INFLUENCE DIAGNOSTICS FOR MEASUREMENT ERROR MODELS

Richard F. Gunst

Department of Statistical Science
Southern Methodist University

SMU/DS/TR-225

June 1988

Influence diagnostics for measurement error models

Richard F. Gunst
Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275, U.S.A.

Summary

Influence diagnostics for measurement error models are suggested by an examination of influence functions. These diagnostics are sensitive to extreme observations along the fitted plane and orthogonal to it, in contrast to least squares diagnostics, which are sensitive to extreme observations in vertical and horizontal directions to the fitted plane.

Same key words: DFBETAS; Cook's D-statistic; Errors in variables; Leverage values; Regression; Studentized residuals.

1. Introduction

Linear regression models $\psi = \pi'\beta$ in which both the response and one or more of the predictor variables $\pi' = (\pi_1, \pi_2, \ldots, \pi_k)$ are measured with error are referred to as linear measurement error models. The published literature on measurement error models is heavily concentrated on finding estimators of the model parameters and on deriving asymptotic properties of the estimators. Very little research has been conducted on other aspects of measurement error model methodology such as influence diagnostics. Fuller (1987) and Kelly (1984) are two exceptions. It is the purpose of this paper to propose outlier diagnostics for linear measurement error models. These diagnostics are restricted to several familiar ones, based on least

squares estimation, that have proven useful for traditional regression models that assume error-free predictors.

Denote the (k+1)-dimensional observable variables by $z_i = \xi_i + w_i$, for i=1,2,...,n, with $z_i' = (y_i \; x_i')$, $\xi_i' = (\psi_i \; \pi_i')$, and $w_i' = (v_i \; u_i')$. The unobservable measurment errors w_i are assumed to be distributed NID(0, Δ_{ww}), where the appropriate elements of Δ_{ww} are zero for predictors that are measured without error. Partition Δ_{ww} conformably with the w_i and denote the components by δ_{vv}, Δ_{uv}, and Δ_{uu}.

2. Estimators

For functional measurement error models, assume that the π_i are constant vectors with the following limits finite and $\Gamma_{\pi\pi}$ positive definite:

$$\mu_\pi = \lim n^{-1} \sum \pi_i \;, \; \Gamma_{\pi\pi} = \lim n^{-1} \sum \pi_i \pi_i' \;.$$

For structural measurement error models assume that the π_i are NID(μ_π , $\Delta_{\pi\pi}$), independently of all the model errors, with $\Delta_{\pi\pi}$ at least positive semidefinite and $\Gamma_{\pi\pi} = \Delta_{\pi\pi} + \mu_\pi \mu_\pi'$ positive definite.

Two types of linear measurement error model estimators are considered. The first estimator assumes that the measurement error covariance matrix Δ_{ww} either is completely known or is estimated by an estimator that is statistically independent of the observed variates z_i. When the error covariance matrix is known or is estimated, a second component of error in the response variable, called equation error by Fuller (1987, Section 2.2), can be included in the model. For these models, let $y_i = \psi_i + v_i + q_i$, where q_i represents the equation error. The equation errors q_i are assumed to be distributed NID(0, δ_{qq}), independently of the measurement errors w_i.

Denote the known or estimated covariance matrix of the measurement errors by S_{ww}. Assume that S_{ww} is an unbiased estimator of Δ_{ww} and is a multiple of a Wishart matrix having d_w degrees of freedom. Note that $S_{ww} = \Delta_{ww}$ and $d_w^{-1} = 0$ if the measurement error covariance matrix is known. Partition S_{ww} conformably with the w_i and denote the components by s_{vv}, S_{uv}, and S_{uu}. Let $M_{xx} = n^{-1} \Sigma x_i x_i'$, and $M_{xy} = n^{-1} \Sigma x_i y_i$. Then an estimator of β is

$$\tilde{\beta} = (M_{xx} - S_{uu})^{-1}(M_{xy} - S_{uv}) \quad . \tag{2.1}$$

A second estimator of β arises when the covariance matrix of the measurement errors is only known up to a multiple. Let $\Delta_{ww} = T_{ww}\, \delta_{ww}$, where T_{ww} is assumed known and δ_{ww} is an unknown constant. Then

$$\tilde{\beta} = (M_{xx} - \hat{\lambda}\, T_{uu})^{-1}(M_{xy} - \hat{\lambda}\, T_{uv}) \ , \tag{2.2}$$

where $\hat{\lambda} = \hat{\delta}_{ww}$ is the smallest root of $|M_{zz} - \lambda\, T_{ww}| = 0$ and T_{ww} has componnents t_{vv}, T_{uv}, and T_{uu}. Provided that the estimators of the model covariance matrices are nonnegative, the estimators (2.1) and (2.2) are maximum likelihood estimators.

Fuller (1987, Sections 2.2 and 2.3, respectively) presents asymptotic normal distributions for first-order Taylor series approximations to the estimators (2.1) and (2.2). Let $\nu = \lim n/d_w < \infty$, with $\nu = 0$ if the covariance matrix is known or known up to a multiple. The covariance matrices of the asymptotic distributions of $n^{1/2}(\tilde{\beta} - \beta)$ for estimators (2.1) and (2.2) can then both be written in the form

$$\begin{aligned}\Delta_{\beta\beta} = \Gamma_{\pi\pi}^{-1} \{ & (\Gamma_{\pi\pi} + \Delta_{uu})(\delta_{qq} + \delta_{ee}) + c\, \Delta_{ue}\, \Delta_{eu} \\ & + \nu\, (\Delta_{uu}\, \delta_{ee} + c\, \Delta_{ue}\, \Delta_{eu})\}\, \Gamma_{\pi\pi}^{-1} \quad ,\end{aligned} \tag{2.3}$$

where $\delta_{ee} = \delta_{vv} - 2\Delta_{vu}\beta + \beta'\Delta_{uu}\beta$, and $\Delta_{ue} = \Delta_{uv} - \Delta_{uu}\beta$. When the error covariance matrix is completely known or is estimated, $c = 1$.

When the error covariance matrix only is known up to a multiple, $c = -1$ and δ_{qq} is set to zero.

3. Influence Functions

Kelly (1984) derives a general expression for the influence function of (2.2) for structural measurement error models when the covariance matrix of the measurement errors is known up to a multiple and the error distribution has finite fourth moments. A derivation of the influence function as in Hampel, Ronchetti, Rousseeuw, and Stahel (1986, p. 230) under the normality assumptions stated in the previous section results in the following form for the influence function:

$$IF_{ML}(d;\beta) = \Gamma_{\pi\pi}^{-1}(x - \delta_{ee}^{-1}\,\Delta_{ue}d)d \quad . \tag{3.1}$$

This influence function is appropriate for the estimator (2.1) if δ_{ee} is replaced by $\delta_{qq} + \delta_{ee}$. In both cases, $d = y - x'\beta$ represents the deviation of the point $(y\ x')$ from the model.

The influence function (3.1) provides an interesting geometric interpretation of the influence of an observation $z = (y\ x')'$ on the maximum likelihood intercept and slope estimators for measurement error models. Partition β' as (β_0, β_m'), where β_0 denotes the intercept coefficient and β_m denotes the vector of $m = k - 1$ regression coefficients for the nonconstant predictors. Partition $x = (1, x_m')'$, $u = (0\ , u_m')'$, and their mean vectors and covariance matrices conformably. Then by partitioning (3.1),

$$IF_{ML}(d;\beta_0) = d - \mu_m' IF_{ML}(\beta_m) \quad ,$$

$$IF_{ML}(d;\beta_m) = \Gamma_{mm}^{-1} r_p r_o \quad , \tag{3.2}$$

where r_o and r_p are, respectively, the orthogonal and the planar components of the influence function relative to the hyperplane of the predictor variables. Specifically,

$$r_o = \delta_{ee.m}^{1/2}\ \delta_{ee}^{-1/2} d$$

$$r_p = \delta_{ee.m}^{-1/2}\ \delta_{ee}^{-1/2}(x_m - \mu_m - \delta_{ee}^{-1}\ \Delta_{me} d) \quad , \qquad (3.3)$$

where $\delta_{ee.m} = \delta_{ee} - \Delta_{em}\Delta_{mm}^{-1}\Delta_{me}$ is the conditional variance of e given u_m and Δ_{me} is the covariance matrix between u_m and e. These influence function expressions remain valid when some of the predictors are measured without error if the appropriate elements in Δ_{ww} are set equal to zero and removed from the expressions for Δ_{me} and Δ_{mm}.

3.1 Least Squares

The influence function (3.1) reduces to the corresponding least squares influence function given in Cook and Weisberg (1982 Section 3.3) or Hinkley (1977) when there are no measurement errors in x:

$$IF_{ML}(d;\beta) = \Gamma_{\pi\pi}^{-1} x d \quad . \qquad (3.4)$$

It is well known that outliers in either the vertical direction or in the horizontal direction, the latter known as leverage points, can severely affect least squares estimates. The effect is quantified by (3.4), where the derivation d imparts the contribution of outliers in the vertical direction and the predictors x impart the contribution of leverage points.

Vertical outliers at the centroid μ_m of the space of the non-constant predictor variables only affect the least squares intercept estimate. Similarly, leverage points that fall on the true regression plane do not affect the intercept or the slope. These conclusions can be confirmed by an examination of the least squares influence function. Sample influence functions have similar properties with the appropriate sample statistics inserted into the influence function expression.

3.2 Measurement Error Models

The influence functions (3.2) depict the influence of an observation z on the maximum likelihood estimators as a function of orthogonal and planar components. Orthogonal outliers along the normal to the regression plane at the centroid μ_m of the space of the nonconstant predictors only affect the intercept estimate. Planar outliers, the equivalent of leverage points, that fall on the regression plane do not affect either the intercept or the slope estimates. These conclusions follow from an examination of the form of the influence function (3.2).

4. Least Squares Influence Diagnostics

Influence diagnostics for measurement error models can be formulated with much the same motivation as those for least squares. The diagnostics investigated in this research are patterned after those of least squares, with adaptations to account for the effects of measurement errors in the predictor variables.

Leverage values h_{ii} (Hoaglin and Welch 1978) for least squares estimators are the diagonal elements of $H = X(X'X)^{-1}X'$, where X represents the n x k matrix of predictor variables. Leverage values are not only important diagnostics of extreme predictor-variable values, they are fundamental to the efficient calculation of other least squares influence diagnostics.

A measure of the effect of extreme predictor variable values on the least squares estimators of the regression coefficients is the statistic DFBETAS:

$$\mathrm{DFBETAS}_{ij} = (\hat{\beta}_j - \hat{\beta}_{j(i)})/\{c_{jj}s_{(i)}^2\}^{1/2} \quad ,$$

where $\hat{\beta}_{j(i)}$ is the least squares estimator of β_j and $s_{(i)}^2$ is the mean squared error, both from the fit without the ith observation, and c_{jj} is the jth diagonal element of $(X'X)^{-1}$. A computationally efficient form of these statistics utilizing leverage values is given in Belsley, Kuh, and Welsch (1980, p. 13). The unscaled vector version of this statistic has an unmistakable similarity to the influence function (3.4):

$$\mathrm{DFBETA}_i = (\hat{\beta} - \hat{\beta}_{(i)}) = (X'X)^{-1}x_i r_i/(1 - h_{ii}) \quad . \tag{4.1}$$

The quantity $r_i/(1 - h_{ii})$ is the deleted residual $y_i - \hat{y}_{i(i)}$.

Following Cook and Weisberg (1982), internally studentized residuals are defined as $t_i = r_i/\{s^2(1-h_{ii})\}^{1/2}$, where r_i is the ith (vertical) residual and s^2 is the mean squared error, both from the least squares fit to the complete data set. Cook's (1977) statistic is

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})'X'X(\hat{\beta}_{(i)} - \beta)/(ks^2) \quad . \tag{4.2}$$

Using (4.1), D_i is also directly related to the influence function (3.4). An alternative form for D_i, which is useful for computations, is

$$D_i = t_i^2 h_{ii}/\{k(1 - h_{ii})\}. \tag{4.3}$$

The influence of vertical outliers (t_i) and leverage points (h_{ii}) on this statistic are evident.

5. Measurement Error Model Influence Diagnostics

One key to the satisfactory implementation of measurement error model influence diagnostics is a suitable definition of a leverage value. The least squares definition is unsatisfactory because it is based on presumably error-free predictor variables. We follow Fuller

(1987, Section 2.2.3) and use estimated predictor-variable values in place of the observed values in the hat matrix:

$$\tilde{H} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}' \quad , \quad \tilde{X}' = (\tilde{x}_1 \, , \, \tilde{x}_2, \, \ldots \, , \, \tilde{x}_n) \quad \text{with}$$

$$\tilde{x}_i = x_i - \tilde{\delta}_{ee}^{-1} \, \tilde{\Delta}_{ue} r_i \quad . \tag{5.1}$$

Note that the measurement error model influence function (3.1) is based on estimated predictors $\tilde{x}_i$ just as the least squares influence function (3.4) is based on the observed predictors x_i. The estimated predictors $\tilde{x}_i$ in (5.1) are proportional to the planar components of the influence functions (3.1) and (3.2).

The diagonal elements of $\tilde{H}$ can be used in the computation of diagnostics similar to those of least squares, with $\tilde{x}_i$ replacing x_i where appropriate. Unlike least squares, however, deleted residuals are not algebraically equal to $r_i/(1-\tilde{h}_{ii})$. There does not appear to be an algebraically simple formula for calculating deleted residuals using the estimators (2.1) or (2.2). Nevertheless, the approximation $r_i/(1-\tilde{h}_{ii})$ has been examined on numerous data sets and found to be quite adequate for use in influence diagnostics.

DFBETAS, studentized residuals, and Cook's statistics can all be calculated using the formulas shown above with appropriate substitution of estimators from the measurement error model fit. Consider, for example, Cook's statistic (4.2). The general form of this statistic is (Cook and Weisberg 1982, Section 3.5.1):

$$D_i = (\tilde{\beta}_{(i)} - \tilde{\beta})'M(\tilde{\beta}_{(i)} - \tilde{\beta})/c \quad . \tag{5.2}$$

For measurement error models, M^{-1} is replaced by an estimator of the covariance matrix of $\tilde{\beta} - \beta$, $\Delta_{\beta\beta}$, and c is replaced by k, the number of predictor variables.

The estimators of $\Delta_{\beta\beta}$ that were investigated for use in (5.2) are based on inserting consistent moment estimators of the covariance terms in (2.3). These estimators (Fuller 1987, Sections 2.2 and 2.3) are based on the delta method and are equivalent to methods based on influence functions and the infinitesimal jackknife (Efron 1981). This type of estimator of the covariance matrix performed well in Kelly's (1984) simulation when the data were normally distributed.

The form of Cook's statistic (5.2) proposed for use as a measurement error model influence diagnostic is

$$D_i = (\tilde{\beta}_{(i)} - \tilde{\beta})' \tilde{\Delta}_{\beta\beta}^{-1}(\tilde{\beta}_{(i)} - \tilde{\beta})/k \quad . \tag{5.3}$$

An approximation to (5.3) using the computational form (4.3) provides excellent agreement on all data sets examined to date. The formula (4.3) is exactly equal to (5.3) when the estimator (2.2) is used and $\nu = 0$.

6. Concrete Compressive Strength Data

Figure 1 displays a scatterplot of the compressive strengths of 41 samples of concrete. The plotted data are measurements of the compressive strengths of each sample taken two and twenty-eight days after pouring. The investigators wish to determine a prediction equation for the strength of concrete twenty-eight days after pouring using the measurements taken two days after pouring.

Overlaid on the plots are two regression fits, least squares and a measurement error model fit. Estimator (2.2) was used for the measurement error model fit, assuming that T_{ww} = diag (1, 0, 1). The zero diagonal element in T_{ww} corresponds to the constant term. It is clear that the least squares and the measurement error model fits are

different, with the measurement error model fit appearing to be greatly affected by observation 21 in the upper right portion of the scatterplot.

Table 1 contains influence diagnostics for several observations. Observation 21 has greater leverage for the measurement error model fit, and great influence on both of the fits, as indicated by the DFBETAS, and Cook's statistics. The internally studentized residuals t_i indicate that this extreme observation is more poorly fit using the least squares fit than the measurement error model fit. Each of these conclusions is supported by the fits in Figure 1 and by the estimated coefficients and fitted responses in Table 2.

Each of the other observations listed in Table 1 appears to be better fit by one of the estimators, attesting to the need for different diagnostics for the least squares and measurement error model fits. In each case, the ability of an estimator to closely fit an observation depends on whether the observation is an outlier in the vertical or the orthogonal directions as indicated by the respective t_i values.

Fuller (1987) uses the $\tilde{x}_i$ in various plots to diagnose model inadequacy. He suggests (p.121) that residuals r_i and $\tilde{x}_i$ could be input to least squares computer programs and used for diagnostic checking. The results presented in this paper provide a theoretical basis for and confirm the appropriateness of that suggestion.

Acknowledgments

Appreciation is expressed to Mr. Richard M. Weed, New Jersey Department of Transportation, for providing the data used in this paper. Critiques of an earlier version of this paper by R.L. Mason, W.R. Schucany, and S. Wang are gratefully acknowledged.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics*. New York: Wiley.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15-18.

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

Efron, B. (1981). Nonparametric estimates of the standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68, 589-99.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics*. New York: Wiley.

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

Hoaglin, D. C. and Welsch, R. (1978). The hat matrix in regression and ANOVA. *Amer. Statistician*, 32, 17-22.

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

Table 1. *Comparison of Least Squares and Measurement Error Model Influence Diagnostics.*

	Least Squares				Measurement Error Model			
Obsn.	Leverage	t_i	D_i	DFBETAS	Leverage	t_i	D_i	DFBETAS
17	.031	-1.849	.055	.158	.080	-1.274	.070	.312
21	.159	4.168	1.640	2.208	.450	2.488	2.538	2.191
22	.197	-.431	.023	.198	.165	1.328	.174	-.545
34	.117	.712	.034	-.230	.059	1.739	.095	-.335
37	.108	1.581	.150	-.492	.034	2.356	.097	-.233

Table 2. *Comparison of Fits.*

(a) Complete Data Set

	Least Squares	Measurement Error Model
Intercept	4,636	61
Slope	0.798	1.514
Fitted $\hat{y}_{21}$*	5,602	6,488

(b) Observation 21 Deleted

	Least Squares	Measurement Error Model
Intercept	3,016	1,731
Slope	0.516	0.945

* Observed y_{21} = 7,695

FIG. 1. CONCRETE COMPRESSIVE STRENGTH DATA