DEPARTMENT OF ECONOMICS

The Nonexistence of Instrumental Variables

P. A. V. B. Swamy, Federal Reserve Board (Retired), USA

George S. Tavlas, Bank of Greece, Economics Research Department, Greece

Stephen G. Hall, University of Leicester, UK

Working Paper No. 09/16
September 2009

The Nonexistence of Instrumental Variables

P.A.V.B. Swamy [a], George S. Tavlas [b,*] and Stephen G. Hall [c,b]

[a] *Board of Governors of the Federal Reserve System (retired), Washington, DC, USA*

[b] *Bank of Greece, 21, El, Venizelos Ave, 102 50 Athens, Greece.* Tel.++30210 320 2370; Fax: ++30210 320 2432. Email: gtavlas@bankofgreece.gr

[c] *Department of Economics, University of Leicester*

Abstract

The method of instrumental variables (IV) and the generalized method of moments (GMM) and their applications to the estimation of errors-in-variables and simultaneous equations models in econometrics require data on a sufficient number of instrumental variables which are (insert space)both exogeneous and relevant. We argue that in general such instruments (weak or strong) cannot exist.

JEL classification: C32, C51

1. Introduction

Researchers are becoming increasingly aware that there are often serious problems with the use of instrumental variable based techniques (both instrumental variable (IV) estimation and versions of generalized methods of moments (GMM) which use instrumental variables). A valid instrument must be uncorrelated with the errors in an equation (exogeneous) and correlated with the explanatory variable (relevant), see Greene (2008, p. 316). The exogeneity condition is criticized in the statistics literature and the relevancy condition is criticized in the econometric literature. Pratt and Schlaifer (1988) point out that without knowing what the errors represent, it is not possible to decide whether or not the exogeneity condition is correct. They further point out that the condition is meaningless if the errors are included in an equation to represent the

* Corresponding author.

variables excluded from the equation. Increasingly, econometricians are finding that when a set of instruments are independent of the error they often have little relevance; this is the problem of 'weak instruments'. In this paper we argue that this should not be a surprising result and that in general it is not possible to find valid instruments. The next section presents a proof of this statement.

2. A General Representation of Misspecification

In general, economic theory suggests relationships between variables, but it does not usually give clear guidance as to the correct functional form or the complete set of variables that are relevant. For example, consider an economic variable, denoted by y_t^*, and its complete set of determinants, denoted by x_{jt}^*, $j = 1, \ldots, L_t$. Here the total number L_t of determinants may be time dependent. Typically, data on y_t^* and on a subset K - 1 of the L_t determinants are available. The remaining L_t - K + 1 determinants are omitted from the model either because they are unobserved or for some other reason. Moreover, these data may contain measurement errors. Let $y_t = y_t^* + v_{0t}$ and $x_{jt} = x_{jt}^* + v_{jt}$, $j = 1, \ldots, K-1$, where the variables without an asterisk are observable, the variables with an asterisk are unobservable, and vs are measurement errors. The theoretical relationship is

$$y_t^* = f_t(x_{1t}^*, ..., x_{L_t t}^*) \qquad (t = 1, \ldots, \mathrm{T}) \tag{1}$$

with unknown functional form.

Without misspecifying the relationship in (1), we can write

$$y_t^* = \alpha_{0t} + \sum_{j=1}^{K-1} \alpha_{jt} x_{jt}^* + \sum_{g=K}^{L_t} \alpha_{gt} x_{gt}^* \tag{2}$$

where the time profiles of the coefficients are determined by the correct functional form of

model (1). These time profiles are unknown, since the correct functional form is unknown. Allowing the coefficients of equation (2) to vary freely defines an infinite class of functional forms, which surely encompasses the correct (but unknown) functional form of (1) as a special case. If spline-, cubic-spline-, P-spline-, or any other-type restrictions are *imposed* on the functional form of model (1), then it can have an incorrect functional form; for examples of spline- and cubic-spline-type restrictions, see Greene (2008, p. 111) and Judge, Griffiths, Hill, Lütkepohl and Lee (1985, p. 803). A main benefit of model (2) is the certainty that the infinite class of functional forms will encompass the correct functional form.

Clearly, the explanatory variables of (2) can be correlated with each other, leading to the well-known problem of multicollinearity. In particular, the $K - 1$ observable determinants (the x_{jt}^*'s) in equation (2) can be correlated with the $L_t - K + 1$ unobserved determinants (the x_{gt}^*'s). To assume otherwise would, in the words of Pratt and Schlaifer (1988), be a "meaningless" assumption. The correlations between the omitted determinants and the observed determinants are implied by

$$x_{gt}^* = \lambda_{0gt} + \sum_{j=1}^{K-1} \lambda_{jgt} x_{jt}^* \qquad (g = K, \ldots, L_t) \tag{3}$$

where λ_{0gt} is a portion of x_{gt}^* remaining after the effects of the x_{jt}^*'s have been removed from x_{gt}^*. Since we do not have data on the $L_t - K + 1$ x_{gt}^* variables, we can eliminate them from equation (2) by substituting (3) into (2), which gives

$$y_t^* = \alpha_{0t} + \sum_{g=K}^{L_t} \alpha_{gt} \lambda_{0gt} + \sum_{j=1}^{K-1} (\alpha_{jt} + \sum_{g=K}^{L_t} \alpha_{gt} \lambda_{jgt}) x_{jt}^* \tag{4}$$

Note that equation (4) shows y_t^* as a function of $K - 1$ included determinants and the reminders of the excluded variables - - i.e., what remains after subtracting the effects on the

excluded variables of the $K - 1$ observable determinants. Equation (4) accounts for both the unknown functional form (since it is derived from equation (2)) and the full set of (time-varying) determinants of y_t^*. It does not, however, account for measurement errors. In this connection, consider model (2) again. It is not in a form that can be estimated. Such a form is derived below.

In terms of the observable variables, equation (2) can be written as

$$y_t = \gamma_{0t} + \sum_{j=1}^{K-1} \gamma_{jt} x_{jt} \tag{5}$$

We call the x_{gt}^*'s "excluded variables" because they are excluded from model (5). The x_{jt}'s are the included explanatory variables. Model (5) coincides with model (2) if

$$\gamma_{0t} = \alpha_{0t} + \sum_{g=K}^{L_t} \alpha_{gt} \lambda_{0gt} + v_{0t} \tag{6}$$

$$\gamma_{jt} = (\alpha_{jt} + \sum_{g=K}^{L_t} \alpha_{gt} \lambda_{jgt})(1 - \frac{v_{jt}}{x_{jt}}) \quad (j = 1, \ldots, K\text{-}1) \tag{7}$$

These equations are derived by establishing the correspondence between equations (4) and (5).[1]

The terms on the right-hand side of equations (6) and (7) provide crucial information. Equation (4) shows that the λ_{0gt}'s, in conjunction with the x_{jt}^*'s, are at least sufficient to determine y_t^*. This is the proof Pratt and Schlaifer (1988, pp. 34 and 50) offer to show that the second term on the right-hand side of equation (6) is a 'sufficient set' of excluded variables; it should be noted that one of the conditions of this proof is that the functional form of model (1) is not misspecified. Pratt and Schlaifer (1988) also show that the condition, $E(\sum_{g=K}^{L_t} \alpha_{gt} \lambda_{0gt} \mid x_{1t}^*, \ldots, x_{K-1,t}^*) = 0$ is meaningful, but the condition that the x_{jt}^*'s be independent of the x_{gt}^*'s themselves is meaningless. They warn against adding an arbitrary error term to a

[1] For the derivation, see Swamy and Tavlas (2007).

linear or nonlinear function of the x_{jt}^*'s and assuming that the x_{jt}^*'s are independent of the error term.

The interpretation of the terms on the right-hand side of equation (7) and their implications are as follows:

- The term α_{jt} is equal to $\partial y_t^* / \partial x_{jt}^*$ (if y_t^* is a continuous function of x_{jt}^*) and corresponds to the bias-free effect of x_{jt}^* on y_t^*, as can be seen from (2). The right sign of α_{jt} is provided by economic theories. The correlation between y_t^* and x_{jt}^* is spurious if $\alpha_{jt} = 0$.
- The term $\sum_{g=K}^{L_t} \alpha_{gt} \lambda_{jgt}$ measures omitted-variables bias. Note that each term in this sum is the product of two coefficients - - the effect of the excluded variable x_{gt}^* on y_t^* (i.e., α_{gt}) and the effect of the included variable x_{jt}^* on the excluded variable x_{gt}^* (i.e., λ_{jgt}). Omitted-variable biases can exist as long as the error terms are present in econometric models.
- The term $(\alpha_{jt} + \sum_{g=K}^{L_t} \alpha_{gt} \lambda_{jgt})(-(v_{jt} / x_{jt}))$ measures measurement-errors bias.[2] These biases exist whenever estimates of some theoretical variables are used as explanatory variables.
- The explanatory variables of model (5) are correlated with their own coefficients because the measurement-error bias component of γ_{jt} is a function of x_{jt}.

[2] The minus sign in the expression reflects the fact that the second parenthetical term on the right-hand side of (7) is one minus the ratio (v_{jt} / x_{jt}).

- Model (5) can be misspecified if the omitted-variable and measurement-error bias (or simply, the specification bias) components of its coefficients in (7) are ignored.

Further discussion of the terms in (7) is given in Hondroyiannis, Swamy and Tavlas (2009).

Having derived the model in (5), which explicitly includes all these forms of misspecification, it is now possible to show why valid instruments cannot be found for this model. Under IV or GMM, we are imposing constant parameters on (5). We can, therefore, re-write (5) as;

$$y_t = \alpha_{0t} + \sum_{g=K}^{L_t} \alpha_{gt}\lambda_{0gt} + \nu_{0t} + \sum_{j=1}^{K-1}(\alpha_{jt} + \sum\nolimits_{g=K}^{L_t} \alpha_{gt}\lambda_{jgt})(1-\frac{\nu_{jt}}{x_{jt}})x_{jt} \tag{8}$$

Now we can illustrate the problem with IV by considering 3 cases.

Case I. (Linear models) By adding and subtracting a constant parameter model we get

$$y_t = \beta_0 + \sum_{j=1}^{K-1}\beta_j x_{jt} + (\alpha_{0t} + \sum_{g=K}^{L_t} \alpha_{gt}\lambda_{0gt} + \nu_{0t} - \beta_0) + \sum_{j=1}^{K-1}((\alpha_{jt} + \sum\nolimits_{g=k}^{L_t} \alpha_{gt}\lambda_{jgt})(1-\frac{\nu_{jt}}{x_{jt}}) - \beta_j)x_{jt} \tag{9}$$

Where the last two terms in (9) becomes the error term in the model. The problem with instrumental variables in this context now becomes apparent; we need to find a variable that is both correlated with x_{jt} , but not correlated with the error term, which itself contains x_{jt}. Such a variable cannot exist. We extend this proof to nonlinear models in Case III below.

Case II. (Linear errors-in-variables model without the error in equation) If

$$\lambda_{0gt} = \lambda_{jgt} = 0 \ \ for\, all\ j, g\ and\ t \tag{10}$$

and

$$\beta_j = \alpha_{jt} \ for\ j = 0,...,K-1 \tag{11}$$

equation (10) implies that there are no omitted variables and (11) implies that the true model has a linear functional form, Under (10) and (11), (9) reduces to an errors-in-variables model and the error term becomes just $\nu_{jt}, j = 0...K-1$. For IV estimation of such a model, we need instruments that are relevant and uncorrelated with the errors (exogenous), see Greene (2008, pp. 327-329). One of the problems here is that the relevancy condition can never be verified because the x_{jt}^{*}'s are rarely if ever observed and assumptions (10) and (11) are highly restrictive.

CaseIII. (Nonlinear models) Note that Cases I and II do not cover nonlinear models. To complete our proof of the nonexistence of valid instruments in Cases I and II, we need to consider the realistic nonlinear case where model (5) with its coefficients satisfying equations (6) and (7) holds. A natural method of identifying the coefficients of model (5) without misspecifying its functional form is to decompose these coefficients into their respective components in (6) and (7). To perform this decomposition, we assume that

$$\gamma_{jt} = \pi_{j0} + \sum_{h=1}^{p-1} \pi_{jh} z_{ht} + \varepsilon_{jt} \quad (j = 0, 1, \ldots, K\text{-}1) \tag{12}$$

where the z_{ht}'s are observable, $E(\varepsilon_{jt} \mid z_{1t},...,z_{p-1,t}) = 0, j = 0, 1, \ldots, K-1$, all *t*, and the ε_{jt}'s may be serially and contemporaneously correlated. It is assumed that in model (5), the x_{jt}'s are *conditionally* independent of their own coefficients given the z_{ht}'s. Changes in policy variables, shift variables representing structural changes in the γ_{jt} and lagged changes in the x_{jt}'s can be used as the z_{ht}'s, as in Hall, Hondroyiannis, Swamy and Tavlas (2009) and Hondroyiannis, Swamy and Tavlas (2009).

We cannot be sure that the equation obtained by substituting equation (12) into equation (5) will have the correct functional form. The only way we can be so sure is by letting *p* tend to infinity so that ε_{jt} converges in probability to zero. It is possible to push ε_{jt} as low as desired with a high probability just by adding additional z_{jt}'s on the right-hand side of equation (12); it does not matter if some of the z_{jt}'s are redundant in the sense that their coefficients in (12) are zero. Equation (12) with infinitely large *p* and without ε_{jt} can completely explain all the variation in γ_{jt} in terms of observable variables. Substituting such an equation into (5) gives an equation with the correct functional form.

Inserting equation (12) into equation (5) gives

$$y_t = \pi_{00} + \sum_{h=1}^{p-1} \pi_{0h} z_{ht} + \sum_{j=1}^{K-1} (\pi_{j0} + \sum_{h=1}^{p-1} \pi_{jh} z_{ht}) x_{jt} + \varepsilon_{0t} + \sum_{j=1}^{K-1} \varepsilon_{jt} x_{jt} \qquad (13)$$

This is an estimable form of model (5). The instrumental variables that are correlated with the x_{jt}'s of model (5) but not with the error terms of model (13) do not exist because these error terms also involve the x_{jt}'s. Therefore, IV estimation of model (13) is not possible. Sometimes it is claimed that in many time-series settings, lagged values of the variables in a model provide natural instrumental variables. The mere fact that the value of $x_{j,t-1}$ was determined before the value of ε_{jt} should not lead one to conclude that $x_{j,t-1}$ is necessarily independent of ε_{jt}. The variable $x_{j,t-1}$ may well have been influenced by a forecast of a variable represented in ε_{jt} or both $x_{j,t-1}$ and ε_{jt} may have been affected by some third variable, as shown by Pratt and Schlaifer (1988, p. 47). Of course even if $x_{j,t-1}$ were independent of the error then this would imply that it was no longer relevant.

3. Conclusion

The instrumental variables that are correlated with the x_{jt}'s of model (5), but not with the error terms of model (13), do not in general exist because these error terms also involve the x_{jt}'s. . These arguments make it clear why practical work with IV methods is plagued by several problems. We would argue that a much better way forward in terms of practical estimation rests on recognition of all the potential sources of misspecification which are present in (5) and starts from a time-varying coefficient model as outlined in Swamy and Tavlas (2001).

Acknowledgments

The views expressed in this paper are the authors' own and are not to be interpreted as those of their respective past or present institutions.

References

Greene, W.H., 2008, Econometric analysis, 6th edn. (Pearson/Prentice Hall, Upper Saddle River, New Jersey).

Hall, S.G., G. Hondroyiannis, P.A.V.B. Swamy and G.S. Tavlas, 2009, The new Keynesian Phillips curve and lagged inflation: A case of spurious correlation? Southern Economic Journal, forthcoming.

Hondroyiannis, G., P.A.V.B. Swamy, and G.S. Tavlas, 2009, The new Keynesian Phillips curve in a time-varying coefficient environment: Some European evidence, Macroeconomic Dynamics 13, 149-166.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T. Lee, 1985, The theory and practice of econometrics, 2nd edn. (John Wiley and Sons, New York).

Pratt, J.W. and R. Schlaifer, 1988, On the interpretation and observation of laws, Journal of Econometrics 39, 23-52.

Swamy P.A.V.B. and G.S.Tavlas, 2001, Random Coefficient Models, Ch 19. In Baltagi B.H.(ed.) A Companion to Theoretical Econometrics, Malden, Blackwell.

Swamy, P.A.V.B. and G.S. Tavlas, 2007, The new Keynesian Phillips curve and inflation expectations: Re-specification and interpretations, Economic Theory 31, 293-306.