also of the model specification because it was the model that led to the selection of the MIIVs. An analyst can test each overidentified equation and this in turn can help to locate problematic parts of the model. Also similarly to the AIV method, the MIIV method typically is used for probability samples with a defined population so that the representativeness of the sample is well described.

One disadvantage of MIIVs is linked to the approximate nature of models. Models by definition are approximations, so it is rare for a model to be exactly correct. This implies that MIIVs might not exactly meet the conditions of IVs. An overidentification test can reveal particularly problematic equations, but this brings us to another disadvantage of MIIVs shared with the AIV method. If an equation is exactly identified, a test of the IV uncorrelatedness with error is not available.

(Quasi) Randomization Instrumental Variables

A third method of IV selection comes from studies in which exposure to an intervention or treatment is randomized. This so-called intention-to-treat variable is recommended as an IV for the treatment variable, where the treatment variable is whether they actually receive the treatment or not. This approach is valuable in that it recognizes that there is not a perfect relation between assignment and actual treatment. Although such an IV might appear in studies that assume equal causal effects for all subjects, much recent social science literature has discussed RIVs in the context of heterogeneous causal effects. I discuss the heterogeneous causal effects in a later section, but here I discuss the characteristics of RIVs in contrast to the AIVs and MIIVs already discussed.

At times, these RIVs are part of a randomized experiment where, for instance, one group might be randomized to a job training program while another is randomly assigned to a control group. Or there might be randomization introduced by "nature" in what are sometimes called natural experiments. Rosenzweig & Wolpin (2000) give several examples of such "natural natural" experiments such as twin births, birth date, gender, and weather events.

Advantages and disadvantages of the RIV method. A primary advantage of RIVs is

that the randomization (or natural experiment) makes correlation with many types of omitted variables unlikely. Most of the variables that compose an error for the outcome variable are unlikely to correlate with a randomization variable. Another advantage of RIVs is that the randomization intention-to-treat variable is likely to be highly correlated with those actually receiving treatment. This lessens the likelihood of weak IVs, a problem that I discuss below. A simpler structural model is another advantage of a randomization design and RIVs. The randomization permits a less elaborate model in that many of the confounders are uncorrelated with the RIV.

Though in many ways the RIV method seems ideal, it too suffers drawbacks. One troublesome assumption is that all of the effect of the intention to treat operates through the treatment variable. In other words, the RIV method assumes that the intention-to-treat variable does not unintentionally manipulate other variables that might affect the outcome. For instance, suppose that selection for a job training program gives hope and confidence for those selected and the opposite for those assigned to the control group. This intent to treat then affects other variables that might affect obtaining a new job that are not part of the job training treatment effect. A related concern is that the randomization might give overconfidence in researchers that discourages them from considering other causal factors that might come into play.

Randomization might also be part of a more artificial environment where people who know that they are part of an intervention or experiment might not behave the same as they would receiving the same treatment if they were not part of the experiment. Hence, the experimental context might create relationships not observed outside of this setting. Also, the subjects in experiments often are not as representative as those in broader probability sample surveys. This raises questions about the population to which the results apply. Finally, it often is the case that the RIV method leads to exact identification so that overidentification tests are not possible. Rosenzweig & Wolpin (2000) suggest that many of these studies unrealistically assume that there is only a single endogenous covariate for the single RIV. More endogenous covariates would cause identification problems.

EVALUATING INSTRUMENTAL VARIABLES

The quality of IV estimates depends on the quality of the IVs on which they rely. Regardless of whether IVs are selected by AIV, MIIV, or RIV methods, IVs require evaluation. IVs should be such that (*a*) IVs ($\mathbf{Z}$) are uncorrelated with ϵ, (*b*) the covariance matrix of IVs ($\mathbf{Z}$) is nonsingular, and (*c*) the rank of the covariance matrix of $\mathbf{Z}$ and $\mathbf{X}$ equals $K_1 + K_2$. One point of great vulnerability is point *a*: that the IVs do not correlate with the equation error. Observational data are often embedded in a complex set of relationships with other variables. To say that the IVs are uncorrelated with the error is to claim that once we take account of the covariates in the model there is no remaining association of the IVs and error. Often theory, logic, or prior research leave considerable uncertainty as to whether such association can be ruled out. As I describe below, under some conditions diagnostic tests are possible for overidentified models. However, if an equation is exactly identified ($K_3 = K_1$), no test is possible.

With regard to point *b* above, information on the nonsingularity of the IVs can be gleaned from the sample covariance matrix of the IVs to see if severe multicollinearity is present. Exact singularity is rare. Moderate to high collinearity is a more likely possibility. Finally, the rank of the covariance matrix of $\mathbf{Z}$ and $\mathbf{X}$ helps in assessing the strength of the partial associations of the IVs with the endogenous $\mathbf{X}_1$ variables. Marginal associations between the IVs ($\mathbf{Z}$) and the covariates ($\mathbf{X}$) lead to "weak instrumental variables," and I summarize some diagnostics for such variables. The next two subsections present overidentification tests and diagnostics for weak IVs. A third section

follows that discusses the optimal number of IVs when there is an excess of candidate IVs.

Overidentification Tests

If the number of covariates that correlate with the disturbance is K_1, then a necessary condition for identification is that there are at least K_1 IVs that are not already covariates in the equation. If each coefficient is identified and the number of IVs is greater than K_1, then the equation is said to be overidentified. That is, there are more IVs than the number needed to identify the equation and its coefficients. The valuable aspect of this excess of IVs is that it permits a test of one of the key assumptions of IVs. A researcher can test the null hypothesis that all IVs are uncorrelated with the equation error versus the alternative hypothesis that at least one IV correlates with the disturbance. Note that failures of the tests tell the researcher that at least one IV correlates with the equation error, but they do not tell you which variables do.

The best known overidentification tests for IVs is over 50 years old and comes from Sargan (1958):

$$T_S = \frac{\hat{\epsilon}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\epsilon}}{\hat{\epsilon}'\hat{\epsilon}/N} \overset{a}{\sim} \chi^2, \qquad 36.$$

where $\hat{\epsilon}$ are the residuals from the IV/2SLS equation, $\boldsymbol{Z}$ are the values of the IVs, and N is the sample size. Asymptotically, the test statistic, T_S, follows a chi square distribution with degrees of freedom (*df*) equal to the number of IVs beyond the minimum needed, that is, the degree of overidentification. A convenient way to calculate the test statistic is

$$NR_u^2 \overset{a}{\sim} \chi^2, \qquad 37.$$

where R_u^2 is the uncentered R^2 of OLS regression of IV/2SLS residuals on IVs. If the original model includes a regression constant and this auxiliary regression includes a constant, then R_u^2 equals the usual R^2 from this auxiliary regression and no further adjustment is needed (Wooldridge 2002, p. 123). Again, the number of excess IVs ($K_3 - K_1$) determines the degrees of freedom.

Basmann (1960) provides another popular overidentification test that is calculated as

$$\frac{\hat{\epsilon}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\epsilon}/(K_3 - K_1)}{\hat{\epsilon}'\boldsymbol{M}_Z\hat{\epsilon}/[N - (K_2 + K_3)]} \overset{a}{\sim} F[K_3 - K_1, N - (K_2 + K_3)], \qquad 38.$$

where $\boldsymbol{M}_Z = \boldsymbol{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, $(K_2 + K_3)$ is the total number of IVs, and $(K_3 - K_1)$ is the number of extra IVs. Asymptotically, it follows an F-distribution with degrees of freedom of $K_3 - K_1$ and $N - (K_2 + K_3)$. The H_o and H_a are the same as the Sargan test. Rejection of H_o means that at least one IV correlates with the error, contrary to the assumption for IVs.

Kirby & Bollen (2009) present several additional variants of the Sargan (1958) and Basmann (1960) tests. In their simulation study, they find that all the test statistics perform well in moderate (e.g., 400) to large samples, but the Sargan (1958) test statistic has the best performance in smaller samples.

These test statistics assume homoscedastic disturbances. For heteroscedastic-consistent test statistics, see Wooldridge (1995), who provides an alternative test statistic for single equation tests. The GMM overidentification test of Hansen (1982) has a heteroscedastic-consistent form that applies to single- or multiequation GMM-IV models (e.g., Hall 2005, Hayashi 2000).

Testing whether the IVs are uncorrelated with the equation disturbance is important regardless of whether AIV, MIIV, or RIV methods led to the IVs. Rejection of H_o means that the IV estimator might not give accurate estimates. The IV overidentification tests are closely linked to the MIIV approach in that it was the model structure that led to the MIIVs for an equation. Rejection of one or more IVs implies rejection of the model specification that led to the MIIV. Hence, the IV overidentification tests are tests of the model specification from the MIIV perspective (Bollen 1996, Kirby & Bollen 2009). (The same is true for the DAG method for observed variable recursive models.) Fortunately, popular statistical software (e.g., Stata, SAS) have one or more of these tests available. All these tests

assume that the equation is overidentified; if exactly identified, these tests are unavailable. Also the tests assume homogeneous rather than heterogeneous causal effects. If the causal effects of a variable differ over individuals, then this could lead to a significant test statistic even when the IVs are uncorrelated with the disturbance. See the section on Heterogeneous Causal Effects.

Weak Instrumental Variables

IVs must not only be uncorrelated with the equation error, but also be associated with the endogenous variable(s) that they predict. A very strong association of an IV with the endogenous variable can improve the standard errors of parameter estimates, but if too strong, it raises questions about whether the IV is uncorrelated with the equation error as is required. Insufficient association of the IVs and the endogenous covariates can increase standard errors and create other problems. Low association between IVs and endogenous variables is known as the weak IVs problem. Weak IVs can occur in several different situations, ranging from the simplest case of a single endogenous covariate and single IV to the more complex situation of a mixture of endogenous and exogenous covariates with multiple additional IVs. I treat weak IVs going from the simplest to more complex models.

One endogenous covariate (x_1), one IV (z). The nature of the weak IVs problem is evident if I return to the simple regression model with an endogenous covariate and a single IV. The simple regression is

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad 39.$$

where $COV(x, \epsilon) \neq 0$ and z is the IV. The IV estimator from Equation 4 is $\beta = \frac{COV(y,z)}{COV(x,z)}$. In the extreme case where the IV(z) is uncorrelated with the covariate (x), the denominator of the IV estimator is zero and β is undefined. Even when the $COV(x, z)$ does not equal but is close to zero, this weak IV creates instability in the sample IV coefficient estimates in that it creates a form of "empirical underidentification."[10]

The preceding assumes that $COV(z, \epsilon) = 0$. If the IV is not perfect so that there is some correlation of the IV and the error, contrary to assumptions, then the IV estimator will not keep its desirable properties. The situation is made worse when the IV variable is weakly correlated with the covariate. A result from Bound et al. (1995, p. 444) helps to illustrate this. They compare the inconsistency of the OLS [plim($\hat{\beta}_{OLS} - \beta$)] and IV [plim($\hat{\beta}_{IV} - \beta$)] estimators in simple regression with a single IV and find that the ratio of the inconsistency is

$$\frac{plim(\hat{\beta}_{IV} - \beta)}{plim(\hat{\beta}_{OLS} - \beta)} = \frac{\rho_{z\epsilon}/\rho_{x\epsilon}}{\rho_{zx}}, \qquad 40.$$

where ρ_{rs} is the correlation between the r and s variables, where r, s indicate the two variables being correlated. Of course, if z is a good IV, then $\rho_{z\epsilon}$ is zero, $plim(\hat{\beta}_{IV}) = \beta$, the numerator is zero, and this ratio is as well. But if $\rho_{z\epsilon}$ and $\rho_{x\epsilon}$ are not zero, then a weak IV (i.e., a low value of ρ_{zx}) can inflate the ratio of inconsistency of the IV versus OLS estimator. Any time $\rho_{z\epsilon}/\rho_{x\epsilon} > \rho_{zx}$, the IV estimator has greater inconsistency than the OLS estimator, and in this sense the researcher would be better off with OLS. For instance, suppose that $\rho_{z\epsilon} = 0.1$, $\rho_{x\epsilon} = 0.4$, and $\rho_{zx} = 0.1$. The ratio of inconsistency is 2.5 so that the inconsistency of $\hat{\beta}_{IV}$ is 2.5 times higher than that of $\hat{\beta}_{OLS}$. The $\hat{\beta}_{IV}$ advantage returns for ρ_{zx} greater than 0.25, but this example illustrates how a weak IV can even make an IV estimator more inconsistent than OLS if its error-IV correlation is not zero.

A natural diagnostic for a weak IV when there is a single covariate and a single IV is simply the correlation of the IV and the covariate (ρ_{zx}) or squared correlation (ρ^2_{zx}), where the latter permits a shared variance interpretation. A significance test of $H_o : \rho_{zx} = 0$ or

[10]This term comes from the latent variable SEM literature, but is not common in the IV literature. It refers to situations in which the population value of a parameter is near but not equal to a value that would underidentify a parameter (Kenny 1979, p. 40).

$H_o : \beta_{xz} = 0$ should accompany the correlation to ensure that the nonzero magnitude is statistically significant (where β_{xz} is the regression coefficient of x on z).

More complicated situations. Suppose I stay with the simple regression of $y = \alpha + x\beta + \epsilon$, but now I have more than one IV. The IV estimator regresses the scalar covariate x on the vector of IVs ($\mathbf{z}$). Bound et al. (1995) find that

$$\frac{plim(\hat{\beta}_{IV} - \beta)}{plim(\hat{\beta}_{OLS} - \beta)} = \frac{COV(\hat{x}, \epsilon)/COV(x, \epsilon)}{R^2_{xz}}, \quad 41.$$

where $COV(\hat{x}, \epsilon)$ is the population covariance of $\hat{x}$ and ϵ with $\hat{x}$ the predicted value of x when regressed on all IVs in $\mathbf{z}$, $COV(x, \epsilon)$ is the population covariance of x and ϵ, and $R^2_{x\mathbf{z}}$ is the population squared multiple correlation from the x on $\mathbf{z}$ regression. This is similar to the previous case of a single covariate and single IV in that even a weak association between the IV and error can lead the $\hat{\beta}_{IV}$ to have greater inconsistency than $\hat{\beta}_{OLS}$ when the IV is weak. But here, rather than the correlation of x and z, the focus is the squared multiple correlation. If the group of IVs has little explanatory power, some of the negative consequences of weak IVs can arise. A natural diagnostic for this problem is to make sure that $R^2_{x\mathbf{z}}$ is sufficiently large and to use the F test of the null hypothesis that all coefficients are zero in this reduced-form equation to make sure that the effects of IVs on the endogenous covariate are statistically significant.

If I have a single endogenous variable ($\mathbf{x}_1$) with other exogenous covariates ($\mathbf{X}_2$), and multiple IVs ($\mathbf{Z}$), the multiple regression equation becomes

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon, \quad 42.$$

where the IVs consist of the exogenous covariates ($\mathbf{X}_2$) and other IVs ($\mathbf{X}_3$) that are not part of the structural model, so that $\mathbf{Z} = [\mathbf{X}_2\ \mathbf{X}_3]$. The $\mathbf{x}_1$ covariate correlates with ϵ whereas $\mathbf{Z}$ ($= [\mathbf{X}_2\ \mathbf{X}_3]$) does not. This equation differs from the previous in that some of the IVs are covariates of the model with direct effects on $\mathbf{y}$.

The $R^2_{x\mathbf{z}}$ for this equation could be high due to $\mathbf{X}_2$ being included among the variables in $\mathbf{Z}$. If I just regressed the endogenous covariate ($\mathbf{x}_1$) on the excluded IVs ($\mathbf{X}_3$), part of the explained variance could be due to the correlations of $\mathbf{X}_2$ and $\mathbf{X}_3$ and it would not give a true sense of the contributions of $\mathbf{X}_3$ that are distinct from $\mathbf{X}_2$ (Nelson & Startz 1990). Instead, a partial R-squared, R^2_p, is useful where the residuals from the regression of $\mathbf{x}_1$ on $\mathbf{X}_2$ are regressed on the residuals of the regression of $\mathbf{X}_3$ on $\mathbf{X}_2$ and the R-squared from this is R^2_p. This gives the variance in the endogenous covariate explained by the $\mathbf{X}_3$ IVs after removing the influences of the $\mathbf{X}_2$ IVs. A low R^2_p suggests weak IVs. A significance test diagnostic derives from $\mathbf{x}_1$ regressed on $\mathbf{X}_2$ and $\mathbf{X}_3$ with a partial F test of all coefficients of $\mathbf{X}_3$ being zero. Staiger & Stock (1997) suggest that an F test statistic less than 10 is a symptom of weak IVs.

The multiple regression model with multiple endogenous and exogenous covariates is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon, \quad 43.$$

with multiple IVs in $\mathbf{Z} = [\mathbf{X}_2\ \mathbf{X}_3]$. This is the most complicated situation in which to diagnose weak IVs. Bowden & Turkington (1984), Hall et al. (1996), Poskitt & Skeels (2002), and Shea (1997) suggest alternative ways to measure weak IVs in this context. Shea (1997, p. 349) proposes a partial R-squared measure that a researcher can construct for each endogenous covariate ($\mathbf{x}_1$) by doing the following:

1. Regress all covariates ($\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$) on $\mathbf{Z}$. Save the fitted values $\hat{\mathbf{X}}$.
2. Regress the endogenous covariate of interest ($\mathbf{x}_1$) on the remaining covariates [$\mathbf{X}$ without $\mathbf{x}_1$] and form predicted value $\hat{\mathbf{x}}_1$. Save the residuals ($\mathbf{x}_1 - \hat{\mathbf{x}}_1$).
3. Regress $\hat{\mathbf{x}}_1$ on the remaining predicted covariates [$\hat{\mathbf{x}}$ without $\hat{\mathbf{x}}_1$] and form a new predicted value $\hat{\hat{\mathbf{x}}}_1$. Save the residuals ($\hat{\mathbf{x}}_1 - \hat{\hat{\mathbf{x}}}_1$).
4. Compute the sample squared correlation between these two sets of residuals ($\mathbf{x}_1 - \hat{\mathbf{x}}_1$) and ($\hat{\mathbf{x}}_1 - \hat{\hat{\mathbf{x}}}_1$).
5. Repeat steps 2 to 4 for each endogenous covariate.

Each endogenous covariate will have a partial R-squared measure. Each partial R-squared that is small is an indicator of weak IVs for the corresponding endogenous covariate and possible estimation problems.

Summary. In the past, most attention to diagnostics given to IVs was whether the IVs were uncorrelated with the equation disturbance as is necessary for proper IVs. Within the last decade or so, methodologists have given much more attention to the issue of IVs that are only weakly related to the endogenous covariates. Weak IVs are particularly troublesome in situations in which the IVs do not completely satisfy the assumption of being uncorrelated with errors. Even small violations can create difficulties when occurring with weak IVs. Diagnostics and tests for weak IVs continue to evolve. See Stock et al. (2002) and Stock & Yogo (2005) for additional diagnostics for weak IVs.

Clearly, weak IV diagnostics should be done in conjunction with the overidentification tests. Violating either condition is damaging; violating both conditions essentially undermines the value of IV estimators.

How Many Instrumental Variables?

A necessary (but not sufficient) condition to identify the coefficients of endogenous covariates is that we have as many IVs that are not already included explanatory variables as there are endogenous covariates. A situation of three endogenous covariates demands at least three IVs that are not already in the equation. But what if we have more than the minimum? How many IVs should we include?

This question goes back many years, as illustrated in Sargan's (1958, p. 400) early work on IVs: "[T]he improvements are usually small after the first three or four instrumental variables have been added. Thus there may be no great advantage in increasing the number of instrumental variables, and from the later discussion it emerges that the estimates have large biases if the number of instrumental variables becomes too large."

Research has not definitively settled this question, but advice is possible based on a series of studies. Several authors have observed that a large number of IVs beyond the minimum can increase the finite sample bias of the IV estimator (e.g., Nagar 1959; Sawa 1969; Mariano 1982; Angrist & Krueger 2001, p. 79). However, Buse (1992) suggests that the key issue is whether the R-squared from the first-stage regression continues to increase as the number of IVs increases. That is, Buse's analytic results suggest that bias in the IV estimator need not increase if the amount of explained variance in the endogenous covariate from the first-stage regression increases in proportion to the number of IVs. A recent simulation study (Bollen et al. 2007) of the MIIV-2SLS IV estimator from Bollen (1996) finds that a large number of IVs tends to bias coefficients mostly in smaller samples (e.g., $N \leq 100$), whereas the excess number of IVs mattered little in big samples.

Combining the evidence tentatively suggests that a large number of IVs (large degree of overidentification) is most harmful when the sample size is small and the incremental contribution to the R-squared of the endogenous covariate regressed on the IVs is small for the added IVs. However, the number of IVs matters less in big samples or when the added IVs continue to improve the R-squared of the first-stage regression. A researcher with a modest sample size should consider choosing one or two IVs above the minimum, whereas a researcher with a large N can include more IVs, though the gains in efficiency are still possible when the added IVs noticeably improve the explained variance in the endogenous covariate. These guidelines require qualification for models with heterogeneous causal effects. As I explain below, each IV might be seen as affecting only a single group, and different IVs affect different groups. In the next section, this perspective is more fully developed.

HETEROGENEOUS CAUSAL EFFECTS

The prior sections assume constant effects of each variable on the dependent variable. For instance, in the simple regression model,

$y_i = \alpha + \beta x_i + \epsilon_i$, the expected effect of x_i on y_i is β regardless of the case. An alternative assumption is that the expected effect of x_i on y_i might differ by case so that

$$y_i = \alpha + \beta_i x_i + \epsilon_i, \qquad 44.$$

where β_i is the expected impact of x_i on y_i for the ith case. That is, the effect of x on y can differ by case. This represents heterogeneous causal effects in contrast to the homogeneous effects I have assumed up to this point. The treatment of random coefficients such as β_i in regression models is a separate literature (e.g., Johnson 1977, 1980; Rubin 1950; Swamy 1970; Theil 1971), most of which falls beyond the scope of this review. I restrict my comments to the situation in which $COV(x_i, \epsilon_i) \neq 0$ and the use of IVs to estimate a model with heterogeneous causal effects. My focus is on research that has sought to synthesize IV estimation from the structural equation tradition of econometrics with the Splawa-Neyman (1923 [1990]) and Rubin (1974) potential outcome (or counterfactual) literature on causal inferences (e.g., Angrist & Imbens 1995, Angrist et al. 1996, Angrist & Pischke 2009, Imbens & Angrist 1994). Morgan & Winship (2007) provide a very useful recent review of this literature.

This literature has provided insight into IVs and their meaning when we allow the causal effects to differ by individual. Much of this literature has concentrated on "treatment effects" when x_i is a dummy variable (1 = treatment, 0 = no treatment). The treatment might be the effects of attending a Catholic school on academic achievement (e.g., Morgan & Winship 2007), the impact of attending a job training program on wages (e.g., Angrist & Pischke 2009, pp. 162–64), or any other dichotomous intervention or treatment variable. Though it is possible to extend this work to multivalued x_i (see, e.g., Angrist & Imbens 1995), I stick with the dichotomous variable situation where these potential outcome ideas are most developed.

The potential outcome perspective requires extensions to the notation. As stated, x_i is a dichotomous treatment variable, and I keep y_i as the outcome measure and z_i as a single dichotomous IV. The potential outcome approach assumes that for each individual there is a potential value $x_i = 1$ and $x_i = 0$, even though in a given sample the ith individual takes just one of these values. The former value is labeled $x_{(1)i}$, the treatment value for the ith individual. The second value for each individual is the no-treatment value named $x_{(0)i}$. Similarly, the potential values of the IV z_i are $z_{(1)i}$ and the second is named $z_{(0)i}$ where, again, in a given sample each individual has only $z_{(1)i}$ or $z_{(0)i}$ observed, with the other unobserved.

To further explain this notation, I use an empirical example presented by Angrist (1990) and Angrist & Pischke (2009). Suppose that interest lies in the causal effect on wages of serving in the military. The "treatment" is military service, where $x_i = 1$ for veterans and $x_i = 0$ for those with no military service. The $x_{(1)i}$ variable is the treatment variable that is observed for those men in the sample who are veterans and is unobserved for those who are nonveterans. Similarly, the $x_{(0)i}$ is the nonveteran status variable that is observed for all men in the sample who did not serve and is unobserved for those who did serve. The IV is a dummy variable created from the draft lottery number of the men in the study. The cohort of men whom they studied were subject to being drafted into the military based on a randomly drawn lottery number assigned according to their birthdays. Those men at or below a cutpoint were eligible to be drafted, whereas those above it were not [there is evidence that the 1970 lottery was not truly randomized (Fienberg 1971); I ignore this issue in this discussion]. Here, those draft eligible have $z_i = 1$ and those not eligible have $z_i = 0$. The $z_{(1)i}$ for each individual is observed only for those men who were draft eligible and was unobserved for those not eligible. In a like fashion, the $z_{(0)i}$ for each individual is unobserved for draft eligible men and observed for ineligible men.

A common assumption in the potential outcome literature is the stable unit treatment value assumption (SUTVA) (Rubin 1978). This assumption is that the potential outcome

for each case i is unrelated to the treatment status of all other individuals. In other words, the values of the variables for the jth case do not affect those for the ith case ($i \neq j$). Wooldridge (2002, p. 604) points out that the more common and stronger assumption of an independent, identically distributed sample from a population implies that SUTVA is met.

This assumption permits a potential outcome definition of the causal effect of z on x for the ith individual. It is $[x_i(z_{(1)i}) - x_i(z_{(0)i})]$, where $x_i(z_{(1)i})$ is the value of x_i for $z_i = 1$ and $x_i(z_{(0)i})$ is the value of x_i for $z_i = 0$. The difference is the causal effect. The causal effect of the IV z on y is $[y_i(z_{(1)i}) - y_i(z_{(0)i})]$, with $y_i(z_{(1)i})$ and $y_i(z_{(0)i})$ being the value of y_i with $z_i = 1$ and $z_i = 0$, respectively (Angrist et al. 1996). The primary obstacle to estimating these quantities is that the potential outcome values of z_i are unobserved, and this prevents direct calculation of these individual causal effects as defined. But with additional assumptions, average or mean causal effects are estimable.

An extremely important assumption is that the IV is assigned randomly so that the probability of z_i being 1 is the same for each i. A fair lottery process as in the draft eligibility example would satisfy this condition because the probability of each individual receiving a number below the cutoff was the same for everyone. The SUTVA and the randomized treatment assumption enable estimation of the "intention-to-treat" mean effects of z on x and of z on y. These are formed by taking the difference in the means of x for those cases with $z_i = 1$ and those with $z_i = 0$ and the difference in the means of y for the same two groups. The intention to treat refers to individuals who were placed in the group to be treated but did not necessarily receive treatment.

Intention-to-treat mean effects are valuable estimates, but it is not unusual to want to know not the causal impact of z but the causal effect of x on y. Additional assumptions are called for to estimate additional causal effects. The exclusion restriction assumption is that $y_i(x, z = 0) = y_i(x, z = 1)$, where x equals 1 or 0. In other words, whichever value x affects y, and once x equals either 1 or 0, the value of z does not matter. It is called an exclusion assumption because z does not directly affect y and is excluded once x is included.

A fourth assumption is that the IV z has a nonzero causal effect on x. The final assumption is the monotonicity assumption, which states that for those cases with a z to x effect the direction of the effect is the same. That is, z does not negatively affect x for some cases and positively affect x for other cases.

With these four assumptions, the IV estimator of

$$\frac{E[y_i|z_i = 1] - E[y_i|z_i = 0]}{E[x_i|z_i = 1] - E[x_i|z_i = 0]}$$

is called the local average treatment effect (LATE) (Imbens & Angrist 1994), which is the treatment effect of x for those whose treatment status can be changed by z. Notice that this is a subgroup of the population. The presence of subgroups is a consequence of the heterogeneous causal effect assumption. Unlike the prior IV sections, here it is not assumed that the same causal effect holds for every individual.

In the current setup of a dichotomous IV and dichotomous treatment, another division is drawn (e.g., Angrist & Pischke 2009). One group consist of the compliers, and these refer to the subpopulation with $z_i = 1$ and $x_i = 1$ or $z_i = 0$ and $x_i = 0$. That is, their treatment statuses are consistent with their IV value. In our example, the compliers consist of two types: those with low lottery numbers and military service and those with high lottery numbers and no military service. "Always takers" make up a second subpopulation in which $x_i = 1$ regardless of whether $z_i = 1$ or $z_i = 0$. In the running example, these are cases of individuals who enter military service regardless of their lottery number. The "never takers" have $x_i = 0$ with $z_i = 1$ or $z_i = 0$. These individuals never serve in the military irrespective of their lottery number. Finally, the "defiers" is a subpopulation that does the opposite of what they are assigned. These are individuals who, if given a high lottery number, would enlist in the military or, if given a low lottery number, would avoid

military service. The monotonicity assumption rules out such a subpopulation, as explained in Angrist et al. (1996).

The LATE estimate is the mean causal effect of the compliers. It does not usually estimate the effect of all those receiving treatment or the average treatment effect on the treated. The average treatment effect of the treated is a combination of the compliers and the always takers, and LATE estimates only the former (Angrist & Pischke 2009, pp. 158–59). This is useful in policy settings where only a certain subgroup responds to the IV. It provides information on the magnitude of change expected for that group.

Another consequence of IV estimates of heterogeneous causal effects is that different IVs can lead to different estimates of average causal effects. An explanation for this is that the individuals who are compliers can differ by IV, so the first IV might lead to treatment for one person but a different IV might not do the same (Angrist & Pischke 2009). This is consistent with the idea of causal heterogeneity. A general causal effect is not well defined. Rather, we expect different causal effects for different individuals with different IVs.

This has implications for the overidentification tests that I discussed in the prior section, Overidentification Tests. In the heterogeneous causal effect model in which the IVs each have different complier subpopulations, a failed overidentification test could be simply a reflection of different complier subpopulations rather than a failure of the IVs to be uncorrelated with the disturbance. Alternatively, failure to reject the usual null hypothesis for the overidentification tests is consistent with homogeneous rather than heterogeneous causal effects (Angrist & Pischke 2009, p. 167).

Weak IVs remain a threat to determining heterogeneous causal effects in that if the IV is only weakly related to the treatment, problems with estimates would be expected. The weak IV issue is closely tied to the fourth assumption that lies behind the LATE estimate. That is, it assumes that z has a nonzero and sufficiently strong causal effect on x.

INSTRUMENTAL VARIABLES IN SOCIOLOGY

The prior sections present a variety of situations in which IVs might be useful and provide diagnostics for evaluating IVs. In this section, I briefly discuss IV methods in practice in sociology. To inform this section, I considered all articles in the *American Sociological Review* (*ASR*), *American Journal of Sociology* (*AJS*), and *Social Forces* (*SF*) that use IVs in their analysis. These journals are widely considered the top three general sociological journals, so reviewing them should provide insight into the sociological practice in the field's most highly rated journals. This examination is limited to articles published between 2000 and 2009, so I can draw on a decade of experience (see sidebar entitled Locating Articles for more information). In total, 57 articles made use of IV methods in their empirical analyses.

Table 1 lists (*a*) the authors and year of study, (*b*) the type of IV application, (*c*) the method of selecting IVs, (*d*) the role of the IV analysis and whether it was the primary analysis or part of a sensitivity analysis, (*e*) whether an overidentification test was used, and (*f*) if weak IVs diagnostics were applied. Complete citations for these studies are in the bibliography.

Table 1 reveals several things. The range of topics that use IVs is broad, from studies of global economic processes to the education of children. This reflects the common occurrence across fields of the problems that IVs are intended to address. However, a set of only 57 out of the nearly 1,500 papers published in *ASR*, *AJS*, and *SF* during this period represents less than 5% of the articles. Feedback relations, omitted variables, measurement error, and sample selectivity surely characterize the majority of sociological studies. Though some solutions need not involve IV methods, the more likely explanation for not seeing even more IVs is that these problems either go unrecognized or are ignored. Education, for instance, is a common covariate in many sociological models that predict individual income.

Yet a good-sized literature in economics treats education as endogenous in explaining income and devises IVs to help estimate its effect (see Rosenzweig & Wolpin 2000, table 1). Scales or indexes are common covariates where no correction is made for the measurement error still present in such measures. Though not the central part of the analysis, this apparent insufficient attention in sociological research to the "endogeneity problem" contrasts with what has been a preoccupation of the economics literature (e.g., Wooldridge 2002, pp. 50–51).

Turning to the 57 studies in **Table 1**, the most common application of IVs is in multiple regression models where the focus is a single dependent variable. Sometimes this is in the context of panel models (e.g., Alderson 2004, Brady et al. 2005), whereas other times it is a spatial model (e.g., Baller & Richardson 2002). Sample selection simultaneous models with one equation to model the sample selection process and another for the main substantive variable is another common IV application (e.g., Meier 2003, Zajac & Westphal 2004). Limited dependent variable models also appear in **Table 1** (e.g., Alon et al. 2001, Lizardo 2006). Other forms of simultaneous equations and latent variable SEM models are less frequent applications for IV methods in these sociological studies.

The selection of IVs is predominantly by AIV, where variables that are not an explicit part of the original structure are brought into the model to identify it. The most common AIV variables are lagged variables in panel data and spatially weighted lagged exogenous variables in spatial models. Just a few studies use RIVs. Clampet-Lundquist & Massey (2008), Kirk (2009), and Ludwig et al. (2008) use intention-to-treat variables as IVs. Regoeczi (2002) specifies a SEM model including latent variable and uses the MIIV method to select IVs based on the model structure. No study used the DAG method of selecting MIIVs. By far, the AIV method is the most common selection method.

The last two columns of **Table 1** address whether the papers applied diagnostics to assess the quality of the IVs. Unfortunately, only a small proportion of the studies reported overidentification tests. The overidentification tests are important in that they test whether all IVs of the analysis are uncorrelated with the equation disturbance as should be true for IVs. Without the test, others cannot know if the IVs and estimates based on them are suitable. Even fewer of the studies addressed the issue of weak IVs. Weak IVs can seriously bias the estimates of IV methods when there is even minor correlation between the IV and error. Without overidentification tests and weak IVs checks, it is difficult to know the trustworthiness of the model estimates.

Of course, if a model is exactly identified, then the overidentification test is unavailable, though weak IV checks still apply. A related concern is that many of the sample selectivity

LOCATING ARTICLES

JSTOR was the primary source for locating articles. The following keywords were searched: instrumental variable, 2SLS, two-stage least squares, two stage least squares, two stage least-squares, 3SLS, three stage least squares, three-stage least squares, three stage least-squares, endogeneity, reciprocal effects, reverse causation, Arellano-Bond, and Heckman selection. The use of "instrument" or "IV" resulted in too many false positives. JSTOR allowed for search of titles, abstracts, and the full text of articles. JSTOR's coverage included *ASR* between 2000 and 2007, *AJS* between 2000 and 2009, and *SF* between 2000 and 2009. Additional steps were taken to identify articles in ASR in 2008 and 2009. For 2008, all articles in ASR were individually read with relevant articles identified by focusing on the abstract, the data and methods sections, the titles, contents, and footnotes of all tables, and any appendices. For 2009, it was possible to search electronic versions of the articles using the same key words as the JSTOR search.

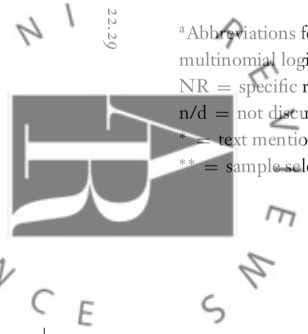
The total search yielded 186 articles, but most of these were false positives. False positives tended to result because the article mentioned a key term but did not use IV methods of analysis, either because the term came up in the literature review of the paper, but was not otherwise used, or it used a non-IV method for its analysis. Of the 186 articles identified in the search, 57 include some form of IV analysis. My thanks to Shawn Bauldry for his work on this literature review.

Table 1 Overview of empirical studies

Source	Application type[a]	IV selection	Role	Overidentification test?	Weak IV check?
Alderson (2004)	Mul. Reg. (panel, Arellano-Bond)	AIV (lagged vars.)	Primary	No	No
Alon et al. (2001)	Lim. Dep. (mult. logit)	AIV	Primary	No	1st stage results
Axinn & Barber (2001)	Lim. Dep. (discrete survival)	AIV	Sensitivity NR	No	No
Baller & Richardson (2002)	Mul. Reg. (spatial model)	AIV (spatial lag)	Primary	No	No
Bandelj (2009)	Mul. Reg. (2SLS)	AIV (lagged vars.)	Sensitivity	Exact	No
Barro & McCleary (2003)	Sim. Eq. (3SLS)	AIV	Primary	Yes	Yes
Brady et al. (2005)	Mul. Reg. (panel, Arellano-Bond)	AIV (lagged vars.)	Sensitivity NR	No	No
Buchmann (2000)	Sim. Eq. (2SLS)	AIV	Primary	(No/maybe exact)	No
Burris (2004)	Mul. Reg. (2SLS)	AIV	Sensitivity	Exact	No
Burris (2005)	Sim. Eq. (sample selection)	AIV	Sensitivity NR	No	No
Castilla (2005)	Sim. Eq. (sample selection)	AIV	Primary	No	1st stage results
Chang (2005)	Sim. Eq. (sample selection)	AIV	Primary	No	1st stage results
Cheng & Powell (2007)	Sim. Eq. (sample selection)	AIV	Sensitivity NR	Exact	No
Clampet-Lundquist & Massey (2008)	Mul. Reg./Lim. Dep. (2SLS)	RIV	Sensitivity NR	No	No
Cunningham & Phillips (2007)	Lim. Dep. (count, spatial model)	AIV (spatial lag)	Primary	No	No
Elman & O'Rand (2004)	Sim. Eq. (endog. switching)	AIV	Primary	No	No
Ferraro & Kelley-Moore (2003)	Sim. Eq. (tobit, sample selection)	AIV	Primary	No	No
Griffin & Bollen (2009)	Mul. Reg./Lim. Dep. (2SLS)	AIV	Sensitivity NR	No	No
Hagan & Foster (2003)	Sim. Eq. (2SLS)	AIV (cross-lag)	Primary	(No/maybe exact)	No
Ingram et al. (2005)	Mul. Reg. (panel, 2SLS)	AIV	Sensitivity	No	No
Jenkins et al. (2006)	Sim. Eq. (spatial model)	AIV (spatial lag)	Primary	Exact	No
Kim & Schneider (2005)	Sim. Eq. (sample selection)	AIV	Primary	Exact	No
Kirk (2009)	Lim. Dep. (2SLS)	RIV	Primary	Yes	Yes
Kiser & Linton (2001)	Mul. Reg. (WLS)	AIV	Sensitivity	No	No
Koçak & Carroll (2008)	Mul. Reg. (FD-IV panel)	AIV (lagged vars.)	Primary	Exact	Yes
Lee (2007)	Mul. Reg. (2SLS)	AIV	Primary	Yes	Yes
Lee & Shihadeh (2009)	Lim. Dep. (count, spatial model)	AIV (spatial lag)	Sensitivity NR	(No/maybe exact)	No
Lizardo (2006)	Lim. Dep. (count)	AIV	Primary	Yes	Yes
Loftus (2001)	Mul. Reg. (2SLS)	AIV	Primary	No	No
Ludwig et al. (2008)	Mul. Reg. (2SLS)	RIV	Primary	No	Yes

Mayer (2001)	Mul. Reg. (2SLS)	AIV	Sensitivity	Exact	No
McCarthy & Casey (2008)	Sim. Eq. (sample selection)	AIV	Primary	No	1st stage results
McDonald & Elder (2006)	Sim. Eq. (sample selection)	AIV	Primary	Exact	No
McManus & DiPrete (2001)	Mul. Reg. (panel model)	AIV (lagged vars.)	Primary	Exact	No
McVeigh et al. (2004)	Mul. Reg. (spatial model)	AIV (spatial lag)	Primary	No	No
Meier (2003)	Sim. Eq. (sample selection)	AIV	Primary	Exact	No
Messner et al. (2004)	Sim. Eq.	AIV	Primary	No	Yes
Morenoff (2003)	Mul. Reg. (spatial model)	AIV (spatial lag)	Primary	No	No
Myers (2000)	Sim. Eq. (sample selection)	AIV	Sensitivity NR	No	No
Noonan et al. (2005)	Sim. Eq. (sample selection)	AIV	Sensitivity NR	Exact	No
Offer & Schneider (2007)	Mul. Reg. (2SLS)	AIV	Primary	Exact	No
Ono (2007)	Sim. Eq. (sample selection)	AIV	Primary	No	1st stage results
Parrado & Flippen (2005)	Sim. Eq. (endog. switching)	AIV	Sensitivity NR	No	No
Paxton & Kunovich (2003)	Mul. Reg. (2SLS)	AIV	Sensitivity NR	No	Yes
Qian et al. (2005)	Sim. Eq. (sample selection)	AIV	Primary	No	No
Regoeczi (2002)	SEM (2SLS)	MIIV	Primary	Yes	Yes
Rosenfeld et al. (2001)	SEM	AIV	Primary	No	1st stage results
Ross et al. (2001)	Mul. Reg. (2SLS)	AIV	Sensitivity	No	No
Rotolo & McPherson (2001)	Mul. Reg. (panel, 2SLS)	AIV	Sensitivity NR	(No/maybe exact)	No
Sanderson & Kentor (2009)	Mul. Reg. (panel, Arellano-Bond)	AIV (lagged vars.)	Primary	No	No
Schram et al. (2009)	Sim. Eq. (sample selection)	AIV	Sensitivity	No	1st stage results
Tomaskovic-Devey & Skaggs (2002)	Sim. Eq. (2SLS)	AIV	Primary	Yes	No
Uggen & Thompson (2003)	Mul. Reg. (2SLS)	AIV	Sensitivity NR	(No/maybe exact)	No
Villarreal & Yu (2007)	Mul. Reg. (2SLS)	AIV	Sensitivity NR	(No/maybe exact)	No
You & Khagram (2005)*	Mul. Reg. (2SLS)	AIV	Primary	Yes	Yes
Young (2009)	Mul. Reg./Sim. Eq. (3SLS, 2SLS, LIML IV)	AIV	Didactic	No	Yes
Zajac & Westphal (2004)	Sim. Eq. (sample selection)	AIV	Sensitivity NR	Exact	No
Zhang et al. (2008)	Sim. Eq. (sample selection)	AIV	Primary	Exact	No

[a] Abbreviations for application type: Mul. Reg. = multiple regression model; Sim. Eq. = simultaneous equations model; Lim. Dep. = limited dependent variable model; Mult. Logit = multinomial logit model. The parentheticals provide additional information about the type of model estimated.
NR = specific results not reported (just a general discussion in the text).
n/d = not discussed (in the "identification" column, this means not enough information is provided to determine the identification status).
* = text mentions an online appendix that provides details about the analysis, but the appendix is no longer available.
** = sample-selection model appears to rely on functional form for identification as the author(s) do not discuss any IVs.



studies included in this review tended to have too few IVs and largely relied on the exogenous covariates of the substantive equation to identify the sample selection term. In essence, the only reason the predicted sample selection term is not perfectly collinear with the exogenous variables is that the sample selection term is a nonlinear function of the exogenous variables. This is a weak basis of identification, and it would be preferable to have additional IVs in the sample selection equation (Winship & Mare 1992).

Another situation in which IVs deserve special attention is when the IVs are lagged variables. Autocorrelation of the residuals of the lagged and contemporary values of endogenous variables can invalidate the lagged values as IVs. Results of autocorrelation diagnostics should be reported.

There are a couple of reasons why overidentification tests and weak IV checks might not be reported. One is that the IV methods might have been part of a sensitivity analysis rather than the primary analysis as indicated in the "Role" column of **Table 1**. A researcher might mention that an IV analysis was done to check whether the results without IV are the same but only report the non-IV analysis. In that situation, lots of details are omitted, and the diagnostics might have been run without having the space to report them.

A second situation in which the overidentification test might not be reported is when heterogeneous causal effects are assumed so that different IVs lead to different treatment effects. However, even in this situation the overidentification test provides interesting information in that Angrist & Pischke (2009, pp. 166–67) suggest that homogeneous causal effects are testable with the usual overidentification tests and multiple IVs. Such a test would be informative even when heterogeneous effects are assumed.

CONCLUSIONS

Observational data, which form the bulk of data available for sociological research, are subject to many problems that create a correlation between the error or disturbance of an equation and one or more of the covariates in a model. Feedback relationships, measurement error, sample selectivity, and omitted variables are all too common in most substantive areas. IV methods provide ways in which these problems can be taken into account when developing estimates. However, the value of the IV methods critically depends on two conditions for IVs: (*a*) being uncorrelated with the equation disturbance and (*b*) not being a weak IV. Threats to fulfilling both conditions are present in many applications. The dominance of the AIV method of selecting IVs contributes to one such threat. Consider the lack of correlation of IV and the equation disturbance. The putative IVs from the AIV method sometimes appear ad hoc in that they are selected after the key equations are already formulated. The nature of the relationship between the AIVs and the rest of the model structure is left unspecified. A more careful examination of the role of these AIVs in the structure of the model could reveal reasons to suspect correlations between the equation error and the AIVs. The MIIV method explicitly shows the role of all MIIVs because the model is formulated in advance. Though there is no guarantee that the MIIVs are uncorrelated with the equation error, they have a stronger theoretical base than the AIVs. Regardless of whether one uses the AIV or MIIV method to select IVs, overidentification tests of the IVs are insufficiently applied. Sometimes this is due to researchers using exactly identified equations so that the overidentification tests are inapplicable, but other times these tests are not used when they could be. These should be routinely applied to overidentified equations.

Weak IVs is a second threat to IV estimators. Recent methodological research has given far more attention to this issue than was true in the past, but my review of the sociological literature reveals that these ideas have not substantially penetrated the published articles. Diagnostics for weak IVs are available, as I summarized in a previous section. There are few reasons not to apply them.

The heterogeneous causality IV methods represent a promising application of IVs. In this area, there appears to be too little testing of whether heterogeneous or homogeneous effects are needed. Rather, heterogeneous effects are assumed without sufficient testing. It is more complicated to test whether the IVs are uncorrelated with the disturbances when the model permits heterogeneous effects. Multiple IVs are often not available. But even when more than one IV is available for a single endogenous covariate, the traditional overidentification tests assume homogeneous causal effects. The traditional tests could be viewed as testing the joint hypothesis that all IVs are uncorrelated with the error and that there are homogeneous causal effects. The alternative hypothesis is either heterogeneous causal effects or error correlated with IVs (Angrist & Pischke 2009, p. 167). Such a test would be informative if the researcher kept in mind the different null and alternative hypotheses being tested when heterogeneous effects are assumed. Weak IV diagnostics also are rarely discussed for heterogeneous effect models but would be valuable.

The number of *ASR*, *AJS*, and *SF* articles using IVs over the past ten years also is informative. Less than 5% of the articles published in the flagship general sociological journals use IVs. Even considering that a significant proportion of articles are not quantitative, this is a low percentage. And this is not because there is an alternative method being used.

Overall, this review reveals that IV methods can be powerful but are underutilized tools to address common problems in sociological research. Even a casual examination of articles in sociology would reveal many situations in which nonnegligible measurement error in covariates, possible omitted variable effects, feedback relations, or sample selectivity are likely but are not discussed. Analyses ignoring these problems are readily permitted and published. Researchers who address these issues and use IVs to do so could benefit from making greater use of diagnostics to assess the quality of the IVs. The diagnostic results might not always please a researcher, but the discipline would nevertheless be better off by addressing these issues. And although we would be right to be concerned about the way IVs are used in sociology, we should be more concerned that they are not used more often.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank Shawn Bauldry for research assistance in the preparation of this review. He, Barbara Entwisle, and Chris Winship provided helpful comments on an earlier draft of this paper. I gratefully acknowledge the support of the National Science Foundation (NSF SES 0617276).

LITERATURE CITED

Adkins LC. 2012. Testing parameter significance in instrumental variables probit estimators: some simulation results. *J. Stat. Comput. Simul.* In press

Alderson AS. 2004. Explaining the upswing in direct investment: a test of mainstream and heterodox theories of globalization. *Soc. Forces* 83:81–122

Alon S, Donahoe D, Tienda M. 2001. The effects of early work experience on young women's labor force attachment. *Soc. Forces* 79:1005–34

Angrist JD. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from Social Security administrative records. *Am. Econ. Rev.* 80:313–36

Angrist JD, Imbens GW. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Stat. Assoc.* 90:431–42

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55

Angrist JD, Krueger AB. 1991. Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* 106:979–1014

Angrist JD, Krueger AB. 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. *J. Econ. Perspect.* 15:69–85

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

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht, Neth.: Kluwer Acad.

Arellano M, Bond S. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Stud.* 58:277–97

Axinn WG, Barber JS. 2001. Mass education and fertility transition. *Am. Sociol. Rev.* 66:481–505

Baller RD, Richardson KK. 2002. Social integration, imitation, and the geographic patterning of suicide. *Am. Sociol. Rev.* 67:873–88

Bandelj N. 2009. The global economy as instituted process: the case of Central and Eastern Europe. *Am. Sociol. Rev.* 74:128–49

Barro RJ, McCleary RM. 2003. Religion and economic growth across countries. *Am. Sociol. Rev.* 68:760–81

Basmann RL. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25:77–83

Basmann RL. 1960. On finite sample distributions of generalized classical linear identifiability test statistics. *J. Am. Stat. Assoc.* 55:650–59

Blau PM, Duncan OD. 1967. *The American Occupational Structure*. New York: Wiley

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

Bollen KA. 1996. An alternative 2SLS estimator for latent variable models. *Psychometrika* 61:109–21

Bollen KA. 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 du Toit, D Sörbom, pp. 119–38. Lincolnwood, IL: Sci. Softw. Int.

Bollen KA, Bauer DJ. 2004. Automating the selection of model-implied instrumental variables. *Sociol. Methods Res.* 32:425–52

Bollen KA, Biesanz JC. 2002. A note on a two-stage least squares estimator for higher-order factor analyses. *Sociol. Methods Res.* 30:568–79

Bollen KA, Guilkey DK, Mroz TA. 1995. Binary outcomes and endogenous explanatory variables: tests and solutions with an application to the demand for contraceptive use in Tunisia. *Demography* 32:111–31

Bollen KA, Kirby JB, Curran PJ, Paxton PM, Chen F. 2007. Latent variable models under misspecification: two stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociol. Methods Res.* 36:46–86

Bollen KA, Kolenikov S, Bauldry S. 2011. *Generalized method of moments—instrumental variable estimators for latent variable models.* Work. Pap. 1–51

Bollen KA, Maydeu-Olivares A. 2007. Polychoric instrumental variable (PIV) estimator for structural equations with categorical variables. *Psychometrika* 72:309–26

Bound J, Jaeger DA, Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90:443–50

Bowden RJ, Turkington DA. 1984. *Instrumental Variables*. Cambridge, UK: Cambridge Univ. Press

Brady D, Seeleib-Kaiser M, Beckfield J. 2005. Economic globalization and the welfare state in affluent democracies, 1975–2001. *Am. Sociol. Rev.* 70:921–48

Brito C. 2010. Instrumental sets. In *Heuristics, Probability, and Causality*, ed. R Dechter, H Geffner, JY Halpern, pp. 295–308. London: College Publ.

Brito C, Pearl J. 2002. *A graphical criterion for the identification of causal effects in linear models.* Presented at Proc. AAAI Conf., Edmonton, Can.

Buchmann C. 2000. Family structure, parental perceptions, and child labor in Kenya: What factors determine who is enrolled in school? *Soc. Forces* 78:1349–78
Burris V. 2004. The academic caste system: prestige hierarchies in PhD exchange networks. *Am. Sociol. Rev.* 69:239–64
Burris V. 2005. Interlocking directorates and political cohesion among corporate elites. *Am. J. Sociol.* 111:249–83
Buse A. 1992. The bias of instrumental variable estimators. *Econometrica* 60:173–80
Cameron AC, Trivedi PK. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge Univ. Press
Castilla EJ. 2005. Social networks and employee performance in a call center. *Am. J. Sociol.* 110:1243–83
Chang ML. 2005. With a little help from my friends (and my financial planner). *Soc. Forces* 83:1469–97
Cheng S, Powell B. 2007. Under and beyond constraints: resource allocation to young children from biracial families. *Am. J. Sociol.* 112:1044–94
Clampet-Lundquist S, Massey DS. 2008. Neighborhood effects on economic self-sufficiency: a reconsideration of the moving to opportunity experiment. *Am. J. Sociol.* 114:107–43
Crow JF. 1978. Wright, Sewall. In *International Encyclopedia of the Social Sciences—Biographical Supplement*. New York: Macmillan
Cunningham D, Phillips BT. 2007. Contexts for mobilization: spatial settings and Klan presence in North Carolina, 1964–1966. *Am. J. Sociol.* 113:781–814
Didelez V, Meng S, Sheehan NA. 2010. Assumptions of IV methods for observational epidemiology. *Stat. Sci.* 25:22–40
Elman C, O'Rand AM. 2004. The race is to the swift: socioeconomic origins, adult education, and wage attainment. *Am. J. Sociol.* 110:123–60
Ferraro KF, Kelley-Moore JA. 2003. Cumulative disadvantage and health: long-term consequences of obesity. *Am. Sociol. Rev.* 68:707–29
Fienberg SE. 1971. Randomization and social affairs: the 1970 draft lottery. *Science* 171:255–61
Geary RC. 1949. Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica* 17:30–58
Goldberger AS. 1964. *Econometric Theory*. New York: Wiley
Goldberger AS. 1972. Structural equation methods in the social sciences. *Econometrica* 40:979–1001
Griffin LJ, Bollen KA. 2009. What do these memories do? Civil rights remembrance and racial attitudes. *Am. Sociol. Rev.* 74:594–614
Grootendorst P. 2007. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Serv. Outcome Res. Methodol.* 7:159–79
Guilkey DK, Mroz TA, Taylor L. 1992. *Estimation and testing in simultaneous equations models with discrete outcomes using cross section data*. Univ. N. C., Chapel Hill, Work. Pap.
Hagan J, Foster H. 2003. S/He's a rebel: toward a sequential stress theory of delinquency and gendered pathways to disadvantage in emerging adulthood. *Soc. Forces* 82:53–86
Hägglund G. 1982. Factor analysis by instrumental variables. *Psychometrika* 47:209–22
Hall AR. 2005. *Generalized Method of Moments*. Oxford: Oxford Univ. Press
Hall AR, Rudebusch GD, Wilcox DW. 1996. Judging instrument relevance in instrumental variables estimation. *Int. Econ. Rev.* 37:283–98
Hansen LP. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54
Hayashi F. 2000. *Econometrics*. Princeton, NJ: Princeton Univ. Press
Holtz-Eakin D, Newey W, Rosen HS. 1988. Estimating vector autoregressions with panel data. *Econometrica* 56:1371–95
Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75
Ingram P, Robinson J, Busch ML. 2005. The intergovernmental network of world trade: IGO connectedness, governance, and embeddedness. *Am. J. Sociol.* 111:824–58
Iwata S. 2001. Recentered and rescaled instrumental variable estimation of tobit and probit models with errors in variables. *Econom. Rev.* 20:319–35

Jenkins JC, Leicht KT, Wendt H. 2006. Subnational economic development policy in the United States, 1971–1990. *Am. J. Sociol.* 111:1122–80

Johnson LW. 1977. Stochastic parameter regression: an annotated bibliography. *Int. Stat. Rev.* 45:257–72

Johnson LW. 1980. Stochastic parameter regression: an additional annotated bibliography. *Int. Stat. Rev.* 48:95–102

Johnston J. 1984. *Econometric Methods*. New York: McGraw Hill

Jöreskog K, Sörbom D. 1993. *LISREL 8: Structural Equation Modeling*. Lincolnwood, IL: Sci. Softw. Int.

Kelejian HH. 1971. Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variable. *J. Am. Stat. Assoc.* 66:373–74

Kenny DA. 1979. *Correlation and Causality*. New York: Wiley

Kim DH, Schneider B. 2005. Social capital in action: alignment of parental support in adolescents' transition to postsecondary education. *Soc. Forces* 84:1181–206

Kirby JB, Bollen KA. 2009. Using instrumental variable tests to evaluate model specification in latent variable structural equation models. *Sociol. Methodol.* 39:327–55

Kirk DS. 2009. A natural experiment on residential change and recidivism: lessons from Hurricane Katrina. *Am. Sociol. Rev.* 74:484–505

Kiser E, Linton A. 2001. Determinants of the growth of the state: war and taxation in early modern France and England. *Soc. Forces* 80:411–48

Koçak Ö, Carroll GR. 2008. Growing church organizations in diverse U.S. communities, 1890–1926. *Am. J. Sociol.* 113:1272–315

Lee C-S. 2007. Labor unions and good governance: a cross-national, comparative analysis. *Am. Sociol. Rev.* 72:585–609

Lee L-F. 1981. Simultaneous equations models with discrete and censored variables. In *Structural Analysis of Discrete Data with Econometric Applications*, ed. CF Manski, D McFadden, pp. 346–64. Cambridge, MA: MIT Press

Lee MR, Shihadeh ES. 2009. The spatial concentration of Southern Whites and argument-based lethal violence. *Soc. Forces* 87:1671–94

Lizardo O. 2006. How cultural tastes shape personal networks. *Am. Sociol. Rev.* 71:778–807

Loftus J. 2001. America's liberalization in attitudes toward homosexuality, 1973 to 1998. *Am. Sociol. Rev.* 66:762–82

Long JS. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage

Ludwig J, Kling JR, Katz LF, Sanbonmatsu L, Liebman J, et al. 2008. What can we learn about neighborhood effects from the moving to opportunity experiment? *Am. J. Sociol.* 114:144–88

Madansky A. 1964. Instrumental variables in factor analysis. *Psychometrika* 29:105–13

Maddala GS. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge Univ. Press

Manski CF. 1988. *Analog Estimation Methods in Econometrics*. New York: Chapman & Hall

Mariano RS. 1982. Analytical small-sample distribution theory in econometrics: the simultaneous equations case. *Int. Econ. Rev.* 23:503–33

Mátyás L, ed. 1999. *Generalized Method of Moments Estimation*. Cambridge, UK: Cambridge Univ. Press

Mayer SE. 2001. How did the increase in economic inequality between 1970 and 1990 affect children's educational attainment? *Am. J. Sociol.* 107:1–32

McCarthy B, Casey T. 2008. Love, sex, and crime: adolescent romantic relationships and offending. *Am. Sociol. Rev.* 73:944–69

McDonald S, Elder GH Jr. 2006. When does social capital matter? Non-searching for jobs across the life course. *Soc. Forces* 85:521–49

McManus PA, DiPrete TA. 2001. Losers and winners: the financial consequences of separation and divorce for men. *Am. Sociol. Rev.* 66:246–68

McVeigh R, Myers DJ, Sikkink D. 2004. Corn, Klansmen, and Coolidge: structure and framing in social movements. *Soc. Forces* 83:653–90

Meier AM. 2003. Adolescents' transition to first intercourse, religiosity, and attitudes about sex. *Soc. Forces* 81:1031–52

Messner SF, Baumer EP, Rosenfeld R. 2004. Dimensions of social capital and rates of criminal homicide. *Am. Sociol. Rev.* 69:882–903

Morenoff JD. 2003. Neighborhood mechanisms and the spatial dynamics of birth weight. *Am. J. Sociol.* 108:976–1017

Morgan MS. 1990. *The History of Econometric Ideas*. Cambridge, UK: Cambridge Univ. Press

Morgan SL, Winship C. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge Univ. Press

Myers SM. 2000. The impact of religious involvement on migration. *Soc. Forces* 79:755–83

Nagar AL. 1959. The bias and moment matrix of the general *k*-class estimators of the parameters in simultaneous equations. *Econometrica* 27:575–95

Nelson CR, Startz R. 1990. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58:967–76

Noonan MC, Corcoran ME, Courant PN. 2005. Pay differences among the highly trained: cohort differences in the sex gap in lawyers' earnings. *Soc. Forces* 84:853–72

Offer S, Schneider B. 2007. Children's role in generating social capital. *Soc. Forces* 85:1125–42

Olsson U. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* 44:443–60

Olsson U, Drasgow F, Dorans NJ. 1982. The polyserial correlation coefficient. *Psychometrika* 47:337–47

Ono H. 2007. Careers in foreign-owned firms in Japan. *Am. Sociol. Rev.* 72:267–90

Parrado EA, Flippen CA. 2005. Migration and gender among Mexican women. *Am. Sociol. Rev.* 70:606–32

Paxton P, Hipp J, Marquart-Pyatt S. 2011. *Nonrecursive Models: Endogeneity, Reciprocal Relationships, and Feedback Loops*. Thousand Oaks, CA: Sage

Paxton P, Kunovich S. 2003. Women's political representation: the importance of ideology. *Soc. Forces* 82:87–113

Pearl J. 2000. The logic of counterfactuals in causal inference. *J. Am. Stat. Assoc.* 95:428–35

Pearl J. 2010. The foundations of causal inference. *Sociol. Methodol.* 40:75–149

Poskitt DS, Skeels CL. 2002. *Assessing instrumental variable relevance: an alternative measure and some exact finite sample theory.* Work. Pap. 862, Dep. Econ., Monash Univ./Univ. Melbourne

Qian Z, Lichter DT, Mellott LM. 2005. Out-of-wedlock childbearing, marital prospects and mate selection. *Soc. Forces* 84:473–91

Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. 2009. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *Am. J. Epidemiol.* 169:273–84

Regoeczi WC. 2002. The impact of density: the importance of nonlinearity and selection on flight and fight responses. *Soc. Forces* 81:505–30

Reiersøl O. 1941. Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9:1–24

Reiersøl O. 1945. Confluence analysis by means of instrumental sets of variables. *Ark. Math. Astron. Fys.* 32:1–119

Richardson DH. 1970. The asymptotic unbiasedness of two-stage least squares. *Econometrica* 38:772

Rivers D, Vuong QH. 1988. Limited information estimators and exogeneity tests for simultaneous probit models. *J. Econom.* 39:347–66

Rosenfeld R, Messner SF, Baumer EP. 2001. Social capital and homicide. *Soc. Forces* 80:283–310

Rosenzweig MR, Wolpin KI. 2000. Natural "natural experiments" in economics. *J. Econ. Lit.* 38:827–74

Ross CE, Mirowsky J, Pribesh S. 2001. Powerlessness and the amplification of threat: neighborhood disadvantage, disorder, and mistrust. *Am. Sociol. Rev.* 66:568–91

Rotolo T, McPherson JM. 2001. The system of occupations: modeling occupations in sociodemographic space. *Soc. Forces* 79:1095–130

Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701

Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58

Rubin H. 1950. Note on random coefficients. In *Statistical Inference in Dynamic Economic Models*, ed. TC Koopmans, pp. 419–21. New York: Wiley

Sanderson MR, Kentor JD. 2009. Globalization, development, and international migration: a cross-national analysis of less-developed countries, 1970–2000. *Soc. Forces* 88:301–36
Sargan JD. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26:393–415
Sawa T. 1969. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *J. Am. Stat. Assoc.* 64:923–37
Schram SF, Soss J, Fording RC, Houser L. 2009. Deciding to discipline: race, choice, and punishment at the frontlines of welfare reform. *Am. Sociol. Rev.* 74:398–422
Shea J. 1997. Instrument relevance in multivariate linear models: a simple measure. *Rev. Econ. Stat.* 79:348–52
Smith RJ, Blundell RW. 1986. An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica* 54:679–85
Sovey AJ, Green DP. 2011. Instrumental variables estimation in political science: a reader's guide. *Am. J. Polit. Sci.* 55:188–200
Splawa-Neyman J. (1923) 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* 5:465–80
Staiger D, Stock JH. 1997. Instrumental variable regression with weak instruments. *Econometrica* 65:557–86
Stock JH, Trebbi F. 2003. Who invented instrumental variable regression? *J. Econ. Perspect.* 17:177–94
Stock JH, Wright JH, Yogo M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econ. Stat.* 20:518–29
Stock JH, Yogo M. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. DW Andrews, JH Stock, pp. 90–108. New York: Cambridge Univ. Press
Swamy PAVB. 1970. Efficient inference in a random coefficient regression model. *Econometrica* 38:311–23
Theil H. 1958. *Economic Forecasts and Policy*. Amsterdam: North-Holland
Theil H. 1971. *Principles of Econometrics*. New York: Wiley
Tomaskovic-Devey D, Skaggs S. 2002. Sex segregation, labor process organization, and gender earnings inequality. *Am. J. Sociol.* 108:102–28
Uggen C, Thompson M. 2003. The socioeconomic determinants of ill-gotten gains: within-person changes in drug use and illegal earnings. *Am. J. Sociol.* 109:146–85
Villarreal A, Yu W-h. 2007. Economic globalization and women's employment: the case of manufacturing in Mexico. *Am. Sociol. Rev.* 72:365–89
Vuong QH. 1984. *Two-stage conditional maximum likelihood estimation of econometric models.* Soc. Sci. Work. Pap. No. 538, Calif. Inst. Technol.
Wilde J. 2008. A note on GMM estimation of probit models with endogenous regressors. *Stat. Pap.* 49:471–84
Winship C, Mare RD. 1992. Models for sample selection bias. *Annu. Rev. Sociol.* 18:327–50
Wooldridge JM. 1995. Score diagnostics for linear models estimated by two stage least squares. In *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, ed. GS Maddala, PCB Phillips, TN Srinivasan, pp. 66–87. Oxford: Blackwell
Wooldridge JM. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press
Wright PG. 1928. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan
Wright S. 1925. Corn and hog correlations. *US Dep. Agric. Bull.* 1300:1–60
Yatchew A, Griliches Z. 1985. Specification error in probit models. *Rev. Econ. Stat.* 67:134–39
You J-S, Khagram S. 2005. A comparative study of inequality and corruption. *Am. Sociol. Rev.* 70:136–57
Young C. 2009. Model uncertainty in sociological research: an application to religion and economic growth. *Am. Sociol. Rev.* 74:380–97
Zajac EJ, Westphal JD. 2004. The social construction of market value: institutionalization and learning perspectives on stock market reactions. *Am. Sociol. Rev.* 69:433–57
Zhang Y, Hannum E, Wang M. 2008. Gender-based employment and income differences in urban China: considering the contributions of marriage and parenthood. *Soc. Forces* 86:1529–60