INFLUENTIAL OBSERVATIONS IN MEASUREMENT ERROR MODELS

Richard F. Gunst
Southern Methodist University
Dallas, Texas 75275

Robert L. Mason
Southwest Research Institute
San Antonio, Texas 78284

SMU/DS/TR-203

Abstract

Influence functions for intercept and slope estimators are used to assess the effects of influential observations on least squares and maximum likelihood estimators for structural measurement error models. Based on the information provided by the influence functions, recommendations are made for the use of diagnostics in the detection of influential observations with measurement error models.

Key Words: Diagnostics, Influence functions, Errors in variables, Least squares, Maximum likelihood.

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275

1. Introduction

Influence functions often are used to assess the effects of influential observations on least squares regression estimators (Cook & Weisberg, 1982). Kelly (1984) derived influence functions for the method of moments intercept and slope estimators in a structural measurement error model having independent measurement errors. In this paper we present influence functions for measurement error models having normally distributed, correlated measurement errors. By re-expressing the influence functions in terms of residual vectors, geometric interpretations of the effects of influential observations are readily apparent.

Linear structural measurement error models are linear models $Y = \alpha + \beta X$ between two stochastic variates (Y,X) in which both variates are measured with error:

$$y_i = Y_i + v_i, \quad x_i = X_i + u_i \quad (i = 1,\ldots,n). \tag{1.1}$$

Assume that X and (v,u) are mutually independent with $X \sim N(\mu_X, \sigma_X^2)$,

$$\begin{pmatrix} v \\ u \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_u^2 \end{pmatrix}\right\}.$$

Let $\lambda = \sigma_v^2/\sigma_u^2$ denote the error variance ratio, $\gamma = \sigma_u^2/\sigma_X^2$ the noise-to-signal ratio for the observable predictor variable x, and $\theta = \rho\lambda^{1/2}$.

Least squares estimators of the intercept and slope parameters are

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}, \quad \hat{\beta} = s_{xy}/s_{xx} \tag{1.2}$$

where s_{yy}, s_{xx}, and s_{xy} denote the sample variances and covariance of the observable variates. Maximum likelihood estimators of the intercept and slope parameters, assuming known values for λ and ρ, are (Reilman, Gunst, & Lakshminarayanan 1985):

$$\tilde{a} = \bar{y} - \tilde{\beta}\,\bar{x}, \qquad \tilde{\beta} = s(\lambda,\theta) + \mathrm{sgn}\{u(\theta)\}\{s(\lambda,\theta)^2 + t(\lambda,\theta)\}^{1/2} \tag{1.3}$$

where

$$s(\lambda,\theta) = (s_{yy} - \lambda s_{xx})/\{2u(\theta)\},$$

$$t(\lambda,\theta) = (\lambda s_{xy} - \theta s_{yy})/u(\theta),$$

and

$$u(\theta) = s_{xy} - \theta s_{xx}.$$

Fuller (1987) provides comprehensive coverage of the estimation of measurement error models, including multivariate and multiple regression models. Properties of least squares and maximum likelihood estimators under measurement error model assumptions are also detailed.

2. Influence Functions

2.1 Least Squares

Under the assumptions accompanying model (1.1) the influence functions for the least squares intercept and slope estimators are

$$IF_{LS}(\alpha) = y_0 - \mu_y - \beta_{LS}(x_0 - \mu_x) - IF_{LS}(\beta)\mu_x$$

and (2.1)

$$IF_{LS}(\beta) = \{(x_0 - \mu_x)(y_0 - \mu_y) - \beta_{LS}(x_0 - \mu_x)^2\}/\sigma_{xx},$$

where $\mu_x = \mu_X$, $\mu_y = \mu_Y = \alpha + \beta\mu_X$, $\beta_{LS} = (\beta + \theta\gamma)/(1 + \gamma)$, and $\sigma_{xx} = \sigma_X{}^2(1 + \gamma)$ is the variance of the observable predictor x. These influence functions are similar to those of Hinkley (1977). While these expressions can be evaluated for any choice of (y_0,x_0), three specific choices simplify the expressions and provide interesting interpretations:

$\underline{x_0 = \mu_x}$

$$IF_{LS}(\alpha) = y_0 - \mu_y \ , \ IF_{LS}(\beta) = 0$$

There is no effect on the least squares slope estimator when an influential observation occurs at the mean of the predictor variable. The intercept estimate is changed by an amount equal to the difference between the response value and the mean of the response variable. The least squares line is moved parallel to the noncontaminated line.

$\underline{y_0 - \mu_y = \beta(x_0 - \mu_x)}$

$$\begin{aligned} IF_{LS}(\alpha) &= -B_{LS}(x_0 - \mu_x) - A_{LS}(x_0 - \mu_x)^2/\sigma_{xx} \\ &= -B_{LS}(x_0 - \mu_x) - B_{LS}(x_0 - \mu_x)^2\mu_x/\sigma_{xx}, \\ IF_{LS}(\beta) &= -B_{LS}(x_0 - \mu_x)^2/\sigma_{xx}, \end{aligned}$$

where A_{LS} and B_{LS} are the asymptotic biases of the least squares estimators,

$$\begin{aligned} A_{LS} &= \alpha_{LS} - \alpha = -B_{LS}\mu_x, \\ B_{LS} &= \beta_{LS} - \beta = -(\beta - \theta)\gamma/(1 + \gamma), \end{aligned}$$

and $\alpha_{LS} = \mu_y - \beta_{LS}\mu_x$, $\beta_{LS} = (\beta + \theta\gamma)/(1 + \gamma)$. When influential observations are true to the unobservable theoretical model, the effects of the original biases in the estimators (A_{LS} and B_{LS}) can be mitigated since the changes in the estimators are opposite the respective biases.

$\underline{y_0 = \mu_y}$

$$IF_{LS}(\alpha) = -\ \beta_{LS}(x_0 - \mu_x) - \beta_{LS}(x_0 - \mu_x)^2\mu_x/\sigma_{xx},$$
$$IF_{LS}(\beta) = -\ \beta_{LS}(x_0 - \mu_x)^2/\sigma_{xx}.$$

An influential observation at the mean of the response variable increases the bias of the least squares slope estimator since the change in the estimator is negatively proportional to the parameter value.

2.2 Maximum Likelihood

The general form for the influence functions of the maximum likelihood intercept and slope estimators are

$$IF_{ML}(\alpha) = y_0 - \mu_y - \beta(x_0 - \mu_x) - IF_{ML}(\beta)\mu_x$$

and (2.2)

$$IF_{ML}(\beta) = \{(\beta - \theta)(y_0 - \mu_y)^2 - (\beta^2 - \lambda)(x_0 - \mu_x)(y_0 - \mu_y) - \beta(\lambda - \beta\theta)(x_0 - \mu_x)^2\}\gamma/\sigma_e^2,$$

where $\sigma_e^2 = \mathrm{var}(e_i) = \mathrm{var}(v_i - \beta u_i) = \sigma_u^2\{(\beta - \theta)^2 + (\lambda - \theta^2)\}$ is the variance of the model error e_i when y_i is expressed as a linear function of x_i. When $\rho = 0$, these expressions reduce to those of Kelly (1984).

$\underline{x_0 = \mu_x}$

$$IF_{ML}(\alpha) = y_0 - \mu_y - (\beta - \theta)\gamma(y_0 - \mu_y)^2\mu_x/\sigma_e^2,$$
$$IF_{ML}(\beta) = (\beta - \theta)\gamma(y_0 - \mu_y)^2/\sigma_e^2.$$

Unlike least squares, the maximum likelihood slope estimator is affected by influential observations at the response mean. The slope estimator is biased proportional to the difference between the true slope parameter and the scaled correlation coefficient, $\theta = \rho\lambda^{1/2}$. When $\rho = 0$, the bias is proportional to the slope coefficient.

$\underline{y_0 - \mu_y = \beta(x_0 - \mu_x)}$

$$IF_{ML}(\alpha) = 0, \quad IF_{ML}(\beta) = 0.$$

Because the maximum likelihood estimators are consistent, an influential observation which is true to the unobservable theoretical model has no effect on the intercept and slope estimators.

$\underline{y_0 = \mu_y}$

$$IF_{ML}(\alpha) = -\beta(x_0 - \mu_x) + \beta\gamma(\lambda - \beta\theta)(x_0 - \mu_x)^2\ \mu_x/\sigma_e^2,$$
$$IF_{ML}(\beta) = -\beta\gamma(\lambda - \beta\theta)(x_0 - \mu_x)^2/\sigma_e^2.$$

The slope bias introduced by an influential observation at the mean of the response variable is opposite of the sign of the true regression coefficient. The fit tends to be "flatter" than the true regression line.

3. Alternative Expressions

The least squares and the maximum likelihood influence functions can be re-expressed in a form in which comparisons of the effects of influential observations are more easily interpreted. The least squares influence functions are expressible in terms of two residual functions:

$$IF_{LS}(\alpha) = r_{y\bullet x} - IF_{LS}(\beta)\mu_x,$$
$$IF_{LS}(\beta) = \sigma_{xx}{}^{-1}r_x r_{y\bullet x}, \qquad (3.1)$$

where

$$r_x = x_0 - \mu_x, \quad r_{y\bullet x} = \{y_0 - \mu_y - \beta_{LS}(x_0 - \mu_x)\}.$$

These residual functions are the differences between the influential response and predictor values and the response and predictor means. In the usual Euclidian representation of the response and predictor variables, these two residual vectors are orthogonal. Figure 1 depicts this representation for a typical influential observation.

The influence functions for the maximum likelihood estimators can be written in a form similar to (3.1) by transforming to polar coordinates after re-expressing the model so that the measurement errors for the response and the predictor variables are uncorrelated and homoscedastic. If one then measures the horizontal (r_h) distance along the true regression line from the origin to the projection of (x_0, y_0) and the perpendicular (r_p) distance from the influential observation to the true regression line, the influence functions can be written as follows:

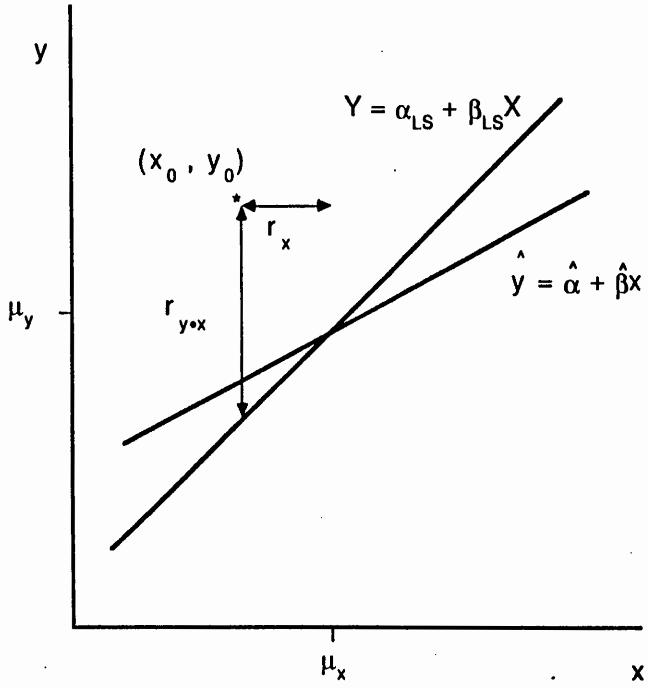
$$IF_{ML}(\alpha) = r_{y\bullet x} - IF_{ML}(\beta)\mu_x,$$
$$IF_{ML}(\beta) = \sigma_X{}^{-2}r_h r_p, \qquad (3.2)$$

where

$$r_h = (\lambda-\theta^2)^{-1/2}\{(\beta-\theta)^2 + (\lambda-\theta^2)\}^{-1/2}[(\beta-\theta)r_{y\bullet x} + \{(\beta-\theta)^2$$
$$+ (\lambda-\theta^2)\}r_x],$$
$$r_p = (\lambda-\theta^2)^{+1/2}\{(\beta-\theta)^2 + (\lambda-\theta^2)\}^{-1/2}r_{y\bullet x}$$

and $r_{y\bullet x}$ is now measured from the true regresion line, $r_{y\bullet x} = y_0 - \mu_y - \beta(x_0 - \mu_x)$.

Fig. 1. Residual Vector Influences on Least Squares Estimators.



y
$Y = \alpha_{LS} + \beta_{LS}X$
(x_0, y_0)
r_x
$\hat{y} = \hat{\alpha} + \hat{\beta}x$
μ_y
$r_{y \cdot x}$
μ_x
x

Fig. 2. Residual Vector Influences on Maximum Likelihood Estimators.

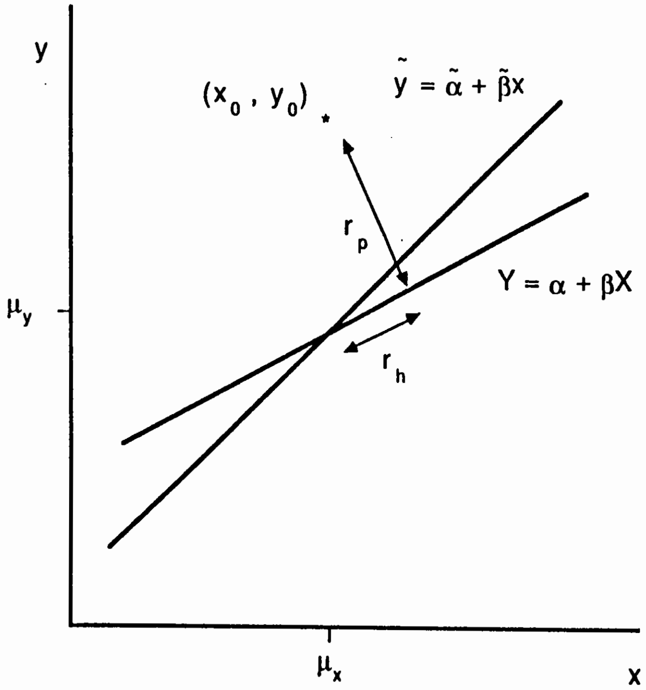


Figure 2 shows this representation for a model with $\lambda=1$ and uncorrelated measurement errors.

In the forms (3.1) and (3.2) it is seen that both of the slope estimators are susceptible to influential observations in all but two directions, those for which the residuals in either term of the product is zero. For the least squares estimator, the two directions are vertical at $x_0 = \mu_x$ and along the (biased) least squares line at $y_0 = \mu_y + \beta_{LS}(x_0 - \mu_x) = \alpha_{LS} + \beta_{LS}x_0$. For the maximum likelihood estimator, the two directions are parallel and orthogonal to the true regression line through the point (μ_x, μ_y).

4. Implications for Diagnostics and Robust Estimation

It is not the intent of this paper to investigate diagnostics or robust estimation for measurement error model estimators. The influence functions do, however, suggest important implications for such procedures. Diagnostics for measurement error model estimators must be sensitive to departures from the true model in all directions except those for which the influence functions are zero. It is clear that diagnostics which only measure departures from assumptions in either the x- or the y-directions are not suitable for maximum likelihood estimators of measurement error model intercept and slope parameters. More sensitive diagnostics would make use of the residual vectors in the influence functions (3.2).

Likewise, robust estimators for measurement error models should be insensitive to local departures from model assumptions in directions that are not parallel to the residual vectors r_h and r_p.

The form of the slope influence functions for least squares and maximum likelihood estimators suggests that satisfactory robust estimators should be based on procedures which weight the two residual components separately. This principle supports the use of bounded influence (Kraker and Welsch, 1982) and similar estimation schemes over alternative robust estimators which only weight response residuals.

References

Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Fuller, W. A. (1987). *Measurement Error Models.* New York: Wiley.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285-92.

Kelly. G. (1984). The influence function in the errors in variables problem, *Ann. Statist.*, 12, 87-100.

Krasker, W. S. & Welsch, R. E. (1982). Efficient bounded-influence regression estimation, *J. Amer. Statist. Assn.*, 77, 595-604.

Reilman, M. A., Gunst, R. F., & Lakshminarayanan, M. Y. (1985). Structural model estimation with correlated measurement errors, *Biometrika*, 72, 669-72.