Journal of Econometrics 28 (1985) 223–230. North-Holland

THE SENSITIVITY OF MLE TO MEASUREMENT ERROR*

David LEVINE

University of California, Los Angeles, CA 90024, USA

Received July 1983, final version received November 1984

The paper explains how to compute a simple summary measure of the sensitivity of maximum likelihood and related non-linear estimates to measurement error in exogenous variables. The proposed measure is a first-order approximation, and its implications for probit and censorship type models are shown to be quite different from ordinary least squares.

1. Introduction

In empirical studies the replacement of theoretical variables with proxies can result in measurement error. Frequently, models are estimated under the assumption of no measurement error in the hope that the resulting errors in inference will be small. It is also important to report how sensitive the estimator is to measurement error – how large is the asymptotic bias under different assumptions about the magnitude of the error and in what direction is the estimator biased?

A useful summary measure that answers both of these questions is the derivative of the asymptotic bias with respect to the variance of the measurement error, evaluated at zero variance. Section 2 of this paper discusses how this derivative can be computed, and why this approach is more tractible than attempting to re-estimate the model explicitly allowing for measurement error. Section 3 specializes to location/scale parameter models, and points out that in models such as probit and normal censorship, unlike the normal linear model, the coefficient of a variable measured with error may *not* be biased down in absolute value. Section 4 analyzes the quality of the approximate correction for bias in a one-variable regression model.

2. A sensitivity measure

Suppose that the probability density of an endogenous variable y conditional on a parameter vector θ, an exogenous variable x^* and other exogenous

*I am grateful to Robert Engle, Clive Granger, Jonathan Hilton, Jerry Hausman, Daniel McFadden, Paul Ruud, Halbert White and two anonymous referees for helpful comments and criticism.

variables z is

$$f(y|\theta, x^*, z). \tag{1}$$

The log-likelihood function is defined as

$$L(\theta, x^*) \equiv \log f(y|\theta, x^*, z), \tag{2}$$

where for notational simplicity y and z are suppressed.

In practice x^* is often not observed but is replaced with a proxy $x = x^* + \sqrt{\lambda}\,\eta$, where η is a random measurement error independent of y, x^* and z with zero mean $\mathrm{E}\eta = 0$ and variance normalized to equal that of x so that $\operatorname{var}(\eta) = \operatorname{var}(x) = \sigma_x^2$. Thus λ is the fraction of the variance of x attributable to measurement error. All random variables are sampled independently from a stationary distribution. The process generating the observable x is taken to be fixed, as is the process generating η. However, the process generating the unobservable x^* necessarily depends on λ; in particular $\operatorname{var}(x^*) = (1-\lambda)\sigma_x^2$. Naturally, when $\lambda = 0$, x^* almost surely equals x.

Throughout this paper I shall use a number of regularity assumptions. These hold for all $0 \leq \lambda < 1$ and all θ in a compact convex parameter space:

(A.1) L and its derivatives to third order with respect to θ and x^* are absolutely integrable.

(A.2) L and its derivatives to third order have bounded absolute moments of some order strictly greater than two.

(A.3) For each λ there is a unique $\theta(\lambda)$ strictly interior to the parameter space which maximizes $\mathrm{E}L(\theta, x^*)$.

(A.4) $I \equiv -\mathrm{E}L_{\theta\theta}(\theta(\lambda), x^*)$ is non-singular.

(A.5) $\mathrm{E}|\eta|^3$ is finite.

Throughout the paper subscripts denote differentiation. Assumption (A.1) guarantees the differentiation and integration can be exchanged when required. Assumption (A.2) guarantees that sample moments almost surely converge uniformly to expectations (and that they are asymptotically normal, although this is not used). Assumption (A.3) is a global identification condition, while (A.4) is a local identification requirement.

If x^* is observed and λ is known, we can form the maximum likelihood estimate θ^{N*} which maximizes $L^{N*}(\theta) \equiv (1/N)\sum_{n=1}^{N} L(\theta, x^{n*})$. The moment assumption (A.2) guarantees that $L^{N*}(\theta)$ almost surely converges uniformly to $\mathrm{E}L(\theta, x^*)$ and thus that θ^{N*} almost surely converges to $\theta(\lambda)$. Furthermore the global identification assumption (A.3) guarantees that $\theta(\lambda)$ is the 'true' parameter vector that generated the data. This is well known and proofs can be found (for example) in Wald (1949) or Frydman (1980).

If we replace the true variable x^* with the proxy x, we can form the quasi-maximum likelihood estimate θ^N which maximizes $L^N(\theta) \equiv (1/N)\sum_{n=1}^N L(\theta, x^n)$. By an analogous argument to the one above, which is detailed in White (1982), θ^N almost surely converges to $\theta(0)$. Unfortunately we are interested in drawing inferences about $\theta(\lambda)$ and, for $\lambda \neq 0$, $\theta(\lambda) \neq \theta(0)$ in general. This is the classical errors-in-variables problem.

There are a variety of approaches to this problem. Theil (1957) gives a very general analysis of specification error in the normal linear model. If additional proxies are available instrumental variables is a possibility. Alternatively we may try to bound $\theta(\lambda)$ as Reiersol (1950) suggests, although this appears impractical in a non-linear setting. With specific assumptions about the form of the distributions of x^* and η, Madansky (1959) has shown that λ may actually be identifiable. In this case the model may be estimated by maximizing either the likelihood conditional on x or the unconditional joint likelihood. Since x is not weakly exogenous, Engle, Hendry and Richard (1983) show that the latter procedure is better. Unfortunately we typically have little confidence in the assumptions about x^* and η required to identify the model, and in any case the computational difficulty involved is great.

If the proxy is very bad (λ is large) it is unlikely that any procedure will yield very reliable inferences about $\theta(\lambda)$ in a finite sample. Typically, however, we use a proxy because *a priori* we think λ is small. In this case it makes sense to think that $\theta(0)$ will be close to $\theta(\lambda)$ and quasi-maximum likelihood estimation is used. A useful supplement to this is sensitivity analysis – how much and how does $\theta(\lambda)$ change when λ is perturbed slightly from zero? This paper provides that analysis.

For small λ we have

$$\theta(\lambda) \approx \theta(0) + \lambda\theta_\lambda(0), \tag{3}$$

where θ_λ is the derivative of θ with respect to λ. Thus $\theta_\lambda(0)$ can be used as a correction factor to approximately correct the estimates derived by replacing x^* with x. This quantity has the advantage that it is easy to compute (and estimate in finite samples) and is easily interpreted by the consumer of empirical work. It is also (as we are about to show) independent of the process generating x^*. This approach is very similar to that taken by Keifer and Skoog (1982) in analyzing omitted variables in non-linear models, and Yatchew and Griliches (1979) in analyzing probit models.

The differentiability assumption (A.1) implies that $\theta(\lambda)$ satisfies the normal equations

$$\mathrm{E}L_\theta(\theta(\lambda), x^*) = 0.$$

Thus by the implicit function theorem

$$\theta_\lambda(0) = -\left[\frac{\partial \mathrm{E}L_\theta}{\partial\theta}\right]^{-1}_{\lambda=0}\left[\frac{\partial \mathrm{E}L_\theta(\theta(0), x^*)}{\partial\lambda}\right]_{\lambda=0}$$

$$= I^{-1}\left.\frac{\partial \mathrm{E}L_\theta(\theta(0), x^*)}{\partial\lambda}\right|_{\lambda=0}. \tag{4}$$

Note that as λ changes so does the distribution of x^*. To compute $\partial\mathrm{E}L_\theta/\partial\lambda$, observe that since $x = x^* + \sqrt{\lambda}\,\eta$,

$$\mathrm{E}L_\theta(\theta(0), x^* + \sqrt{\lambda}\,\eta) = 0, \tag{5}$$

independent of the value of λ. Thus if we differentiate (5) with respect to λ at $\lambda = 0$, we find

$$\left.\frac{\partial \mathrm{E}L_\theta(\theta(0), x^*)}{\partial\lambda}\right|_{\lambda=0} = -\left.\frac{\partial \mathrm{E}L_\theta(\theta(0), x^0 + \sqrt{\lambda}\,\eta)}{\partial\lambda}\right|_{\lambda=0}, \tag{6}$$

provided the right-hand side exists. Here x^0 is a random variable independent of η which is drawn from the x^* distribution (or more accurately the joint distribution of x^* and z) when $\lambda = 0$. In other words, on the left-hand side we let the distribution of x^* vary with λ and hold the weight on η fixed at zero; on the right-hand side we hold the distribution of x^* fixed at $\lambda = 0$ and allow the weight on η to vary. To evaluate the right side of (6) we use a Taylor series expansion to find for an appropriate choice of the matrix $\bar{L}_{\theta xxx}$,

$$\frac{\partial}{\partial\lambda}\mathrm{E}L_\theta(\theta(0), x^0 + \sqrt{\lambda}\,\eta) \tag{7}$$

(A) $\quad = \dfrac{\partial}{\partial\lambda}\mathrm{E}\{L_\theta(\theta(0), x^0)$

(B) $\quad + L_{\theta x}(\theta(0), x^0)\sqrt{\lambda}\,\eta$

(C) $\quad + L_{\theta xx}(\theta(0), x^0)(\lambda/2)\eta^2$

(D) $\quad + \bar{L}_{\theta xxx}(\lambda^{3/2}/6)\eta^3\}$

(E) $\quad = \tfrac{1}{2}\mathrm{E}L_{\theta xx}(\theta(0), x^0)\sigma_x^2.$

Line (A) vanishes by assumption (A.3), line (B) by $\mathrm{E}\eta = 0$, line (C) and (E) are the same since $\mathrm{E}\eta^2 = \sigma_x^2$, and line (D) vanishes sinces $\lambda^{3/2}/\lambda = \lambda^{1/2}$ vanishes at

$\lambda = 0$ while $\bar{L}_{\theta xxx}$ remains bounded by Taylor's theorem. Thus since x^0 and x are almost surely the same, we have computed

$$\theta_\lambda(0) = -\tfrac{1}{2} I^{-1} \mathrm{E} L_{\theta xx}(\theta(0), x)\sigma_x^2. \tag{8}$$

Since by (A.2) the sample moments for I and $L_{\theta xx}$ converge uniformly in probability to the true moments, it follows from Amemiya's lemma and plim $\theta^n = \theta(0)$ that if $\hat{\sigma}_x^2$ is consistent for σ_x^2 then the estimator

$$\theta_\lambda^N = \frac{1}{2}\left[\frac{\sum_{n=1}^{N} L_{\theta\theta}(\theta^N, x^n)}{N}\right]^{-1}\left[\frac{\sum_{n=1}^{N} L_{\theta xx}(\theta^N, x^n)}{N}\right]\hat{\sigma}_x^2 \tag{9}$$

is consistent for $\theta_\lambda(0)$.

It should be noted that the preceding derivation applies not only to MLE, but to any estimator defined by equating sample moments of functions of the data and parameters to zero: non-linear least squares, NL2SLS and NL3SLS all have this form.

3. Location/scale parameter models

Now consider the special case of a location/scale parameter model in which the log-likelihood function is

$$L(\beta, \sigma, Z) = -\log \sigma - H(Z\beta/\sigma). \tag{10}$$

Here $\theta = (\beta, \sigma)$ where β is a k-vector of slope coefficients, σ is a scale parameter, and Z is k-dimensional row vector of exogenous variables. The first variable Z^1 is presumed to be measured with error. The normal linear model, probit, logit and the censored normal linear model all have likelihood functions of this form.

Define weights

$$\begin{aligned} W^0 &\equiv \sigma_x^2 \mathrm{E} H''(Z\beta/\sigma)/\sigma^2, \\ W^j &\equiv \sigma_x^2 \beta^1 \mathrm{E} Z^j H'''(Z\beta/\sigma)/2\sigma^3, \qquad j = 1, \ldots, k, \\ W^{k+1} &\equiv -\sigma_x^2 \beta^1 \{\mathrm{E} Z\beta H'''(Z\beta/\sigma) + 2\sigma^3 W^0\}/2\sigma^4. \end{aligned} \tag{11}$$

Let σ_{ij} be the asymptotic standard errors of maximum likelihood without measurement error: the entries in the matrix $-I^{-1}$. Algebraic manipulation of

(8) then shows that the correction factor for β^i is

$$\beta^i_\lambda = \beta^1 \left[W^0 \sigma_{1i} + \sum_{j=1}^{k} W^j \sigma_{ij} + W^{k+1} \sigma_{i,k+1} \right]. \tag{12}$$

The first term in (12) $[\beta^1 W^0 \sigma_{1i}]$ should be thought of as the 'normal' effect. In the normal linear model $W^j = 0$ for $j = 1, \ldots, k$ and $\sigma_{i,k+1} = 0$, so only this term matters. Also, $W^0 = \sigma_x^2/\sigma^2$, so β^1 is biased down in absolute value, and other coefficients are biased up or down depending on their (asymptotic) correlation with β^1. In non-normal models with a constant term $W^0 > 0$ is part of the second-order conditions for a maximum, so the first term again tends to bias β^1 down to absolute value.

The second term in (12) $[\sum_{j=1}^{k} \beta^1 W^j \sigma_{ij}]$ should be thought of as the 'non-linear' effect. In the normal linear model, the normal equations for $\beta^2, \ldots, \beta^k$ are linear in Z^1 and are thus unaffected by measurement error which operates through the *second* derivative $(L_{\theta xx})$ of the normal equations with respect to the proxy. In non-normal cases, the normal equations are non-linear in Z^1 and thus *are* affected by measurement error. The second term measures the consequences of this effect.

The third term in (12) $[\beta^1 W^{k+1} \sigma_{i,k+1}]$ should be thought of as the scale effect. Measurement error significantly biases estimation of the scale parameter σ since random variation in the endogenous variable is confounded with measurement error. In the normal model, block diagonality insures that $\sigma_{i,k+1} = 0$ – that failure to estimate σ correctly doesn't affect estimates of the slope parameters. Otherwise, when $\sigma_{i,k+1} \neq 0$, the error in estimating σ feeds back to bias the slope parameters. In censorship models estimates of slope parameters hinge critically on the estimated scale parameter and the third term is a potentially serious source of error.

In OLS the coefficient of the proxy is biased down in absolute value as are positively correlated coefficients with the same sign; in general, the direction the estimate must be adjusted is the sign of the coefficient of the proxy times the sign of the correlation with the proxy. This result, on which so much of our intuition is based, is *wrong* in non-normal models. As shown, there are two additional effects – the non-linear effect and the scale effect – which must be consi lered to sign the bias due to measurement error.

4. Simple regression

The adjustment factor $\theta_\lambda(0)$ in (8) and (12) enables an *approximate* adjustment to the estimator $\theta(0)$ derived by replacing x^* with x in the likelihood function. How good is this approximation? In the case of one variable regression, an exact correction can be computed to compare with the approximation.

Suppose that the endogenous variable is generated by

$$y = \theta(\lambda)x^* + \varepsilon, \tag{13}$$

where $\varepsilon \sim N(0, \sigma^2)$ and $x^* \sim N(0, (1-\lambda)\sigma_x^2)$. The estimator derived by doing OLS using $x = x^* + \sqrt{\lambda}\,\eta$ in place of x^* is

$$\theta(0) = \mathrm{E}xy/\mathrm{E}x^2. \tag{14}$$

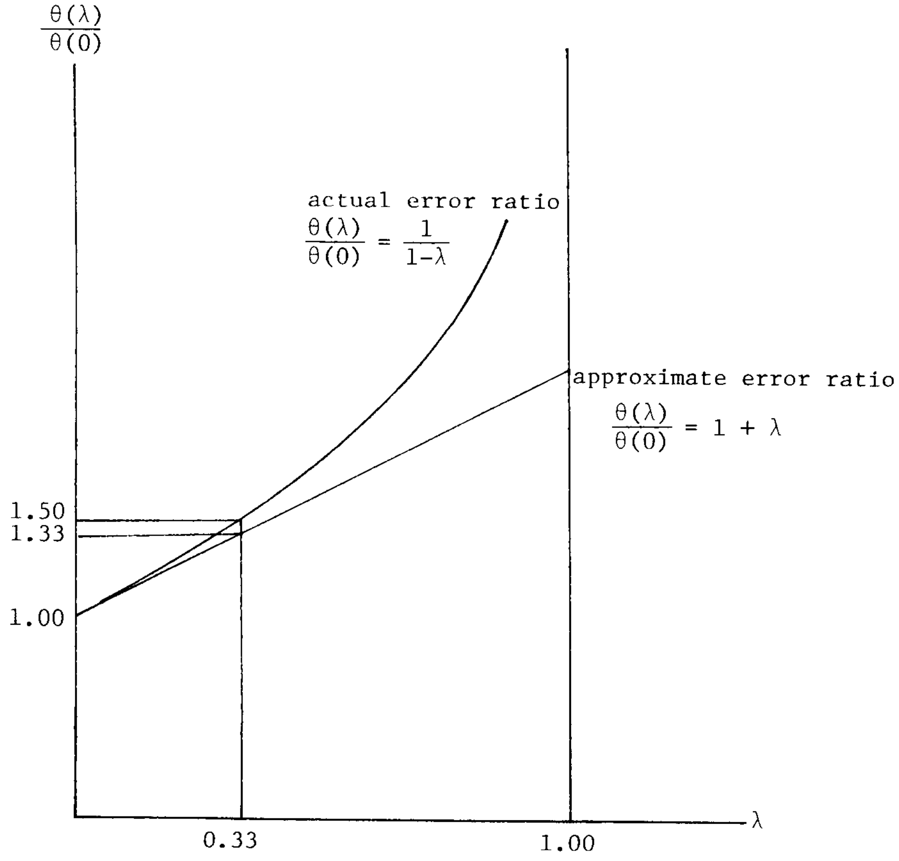

Fig. 1. Approximation error in simple regression.

A direct computation shows that

$$\theta(\lambda) = \theta(0)[1-\lambda]^{-1}, \tag{15}$$

while the approximate value $\theta^a(\lambda)$ computed as $\theta^a(\lambda) = \theta(0) + \lambda\theta_\lambda(0)$ is computed from (12) as

$$\theta^a(\lambda) = \theta(0)[1+\lambda]. \tag{16}$$

Naturally (16) is simply the tangent line to (15) at $\lambda = 0$. As sketched in fig. 1, the quality of the approximation depends on λ – the fraction of the variance of x accounted for by measurement error. If the variance of x is almost entirely due to measurement error, the approximation is quite bad. However, even with a third of the variance of x due to measurement error, the approximation eliminates two-thirds of the bias.

References

Engle R., D. Hendry and J. Richard, 1983, Exogeneity, Econometrica, March.

Frydman, R., 1980, A proof of the consistency of maximum likelihood estimators of non-linear regression models with autocorrelated errors, Econometrica, May.

Keifer, N. and G. Skoog, 1982, Local asymptotic specification error analysis (Cornell and Chicago).

Madansky, A., 1959, The fitting of straight lines when both variables are measured with error, Journal of the American Statistical Association.

Reiersol, O., 1950, Identifiability of a linear relationship between variables which are subject to error, Econometrica, 315–389.

Theil, H., 1957, Specification errors and the estimation of economic relationships, Review of the International Statistics Institute.

Wald, A., 1949, A note on the consistency of MLE, Annals of Mathematical Statistics.

White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica, Jan.

Yatchew, A. and Z. Griliches, 1979, Specification errors in probit and logit models, Harvard Institute of Economics discussion paper no. 717.