# The Sensitivity of MLE to Measurement Error[*]

by

David Levine

UCLA
Department of Economics
Working Paper #251
June 1982

[*] I must thank Jonathan Hamilton, Jerry Hausman and Paul Ruud.

# The Sensitivity of MLE to Measurement Error*

by

David Levine

November 1981

951

*I must thank Jonathan Hamilton, Jerry Hausman and Paul Ruud.

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 two 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 three 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 <u>not</u> be biased down in absolute value. Section four 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 $\theta$, an exogenous variable x* and other exogenous variables z is

$$(2.1) \qquad f(y|\theta,x^*,z).$$

The log-likelihood function is defined as

$$(2.2) \qquad L(\theta,x^*) \equiv \log f(y|\theta,x^*,z)$$

where for notational simplicity y and z are supressed. I shall suppose that L satisfies two regularity conditions:

(A1) L and its derivatives to third order with respect to $\theta$ and x* are absolutely integrable;

(A2) $I \equiv -EL_{\theta\theta}(\theta_o,x)$ is non-singular.

Here $\theta_o$ is the actual parameter value generating the data, while subscripts denote differentiation. Assumption (A1) guarantees differentiation and integration can be exchanged when required, while (A2) insures that the model is locally identified. Under these assumptions $\theta_o$ is a locally unique solution of the normal equations

$$(2.3) \qquad EL_\theta(\theta_o,x^*) = 0$$

In practice, x* is often not observed, but replaced with a proxy $x=x^* + \sqrt{\lambda}\eta$. Here $\lambda$ is a non-negative constant and $\eta$ is a random variable independent of of y, x* and z, and normalized so that $E\eta=0$ and $E\eta^2=1$. I also assume that $E|\eta|^3$ exists.

The case I wish to consider is when the scale factor $\lambda$ is a priori believed small. The problem is one of sensitivity analysis: how do inferences about $\theta_o$ depend on $\lambda$? One solution is to assume that a specific function $g(x^*|x,\lambda,z)$ is the probability density of x* conditional on the proxy x, the scale factor $\lambda$, and z. In this case $\theta_o$ may be found by maximizing

$$(2.4) \qquad E\log[\int f(y|\theta,x^*,z)g(x^*|x,\lambda,z)dx^*].$$

If $\lambda$ is large this is the only alternative, and indeed it is possible for some functions g that $\lambda$ is identifiable. However, we typically have limited confidence in the specification of g, and if the proxy is very noisy ($\lambda$ large) it is unlikely that we can draw reliable inferences about $\theta_o$ in a finite sample. Furthermore, maximizing (2.4) is often impractical.

If $\lambda$ is small, there is an alternative technique for drawing inferences about $\theta_o$. Let the function $\theta(\lambda)$ be implicitly defined as the solution to

$$(2.5) \qquad EL_\theta(\theta,x^*+\sqrt{\lambda}\eta) = 0,$$

that is, the estimator derived by replacing x* with the proxy x. As noted above, $\theta(0)=\theta_o$. Define the vector of derivatives

$$(2.6) \qquad \theta_\lambda \equiv -\frac{\partial\theta}{\partial\lambda}\bigg|_{\lambda=0}.$$

Then for small $\lambda$ $\theta(\lambda) \approx \theta_o-\lambda\theta_\lambda$, or equivalently

(2.7) $$\theta_o \approx \theta(\lambda) + \lambda\theta_\lambda .$$

Thus, $\theta_\lambda$ can be used as a correction factor to correct the estimates derived by replacing x* with the proxy x. This quantity has the advantage that it is easy to compute (and to estimate in finite samples) and is easily interpreted by the consumer of empirical work. Also, as we are about to show, it is independent of the specification of g.

To evaluate $\theta_\lambda$ apply the implicit function theorem to (2.5) to find

(2.8) $$\theta_\lambda \equiv - \frac{\partial\theta}{\partial\lambda}\bigg|_{\lambda=0}$$

$$= \left[\frac{\partial EL_\theta}{\partial\theta}\bigg|_{\lambda=0}\right]^{-1} \frac{\partial EL_\theta}{\partial\lambda}\bigg|_{\lambda=0}$$

$$= - I^{-1} \frac{\partial EL_\theta}{\partial\lambda}\bigg|_{\lambda=0.}$$

Then, to evaluate $\partial EL_\theta/\partial\lambda|_{\lambda=0}$ expand (2.5) in a Taylor series to find for an appropriate choice of $\bar{x}$

(2.9) $$\frac{\partial}{\partial\lambda} EL_\theta(\theta_o, x^* + \sqrt{\lambda}\eta)$$

$$= \frac{\partial}{\partial\lambda} E\Big\{ L_\theta(\theta_o, x^*) \tag{A}$$

$$+ L_{\theta x}(\theta_o, x^*)\sqrt{\lambda}\eta \tag{B}$$

$$+ L_{\theta xx}(\theta_o, x^*)\ (\lambda/2)\eta^2 \tag{C}$$

$$+ L_{\theta xxx}(\theta_o, \bar{x})\ (\lambda^{3/2}/6)\eta^3 \Big\} \tag{D}$$

$$= (1/2)\, E\, L_{\theta xx}(\theta_o, x^*) . \tag{E}$$

Line (A) vanishes by (1.3), line (B) by $E\eta=0$, line (C) and (E) are the same since $E\eta^2=1$ and line (D) vanishes since $d\lambda^{3/2}/d\lambda$ vanishes at $\lambda=0$. Thus, we have computed

$$(2.10) \qquad \theta_\lambda = -(1/2)I^{-1}\ EL_{\theta xx}(\theta_{o,}x^*).$$

In a finite sample (2.10) can be consistently estimated if $\hat{\theta}$ is a consistent root of the sample normal equations when $\lambda=0$ and the sample moments for I and $L_{\theta xx}$ converge uniformly in probability to the true moments. In this case, it follows from Amemiya's lemma that the estimator

$$(2.11) \qquad \hat{\theta}_\lambda = (1/2)\left[\frac{\overset{N}{\Sigma}L_{\theta\theta}(\hat{\theta},x_n)}{N}\right]^{-1}\left[\frac{\overset{N}{\Sigma}L_{\theta xx}(\hat{\theta},x_n)}{N}\right]$$

is consistent for $\theta_\lambda$.

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.

Finally, as a matter of reporting, it is sensible to report $\text{var}(x)\theta_\lambda/100$ which is scale free. When multiplied by the variance of measurement error as a percentage of the total variance of x, it gives the approximate correction to the coefficients. Since our priors are in terms of percent measurement error rather than absolute measurement error, this coefficient is easiest to interpret.

## 3. Location/Scale Parameter Models

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

$$(3.1) \qquad L(\beta,\sigma,Z) = -\log \sigma - H(Z\beta/\sigma).$$

Here $\theta = (\beta,\sigma)$ where $\beta$ is a k-vector of slope coefficients, $\sigma$ 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

$$(3.2) \qquad \begin{aligned} W_o &\equiv EH''(Z\beta/\sigma)/\sigma^2 \\ W_j &\equiv \beta_1 \, E \, Z_j H'''(Z\beta/\sigma)/2\sigma^3 \qquad j = 1,\ldots,k \\ W_{k+1} &\equiv -\beta_1 \left\{ EZ\beta H'''(Z\beta/\sigma) + 2\sigma^3 W_o \right\} / 2\sigma^4. \end{aligned}$$

Let $\sigma_{ij}$ be the asymptotic standard errors of maximum likelihood without measurement error: the entries in the matrix $-I^{-1}$. Algebraic manipulation of (2.10) then shows that the correction factor for $\beta_i$ is

$$(3.3) \qquad \beta_{\lambda i} = \beta_1[W_o\sigma_{1i} + \Sigma_{j=1}^{k} W_j\sigma_{ij} + W_{k+1}\sigma_{i,k+1}].$$

The first term in (3.3) $[\beta_1 W_o \sigma_{1i}]$ 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, $W_o = 1/\sigma^2$, so $\beta_1$ is biased down in absolute value, and other coefficients are biased up or down depending on their (asymptotic) correlation with $\beta_1$. In non-normal models with a constant term $W_o>0$ is part of the second order conditions for a maximum, so the first term again tends to bias $\beta_1$ down in absolute value.

The second term in (3.3) $[\Sigma_{j=1}^{k}\beta_{,}W_j\sigma_{ij}]$ 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 <u>second</u> 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 <u>are</u> affected by measurement error. The second term measures the consequences of this effect.

The third term in (3.3) $[\beta_1 W_{k+1}\sigma_{i,k+1}]$ should be thought of as the scale effect. Measurement error significantly biases estimation of $\sigma$ 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 $\sigma$ correctly doesn't affect estimates of the slope parameters. Otherwise, when $\hat{\sigma}_{i,k+1} \neq 0$, the error in estimating $\sigma$ 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 <u>wrong</u> in non-normal models. As shown, there are two additional effects -- the non-linear effect and the scale effect -- which must be considered to sign the bias due to measurement error.

## 4. Simple Regression

The adjustment factor $\theta_\lambda$ in (2.10) and (3.3) enables an *approximate* adjustment to the estimator $\theta$ 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

$$(4.1) \qquad y = \theta_o x^* + \varepsilon$$

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

$$(4.2) \qquad \theta = Exy/Ex^2.$$

A direct computation shows that

$$(4.3) \qquad \theta_o = \theta[1-(\lambda/m^2)]^{-1}$$

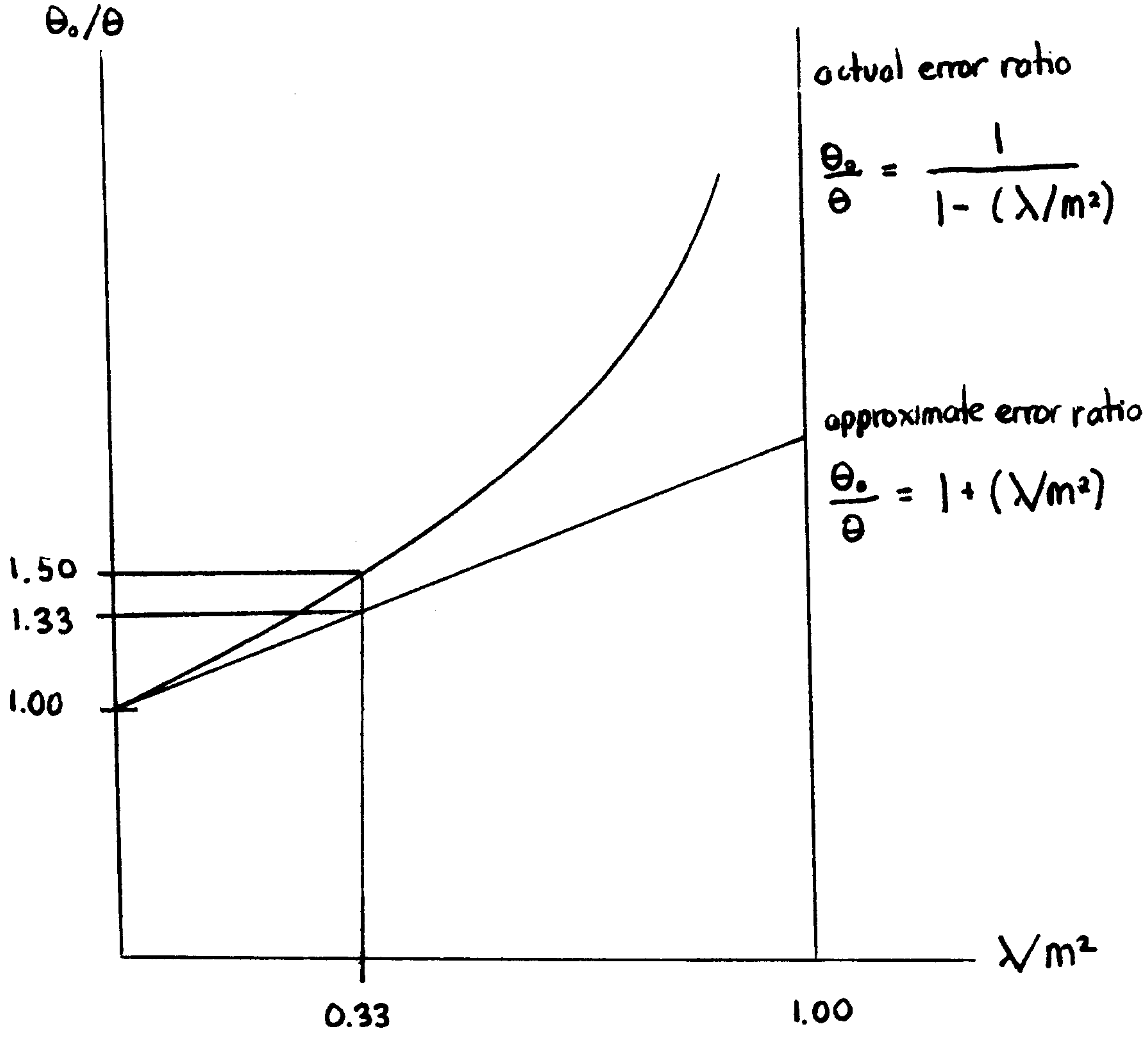
while the approximate value $\theta_o^a$ computed as $\theta_o^a = \theta + \lambda\theta_\lambda$ is computed from (3.3) as

$$(4.4) \qquad \theta_o^a = \theta[1+(\lambda/m^2)].$$

Naturally (4.4) is simply the tangent line to (4.3) at $\lambda=0$. As sketched in the attached figure, the quality of the approximation depends on $(\lambda/m^2)$ -- 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.

Figure

Approximation Error in Simple Regression


$\theta_o/\theta$
actual error ratio
$\frac{\theta_o}{\theta} = \frac{1}{1-(\lambda/m^2)}$
approximate error ratio
$\frac{\theta_o}{\theta} = 1+(\lambda/m^2)$
$\lambda/m^2$