ELSEVIER

Journal of Econometrics 83 (1998) 213–237

JOURNAL OF Econometrics

Model specification and endogeneity

Alice Nakamura[a,*], Masao Nakamura[b]

[a] *Faculty of Business, University of Alberta, Edmonton, Alb., Canada T6G 2R6*
[b] *Faculty of Commerce, University of British Columbia, Vancouver, BC, Canada V6T 1Z2*

Received 1 March 1995; received in revised form 1 June 1996

Abstract

This paper considers the treatment of endogenous explanatory variables in the work of the Cowles Commission and in Carl Christ's classic 1966 textbook, and certain problems that arise when this approach is followed in areas such as the study of female labor supply where a priori knowledge is sparse or uncertain. The motivations for, and evidence against the use of, mixed estimation approaches involving exogeneity pretests are explored. The paper concludes with a consideration of complementary and alternative empirical approaches, including greater use of predictive evaluation as suggested by Christ. © 1998 Elsevier Science S.A.

Keywords: Endogeneity; endogeneity testing; Instrumental variables; Female labor supply
JEL classification: C14; C30

1. Introduction

The Cowles Commission for Research in Economics, in its Chicago era, helped to put the concept of endogeneity into place theoretically as a corner-stone

* Corresponding author.

This paper was prepared for 'Macroeconometrics and Econometric Methods: A Conference in Honor of Professor Carl F. Christ', The Johns Hopkins University, 21–22 April 1995. The research was partially supported by grants from the Social Sciences and Humanities Research Council of Canada. The authors thank Don Andrews, Patrick Asea, Erwin Diewert, R.W. Farebrother, Lawrence Klein, Jan Kmenta, Hiroki Tsurumi, and particularly John Cragg as well as an anonymous referee for helpful comments on earlier versions of various parts of this paper. We also thank Ken Nakamura, a medical student, for help with the section on research methods in epidemiology. All remaining errors and misinterpretations are our responsibility.

PII S0304-4076(97)00070-5

of econometric identification, estimation, and inference.[1] Carl F. Christ deserves considerable credit for the fact that the Cowles Commission approach to these problems became widely accepted. His classic 1966 textbook, *Econometric Models and Methods*, made these advances more accessible to interested researchers and subsequent generations of graduate students including our own.

Christ's motivation, and that of the Cowles Commission, for defining and drawing out the consequences of the endogeneity of variables in econometric models was to further behavioral understanding. This paper examines how an emphasis on endogeneity has spurred the development and use of instrumental variables estimation and endogeneity tests in areas such as labor economics, and how concerns about endogeneity are (and are not) helping to improve behavioral understanding.

In Sections 2 and 3, we review certain basics of why the concept of endogeneity is important, and distinguish two types of operationally different endogenous variables: those of primary interest in the context of a particular research project, and endogenous control variables. Section 4 presents an introduction to endogeneity testing from the applied perspective of research on female labor supply. Section 5 examines formal statistical properties of the endogeneity tests introduced in Section 4. Section 6 looks back at the original motivations spelled out by Christ for designating some variables as endogenous, and considers possible costs of a preoccupation with endogeneity testing and the instrumentation of right hand variables suspected of being endogenous. Alternative and supplementary research strategies are considered. Section 7 concludes.

2. Endogeneity in traditional simultaneous equations models

Economics models in consumer demand analysis and many areas of macroeconomics have traditionally involved multiple dependent variables theorized to be causally and simultaneously interrelated. This is the context within which concerns about identification and endogeneity evolved, stimulating the development of simultaneous equations estimation methods. It is within this context that Christ introduces the concept of endogeneity. Following the approach of the Cowles Commission (Koopmans and Hood, 1953, pp. 117–120), Christ defines an endogenous variable in a linear simultaneous equations model as the complement of an exogenous variable:

> *An exogenous variable in a stochastic model is a variable whose value in each period is statistically independent of the values of all the random disturbances in the model in all periods.* ... Exogenous variables may be random, or may be

[1] For a retrospective overview of the Cowles Commission's contributions to econometric methods and practice, see Christ (1994).

deliberately set by some agency, as by government. Variables that are not exogenous are *endogenous.* (Christ 1966, pp. 156–157)

In Christ's treatment of endogeneity, two context-specific types of endogenous variables are lumped together under this label.[2] We define and discuss the first in this section and the second in the second subsection of the following section. The first type are the variables that are the direct focus of research interest. Christ (1966, pp. 12, 13) provides an example:

> ... imagine a theory that claims to be able to make a conditional prediction of national income for next year, given the magnitudes for next year of investment, government purchases, and tax receipts. Then in such a theory national income is an *endogenous variable.*...

We will refer to endogenous variables for which equations must be developed to fulfill primary research objectives as *primary endogenous variables.* In Christ's textbook, as in the work of the Cowles Commission, situations are contemplated in which there is *true* a priori information on the interrelationships among the primary endogenous variables of a model.

Single equation coefficient estimates for the directly included endogenous variables will pick up not only the direct effects of these variables but also spurious effects due to the correlations of these variables with model disturbance terms. This is the *endogeneity bias problem.* When equations with directly included endogenous explanatory variables are estimated by ordinary least squares (OLS) – the context in which concerns about endogeneity bias problems were first raised – the resulting endogeneity problem is also called the *OLS bias problem* (Christ, 1966, pp. 453–464). The use of IV or some other simultaneous equations estimation method is how researchers often try to deal with these bias problems.[3]

The main acknowledged cost of using a simultaneous equations estimation method is a *loss of efficiency.* The IV approach involves separating a variable suspected of being endogenous, such as the wage variable in a labor supply model, into the portion explained by an auxiliary IV equation and the anxiliary regression residuals. Suppose the objective is to estimate β_w, the coefficient of the wage variable, w, in a labor supply equation. If e_w denotes the residuals from an auxiliary instrumental wage equation, the loss of efficiency in estimating β_w comes from the addition of $\beta_w e_w$ to the error term for the labor supply equation when the predicted wage variable from the auxiliary equation, $\hat{w}$, is substituted

---

[2] Papers by Geweke (1987) led us to think about this distinction.

[3] Although it is rarely mentioned in studies using instrumental variables, in finite samples, IV estimates can also be biased even when the instruments are exogenously determined. See Pagan and Jung (1993), Angrist and Krueger (1995) and Buse (1992).

for w in the labor supply equation. When the R^2 for the auxiliary wage equation is low, this loss of efficiency can be substantial.

Nevertheless, with primary endogenous variables in situations where there is appropriate a priori information and sufficient data, basic research objectives and the statistical advantage of avoiding or lessening endogeneity bias problems both point in the direction of using a simultaneous equations estimation method such as IV.

3. Dealing with endogeneity outside the textbook context

3.1. Interrelated primary endogenous variables

In modern empirical research, many of the models used resemble the traditional textbook simultaneous equations models in that they involve sets of interrelated primary endogenous variables. For instance, models of female labor supply are often specified to involve equations for both the wages women receive and their annual hours of work, with hours of work assumed to depend on the wage variable. However, in this and many other areas of applied economic research there is little in the way of strongly held a priori knowledge.[4] As a consequence, there is considerable disagreement concerning the choice of variables to be included in the equations to be estimated. Low R^2 values for the auxiliary IV equations for variables believed to be endogenous are common, and the associated efficiency losses can be large.

Moreover, in areas where a priori knowledge is sparse or uncertain, two other potential costs of instrumentation can be important. The first is that some of the variables included on the right hand side of the auxiliary IV equations may not be exogenous either. This can result in *endogeneity bias problems even after instrumentation*. In fact, an endogeneity bias problem can be worsened by instrumentation if the explanatory variables included in an auxiliary IV equation pick up components of variation in a variable suspected of being endogenous that are, in fact, correlated with the true equation disturbance term and account for very little of the truly exogenous variation in this variable. An example may help clarify this point.

Consider an office situation in which the salaries of the secretaries differ depending on labor supply related attributes such as whether they work full time or part time and their seniority, as well as on whether the personnel director happened to be there at the time of the initial hiring or the acting personnel director was in charge who makes systematically different decisions about starting wage rates. The secretary-to-secretary differences in who did the hiring are a source of purely exogenous wage variation, from a worker perspective, that

[4] There are other important problems we do not deal with including sample selectivity.

would not be captured by any of the variables commonly included in auxiliary IV wage equations. Rather, with an IV approach this exogenous wage variation would probably end up as part of the residual wage contribution to the error term for the labor supply equation. On the other hand, some truly endogenous components of wage variation may well be picked up by schooling variables in an instrumental wage equation, because of the effects of unobserved ability and taste factors. This, in turn, could cause endogeneity bias problems despite the use of IV estimation.

A second concern when there is little a priori knowledge is that instrumentation will result in a *loss of relevance* because the auxiliary equations fail to capture the types of variation in the right-hand side endogenous variables that are important from an applications perspective. This concern is discussed more fully in Section 6.1. This worry, coupled with concerns that the available instruments may not be truly exogenous and with the knowledge that there will always be efficiency losses from instrumentation, has motivated some researchers including ourselves to seek ways of verifying the *existence*, or of determining the *seriousness*, of potential endogeneity problems before deciding on whether or not to use IV estimation. More specifically, these concerns have led some researchers to be interested in a pretest estimation strategy, where the choice of whether to stick with single equation estimation results or to use IV or some other simultaneous equations estimation method would depend on empirical pretest evidence concerning the existence or the likely severity of the suspected endogeneity bias problems.

3.2. *Endogenous control variables*

In the work of the Cowles Commission and in Christ's textbook, the terms exogenous and endogenous are given rigorous statistical definitions. There can be problems of endogeneity bias whenever included explanatory variables fail to satisfy the statistical definition of exogeneity. We have discussed the wage variable in models of female labor supply as an example of a primary endogenous variable for which a behavioral, or at least a forecasting, equation is needed to meet the stated research objectives as well as to deal with possible endogeneity bias problems. However, in many applied research areas, many of the explanatory variables suspected of being endogenous are *not* of direct research interest.

Explanatory variables suspected of being endogenous but which are not of direct interest in the given research context will be termed *endogenous control variables*. They are included to control for effects that might otherwise obscure the behavioral responses of prime interest.

4. Evolution of endogeneity testing in labor economics

In this section we trace the evolution of the main endogeneity tests that have been used in cross-sectional and panel data labor economics

studies.[5] Interest in endogeneity pretests in research on female labor supply was stimulated by frustration with the quality of the instrumental equations for the explanatory variables suspected of being endogenous.

For example, in a series of papers, Nakamura and Nakamura produced labor supply wage elasticity estimates for married women that differed greatly from the published results of others for women. The R^2 values for the auxiliary wage equations in the Nakamura and Nakamura studies rarely exceeded 0.34 and were mostly in a range of 0.06 to 0.14. This raised questions about whether the weak auxiliary wage equations were the cause of the unusual elasticity estimates and whether the *extent* of the endogeneity problems for the wage variable justified the inevitable losses in efficiency due to the substitution of poor instrumental wage variables.[6]

---

[5] There are other approaches to endogeneity testing besides the sort that has found application in labor economics, including methods for which it is necessary to have observations over time. Also, Reynolds (1982) takes a Baysian posterior odds approach to the problem of endogeneity testing. Discussions and results concerning alternative possible definitions of endogeneity, causality, and identifiability can be found in Simon (1953), in Zellner (1984) and in Cragg and Donald (1993).

[6] Killingsworth and Heckman (1986, pp. 134–135) sum up the theoretical arguments that were used to explain why the female uncompensated wage elasticity of labor supply might be expected to be considerably more positive than the male elasticity:

> The first step is to apply to commodity demands the discussions of input demand of Hicks (1965, pp. 242–246), Marshall (1920, pp. 386, 852–853), and Pigou (1946, p. 682): the elasticity of demand for a good (in this case, leisure) with respect to its price (in this case, the wage rate) will be greater, the greater is the availability of alternatives to that good. The next step (Mincer, 1962) is to observe that women in effect have more alternative uses for their time – market work, home work and leisure – than do men, who for the most part divide their time between only two uses, market work and leisure. In other words, the substitution towards market work that men undertake when their wage rises is primarily a substitution away from leisure, whereas a wage increase leads women to substitute away from both leisure and home work.

On our elasticity estimates, Killingsworth and Heckman (1986, pp. 185, 193) write:

> The main exception to these generalizations concerns the results of studies of US and Canadian data by Nakamura and Nakamura (1981b), Nakamura, Nakamura and Cullen (1979) and Robinson and Tomes (1985). Here, the uncompensated elasticity of labor supply with respect to wages is negative. ... It is tempting simply to dismiss such results as mere anomalies. ...

Commenting on the evolution of this empirical literature and a later 1981 survey article by Heckman, Killingsworth and MaCurdy, Berndt (1991, pp. 634–634) writes:

> Not all labor econometricians agree with Heckman, Killingsworth, and MaCurdy's assessment. In particular, already in the late 1970s, Alice Nakamura and Masao Nakamura reported results of a second-generation study that found female labor supply to be basically unresponsive to changes in wage rates, similar to the findings reported by others for males.... Additional supporting findings were presented later in Nakamura and Nakamura (1981b, 1985a,b). This controversy about whether males and females respond differently to wage rate changes is of considerable interest....

Doubts were also raised about whether the instrumental wage variables were capturing the *relevant* wage variations, from the perspective of labor supply behavior. The explanatory variable that accounted for most of the explained variation in the auxiliary wage equations was years of schooling: a variable that remains fixed in value for most adults over long numbers of years. Doubts were raised also as to whether the education and certain other variables in the auxiliary wage equations were truly exogenous. It is problems of these sorts that stimulated the interest of applied labor economists in endogeneity tests.

In 1973 and 1974 papers, Wu presented tests for endogeneity based on his test statistics T_1, T_2, T_3 and T_4. Hausman (1978) presented another endogeneity test – one that appeared to be far more convenient to implement than those of Wu. Nakamura and Nakamura (1981a) showed that the OLS-IV specialization of Hausman's endogeneity test approach is identical to Wu's T_2 test for the linear class of models specified in Wu's (1973, 1974) papers.[7] Nakamura and Nakamura also showed the exact relationships of the Hausman and Wu tests to the endogeneity tests of Durbin (1954), Revankar and Hartley (1973), and Revankar (1978).

Basic statistical properties of the endogeneity tests cited above are explored in Section 5. This material is important because these endogeneity tests are still widely used, and some findings that are based on them are important. For example, assertions that variables for past work experience must be instrumented in models for female labor supply are usually backed up by citations to an influential paper by Mroz (1987) that makes extensive use of Wu–Hausman type tests.[8] Mroz's objectives in his paper were to address endogeneity concerns raised by others, and to try to narrow the wide range of estimates in the published literature for the income and the substitution effects of the wage variable on the labor supply of married women. Mroz explains:

> Everyone familiar with the past ten years' research on empirical models of female labor supply is aware of the wide range of estimated income and substitution effects. ... The estimates presented ... demonstrate the sensitivity of the wage and income coefficients to minor variations in the variables used to instrument the wage rate. ... Notice that estimates using the set of instrumental variables with the wife's market experience. ... yield larger wage responses than the rows without this set of instruments. ... This suggests a possible specification error. To test for such errors, we apply variants of the specification tests proposed by Durbin (1954), Wu (1973), Hausman (1978), and White (1982). (Mroz 1987, pp. 765–773)

---

[7] Much of the proof in Nakamura and Nakamura (1981a) was provided in an anonymous *Econometrica* referee report as a replacement for a more cumbersome proof in the original paper. We are grateful now to be able to acknowledge this help from Adrian Pagan.

[8] For specifics of his endogeneity tests, see Mroz (1987, pp. 773–774, 796–798).

Methodological differences in econometrics are of much greater interest when the differences in perspective lead to important differences in reported applied findings. This is certainly the case with the use of endogeneity tests in research on female labor supply!

5. The properties of the Wu–Hausman endogeneity test

To understand whether endogeneity testing provides a general solution to the problem of deciding when to instrument right-hand variables suspected of being endogenous, it is necessary to examine the formal statistical properties of the tests.

As already noted, Wu (1973, 1974) presents the endogeneity test statistics T_1, T_2, T_3 and T_4. He finds his T_2 test to be the best. Hausman (1978) presents two asymptotically equivalent statistics which differ only in that one uses an IV estimator of the standard error of the regression while the other uses the estimator obtained from application of OLS. Hausman establishes most of his theoretical results for the former of these. We have shown that this statistic of Hausman's, which is Durbin's (1954) statistic, is identical to Wu's T_3 or T_4 depending on specifics of the estimator for the standard error of the regression. The second of Hausman's statistics (Hausman 1978, p. 1259) is the one that has been used most in applied studies, and that we have shown to be identical to Wu's T_2. We refer to it as the *Wu–Hausman statistic* since it was Hausman who presented it in a form that led to its widespread use but Wu who presented it first.

Both Wu (1973) and Hausman (1978) examine the asymptotic, local power properties of their various test statistics, where by local what is meant is that the departure from the null hypothesis tends to zero as the sample size (n) goes to infinity. However, the alternatives that are relevant in applied settings are rarely local in nature. More relevant comparisons of the test statistics can be carried out using the exact distributional results of Kariya and Hodoshima.

Under the assumption of normal disturbance terms, Kariya and Hodoshima (1980) derive the exact conditional distributions of the Wu–Hausman statistic and of the closely related Revankar (1978) statistic. They show that both these test statistics obey noncentral conditional F distributions. The distribution for the Revankar statistic, RV, is characterized by a single noncentral parameter, δ_1. The Wu–Hausman statistic, T_2, is shown to obey a doubly noncentral F distribution with the noncentral parameters δ_1, as for the distribution of the Revankar statistic, and δ_2. One hindrance to using these results of Kariya and Hodoshima is that their expressions for δ_1 and δ_2 are difficult to interpret. Building on their results, in the following subsection and the appendix we derive alternative expressions that can be more easily understood. We use these alternative expressions to examine properties of the Wu–Hausman and related endogeneity tests.

5.1. The conditional distributions of the Wu–Hausman and Revankar statistics

Consider the general linear two-equation model with additive disturbances given by[9]

$$y_1 = \alpha_1 y_2 + Z_1\alpha_2 + \lambda_1 u_1 \tag{1a}$$

and

$$y_2 = Z_1\beta_1 + Z_2\beta_2 + v_2, \tag{1b}$$

where y_1 and y_2 are $(n \times 1)$ vectors of observations on two endogenous variables; Z_1 and Z_2 are $(n \times K_1)$ and $(n \times K_2)$ matrices of observations on K_1 exogenous variables included and K_2 exogenous variables excluded from the structural equation (1a) for the primary endogenous variable y_1; $v_2 = \lambda_2 u_1 + \lambda_3 u_2$ where u_1 and u_2 are $(n \times 1)$ vectors of random disturbances which are independently normally distributed with means of 0 and variances of 1; and α_1, α_2, β_1, β_2, λ_1, λ_2 and λ_3 are unknown parameters or parameter vectors of appropriate dimensions. The population correlation between $\lambda_1 u_1$ and v_2 is $\rho = [\lambda_2^2/(\lambda_2^2 + \lambda_3^2)]^{1/2}$.

Let the reduced form equation for y_1 be

$$y_1 = Z_1\Pi_1 + Z_2\Pi_2 + v_1, \tag{2}$$

where $v_1 = \lambda_1 u_1 + \alpha_1 v_2$. Then we have: $\Omega_{11} = \mathrm{Var}(v_1) = \lambda_1^2 + 2\lambda_1\lambda_2\alpha_1 + \alpha_1^2(\lambda_2^2 + \lambda_3^2)$, $\Omega_{22} = \mathrm{Var}(v_2) = \lambda_2^2 + \lambda_3^2$, $\Omega_{12} = \Omega_{21} = \mathrm{Cov}(v_1, v_2) = \lambda_1\lambda_2 + \alpha_1(\lambda_2^2 + \lambda_3^2)$, and $\Omega_{11.2} = \{\Omega_{22}\Omega_{11} - (\Omega_{12})^2\}/\Omega_{22} = \lambda_1^2\lambda_3^2/(\lambda_2^2 + \lambda_3^2)$. We denote by S_{11}, S_{12}, S_{22}, $\hat{\beta}_2$ and $\hat{\Pi}_2$ the OLS estimators of Ω_{11}, Ω_{12}, Ω_{22}, β_2 and Π_2, respectively. The number of included endogenous explanatory variables in the structural equation (1a) for y_1 is $G_2 = 1$. Eq. (1b) is a reduced form instrumental equation for y_2. The variable y_2 could be either a second primary endogenous or an endogenous control variable. The structural equation (1a) for y_1 is assumed to be identified. (The variance of the IV or two-stage least squares estimator for α_1 in (1a) exists if $K_2 - 1 \geqslant 2$.)

The Wu–Hausman and related endogeneity tests can be viewed as tests of either H_0: $\mathrm{Cov}(\lambda_1 u_1, v_2) = 0$ or of H_0': $\rho = 0$.[10] $\mathrm{Cov}(\lambda_1 u_1, v_2)$ is the population covariance between $\lambda_1 u_1$, the disturbance term of the structural equation for y_1, and v_2, the disturbance term of the auxiliary reduced form equation for y_2. This covariance is the numerator of ρ, the population correlation between $\lambda_1 u_1$ and v_2. So ρ will be zero if and only if the covariance is zero.

For the given model, the Wu–Hausman statistic may be written as[11]

$$T_2 = c_1(Q^*/Q_2) \tag{3}$$

---

[9] It is possible as well to prove these results for a general multi-equation system.

[10] See Nakamura and Nakamura (1985c, pp. 215–216, 220–226) and Nakamura et al. (1990, pp. 100–103) on the choice of the null hypothesis in endogeneity testing.

[11] See Nakamura and Nakamura (1981a) for the identity of Wu's T_2 and the Hausman statistic. See Kariya and Hodoshima (1980, p. 47, Eqs. (3.16) and (3.18)) for expressions in (3) and (4).

where $c_1 = (n - K_1 - 2G_2)/G_2$, $Q^* = (b_1 - b_2)'[(y_2'A_2y_2)^{-1} - (y_2'A_1y_2)^{-1}]^{-1}(b_1 - b_2)$, $b_i = (y_2'A_iy_2)^{-1}y_2'A_iy_1$ for $i = 1, 2$, $A_1 = I - Z_1(Z_1'Z_1)^{-1}Z_1'$, $A_2 = Z(Z'Z)^{-1}Z' - Z_1(Z_1'Z_1)^{-1}Z_1'$, $Z = (Z_1, Z_2)$, and $Q_2 = S_{11.2} + q_2$, and where $q_2 = \hat{\Pi}_2'[A_{22}^{-1} - A_{22}^{-1}\hat{\beta}_2(\hat{\beta}_2A_{22}^{-1}\hat{\beta}_2)^{-1}\hat{\beta}_2'A_{22}^{-1}]\hat{\Pi}_2$, $A_{22} = [Z_2'Z_2 - Z_2'Z_1(Z_1'Z_1)^{-1}Z_1'Z_2]^{-1}$, and $S_{11.2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$. Using this same notation, Revankar's statistic may be written as

$$\mathrm{RV} = c_3(Q^*/S_{11.2}) \tag{4}$$

where[12] $c_3 = (n - K - G_2)/G_2$ and $K = K_1 + K_2$.

Kariya and Hodoshima (1980, p. 48, Eq. (3.31)) show that the distributions of the Wu–Hausman statistic, T_2, and Revankar's RV, conditional on the values for $\hat{\beta}_2$ and S_{22}, are given by

$$T_2 \mid (\hat{\beta}_2, S_{22}) \sim F''(\delta_1, \delta_2 : G_2, n - k_1 - 2G_2) \tag{5a}$$

and

$$\mathrm{RV} \mid (\hat{\beta}_2, S_{22}) \sim F'(\delta_1 : G_2, n - K - G_2). \tag{5b}$$

F'' and F' are the doubly noncentral and noncentral F distributions, respectively.

In the appendix, we show that δ_1 in (5a) and (5b) can be expressed as

$$\delta_1 = (\rho^2/\lambda_3^2)\, S_{22}(R^2_{y_2 - Z_1\beta_1 \cdot Z_2})\ (\mathrm{C.F.}), \tag{6a}$$

or, for a large n, as

$$\delta_1 \cong [\rho^2/(1 - \rho^2)](n - k)\,(R^2_{y_2 - Z_1\beta_1 \cdot Z_2}). \tag{6b}$$

In (6a), ρ^2 is the square of the population correlation between the true error terms for the structural equation (1a) for y_1 and for the auxiliary equation (1b) for y_2; the parameter λ_3^2 is the population variance of the component of the true error term for the equation for y_2 that is independent of the true error term for equation (1a) for y_1; the statistic S_{22} is the OLS estimator of the population variance of the true error term for (1b); and $R^2_{y_2 - Z_1\beta_1 \cdot Z_2}$ is the R^2 from the regression of $y_2 - Z_1\beta_1$ on Z_2, where Z_1 denotes the K_1 exogenous variables included in both (1a) and (1b) while Z_2 denotes the K_2 exogenous variables included in (1b) but excluded from (1a). C.F. approaches 1 as n goes to infinity.

---

[12] $S_{11.2}$ in the denominator of RV as given in (4) is the residual sum of squares from the regression of the OLS residuals from the reduced form equation for y_1 on the OLS residuals from the reduced form equation for y_2. Q_2 in the denominator of T_2 as given in (3) is the residual sum of squares from the regression of y_1 on y_2, Z_1 and the OLS residuals from the reduced form equation for y_2. Thus, q_2 is the amount by which the residual sum of squares from the regression of y_1 on y_2, Z_1 and the OLS residuals from the reduced form equation for y_2 exceeds the residual sum of squares from the regression of the OLS residuals from the reduced form equation for y_1 on the OLS residuals from the reduced form equation for y_2. The quantity Q^* may also be interpreted as the difference between the residual sum of squares when OLS is applied to (1a) and the residual sum of squares when y_1 is regressed on y_2, Z_1 and the OLS residuals from the reduced form equation for y_2.

For the second noncentral parameter for the distribution of the Wu–Hausman statistic, it is shown in the appendix that

$$\text{plim}_{n\to\infty}(\delta_2/n) = 0, \tag{7}$$

and that the Wu–Hausman and Revankar tests are consistent tests. A correct proof of the consistency of these tests does not seem to have been given before in the literature.

5.2. *Finite sample comparisons of the Wu–Hausman and Revankar tests*

From the appendix expression (A.4) it can be seen that the second noncentral parameter for the Wu–Hausman statistic, δ_2, will always be positive when $K_2 > G_2$. The T_2 test will tend to be more powerful than the RV test due to the effect of the associated degrees of freedom, but the RV test will tend to be more powerful than the T_2 test due to the effects of δ_2 on the distribution for T_2.[13] However, from expression (A.5) in the appendix, it can be seen that δ_2 cannot increase faster in probability than δ_1 as n increases. This suggests that the properties of the finite sample distribution of the Wu–Hausman statistic are primarily determined by δ_1: a conjecture that is supported by Monte Carlo results.[14] These Monte Carlo experiments also show that the net effects of δ_2 and of the difference in degrees of freedom on the relative powers of the Wu–Hausman and Revankar tests are typically small. When Eq. (1a) for y_1 is just identified, the T_2 and RV test statistics are identical, and both have distributions that depend only on δ_1.

Using methods parallel to our analysis of the Wu–Hausman and Revankar tests, it can be shown that the distributions of all of the endogeneity test statistics that were mentioned in Section 4 are primarily determined by δ_1. These results focus attention on expressions (6a) and (6b).

5.3. *Poor power for endogeneity tests when the auxiliary instrumental equation is weak*

A result that emerges from the expressions for δ_1 and the role of this noncentral parameter in determining the power properties of the Wu–Hausman and related endogeneity tests is that their power will be higher the higher the proportion is of the variability in the included endogenous variable that is explained by the exogenous variables excluded from the structural equation.

---

[13] Because of the properties of the doubly noncentral F distribution (see Lehmann, 1959), the conditional power of the Wu–Hausman test is a strictly increasing function of δ_1 but a strictly decreasing function of δ_2 to the extent that δ_1 and δ_2 can take on independent values.

[14] We are referring to our own unpublished Monte Carlo results and to Thurman (1986).

More specifically, the power is an increasing function of $R^2_{y_2-Z_1\beta_1\cdot Z_2}$ which appears in (6a) and (6b). We can express this partial R^2 as

$$R^2_{y_2-Z_1\beta_1\cdot Z_2}=\frac{R^2_{y_2\cdot Z_1,Z_2}-R^2_{y_2\cdot Z_1}}{1-R^2_{y_2\cdot Z_1}}$$

where $R^2_{y_2\cdot Z_1}$ is the usual overall R^2 for the regression of y_2 on Z_1, and $R^2_{y_2\cdot Z_1,Z_2}$ is the R^2 for the regression of y_2 on Z_1 and Z_2. The minimum value that $R^2_{y_2-Z_1\beta_1\cdot Z_2}$ can take on is zero, in which case $R^2_{y_2\cdot Z_1,Z_2}=R^2_{y_2\cdot Z_1}$ and (1a) is unidentified. The maximum possible value for $R^2_{y_2-Z_1\beta_1\cdot Z_2}$ is $R^2_{y_2\cdot Z_1,Z_2}$, which is the R^2 for the instrumental equation for y_2. When this R^2 is low, the power of the Wu–Hausman, or any of the other closely related endogeneity tests, will tend to be low.[15]

Weak auxiliary equations are commonplace in labor economics research based on micro level data. (See Revankar and Yoshino (1990) for a macro data example where this is so as well.) Mroz (1987, p. 770) reports R^2 values for instrumental wage equations ranging from 0.15 to 0.23. The R^2 values for auxiliary equations for child status variables are often even lower. In situations like these where a pretest for evaluating a potential endogeneity problem is particularly needed, acceptances of a null hypothesis of no endogeneity are likely to be Type II errors. The use of a Wu–Hausman type endogeneity test provides the appearance, but not the reality, of rigorous investigation of whether it is reasonable to use OLS rather than IV estimation results.

5.4. *Endogeneity tests and endogeneity bias*

Because the distribution of T_2 is primarily determined by δ_1, we see from (6a) and (6b) that the power of the Wu–Hausman test will rise as ρ^2 rises, where

$$\rho^2=\frac{[\mathrm{Cov}(\lambda_1u_1,\ \lambda_2u_1+\lambda_3u_2)]^2}{\mathrm{Var}(\lambda_1u_1)\ \mathrm{Var}(\lambda_2u_1+\lambda_3u_2)}=\frac{\lambda_2^2}{\lambda_2^2+\lambda_3^2}. \tag{8}$$

Recall that λ_1u_1 is the true error term for Eq. (1a) for y_1, and $v_2=\lambda_2u_1+\lambda_3u_2$ is the error term for Eq. (1b) for y_2. Nonzero values of ρ^2 are the root source of the OLS bias problem. However, the value of ρ^2 is not the sole determinant of the *size* of the OLS bias. Using the result given in Nakamura and Nakamura (1985c, p. 215, Eq. (4)) and the given properties of u_1 and u_2, the population value of the OLS bias can be expressed as

$$\begin{aligned}B&=\mathrm{plim}(b_1-\alpha_1)\\&=\frac{[\mathrm{Cov}(\lambda_1u_1,\lambda_2u_1+\lambda_3u_2)]^2}{\mathrm{plim}(1/n)\ (y_2'A_1y_2)}=\frac{\lambda_1\lambda_2}{\mathrm{plim}(1/n)\ (y_2'A_1y_2)},\end{aligned} \tag{9}$$

[15] On this topic, see, for example, Cragg and Donald (1993).

where b_1 is the OLS estimator of α_1 in (1a), and where $y_2'A_1y_2$ is the sum of squared residuals from the regression of y_2 on Z_1. It can be shown that B depends on λ_1, λ_2 and λ_3 whereas ρ depends on λ_2 and λ_3 but not λ_1 (as is also the case for both δ_1 and δ_2).

From the expressions for ρ and B, we see that $\rho = 0$ if and only if $\lambda_2 = 0$, and when $\lambda_2 = 0$ we also have $B = 0$. Hence H_0^*: $B = 0$ will always be true when H_0': $\rho = 0$ and H_0: $\mathrm{Cov}(\lambda_1 u_1, v_2) = 0$ are true. But the *power* of the Wu–Hausman test will not be affected by changes in the value of λ_1, and hence in the (nonzero) value of B, so long as λ_2 and λ_3 are fixed. Monte Carlo evidence reveals that when ρ and also the explanatory power of the instrumental equation are low, the power of the test of H_0^*: $B = 0$ (and also of H_0': $\rho = 0$ and of H_0: $\mathrm{Cov}(\lambda_1 u_1, v_2) = 0$) can be low despite values of B that are arbitrarily large relative to α_1, the true coefficient of the endogenous explanatory variable. (See Nakamura and Nakamura, 1985c, pp. 220–221.) Unfortunately, it is B rather than ρ that is of real interest in most applied contexts where there are concerns about the possible endogeneity of included explanatory variables.[16]

The importance of this is that tests of endogeneity are usually applied when researchers have strong a priori reasons for believing that ρ, and hence B, are nonzero. What most applied researchers who use Wu–Hausman type tests are trying to do is heed Durbin's (1954, p. 27) advice: "Since the use of an instrumental variable involves a certain loss of efficiency one should feel rather cautious about using it until the extent of the bias in the ordinary least-squares estimators has been investigated". Christ (1966, pp. 157–158) offers similar advice:

> ... there is *no* point in the enlargement of most models at which a convincing stand can be made against such arguments for the addition of another equation – unless it is the point where all possible variables have already been included, and of course the model would then be utterly unmanageable. What the economist should do in practice, therefore, in my opinion, is to stop adding equations and variables when he believes that the variables he chooses to call exogenous meet the definition *closely enough* so that the errors incurred through the discrepancy are small in comparison with the degree of accuracy that he thinks is desirable for his purpose (or is attainable). This is necessarily a somewhat arbitrary decision.

If an exogeneity test is to be used as a basis for deciding when B is 'close enough' to zero, it is important that the power of the test reliably rises as the departure of B from zero increases in size. This is not the case for the Wu–Hausman and the other related tests of endogeneity.

---

[16] Hausman and Taylor (1981) note: 'It appears in practice ... that H_0 is frequently tested in situations where we can infer from the subsequent actions taken that the hypothesis H_0^* was intended ... (p. 13). However, they do not consider the differences in power that we do for tests of H_0^*: $B = 0$ versus H_0: $\mathrm{Cov}(\lambda, u_1, v_2) = 0$ or H_0': $\rho = 0$.

The perverse power properties of these endogeneity tests are rooted in problems with properly standardizing the difference $(b_1 - b_2)$ in the numerator of expressions (3) and (4) for T_2 and RV, respectively. As defined following (3), b_2 is an IV estimator of α_1 in (1a). This estimator is consistent whether or not the associated explanatory variable is endogenous; thus, under the maintained assumptions of the model, $\text{plim}(b_2) = \alpha_1$. Since b_1 is the OLS estimator of α_1, $(b_1 - b_2)$ is a point estimator of $B = \text{plim}(b_1 - \alpha_1)$, the population value of the OLS bias defined in (9). This is so under the null hypothesis H_0: $\text{Cov}(\lambda_1 u_1, v_2) = 0$ which is equivalent to H_0': $\rho = 0$ and to H_0^*: $B = 0$. It is also so under the corresponding alternative hypotheses. However, the variance estimators for the difference $(b_1 - b_2)$ that are used in forming the various endogeneity test statistics are only consistent under the null hypotheses H_0, H_0' and H_0^* (see Durbin (1954, p. 29) or Hausman (1978)).

In labor economics, tests of endogeneity are being used as though the difference $(b_1 - b_2)$ were being properly standardized *even if there is some endogeneity*. In particular, small values of endogeneity test statistics are often interpreted informally as indications that whatever endogeneity problems there are are not severe. This inference is inappropriate. The Wu–Hausman and related tests are tests of the *existence* of an endogeneity bias; they do not provide a metric for the *seriousness* of an existing endogeneity bias problem.

5.5. *The hazard of endogenous instruments*

A further result that emerges from examination of expressions (3) and (4) is that the T_2 and RV tests are only valid when suitable instrumental variables can be found – suitable in at least the minimal sense that these instruments are independent of the model disturbance terms. Otherwise, it may not be true that $\text{plim}(b_2) = \alpha_1$, and $(b_1 - b_2)$ may be an inconsistent estimator of B. Pretesting of the Wu–Hausman variety replaces a priori speculation about the possible endogeneity of the explanatory variable that is the object of the pretest with a priori speculation about the possible endogeneity of the variables in the whole set of instruments to be used in carrying out the endogeneity test. We must ask what is gained by this.

5.6. *Rejection of a pretest mixed estimation approach*

The properties of Wu–Hausman type tests lead us to reject a mixed estimation strategy, with the decision to use OLS versus IV results depending on the outcome of an exogeneity test. A first reason is that, as shown in Section 5.3, when the efficiency losses from using IV are large, the power of Wu–Hausman type pretests is likely to be low. The tests are weak when they are needed most. Second of all, as was shown in Section 5.4, these are tests for the existence, not the seriousness, of endogeneity bias. They do not provide an appropriate metric

for judging the *seriousness* of an endogeneity bias problem, which is what applied researchers are usually concerned about when trying to choose between OLS and IV results. A third issue that was noted in Section 5.5 above is that Wu–Hausman type tests are of no use in situations where a researcher is concerned that, in addition to efficiency losses, instrumentation may not fully solve the problem of endogeneity bias because the available instruments are not truly exogenous.

6. Dealing with endogeneity without pretests

Without endogeneity pretesting, we are back in the position of having to make variable by variable judgements as to which of the explanatory variables in a model should be instrumented without the benefit of any generally accepted and replicable decision rule. In this situation, some recommend *always* instrumenting explanatory variables suspected of being endogenous. These researchers argue that appropriate instruments can almost always be found, and that bias and inconsistency problems are inherently more serious than efficiency losses. As evidence that appropriate instruments can be found, supporters of always instrumenting endogenous variables often cite the growing literature on natural experiments.

6.1. Rejection of an 'always instrument' policy

We reject the 'always instrument' policy. A natural experiment rarely provides exogenous variation for more than *one* of the included explanatory variables in a model. Natural experiments cannot solve all the endogeneity problems of, say, a labor economist interested in estimating an equation for annual hours of work that contains a wage variable, a work experience variable, an education variable, a husband's income variable, and several child status variables all of which are probably endogenous to some degree under normal circumstances. Even variables like years of age or an indicator for being nonwhite that are often said to be unassailably exogenous are not really so. No one works more or less simply because another birthday has passed by, or because of being nonwhite. In a labor supply model, variables for age and race are serving as proxies for related circumstances and behavior, some of which are probably endogenous. Thus we reject the 'always instrument' policy because we regard it as infeasible.

Second of all, with endogenous control variables we reject an 'always instrument' policy because of the resulting diversion of effort. For example, the decision to instrument the child status variables in a labor supply equation means that attention must be diverted away from trying to understand how children affect the labor supply behavior of their mothers to the side issue of how to specify auxiliary child status equations. A related problem of an always

instrument policy is that it encourages applied researchers to limit the variables they include in their models so as to avoid difficult instrumentation problems. Also, efforts to find truly exogenous instruments have the unfortunate tendency to push researchers toward the use of substantively irrelevant instruments.

A third reason we oppose an 'always instrument' policy is that there is usually little real evidence that the instruments that are used are exogenous themselves. This was one of our reasons as well for rejecting a Wu–Hausman pretest estimation strategy.

A related fourth reason we oppose an 'always instrument' policy has to do with the simplified nature of economic models. In instrumentation, variables are usually viewed as homogeneous. Consider a wage variable, w, which appears in a labor supply equation with a true coefficient β_w. As before, let $\hat{w}$ denote the predicted values of w from an auxiliary IV equation. Usually we assume that if $\hat{w}$ is substituted for w in the labor supply equation, then β_w is the true coefficient of the predicted wage variable too. But $\hat{w}$ will only reflect those components of the variation in w captured by the explanatory variables in the auxiliary wage equation.

People's actual wages change for reasons such as improvements in qualifications leading to merit increments or promotions, or switches to better jobs; geographical moves; job loss; seniority wage contract provisions; personal problems requiring reductions in job responsibilities; and so on. In addition, wages differ from worker to worker because of inter-employer differences in compensation practices, differing economic conditions at the time of initial hiring, statistical discrimination, and other such factors. Different choices of instruments will pick up different components of the variation in a wage variable even if all the instruments are truly exogenous, and the 'true' coefficients associated with these different components may differ. Because of this, our general belief is that multiple instruments should be experimented with and reported for important right-hand side endogenous variables, and that the results for direct estimation should be reported as well. For less important explanatory variables – endogenous control variables for which the researcher is unable or unwilling to spend the effort to check out relevant alternative instruments – we are against instrumentation. In general, we are against claims of parameter consistency based on poorly supported instrumentation attempts.

6.2. Possible lessons from research in epidemiology

Economics and epidemiology have a great deal in common in terms of pervasive possibilities for endogenous feedback and co-determination of outcome variables. The difficulties are similar, but the research approaches adopted for dealing with these difficulties differ.

Consider the early medical case history findings that seemed to indicate an association between high cholesterol levels and cardiovascular disease. There

were suggestions that these results might be partially, or even entirely, due to the fact that those with higher cholesterol levels also, on average, weigh more, exercise less, and smoke more. These concerns might lead an applied econometrician to begin thinking about regressing some indicator of cardiovascular disease on variables for body weight, amount of exercise per week, and cigarettes smoked per week as well as a cholesterol level measure, though the lack of theories about the underlying behavioral mechanisms would hamper this approach. The econometrician would also be troubled by suggestions that factors that had been ignored in data collection, such as tension or hostility levels and other aspects of eating behavior, could affect the likelihood and severity of cardiovascular disease. Without information on these sorts of factors, concerns might arise about correlations between the equation error term and some of the included explanatory factors affecting cardiovascular disease – much as labor economists such as Schultz (1980) and Mroz (1987) have worried that the taste for work in the true disturbance term for a labor supply equation may be correlated with the child status and various other included explanatory variables.

Why is it that medical journals are *not* filled with cardiovascular multiple regression models, and with studies reporting Wu–Hausman type test results to ascertain whether certain explanatory variables in cardiovascular disease equations are endogenous? Why are attempts to instrument potentially endogenous explanatory variables hard to find in medical research journals?

In exploring potential causal factors, five distinct questions arise: (1) Is there any detectable effect? (2) How big does the effect seem to be? (3) How big would the effect have to be to be of applied relevance? (4) Is it possible that the apparent effect – or the apparent smallness or lack of an effect – is partially, or even wholly, due to other causal factors? (5) Are there synergistic interactions with other causal factors?

Our impression from the epidemiological literature on the effects of cholesterol on cardiovascular disease is that there is heavy reliance on analysis of variance methods for detecting differences among various groupings of the available data. It seems that a matched samples approach is often used in grouping observations, and that the groupings used in successive studies are progressive in that increasing numbers of factors are taken into account in the matching process. *Progressively expanded and retargetted data collection efforts seem to be an integral part of this research approach.* Replication of estimation results and prediction also appear to be fundamental to the process by which findings becoming 'established'.

Our impression is that a progressive research strategy is embarked on in epidemiology without pretending to have fully specified models of the relevant behavioral processes, and without the custom of making unsubstantiated claims about the independence of explanatory factors versus unobserved true error terms. This practice stands in contrast to the stated a priori theory needs of

empirical economists as portrayed by the Cowles Commission and Christ (1994, p. 33):

> The Cowles view was that to understand a particular aspect of economic behavior, such as the price of food, or aggregate personal consumption, one wanted a system of equations capable of describing it. These equations should contain all relevant observable variables, be of known form (e.g. linear, log-linear, quadratic), and have estimatable coefficients. ... Little attention was given to how to choose the variables and the form of the equations; it was thought that economic theory would provide this information in each case. ... The main effort was directed to estimating the equations once they had been formulated.

The Cowles Commission research strategy seems to be to use theory and other a priori knowledge to provide answers to questions (4) and (5) above, and to build these answers into the models to be used in carrying out empirical investigations of questions (1) and (2). (Question (3) above is largely ignored in this approach.) This empirical strategy would be appropriate and effective in labor economics *if only we had the needed a priori knowledge.*

6.3. Prediction as one means for choosing among alternative estimation results

Most of the approaches we would recommend for dealing with potentially endogenous explanatory variables – these include direct estimation as well as estimation with alternative instrument sets, replication with different existing data sets, and ongoing new data collection to permit the inclusion of new control variables – will lead to competing sets of estimation results. How should we choose among them?

Noting the possibility of poor a priori information, Christ (1966) writes:

> On many occasions throughout this book it has been emphasized that in practice the so-called a priori information is far from certain and that it is important to have some means of testing and evaluating this information. (p. 531)
>
> In this situation the economist ... can try out several different theoretically reasonable forms, in a sort of experimental fashion confronting each one with relevant data, and then choose among them after he sees how well they fit the data. ...
>
> The danger lies in the possibility of being too clever or too persistent, and finding an equation that fits the available data well enough but is nevertheless wrong because it describes temporary or accidental features. ... The best protection we can have against this danger is to test our equations against data that could not have influenced the choice of the equations. (pp. 8–9)

> In cross-section studies ... the sample is typically very much larger than in time series studies. We can therefore divide an available sample into two parts, each containing hundreds or thousands of observations, one part to be used initially to help suggest the form of the model and the other part to be used later as a test of the predictive ability of the model chosen. (p. 548)

In our studies of female labor supply behavior we experiment with alternative model specifications, and replicate our results for Canadian census data, US census data, and US Panel Study of Income Dynamics (PSID) data for multiple time periods. Also, in most of our studies, we make use of data samples for multiple demographic groups, and we report direct OLS results and also IV results for alternative instrument sets. Our uncompensated wage elasticity estimates for married women consistently fall in a far narrower range than the range of −0.89 to 15.24 reported by Killingsworth (1983, Table 43, pp. 194–199) for so-called second generation studies. In our book *The Second Paycheck* (Nakamura and Nakamura, 1985a), based on PSID data, we provide the elasticity estimates for a number of different estimation approaches and demographic groups, as summarized in Table 1. These are then compared with the elasticity estimates of others:

> [E]stimates of the uncompensated wage elasticity of hours of work for wives from experimental data are found from Killingsworth (1983, Table 6.2, pp. 398–399) to range from −0.36 to 0.94. ... the estimates for wives from experimental data span roughly the same range as the estimates for men from both nonexperimental and experimental data (−0.38 to 0.28). Using nonexperimental data, Nakamura, et al. (1979, p. 800) obtain uncompensated wage elasticities of hours of work for married women who work of −0.320 to 0.299; Nakamura and Nakamura (1981b, p. 483) report values of −0.495 to 0.654; and values of −0.197 to −0.030 are reported in Nakamura and Nakamura (1983, p. 246). ... Also, using yet another data set for Canada with an improved wage variable Chris Robinson and Nigel Tomes (1985) have obtained results for married women that support our findings. (Nakamura and Nakamura, 1985a, pp. 181–183)

Having documented the effects of differences in estimation methods on the uncompensated wage elasticity, a wide variety of in-sample and out-of-sample predictive evaluations involving cross-sectional, year-to-year, and cumulative comparisons are also presented.[17] Interested readers are referred to Nakamura

---

[17] We have increasingly focused on year-to-year changes and comparisons (see Nakamura and Nakamura, 1992, 1994) because we agree with Manski (1995, p. 135):

> We have seen that the reflection problem can make it rather difficult to draw conclusions about the nature of social effects from observations of the equilibrium outcomes experienced by a population.... Observation of the dynamics of social processes can sometimes open new possibilities for inference.

Table 1
Uncompensated wage elasticities of hours of work evaluated at the mean hours of work

	Wives		Unmarried women	
	21–46	47–64	21–46	47–64
Working women who worked in year t−1				
Using IV wage estimates and including lagged hours in the labor supply equation	0.06	0.04	0.12	0.06
Using OLS wage estimates and including lagged hours in the labor supply equation	−0.10	−0.08	−0.03	−0.03
Using IV wage estimates without including lagged hours in the labor supply equation	−0.10	−0.29	−0.02	0.03
Working women who did not work in year t−1				
Using IV wage estimates	−0.59	−0.14	0.21	−0.94
Using OLS wage estimates	−0.10	−0.32	−0.02	−0.35

Source: The figures shown are from tables in Nakamura and Nakamura (1985a, pp. 187–189).

and Nakamura (1985a), as well as Nakamura and Nakamura (1992) and Nakamura et al. (1990); and also to Nakamura and Walker (1994) for discussions of predictive approaches to model choice.

The key insight that underlies our own use of predictive analyses is that if the estimated equations cannot capture the main features of the distributions of the dependent variables, both in-sample and out-of-sample, 'it is likely that at least one wrong assumption was made somewhere' (Christ, 1966, p. 158). Heckman (1985) in his Foreword to our book *The Second Paycheck* provides the following generous assessment of our methods of model choice and validation – an assessment that we repeat because it is strongly supportive of an emphasis on prediction and on the use of judgment in enlarging economic models. Heckman writes:

> This book advocates and implements a novel approach to model verification. The approach pursued in many recent studies of labor supply has been to arrive at final empirical specifications for a single demographic group by means of a battery of '*t*' and '*F*' tests on the coefficients of candidate variables. The problem of pretest bias is conveniently ignored. Only rarely ... do analysts ask how well fitted micro relationships explain other aspects of labor supply such as the aggregate time series movement. Focusing on one demographic group in isolation, these studies present a bewildering array of findings that have thus far eluded synthesis.
>
> This book does not adopt the conventional '*t*' ratio methodology. The authors estimate the same models for a variety of age, marital status, and sex

groups and look for commonalities of findings across groups. They look for consistency in the impact of explanatory variables on different dimensions of labor supply. Models are simulated both within samples and out of samples. ...

Given the lack of basic knowledge about empirical regularities of labor force dynamics, the approach taken by the authors appears the most scientifically promising one. (pp. xi–xii)

7. Conclusions

In many areas of applied economic research, the ordinary least squares (OLS) estimates of key relationships are suspect because of potential endogeneity bias problems. However, instrumental variables (IV) estimates can be suspect as well because of large standard errors and erratic parameter estimates, and because of other concerns such as the potential endogeneity or lack of relevant variation in the variables included in the auxiliary IV equation(s). In cases like these, applied researchers have turned to endogeneity tests such as those of Wu and of Hausman in the quest for a statistically sound and replicable basis for choosing between OLS and IV estimation results. We conclude that this hope is often unrealistic.

Lacking a general statistical decision rule for choosing between OLS and IV or other simultaneous equation estimation results, we consider other research strategies that might help. These include strategies for forming hypotheses about potentially important omitted factors: hypotheses that can then guide further data collection efforts. These also include predictive evaluations of alternative sets of estimation results. These are research strategies that can help direct attention toward empirical models that can explain key features of the observed behavior; they are not strategies that will necessarily lead us to decide on a *single* best estimated model.

Appendix

To derive interpretable expressions for δ_1 and δ_2, we begin with the following expressions for these parameters that Kariya and Hodoshima give in their derivation of the exact conditional distributions of the Wu–Hausman and Revankar statistics:

$$\delta_1 = \eta' M(I + W)^{-1} M\eta/\Omega_{11.2} \tag{A.1}$$

and

$$\delta_2 = \eta N(I + W)^{-1} N\eta/\Omega_{11.2} = \eta' N\eta/\Omega_{11.2}, \tag{A.2}$$

with $\eta = C(\Pi_2 - \beta_2\Omega_{22}^{-1}\Omega_{12})$, $M = C\hat{\beta}_2(\hat{\beta}_2'C'C\hat{\beta}_2)^{-1}\hat{\beta}_2'C'$, $N = I - M$, $W = C\hat{\beta}_2 S_{22}^{-1}\hat{\beta}_2'C'$, and $\Omega_{11.2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$, and with C defined as a non-singular $(K_2 \times K_2)$ matrix such that $C^2 = Z_2'[I - Z_1(Z_1'Z_1)^{-1}Z_1']Z_2$. The values for $\hat{\beta}_2$ and S_{22} on which the distributions of T_2 and RV are conditional can be readily obtained in any real application by estimating the reduced form equation for y_2. We view the conditional nature of these distributions as an advantage. However, the expressions given in (A.1) and (A.2) are difficult to interpret because it is not obvious how M, N and W relate to the salient properties of a given model.

To express δ_1 in (A.1) in more meaningful terms, we let $a = C\hat{\beta}_2$, $\alpha = C\beta_2$, and $m = \alpha_1 - \Omega_{22}^{-1}\Omega_{21}$. If the structural equation for y_1 is identified (i.e., $K_2 \geqslant G_2$), then $\beta_2\alpha_1 = \Pi_2$ and

$$\eta = C(\Pi_2 - \beta_2\Omega_{22}^{-1}\Omega_{21}) = C\beta_2(\alpha_1 - \Omega_{22}^{-1}\Omega_{21}) \tag{A.3}$$

which implies that $\eta = \alpha m$. We also have $(I + W)^{-1} = (I + aa'/S_{22})^{-1} = I - [aa'/(S_{22} + a'a)]$, $a'(I + W)^{-1}a = a'S_{22}a/(S_{22} + a'a)$, and $M = a(a'a)^{-1}a'$. Substituting into (A.1), we get $\delta_1 = (m^2/\Omega_{11.2})\alpha' a(a'a)^{-1}\,[a'S_{22}a/(S_{22} + a'a)]$ $(a'a)^{-1}a'\alpha = (m^2/\Omega_{11.2})(S_{22})(\alpha'a/a'a)^2[a'a/(S_{22} + a'a)]$. Since ρ is the correlation coefficient between $\lambda_1 u_1$ and v_2, and since $\rho^2 = \lambda_2^2/(\lambda_2^2 + \lambda_3^2)$, we have $m^2/\Omega_{11.2} = \rho^2/\lambda_3^2$. The term $(\alpha'a/a'a)^2$ is essentially a finite sample correction factor that we will denote by C.F. The term $a'a/(S_{22} + a'a)$ is the R^2 from the regression of $y_2 - Z_1\beta_1$ on Z_2, which we denote by $R^2_{y_2 - Z_1\beta_1 \cdot Z_2}$. This term is the ratio of the variation of $y_2 - Z_1\beta_1$ explained by Z_2 to the total variation of $y_2 - Z_1\beta_1$; that is, it is the proportion of the total variation of y_2 explained by Z_2 after removing the effects of Z_1. Using this notation, δ_1 can be expressed as in (6a) in the text. Moreover, since $S_{22}/(n-k)$ tends toward $\Omega_{22} = \lambda_2^2 + \lambda_3^2$ and C.F. tends toward 1 as n goes to infinity, with $\text{plim}_{n\to\infty}\, a = \alpha$, we also have the approximate expression (6b).

Using the same notation, δ_2 in (A.2) can be rewritten in a more interpretable exact form as

$$\begin{aligned}
\delta_2 &= (m^2/\Omega_{11.2})\alpha'[I - a(a'a)^{-1}a']\alpha \\
&= (m^2/\Omega_{11.2})[\{(\alpha'\alpha)(a'a) - (\alpha'a)^2\}/(a'a)] \\
&= (\rho^2/\lambda_3^2)(n^2/n)[\{(\alpha'\alpha/n)(a'a/n) - (\alpha'a/n)^2\}/(a'a/n)].
\end{aligned} \tag{A.4}$$

Under standard assumptions, C^2/n converges in probability to a constant matrix. From this property and the consistency of the estimator of β_2 it follows that

$$\text{plim}_{n\to\infty}\,(\delta_2/n) = \frac{\rho^2}{\lambda_3^2}\left(\frac{1}{a^*}\right)[(a^*)(a^*) - (a^*)^2] = 0, \tag{A.5}$$

where $a^* = \text{plim}_{n\to\infty}\,(\sum_{i=1}^{K_2} a_i^2/n)$.

Since N defined following Eq. (A.2) is idempotent, we get from (A.4) and (A.5) that $0 = \Omega_{11.2}\,\text{plim}_{n\to\infty}(\delta_2/n) = \Omega_{11.2}\,\text{plim}_{n\to\infty}(\eta'N'N\eta/n) = \Omega_{11.2}$ $(\text{plim}_{n\to\infty}\,\eta'N'/\sqrt{n})(\text{plim}_{n\to\infty}\,N\eta/\sqrt{n})$, or

$$\text{plim}_{n\to\infty}(N\eta/\sqrt{n}) = 0. \tag{A.6}$$

We also have

$$N\eta = NC(\Pi_2 - \beta_2\Omega_{22}^{-1}\Omega_{21}) = NC\Pi_2 - NC\beta_2\Omega_{22}^{-1}\Omega_{22}. \tag{A.7}$$

Since $NC\beta_2\Omega_{22}^{-1}\Omega_{21} = [I - (C\hat{\beta}_2(\hat{\beta}_2C'C\hat{\beta}_2)^{-1}\hat{\beta}_2'C']C\beta_2\Omega_{22}^{-1}\Omega_{21}$, it follows from $\text{plim}_{n\to\infty}\hat{\beta}^2 = \beta_2$ that

$$\text{plim}_{n\to\infty}\,NC\beta_2\Omega_{22}^{-1}\Omega_{21} = 0. \tag{A.8}$$

Thus, from (A.6)–(A.8), we get

$$\text{plim}_{n\to\infty}(N\eta/\sqrt{n}) = \text{plim}_{n\to\infty}(NC\Pi_2/\sqrt{n}) = 0. \tag{A.9}$$

We can write $q_2 = \hat{\Pi}_2C'N'NC\hat{\Pi}_2$, where q_2 is defined following (3) in the text and where it can be seen from (3) and (4) in the text that q_2 is the only difference other than degrees of freedom between the Wu–Hausman and Revankar statistics. Thus we have

$$q_2/n = (\hat{\Pi}_2'C'N'/\sqrt{n})(NC\hat{\Pi}_2/\sqrt{n}). \tag{A.10}$$

Using $\text{plim}_{n\to\infty}\hat{\Pi}_2 = \Pi_2$, we get from (A.9) that

$$\text{plim}_{n\to\infty}(NC\eta_2/\sqrt{n}) = \text{plim}_{n\to\infty}(NC\Pi/\sqrt{n}) = 0. \tag{A.11}$$

Thus, property (7) in the text follows from (A.10) and (A.11).

Kariya and Hodoshima (1980, p. 50) note the result given in (7) without proof. Tsurumi and Shiba (1982) discuss this result as it relates to the identification problem. This result seems to be implied by Revankar (1978, Proposition 4, p. 175), but his proof is based on incorrect lemmas by Wu, as explained, by Hausman (1978, fn. 7, p. 1257): 'Wu's derivation of the (non-local) limiting distribution of the test statistic under the alternative hypothesis in equation (3.12) of his paper (1973) seems incorrect since application of the central limit theorem on p. 748 requires the sum of random variables with zero mean.... Interpreted locally Wu's results seem valid since only the usual least squares variance term v_1 is needed'.

The above results imply the following large sample properties. First, result (7) that $\text{plim}_{n\to\infty}(q_2/n) = 0$ means that $\text{plim}_{n\to\infty}T_2 = \text{plim}_{n\to\infty}\text{RV}$. Thus for large n both T_2 in (3) and RV in (4) essentially obey the same conditional distribution $F'(\delta_1 : G_2, \infty)$ and hence have essentially the same power. Asymptotically there is no difference in power between the Wu–Hausman and Revankar tests. Secondly, it is seen from (6a) in the text that $\text{plim}_{n\to\infty}\delta_1 = \infty$ since $\text{plim}_{n\to\infty}S_{22} = \infty$. Hence both the Wu–Hausman and Revankar tests have the power of one for large values of n. This proves that both tests are consistent tests.

Note that when the structural equation for y_1 is just identified so that $K_2 = G_2$, then $N = I - M = 0$ and from (A.2) we see that δ_2 is identically equal to zero. Also, since $K = K_1 + K_2$ by definition, when $K_2 = G_2$ the degrees of freedom for the two tests are the same. Hence, the two tests will be equally powerful. This is what we would expect, since in this case $q_2 = 0$ and $c_1 = c_3$, and hence the Wu–Hausman and Revankar statistics are identical.

References

Angrist, J.D., Krueger, A.B., 1995. Split-sample instrumental variables estimates of the return to schooling. Journal of Business and Economic Statistics 13, 225–235.

Berndt, E.R., 1991. The Practice of Econometrics: Classic and Contemporary. Addison Wesley, Reading, MA.

Buse, A., 1992. The bias of instrumental variables estimators. Econometrica 60, 173–180.

Christ, C.F., 1966. Econometric Models and Methods. Wiley, New York.

Christ, C.F., 1994, The Cowles Commission's contributions to econometrics at Chicago, 1939–1955. Journal of Economic Literature 32, 30–59.

Cragg, J.C., Donald, S.G., 1993. Testing identifiability and specification in instrumental variable models. Econometric Theory 9, 222–240.

Durbin, J., 1954. Errors in variables. Review of the International Statistical Institute 22, 23–32.

Geweke, J., 1987. Endogeneity and exogeneity. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: a Dictionary of Economics, vol. 2. Macmillan, London, pp. 134–136.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica 46, 1251–1271.

Hausman, J.A., Taylor, W.E., 1981. Comparing specification tests and classical tests, Bell Laboratories economics discussion paper.

Heckman, J.J., 1985. Foreword. In: Nakamura, A., Nakamura M. (Eds.), The Second Paycheck: a Socioeconomic Analysis of Earnings. Academic Press, Orlando, FL, pp. xi–xii.

Heckman, J.J., Killingsworth, M.R., MaCurdy, T.E., 1981. Empirical evidence on static labour supply models: a survey of recent developments. In: Hornstein, Z., Grice, J., Webb, A. (Eds.) The Economics of the Labour Market, Her Majesty's Stationery Office, London, pp. 73–122.

Hicks, J.R., 1965. The Theory of Wages, 2nd ed. Macmillan, London.

Kariya, T., Hodoshima, J., 1980. Finite sample properties of the tests of independence in structural systems and LRT. Economic Studies Quarterly 31, 45–56.

Killingsworth, M.R., 1983. Labor Supply. Cambridge University Press, New York.

Killingsworth, M.R., Heckman, J.J., 1986. Female labor supply: a survey. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. I. North-Holland, Amsterdam, pp. 103–204.

Koopmans, T.C., Hood, W.C., 1953. The estimation of simultaneous linear economic relationships. In: Hood, W.C., Koopmans, T.C. (Eds.), Studies in Econometric Methods, Wiley, New York, pp. 112–199.

Lehmann, E.L., 1959. Testing Statistical Hypotheses. Wiley, New York.

Manski, C.F., 1995. Identification Problems in the Social Sciences. Harvard University Press, Cambridge, MA.

Marshall, A., 1920. Principles of Economics, 8th ed. Macmillan, New York.

Mincer, J., 1962. Labor force participation of married women: a study of labor supply. In: Aspects of Labor Economics. National Bureau of Economic Research, Princeton University Press, Princeton, NJ, pp. 63–97.

Mroz, T.A., 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica 55, 765–799.

Nakamura, A., Nakamura, M., 1981a. On the relationships among several specification error tests presented by Durbin, Wu and Hausman. Econometrica 49, 1583–1588.

Nakamura, A., Nakamura, M., 1981b. A comparison of the labor force behavior of married women in the United States and Canada, with special attention to the impact of income taxes. Econometrica 49, 451–489.

Nakamura, A., Nakamura, M., 1983. Part-time and full-time work behavior of married women: a model with a doubly truncated dependent variable. Canadian Journal of Economics, 229–257.

Nakamura, A., Nakamura, M., 1985a. The Second Paycheck: a Socioeconomic Analysis of Earnings. Academic Press, Orlando, FL.

Nakamura, A., Nakamura, M. 1985b. Dynamic models of the labor force behavior of married women which can be estimated using limited amounts of past information. Journal of Econometrics 27, 273–298.

Nakamura, A., Nakamura, M., 1985c. On the performance of tests by Wu and by Hausman for detecting the ordinary least squares bias problem. Journal of Econometrics 29, 213–227.

Nakamura, A., Nakamura, M., 1992. The econometrics of female labor supply and children. Econometric Reviews 11, 1–71.

Nakamura, A., Nakamura, M., 1994. Predicting female labor supply: effects of children and recent work experience. Journal of Human Resources 29, 304–327.

Nakamura, A., Nakamura, M., Duleep, H.O., 1990. Alternative approaches to model choice. Journal of Economic Behavior and Organization 14, 97–125.

Nakamura, A., Walker, J.R., 1994. Learning from empirical evidence. Journal of Human Resources 29, 223–247.

Nakamura, M., Nakamura, A., Cullen, D., 1979. Job opportunities, the offered wage, and the labor supply of married women. American Economic Review 69, 787–805.

Pagan, A.R., Jung, Y., 1993. Understanding some failures of instrumental variable estimators, Working paper, Australian National University.

Pigou, A.C., 1946. The economics of welfare, 4th ed. Macmillan, London.

Revankar, N.S., 1978. Asymptotic relative efficiency analysis of certain tests of independence in a structural system. International Economic Review 19, 165–179.

Revankar, N.S., Hartley, M.J., 1973. An independence test and conditional unbiased predictions in the context of simultaneous equation systems. International Economic Review 14, 625–631.

Revankar, N.S., Yoshino, N., 1990. An expanded equation approach to weak-endogeneity tests in structural systems and a monetary application. Review of Economics and Statistics 72, 173–177.

Reynolds, R.A., 1982. Posterior odds for the hypothesis of independence between stochastic regressors and disturbances. International Economic Review 23, 479–490.

Robinson, C., Tomes, N., 1985. More on the labour supply of Canadian women. Canadian Journal of Economics 18, 156–163.

Schultz, T.P., 1980. Estimating labor supply functions for married women. In: Smith, J. (Ed.), Female labor supply. Princeton University Press, Princeton, NJ, pp. 25–89.

Simon, H., 1953. Causal ordering and identifiability. In: Hood and Koopmans, Eds., Studies in Econometric Methods, Wiley, New York, pp. 49–74.

Thurman, W.N., 1986. Endogeneity testing in a supply and demand framework. Review of Economics and Statistics 68, 638–646.

Tsurumi, H., Shiba, T., 1982. Bayes factor and likelihood ratio interpretation of the F statistics for testing exogeneity. Econometric Society Meetings, New York, December.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.

Wu, D., 1973. Alternative tests of independence between stochastic regressors and disturbances. Econometrica 41, 733–750.

Wu, D., 1974. Alternative tests of independence between stochastic regressors and disturbances: finite sample results. Econometrica 42, 529–546.

Zellner, A., 1984. Causality and Econometrics. In: Zellner, A. (Ed.), Basic Issues in Econometrics. University of Chicago Press, Chicago, IL, pp. 35–74.