Measurement error in a single regressor

Erik Meijer* and Tom Wansbeek

June 22, 2000

SOM-theme F Interactions between consumers and firms

Abstract

For the setting of multiple regression with measurement error in a single regressor, we present some very simple formulas to assess the result that one may expect when correcting for measurement error. It is shown where the corrected estimated regression coefficients and the error variance may lie, and how the t-value behaves.

Keywords: attenuation, t-statistic, CALS

JEL classification: C21; C52

We thank Paul Bekker, Theo Dijkstra, and Ton Steerneman for their valuable comments.

* Department of Econometrics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands Tel.: +31 50 363 3793; fax: +31 50 363 3720; e-mail: e.meijer@eco.rug.nl.

1 Introduction

In the applied econometrics literature of the cross-sectional type, one often meets the case of a regression equation being estimated where the regressor set includes a number of control variables to account for population heterogeneity in addition to one regressor focal to the topic under study. This variable is often theory-based and is not directly observable, for conceptual or practical reasons. Examples are, among many others, consumption models relating consumption to permanent income, labor supply models relating hours to wage per hour, models assessing the policy impact of central bank independence, and investment models relating investment to Tobin's q. The case of multiple regression with measurement error in a single but interesting regressor may hence be considered almost generic.

It is generally appreciated that in such cases OLS gives an inconsistent result. Most textbooks discuss this and come up with solutions. For example, when the measurement error variance is known, the OLS results can be adapted to construct a consistent estimator. However, the approach most frequently mentioned is the use of instrumental variables. Instruments can sometimes be found by further scrutinizing the available data set but they can also be constructed from the data already employed for estimating the model. For example, when the variables are non-normal, the model $y = x\beta + u$, with x subject to measurement error, can be consistently estimated with $x * y$ as instrument, as was shown by Lewbel (1997), who elaborated on this simple idea to suggest a wide set of IV's.

Be it as it is, before entering on such a course it may be worthwhile to know what to expect when applying methods that correct for measurement error in a single regressor in a multiple-regression setting; see e.g. Wansbeek and Meijer (in press) for an overview. To that end we group, in this note, a number of partly new results that are very simple to apply and that are intended to provide guidance to the applied researcher. These results build on the output of OLS estimation of the model and show how the results change when different values of the measurement error variance (ϕ) are considered.

Throughout, we will assume that the measurement error is stochastically independent from the value of the underlying explanatory variable, from the other explanatory variables, and from the equation error. This is known as the standard errors-in-variables situation. Krasker and Pratt (1986, 1987) showed that if this is not the case, then even in the limit we can frequently not be sure of the signs of regression coefficients.

The set-up is as follows. After summarizing, in section 2, the basic results for regression with measurement error in multiple regression, we narrow the results down for the case of measurement error in a single variable in section 3. It is shown where the estimate of the regression coefficients and of the residual error variance may lie depending on ϕ. The situation is depicted graphically in section 4. A consistent estimator of the asymptotic variance is given in section 5.

As stated above, the mismeasured variable is often the core variable in the research, and that makes an investigation of the correction procedure for its coefficient estimate especially interesting. As section 6 shows, increasing values of ϕ correspond with increasing values of the estimate of its asymptotic variance. This increase outpaces the increase in coefficient estimate and the t-value is monotonously decreasing.

The relationship between the measurement error variance and the t-value is given explicitly below. This relationship can be of direct use in applied work, where t-values are assigned a dominant role, since the impact of ϕ on the t-value can be assessed directly. In particular, it can be seen at what level of noise in a variable it 'disappears into insignificance'.

2 Properties of the measurement error model

The standard linear multiple regression model can be written as

$$y = \Xi\beta + \varepsilon, \tag{1}$$

where y is an observable N-vector and ε an unobservable N-vector of random variables, assumed i.i.d. with zero expectation and variance σ_ε^2. The g-vector β is fixed but unknown. The $N \times g$-matrix Ξ contains the regressors, assumed independent of ε.

If there are errors of measurement in the explanatory variable, Ξ is not observable. Instead, we observe the matrix $X = \Xi + V$, where V $(N \times g)$ is a matrix of measurement errors. Its rows are assumed to be i.i.d. with zero expectation and covariance matrix Ω $(g \times g)$ and uncorrelated with Ξ and ε.

Define the sample second-moment matrices of Ξ and X:

$$K_N \equiv \frac{1}{N}\Xi'\Xi; \qquad A_N \equiv \frac{1}{N}X'X.$$

Note that A_N is observable but K_N is not.

We can interpret (1) in two ways. It is either a *functional* or a *structural* model. Under the former interpretation, we do not make explicit assumptions regarding the distribution of Ξ, but consider its elements as unknown fixed parameters. Under the

latter interpretation, the elements of Ξ are supposed to be random variables. The assumption $\operatorname{plim} K_N = K$, with K a positive definite $g \times g$-matrix, is meant to cover both cases. As a consequence, $A \equiv \operatorname{plim} A_N = K + \Omega$.

Let $b \equiv (X'X)^{-1}X'y$ and $s_\varepsilon^2 \equiv \frac{1}{N}y'(I_N - X(X'X)^{-1}X')y$ be the usual estimators of β and σ_ε^2, neglecting measurement error. As is well-known, these estimators are inconsistent:

$$\begin{aligned}\operatorname{plim} b &= A^{-1}(A - \Omega)\beta \\ \operatorname{plim} s_\varepsilon^2 &= \sigma_\varepsilon^2 + \beta'\Omega(\operatorname{plim} b).\end{aligned}$$

If Ω is known, Slutsky's theorem implies that

$$\hat{\beta} \equiv (A_N - \Omega)^{-1}A_N b \tag{2}$$

$$\hat{\sigma}_\varepsilon^2 \equiv s_\varepsilon^2 - \hat{\beta}'\Omega b \tag{3}$$

are consistent. The asymptotic distribution of these consistent estimators for both the functional and the structural model is given by

$$\sqrt{N}\begin{bmatrix} \hat{\beta} - \beta \\ \hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2 \end{bmatrix} \xrightarrow{d} \mathcal{N}_{g+1}\left(0, \begin{bmatrix} \gamma K^{-1}AK^{-1} + \omega\omega' & -2\gamma\omega \\ -2\gamma\omega' & 2\gamma^2 \end{bmatrix}\right), \tag{4}$$

where $\gamma \equiv \sigma_\varepsilon^2 + \beta'\Omega\beta$ and $\omega \equiv K^{-1}\Omega\beta$, cf. Kapteyn and Wansbeek (1984).

3 Measurement error in a single regressor

We now consider the case where there is measurement error in a single regressor only. Without loss of generality we take this to be the first regressor. Then

$$\Omega = \phi e_1 e_1', \tag{5}$$

with e_1 the first unit vector, and ϕ the variance of the measurement error in the first regressor. To elaborate this case, the following notation is used:

$$\begin{aligned} a &\equiv A_N^{-1}e_1 \\ \alpha &\equiv e_1'A_N^{-1}e_1 \\ \theta &\equiv \frac{1}{1 - \phi\alpha}. \end{aligned}$$

Hence $A_N a = e_1$ and

$$\theta = 1 + \theta\phi\alpha \tag{6a}$$

$$\theta^2 = 1 + \theta\phi(\theta + 1)\alpha \tag{6b}$$

$$I_g + \theta\phi a e_1' = (A_N - \phi e_1 e_1')^{-1} A_N \tag{6c}$$

$$\theta a = (A_N - \phi e_1 e_1')^{-1} e_1; \tag{6d}$$

(6c) and (6d) follow by multiplying both sides by $A_N - \phi e_1 e_1'$. We assume that the data relate to the values of ϕ deemed relevant such that $\theta > 0$; this holds anyhow in the limit.

Substitution of (5) in (2) using (6c) gives

$$\begin{aligned}\hat{\beta} &= (A_N - \phi e_1 e_1')^{-1} A_N b = (I_g + \theta\phi a e_1')b \\ &= b + (\theta\phi b_1)a = b + \lambda a,\end{aligned} \tag{7}$$

with b_1 the first element of b and

$$\lambda \equiv \theta\phi b_1. \tag{8}$$

In particular, the first element $\hat{\beta}_1$ of $\hat{\beta}$ can be written as

$$\hat{\beta}_1 = (1 + \theta\phi\alpha)b_1 = \theta b_1. \tag{9}$$

We assume $b_1 > 0$ without loss of generality, hence $\hat{\beta}_1 \geq 0$. So (9) gives the correction for the downward bias when estimating β_1 by OLS if there is measurement error, and (7) shows, for all elements of β jointly, that this correction is along a line in β-space, from b in the direction given by the first column of the inverse of $X'X$.

We now consider the estimation of σ_ε^2. Combining (3), (5), (8), and (9) gives

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 &= s_\varepsilon^2 - \phi\hat{\beta}' e_1 e_1' b = s_\varepsilon^2 - \phi\hat{\beta}_1 b_1 \\ &= s_\varepsilon^2 - \phi\theta b_1^2 = s_\varepsilon^2 - \lambda b_1\end{aligned} \tag{10}$$

as a consistent estimator given a value for ϕ.

Restricting this estimator to nonnegative values imposes an upper limit on the values of ϕ that one may wish to consider. Putting (10) to zero and solving gives

$$\phi_{\max} = \frac{1}{\alpha + b_1^2/s_\varepsilon^2} \tag{11}$$

since then

$$\theta_{\max} = \frac{1}{1 - \phi_{\max}\alpha} = 1 + \frac{\alpha}{b_1^2/s_\varepsilon^2},$$

hence

$$\lambda_{\max} = \theta_{\max}\phi_{\max}b_1 = \frac{s_\varepsilon^2}{b_1}, \tag{12}$$

which solves (10). This gives

$$\hat{\beta}_{1\,\max} = \theta_{\max}b_1 = b_1 + \frac{\alpha}{b_1}s_\varepsilon^2$$

as the upper bound on the estimator of β_1.

4 A graphical illustration

Adapting from Bekker, Kapteyn, and Wansbeek (1984) we can illustrate the above graphically. Let us first consider the case of general $\Omega \geq 0$, and assume that the values considered for Ω satisfy $\Omega < A_N$. (Again, this holds anyhow in the limit.) Hence

$$(A_N - \Omega)^{-1} - A_N^{-1} \geq 0, \tag{13}$$

with equality holding only if $\Omega = 0$. An implication of (2) is

$$\hat{\beta}' A_N b = b' A_N (A_N - \Omega)^{-1} A_N b.$$

Then, on using (13),

$$(\hat{\beta} - b)' A_N b = b' A_N \left((A_N - \Omega)^{-1} - A_N^{-1}\right) A_N b \geq 0,$$

with equality holding only if $b = 0$. So, whatever Ω may be, the corrected estimators lie in β-space beyond the plane $(\hat{\beta} - b)' A_N b = 0$, as seen from the origin. This is the plane through b perpendicular to $A_N b$.

We now consider the implication of $\hat{\sigma}_\varepsilon^2 \geq 0$ or

$$\hat{\beta}'\Omega b \leq s_\varepsilon^2, \tag{14}$$

cf. (3). Rewriting (2) as $\hat{\beta}'\Omega = (\hat{\beta} - b)' A_N$ and substituting this in (14) gives

$$(\hat{\beta} - b)' A_N b \leq s_\varepsilon^2. \tag{15}$$

So the measurement-error corrected estimator cannot lie beyond the plane $(\hat{\beta} - b)' A_N b = s_\varepsilon^2$. This plane is parallel to the plane $(\hat{\beta} - b)' A_N b = 0$, further away from the origin. The situation is illustrated in figure 1. The set of $\hat{\beta}$'s compatible with $\Omega = \phi e_1 e_1'$ is given by the line segment between the two planes, ranging from b to $b + \lambda_{\max} a$.

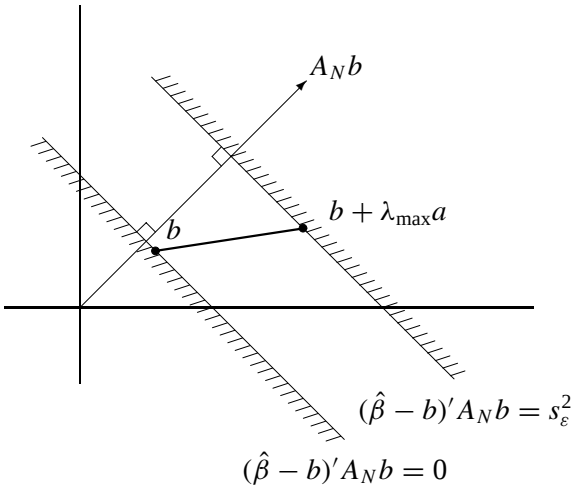

Figure 1: Admissible values of $\hat{\beta}$.

5 Estimating the asymptotic variance

For Ω as in (5) with given ϕ, we can derive the asymptotic variance of the estimators (9) and (10) by elaborating (4) for this special case. More interestingly, we elaborate consistent estimation of this variance for $\hat{\beta}$ based on the consistent estimators for β and σ_ε^2. We consider the various terms in (4) in turn.

First, using (6a) and (8), a consistent estimator of $\gamma = \sigma_\varepsilon^2 + \beta'\Omega\beta = \sigma_\varepsilon^2 + \phi\beta_1^2$ is given by

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 + \phi\hat{\beta}_1^2 &= s_\varepsilon^2 - \lambda b_1 + \theta^2\phi b_1^2 = s_\varepsilon^2 - \lambda b_1 + \theta\lambda b_1 \\ &= s_\varepsilon^2 + (\theta - 1)\lambda b_1 = s_\varepsilon^2 + \theta\phi\alpha\lambda b_1 = s_\varepsilon^2 + \lambda^2\alpha.\end{aligned}$$

Next, using (6d) and (9), $\omega = K^{-1}\Omega\beta = \phi(A - \phi e_1 e_1')^{-1} e_1 e_1'\beta$ is consistently estimated by

$$\phi(A_N - \phi e_1 e_1')^{-1} e_1 e_1'\hat{\beta} = \theta\phi\hat{\beta}_1 a = \theta^2\phi b_1 a = \theta\lambda a,$$

and, using (6a)—(6c), $K^{-1}AK^{-1} = (A - \phi e_1 e_1')A(A - \phi e_1 e_1')$ is estimated by

$$\begin{aligned}(A_N - \phi e_1 e_1')^{-1} A_N (A_N - \phi e_1 e_1')^{-1} &= (I_g + \theta\phi a e_1')A_N^{-1}(I_g + \theta\phi a e_1') \\ &= A_N^{-1} + \theta\phi(\theta + 1)aa'.\end{aligned}$$

Putting these results together gives for $\hat{\beta}$:

$$\widehat{(\text{asy.var.}\hat{\beta})} = (s_\varepsilon^2 + \lambda^2\alpha)(A_N^{-1} + \theta\phi(\theta+1)aa') + \theta^2\lambda^2 aa', \tag{16}$$

which of course reduces to $s_\varepsilon^2 A_N^{-1}$ when there is no measurement error.

6 The t-value for the first regression coefficient

For the estimator of the coefficient of the first regressor, taking the upper-left element in (16) gives

$$\begin{aligned}\widehat{(\text{asy.var.}\hat{\beta}_1)} &= (s_\varepsilon^2 + \lambda^2\alpha)(\alpha + \theta\phi(\theta+1)\alpha^2) + \theta^2\lambda^2\alpha^2 \\ &= \theta^2\alpha(s_\varepsilon^2 + 2\lambda^2\alpha).\end{aligned}$$

The t-statistic if there is no measurement error is

$$t_0 = \frac{b_1\sqrt{N}}{\sqrt{s_\varepsilon^2\alpha}},$$

so

$$s_\varepsilon^2 = \frac{b_1^2 N}{t_0^2\alpha}. \tag{17}$$

The t-statistic corresponding with a measurement error of size ϕ is, using (9) and (17), given by

$$\begin{aligned}t_\phi &= \frac{\hat{\beta}_1\sqrt{N}}{\sqrt{\widehat{(\text{asy.var.}\hat{\beta}_1)}}} \\ &= \frac{\theta b_1\sqrt{N}}{\theta\sqrt{\alpha}\sqrt{s_\varepsilon^2 + 2\lambda^2\alpha}} \\ &= \frac{t_0}{\sqrt{1 + \frac{2}{N}h_\phi^2}},\end{aligned} \tag{18}$$

with $h_\phi \equiv \lambda\alpha t_0/b_1 = (\theta-1)t_0$. After some straightforward calculations there appears to hold

$$\frac{\partial t_\phi}{\partial\phi} = -\frac{2}{N}\phi\left(\theta t_\phi\right)^3\alpha^2 < 0.$$

So t_ϕ has derivative 0 in $\phi = 0$ and $-\alpha\sqrt{N/2}$ in $\phi = 1/\alpha$, and decreases monotonically. Figure 2 illustrates the behavior of t_ϕ , for $\alpha = 1$ and $N = 100$, for various values of t_0. Substitution of (17) in (11) gives

$$\phi_{\max} = \frac{1}{\alpha\left(1 + t_0^2/N\right)},$$

so the relevant parts of the curves in Figure 2 end to the left of $\phi = 1/\alpha = 1$, and the minimal t-values are above 0.

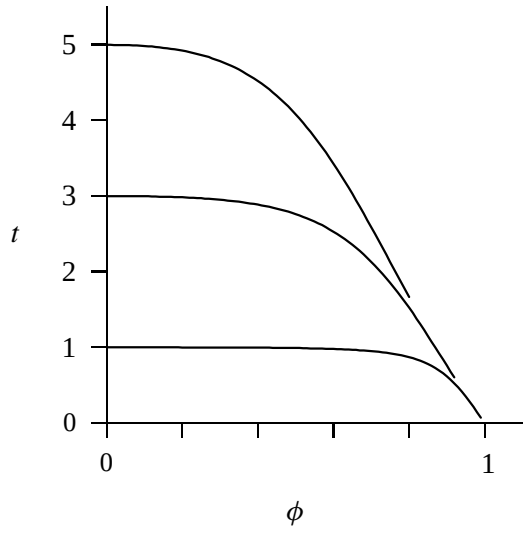

Figure 2: t_ϕ as a function of ϕ and t_0.

Sometimes it is of interest to see which level of ϕ corresponds with a given t-value. Let this value be t_*. This corresponds with $\phi = \phi_*$ satisfying

$$t_* = \frac{t_0}{\sqrt{1 + \frac{2}{N}h^2_{\phi_*}}}, \tag{19}$$

cf. (18). Let

$$w \equiv \left\{\frac{N}{2}\left(\left(\frac{t_0}{t_*}\right)^2 - 1\right)\right\}^{1/2},$$

then (19) implies $h_{\phi_*} = w$ or

$$\phi_* = \frac{1}{\alpha}\frac{w}{w + t_0}.$$

In particular, one may be interested to see, by putting t_* at the conventional level of 2, what level of noise in a variable makes its estimated coefficient disappear into insignificance.

References

Bekker, P. A., Kapteyn, A., and Wansbeek, T. J. (1984). Measurement error and endogeneity in regression: bounds for ML and 2SLS estimates. In T. K. Dijkstra (Ed.), *Misspecification analysis* (pp. 85–103). Berlin: Springer.

Kapteyn, A., and Wansbeek, T. J. (1984). Errors in variables: Consistent Adjusted Least Squares (CALS) estimation. *Communications in Statistics — Theory and Methods*, *13*, 1811–1837.

Krasker, W. S., and Pratt, J. W. (1986). Bounding the effects of proxy variables on regression coefficients. *Econometrica*, *54*, 641–655.

Krasker, W. S., and Pratt, J. W. (1987). Bounding the effects of proxy variables on instrumental-variables coefficients. *Journal of Econometrics*, *35*, 233–252.

Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R & D. *Econometrica*, *65*, 1201–1213.

Wansbeek, T. J., and Meijer, E. (in press). Measurement error and latent variables. In B. H. Baltagi (Ed.), *Companion in theoretical econometrics.* Oxford, UK: Basil Blackwell.