A Test for Correlation between Signal and Noise within the Errors in Variables Model

Ramses H. Abul Naga*
DEEP, Université de Lausanne, CH-1015 Lausanne, Switzerland.

May 3, 2002

Abstract

When testing for measurement error, the vector of contrasts is the difference between the OLS and IV solutions. When testing for correlated measurement error, the OLS estimator must be replaced by a statistic which achieves consistency under the null hypothesis of uncorrelated measurement error. We propose one such estimator when one amongst several regressors is assumed to be measured with noise, and derive the related Hausman-type test.

Keywords: errors in variables, correlated measurement error, consistent adjusted least squares, Hausman tests.

JEL codes: C_1, C_4.

*This research was funded by the Swiss National Science Foundation under the Athena programme (grant no. 1217-53567.98). I wish to thank Alberto Holly, Mario Jametti and Jaya Krishnakumar for discussions. I am responsible for any errors or omissions.

1. Introduction

The errors in variables model features as a center piece of many micro-econometric applications. Variables such as managerial ability, land quality, human capital and permanent income are useful theoretical concepts which nonetheless do not lend themselves to direct measurement. Such concepts are often introduced in empirical systems as latent variables, and variants of two-stage least squares / instrumental variables (2SLS/IV) procedures are used to identify the related models.

A key identifying assumption underlying the errors in variables model is that signal and noise are uncorrelated. Such an assumption has long been questioned in empirical work (e.g. Bowles and Nelson, 1974) and has been found untenable in some specific contexts, such as the reporting of earnings (Bound et al., 1994). In the latter case, comparing company records with self reported pay, the authors concluded that earnings are *mean-reverting*, taken to signify, among other things, that signal and noise are negatively correlated (see also Heckman, 1993).

The presence of correlated measurement error does not on its own cause 2SLS/IV procedures to break down. However, it exacerbates the under-identification problem inherent to the errors in variables model. In turn, this implies that the estimation of a system of equations by a full information procedure can result in severe specification biases when measurement is wrongly assumed to be uncorrelated with the unobserved latent variables.

It is therefore important in empirical work to make allowance for this more general form of measurement error, and to devise tests for detecting it. This is what we set out to do here. In section 2 we examine the bias of the OLS solution in presence of correlated measurement error. The sense in which the OLS estimator of the slope coefficients is subject to attenuation is shown to be all the more vague as signals and errors are allowed to correlate. In section 3 we specialize our discussion to a situation where only one amongst several regressors is assumed to be measured with noise. As argued by Meijer and Wansbeek (2000), this is the case most frequently encountered in empirical work. When testing for measurement error (Wu, 1973, Hausman, 1978), the vector of contrasts is the difference between the OLS and IV solutions. When testing for correlated measurement error, the problem addressed in this paper, the OLS estimator must be replaced by a statistic which achieves consistency under the null hypothesis of uncorrelated measurement error. We propose one such estimator for the model of section 3, and derive in section 4 the related Hausman-type test. Section 5 contains concluding comments.

2. The OLS estimator

We are interested in the estimation of a model

$$y = H\beta + \varepsilon \tag{1}$$

where y is an $n-$dimensional vector, and H is an $n \times k$ matrix of unobserved regressors. We assume all variables are centered about their respective means, and furthermore, $E(\varepsilon|H) = 0$, $E(\varepsilon^2|H) = \sigma_\varepsilon^2 I$ and $cov(H) = E(H'H) = \Sigma_H$, a matrix of full rank.

While H is unobserved, we have a matrix X of proxies such that

$$X = H + V \tag{2}$$

where $E(V'V) = \Omega$. We shall assume here that Ω is a full rank matrix, as would be the case when all variables are subject to measurement error. We shall part with this assumption in section 3 below, when we specialize our discussion to the case where only one regressor is measured with noise.

Unlike in the classical treatment of the errors in variables model (Wansbeek and Meijer, 2000, ch.2), it is assumed that H, the signal, and V the noise in X, are correlated: $E(H'V)=\Delta$. Accordingly, if Σ_X denotes $E(X'X)$, we have

$$\Sigma_X = \Sigma_H + \Delta + \Delta' + \Omega \tag{3}$$

If we assume that both $\Sigma_H + \Delta$ and $\Delta' + \Omega$ are invertible matrices, we may obtain the following expression for the inverse of Σ_X:

$$\Sigma_X^{-1} = (\Sigma_H + \Delta)^{-1} - (\Sigma_H + \Delta)^{-1}\left[(\Delta' + \Omega)^{-1} + (\Sigma_H + \Delta)^{-1}\right]^{-1}(\Sigma_H + \Delta)^{-1} \tag{4}$$

In contrast with (1), the measurement model takes the form

$$\begin{aligned} Y &= XB + U && \text{(5a)} \\ U &= \varepsilon - V\beta && \text{(5b)} \end{aligned}$$

Let $b = (X'X)^{-1}X'y$ and $s_\varepsilon^2 = y'[I - X(X'X)^{-1}X']y/n$ denote the OLS estimators of β and σ_ε^2. Then,

$$\begin{aligned} p\lim(b) &= \beta + \Sigma_X^{-1}\left[p\lim\left(\frac{H'\varepsilon - H'V\beta + V'\varepsilon - V'V\beta}{n}\right)\right] \\ &= \Sigma_X^{-1}(\Sigma_H + \Delta)\beta \end{aligned} \tag{6}$$

For the regression standard error we write $s_\varepsilon^2 = y'y/n - b'X'Xb/n$ and use (6) to obtain $p\lim(s_\varepsilon^2) = \sigma_\varepsilon^2 + \beta'[(\Sigma_H - (\Sigma_H + \Delta)\Sigma_X^{-1}(\Sigma_H + \Delta)]\beta$. Using the inverse formula (4) for Σ_X,we arrive at

$$p\lim(s_\varepsilon^2) = \sigma_\varepsilon^2 + \beta'(\Delta' + \Omega)\, p\lim(b) - \beta'\Delta\beta \tag{7}$$

Setting $\Delta = 0$ in (6) and (7), we obtain the standard probability limit formulae for b and σ_ε^2 in the classical errors in variables model. As Δ may contain negative diagonal elements (a case which Bound et al., 1994 call *mean-reverting* measurement error), as well as positive elements, it may be noted from (6) that the sense in which measurement error causes attenuation becomes all the more vague under this more general model where signal and noise are allowed to correlate. Likewise, it may be observed from (7) that in presence of $\Delta \neq 0$, it is no longer necessarily the case that s_ε^2 over-estimates σ_ε^2.

3. The single regressor case

In most econometric applications, a single regressor is assumed to be measured subject to error. As is well known, even then, the model [1–2] is not identifiable without the introduction of auxiliary information. This more specialized model we consider here on may be written in the form

$$\begin{aligned} y &= h_1\beta_1 + X_2\beta_2 + \varepsilon & (1') \\ x_1 &= h_1 + v_1 & (2') \end{aligned}$$

To achieve identification, it is usually assumed that a set of m instruments Z_1 for h_1, are available:

$$h_1 = Z_1\pi + \theta \tag{8}$$

such that $m \geq 1$ and Z_1 satisfies the orthogonality requirements $E(Z_1'v_1) = E(Z_1'\theta) = E(Z_1'\varepsilon) = 0$. However, as h_1 and v_1 correlate in (1'), $E(h_1'v_1) = \delta_{11}$, it must be the case that $E(\theta'v_1) \neq 0$, and is also equal to δ_{11}. We may then write the reduced form corresponding to [1'–2'] and (8) as

$$\begin{aligned} y &= Z_1\pi\beta_1 + X_2\beta_2 + \theta\beta_1 + \varepsilon & (9) \\ x_1 &= Z_1\pi + \theta + v_1 & (10) \end{aligned}$$

It may be noted from the above system that π, β_1 and β_2 are identified. However, writing the covariance matrix for the reduced form disturbances for the ith observation:

$$\begin{bmatrix} \beta_1^2 E(\theta^2) + \sigma_\varepsilon^2 & E(\theta^2)\beta_1 + \delta_{11} \\ E(\theta^2)\beta_1 + \delta_{11} & E(\theta^2) + E(v_1^2) + 2\delta_{11} \end{bmatrix} \tag{11}$$

it may be observed that $E(\theta^2)$, $E(v_1^2)$ and σ_ε^2 are only identified when $\delta_{11} = 0$. This null hypothesis, that of uncorrelated measurement error, is what we shall be interested in testing below.

Let $j_1 = [1, 0, ..., 0]'$ be a $k-$dimensional vector, and let $\omega_{11} = E(v_1^2)$. In the context of the model (1'–2'), $\Omega = \omega_{11} j_1 j_1'$ and $\Delta = \delta_{11} j_1 j_1'$. Then the covariance matrix of (3) specializes to

$$\Sigma_X = \Sigma_H + (\omega_{11} + 2\delta_{11}) j_1 j_1' \tag{12}$$

If $X = [x_1 \quad X_2]$ and $Z = [Z_1 \quad X_2]$ and $P_Z = Z(Z'Z)^{-1}Z'$, then the instrumental variables estimator $\hat{\beta} = (X'P_Z X)^{-1} X' P_Z y$ will remain consistent both under the null, $H_o : \delta_{11} = 0$, and the alternative hypothesis of correlated measurement error. The maximum likelihood estimator of the system [9-10] will be best asymptotic normal under H_o, and inconsistent under the alternative $\delta_{11} \neq 0$.

4. A test

A Hausman test for the assumption $\delta_{11} = 0$ can therefore be based on the comparison of the ML and IV estimates of β. It should be noted however that the ML estimator of β is derived under the additional assumption that the data are drawn from a multivariate normal distribution (see Jőreskog and Goldberger, 1975)[1]. It may therefore be of interest to test for correlated measurement error *per se*, without appealing to the normality assumption.

An alternative route therefore consists in deriving another estimator of β which does not call upon distributional assumptions. Let $\tilde{\omega}_{11}$ denote any consistent estimator of ω_{11} when the null hypothesis $\delta_{11} = 0$ is true. Under H_o,

$$\tilde{\Sigma}_H = X'X/n - \tilde{\omega}_{11} j_1 j_1' \tag{13}$$

[1]The ML estimator of the system [9-10] differs from that of the standard simultaneous equations model in that β is estimated exploiting joint restrictions on the reduced form parameters and the covariance matrix of the disturbances. It obtains in implicit form, as the solution to an eigen-value / eigen-vector problem.

consistently estimates Σ_H. If b is the OLS estimator, then

$$\tilde{b} = \frac{1}{n}\tilde{\Sigma}_H^{-1} X'Xb \tag{14}$$

consistently estimates β when the null hypothesis $\delta_{11} = 0$ is correct. The above estimator is a feasible form for the *consistent adjusted least squares* estimator of β (see Wansbeek and Meijer, 2000; ch. 5).

Following Hausman (1978) we may now define the vector of contrasts $\hat{\beta} - \tilde{b}$, with covariance matrix $cov(\hat{\beta} - \tilde{b})$. When the null hypothesis is valid, the test statistic

$$T = (\hat{\beta} - \tilde{b})'[\widehat{cov(\hat{\beta} - \tilde{b})}]^{-}(\hat{\beta} - \tilde{b}) \tag{15}$$

is asymptotically a χ^2_k variate, where A^- denotes a generalized inverse for a given matrix A and $\widehat{cov(\hat{\beta} - \tilde{b})}$ is any consistent estimator of $cov(\hat{\beta} - \tilde{b})$. For the test to become operational, it remains to derive an explicit form for the covariance matrix $cov(\hat{\beta} - \tilde{b})$. As $\tilde{b}$ however is not the best asymptotic normal estimator, we cannot readily assume that $cov(\hat{\beta} - \tilde{b})$ is equal to the difference $cov(\hat{\beta}) - cov(\tilde{b})$. Hence, below we define $B = \begin{bmatrix} \tilde{b} \\ \hat{\beta} \end{bmatrix}$ and derive the covariance matrix of this vector.

If $\hat{e}_1$ and $\hat{e}_2$ are the vectors of residuals pertaining to the reduced forms (9) and (10), then $\tilde{\omega}_{11} = (\hat{e}_2'\hat{e}_2 - \hat{e}_2'\hat{e}_1/\hat{\beta}_1)/n$ consistently estimates ω_{11} under the null hypothesis $\delta_{11} = 0$, as may be verified via inspectation of the covariance matrix (11). Under the alternative $\delta_{11} \neq 0$, $\tilde{\omega}_{11}$ converges in probability to $\omega_{11} + \delta_{11}(2 - 1/\beta_1)$. Under H_1 therefore

$$p\lim(\tilde{b}) = \left[\Sigma_X - (\omega_{11} + \delta_{11}(2 - 1/\beta_1))\, j_1 j_1'\right]^{-1} (\Sigma_H + \delta_{11} j_1 j_1')\, \beta \tag{16a}$$

By substituting (12) for Σ_X, this expression further simplifies to

$$p\lim(\tilde{b}) = \left[\Sigma_H + \frac{\delta_{11}}{\beta_1} j_1 j_1'\right]^{-1} (\Sigma_H + \delta_{11} j_1 j_1')\, \beta \neq \beta \tag{16b}$$

as required for the test to detect departures from the null hypothesis of interest.

If $M_Z = I - P_Z$, the expression for $\tilde{\omega}_{11}$ may further be written as $\tilde{\omega}_{11} = [x_1' M_Z x_1 - x_1' M_Z (y - X_2\hat{\beta}_2)/\hat{\beta}_1]/n$, so as to highlight the dependency of $\tilde{b}$ on the instrumental variables estimator $\hat{\beta}$. If we let $B = \begin{bmatrix} \tilde{b} \\ \hat{\beta} \end{bmatrix}$, then we evaluate the covariance matrix of B using the delta method as $cov(B) = \begin{bmatrix} J_\beta \\ I_k \end{bmatrix} cov(\hat{\beta}) \begin{bmatrix} J_\beta' & I_k \end{bmatrix}$,

where J_β is the Jacobian of the transformation going from $\hat{\beta}$ to $\tilde{b}$. For the purpose of obtaining J_β, it may be easier to invert $\tilde{\Sigma}_H$ of (13), so as to express $\tilde{b}$ in the form

$$\tilde{b} = \left[I_k - \left(j_1'(X'X)^{-1} j_1 + \frac{1}{\tilde{\omega}_{11}} \right) (X'X)^{-1} j_1 j_1' \right] b \tag{17}$$

for which J_β obtains as

$$J_\beta = (1 + j_1'(X'X)^{-1} j_1)^{-2} (X'X)^{-1} j_1 j_1' \left[\hat{e}_2'\hat{e}_2/\hat{\beta}_1^2 \quad \hat{e}_2'X_2/\hat{\beta}_1 \right] b \tag{18}$$

It follows then that $cov(\hat{\beta} - \tilde{b})$ in (15) is the matrix

$$cov(\hat{\beta} - \tilde{b}) = J_\beta cov(\hat{\beta}) J_\beta' + cov(\hat{\beta}) - J_\beta cov(\hat{\beta}) - cov(\hat{\beta}) J_\beta' \tag{19}$$

with J_β as defined in (18).

5. Conclusions

In presence of correlated measurement error, the claim that the OLS estimator is subject to an attenuation bias becomes all the more imprecise as the correlation between signal and noise can be either negative or positive. When the researcher possesses auxiliary information on the error-ridden regressors, the identification problem can be partly alleviated under the assumption that measurement error is uncorrelated with the signal, as is assumed in most discussions of the errors-in-variables model. The purpose of this paper was to propose a specification test for this assumption, in the context of a model where only one regressor is assumed to be subject to measurement error.

6. References

Bound J., C. Brown, G. Duncan and W. Rodgers (1994): "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data", *Journal of Labor Economics* 12, 345-368.

Bowles S. and V. Nelson (1974) "The 'Inheritance of IQ' and the Intergenerational Reproduction of Economic Inequality", *Review of Economics and Statistics* 56, 39-51.

Hausman J. (1978): "Specification Tests in Econometrics", *Econometrica* 46, 1251-1271.

Heckman J. (1993): "What Has Been Learned About Labor Supply in the Past Twenty Years?" *American Economic Review* 83 (Papers and Proceedings), 116-121.

Jöreskog K. and A. Goldberger (1975): "Estimation of a Model with Multiple Indicators and Multiple Causes on a Single Latent Variable", *Journal of the American Statistical Association* 70, 631-639.

Meijer E. and T. Wansbeek (2000): "Measurement Error in a Single Regressor", *Economics Letters* 69, 277-284.

Wansbeek T. and E. Meijer (2000): *Measurement Error and Latent Variables in Econometrics*, Amsterdam, North Holland.

Wu D.-M. (1973): "Alternative Tests of Independence between Stochastic Regressors and Disturbances", *Econometrica* 41, 733-750.