NBER WORKING PAPER SERIES

# ROBUST LINE ESTIMATION WITH ERRORS IN BOTH VARIABLES

Michael L. Brown

Working Paper No. 83

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

May 1975



* NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

## Abstract

The estimator holding the central place in the theory of the multivariate "errors-in-the-variables" (EV) model results from performing orthogonal regression on variables rescaled according to the covariance matrix of the errors [7]. Our first principal finding, via Monte Carlo on the univariate model, essentially relegates this estimator to use only in large samples on very well-behaved data, i.e., with no trace of outlier contamination. A modification, requiring a robust preliminary slope, is proposed that essentially sets out the generalization to EV of the w-estimator in regression. It is demonstrated that the modification is robust to outlier contamination even in small samples, given a sufficiently good preliminary estimator. A candidate for a preliminary slope estimator based on the data is proposed and its performance under simulation examined. Least-absolute residuals estimation in EV is cited as an alternative candidate.

# Contents

## 1. Motivation from the Robust Regression

Huber [3] addressed the question of generalizing robust estimates of a location parameter to the problem of estimating robustly the coefficients $\{\beta_j\}_{j=1}^{n}$ in a multivariate regression model

$$y_i = \sum_{j=1}^{p} \beta_j X_{ij} + v_i \tag{1}$$

with contaminated errors $v_i$. He asserts that "[in] the classical least squares theory... the matrix $X = (X_{ij})$ is thought to derive from a fixed and rigorous mathematical model. In statistics, it is more customary to treat the coefficients $X_{ij}$ as independent variables, possibly also subject to errors. Next to nothing is known about how to robustize regression procedures with respect to errors in the $X_{ij}$." (underlines ours). The underlined statement is the problem we consider.

## 2. The Simplest EV Model

### 2.1 Classical assumptions for the univariate model

We set p=1 in (1) and, to quantify the phrase "errors in the $X_{ij}$", we introduce an error u into X, so that we no longer observe $X_i$ directly but rather $x_i$, where

$$x_i = X_i + u_i, \quad i=1,\ldots,n . \tag{2}$$

To fill out the EV model specification, we assume

$$y_i = Y_i + v_i \tag{3}$$

with the "true" values exactly linearly related:

$$Y_i = \beta X_i \quad . \tag{4}$$

(For the moment we assume a constant term $\alpha$ equal to 0 for simplicity.) Further,

$$E(v_i) = 0 \tag{5a}$$

$$E(u_i) = 0 \quad . \tag{5b}$$

$$\mathrm{cov}(u_i,v_i) = 0 \quad . \tag{6}$$

We suppose

$$X_i \sim N(0,\sigma_X^2)$$

(a "structural relation") and

$$\mathrm{cov}(X_i, v_i) = 0 \tag{7a}$$

$$\mathrm{cov}(X_i, u_i) = 0 \quad . \tag{7b}$$

(Note: Our discussion applies equally well to the "functional relation", where the $X_i$'s arise in some deterministic fashion rather than as realizations of a random variable X.)

Our given data consist of the n observations

$$\{(x_i, y_i)\}_{i=1}^{n} \quad , \tag{8}$$

with successive observations taken independently.

Notice that $x \equiv X (u \equiv 0)$ is the case of regression with a line through the origin.

2.2 Contaminated error distributions

We define a contaminated EV model with u and v samples from the contaminated normal distribution. This means that

$u_i$ is drawn from $N(0,\sigma_u^2)$ with probability $(1-\gamma_u)$ (9a)

and from $N(0,h_u^2)$ with probability $\gamma_u$ (9b)

where $h_u >> \sigma_u$ and similarly for v.

2.3 Identifiability problem in classical EV

Lindley [5] and others have pointed out that, in the classical case of gaussian errors,

$$\gamma_u = .00 \qquad h_u = 0 \tag{10a}$$

$$\gamma_v = .00 \qquad h_v = 0 \tag{10b}$$

the ML estimators of the parameters $\beta$, $\sigma_u^2$, $\sigma_v^2$ cannot all be consistent. Kendall and Stuart [4] observe that "we must make an assumption about the error variances.... [Assuming]

$$\underline{\sigma_v^2/\sigma_u^2 \quad \text{known}}$$

...is the classical means of resolving the unidentifiability problem."

Throughout, we assume

$$\lambda \equiv \sigma_v^2/\sigma_u^2 \tag{11}$$

is known.

3. Form of ML Estimator for Classical EV

It is convenient to derive the ML estimator of $\beta$, BML say, for the functional relation; its form remains unchanged for the structural relation. The likelihood function is

$$\frac{1}{\sigma_u^n \sigma_v^n} \exp\left\{-\frac{1}{2\sigma_v^2} \sum_i (y_i - Y_i)^2 - \frac{1}{2\sigma_u^2} \sum_i (x_i - X_i)^2\right\}, \text{ which} \tag{12}$$

allows us to characterize the ML estimator of $\beta$ as

$$\text{BML} = \underset{\tilde{\beta};\, \tilde{X}_1, \ldots, \tilde{X}_n}{\min^{-1}} \left\{ \sum_{i=1}^{n} \left[ \left( \frac{y_i - \tilde{\beta}\tilde{X}_i}{\sqrt{\lambda}} \right)^2 + (x_i - \tilde{X}_i)^2 \right] \right\}, \tag{13}$$

where the symbol to the left of the braced quantity means "that value of $\tilde{\beta}$ for which the braced function of $(\tilde{\beta};\tilde{X}_1,\ldots,\tilde{X}_n)$ is a minimum, for $-\infty < \tilde{\beta} < \infty$, $-\infty < \tilde{X}_i < \infty$, $i=1,\ldots,n$." We see that transforming $y$ to $y/\sqrt{\lambda}$ and $\beta$ to $\beta/\sqrt{\lambda}$ makes the effective $\lambda$ equal to 1, so we assume

$$\lambda = 1 \tag{14}$$

from now on without loss of generality.

Setting $\frac{\partial}{\partial \tilde{X}_i}\{\ \} = 0$ where "{ }" is the braced quantity from the $\min^{-1}$ condition yields

$$\hat{X}_i = \frac{x_i+(\text{BML})y_i}{1+(\text{BML})^2} \tag{15}$$

Setting $\frac{\partial}{\partial \tilde{\beta}}\ \{\ \} = 0$ from the $\min^{-1}$ condition implies that

$$\text{BML} = \frac{\sum_{i=1}^{n} \hat{X}_i y_i}{\sum_{i=1}^{n} \hat{X}_i^2} \tag{16}$$

where $\hat{X}_i$ of (16) is itself a function of BML. Thus BML is of precisely the same form as BMLR, the ML estimator in regression, except that in place of the known $X_i$ in regression are estimates $\hat{X}_i$. (Note: Because the various approaches to the identifiability problem in EV are distinguished essentially by how they obtain the $\hat{X}_i$, it is in our view unfortunate that the $X_i$ in EV bear the name incidental parameters.)

"Unwrapping" the implicit characterization (16) yields

$$\text{BML} = \frac{\delta + \sqrt{\delta^2 + 4s_{12}^2}}{2s_{12}} \tag{17}$$

where

$$s_{11} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad s_{12} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i y_i \qquad s_{22} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

$$\delta \equiv s_{22} - s_{11}$$

BML is formed by minimizing the sum of squared-"residuals" taken perpendicular to the estimated line. I.e., BML is the orthogonal regression estimator on rescaled variables; see e.g., Malinvaud [7] Chaps. 1, 10.

## 4. BML as the "Best" ML/LS Estimator

The classical ML estimators that apply throughout the range of model parameters are those which assume knowledge of

$$\sigma_u^2 \qquad (\hat{\beta}_u, \text{ say});$$

or

$$\sigma_v^2 \qquad (\hat{\beta}_v, \text{ say});$$

or

$$\lambda = \frac{\sigma_v^2}{\sigma_u^2} \qquad \text{BML} \quad .$$

or <u>both</u> or $\sigma_u^2$ and $\sigma_v^2$ ($\hat{\beta}_{u,v}$, say). See Madansky [6].

Kendall and Stuart summarize a result of Birch that justifies calling BML the "best" of $\hat{\beta}_u$, $\hat{\beta}_v$, BML. Birch demonstrated that

$$BML = \hat{\beta}_{u,v}$$

except under conditions (the violation of certain inequalities relating sample product-moments to model parameters) which Kendall and Stuart state "seem unlikely [to hold] in practice". While we regard this last claim as perhaps a bit optimistic (because the violation occurred about 5-10% of the time in our simulations), this fact nevertheless allows us to understand why BML is at least as good an estimate of $\beta$, uniformly over the range of model parameters, as the better of the other two ML estimates which each require knowledge of exactly one of $\sigma_u^2$ or $\sigma_v^2$. Our simulations confirmed this property when both u and v are normally distributed.

Madansky gives a remarkable survey of the history of this estimator, asserting that the form of BML "has appeared independently innumerable times with the earliest appearance in 1879...".

It is also of interest to recall Malinvaud's laudatory remarks about BML. He derives an approximation to BML's asymptotic variance which, he says, "allows us to verify that, in the case [of one X-variate --- M.L.B.] and probably generally, the weighted regression has good asymptotic efficiency at least when the dispersion of the errors is small relative to that of the true variables, and when the errors are normally distributed". Comparing this variance with the minimum variance lower bound, he concludes that "the asymptotic efficiency is very near 1 if $[\frac{\text{variance of errors}}{\text{variance of true } X}]$ is small". Later: "[these] results... apply only to the asymptotic distribution of the [estimator BML]. Unfortunately there seems to exist no study of the properties of this [estimator] for finite samples."

## 5. Monte Carlo Results for BML

### 5.1 Specification

We consider first the performance of the ML/LS estimator BML under contamination; for comparison we include

$$BMLR = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{18}$$

BMLR is the usual ML/LS estimator that would be appropriate if x were error-free (i.e. $u \equiv 0$, $x \equiv X$).

We choose model parameters which, except for the contamination in y only, are symmetric in x and y:

$n = 20;100$

$\beta = 1.$

$\sigma_X^2 = 1.$

$\sigma_u^2 = 0.50^2 \qquad \delta_u = .00 \qquad h_u = .0$

$\sigma_v^2 = 0.50^2 \qquad \delta_v = .05 \qquad h_v$ varying

Note that this implies $\lambda = 1$.

We are thus considering both the small-sample and the large-sample performance of BML. When n = 20, e.g., an average of (.05) (20) = 1 out of 20 observations is drawn from $N(0,h_v^2)$ and the rest from $N(0,.50^2)$. Notice that $h_v = 0.50$ corresponds to classical EV.

Table 1. $\gamma_V = 0.05; \beta=1.$

100 replications

| n | $h_V$ | MSE(BMLR) | MSE(BML) |
|---|---|---|---|
| 20 | 0.50 | .0499 | .0391 |
| 20 | 1.00 | .0501 | .0474 |
| 20 | 1.25 | .0506 | .0559 |
| 20 | 1.50 | .0513 | .0694 |
| 20 | 2.0 | .0533 | .1261 |
| 20 | 3.0 | .0602 | 2.450 |
| 20 | 10.0 | .2120 | 246. |
| 100 | 1.50 | .0395 | .0119 |
| 100 | 3.0 | .0401 | .105 |
| 100 | 10.0 | .0580 | 26.9 |

5.2 Comments on Table 1

It is in our view difficult to overstate the gravity - and the irony! - of the results of Table 1, which tables the quantity

$$MSE(\cdot) = \frac{1}{100} \sum_{i=1}^{100} (\cdot_i - 1.)^2$$

At n=20 we see that when just one observation is drawn from $N(0,1.25^2)$ instead of from $N(0,.50^2)$, BML is a poorer estimate than BMLR, which ignores the error in x! Contamination this light would almost always be indistinguishable from pure gaussian sampling. By the time $h_V$ becomes "very noticeably" large, BML's distribution has acquired outrageously heavy tails, while BMLR has kept relatively stable. (See the $h_V$=10 entries.) The transition at n=100 occurs between $h_V$=1.5 and 3., which is still more than small enough to go unnoticed in a real sampling situation. BML's performance is thus so shockingly poor as to make the ML/LS estimator appropriate for the regression model, BMLR -- whose non-robustness properties are by now notorious -- look almost well-behaved by comparison! For this reason, we come to the ironic conclusion that those investigators who have "looked the other way" when the possibility arose of error in the independent variable in their "regression" models, and used BMLR "by default", instead of BML which requires knowledge of $\lambda$, probably made the better choice. But this is a choice between Scylla and Charybdis; the next section proposes a way out of this strait.

## 6. "Robustizing" BML : w-Estimation

### 6.1 Introduction to w-Estimation in Regression (R)

The w-estimator in R, BWR, with preliminary slope $\check{\beta}$, robust scale-measure $\underline{s}$ of residuals, and "psi-function" $\psi$, is defined as

$$\text{BWR} = \min_{\tilde{\beta}}^{-1} \sum_{i=1}^{n} [\check{w}_{yi} \cdot (y_i - \tilde{\beta} X_i)^2] \ , \qquad (19)$$

where

$$\check{r}_{yi} = y_i - \check{\beta} X_i \qquad (20)$$

is the residual from the preliminary slope $\check{\beta}$, and

$$\check{w}_{yi} = \frac{\psi(\check{r}_{yi}/\check{s}_y)}{\check{r}_{yi}/\check{s}_y} \qquad (21)$$

is a weight superposed onto the $i^{th}$ residual serving to "damp the influence of" the $i^{th}$ point on BWR to the extent that it is diagnosed as an outlier. Notice that when $\psi(r) \equiv r$, BWR=BMLR. See Beaton and Tukey [ 2 ] and Andrews et. al. [1].

## 6.2 Proposed w-Estimator in EV

We propose the following natural generalization BW of BWR to EV. BW involves two weights $\check{w}_{xi}$, $\check{\check{w}}_{yi}$ each intended to "weight-down an outlying coordinate" of the $i^{th}$ point, superposed onto the ML/LS estimator in EV, $\beta^{EV}$: for $\psi$, $\check{\beta}$, and s with the meanings as in R, we set

$$BW \equiv \min_{\tilde{\beta};\tilde{X}_1,\ldots,\tilde{X}_n}^{-1} \sum_{i=1}^{n} [\check{\check{w}}_{yi} \cdot (y_i-\tilde{\beta}\tilde{X}_i)^2 \tag{22}$$

$$+ \check{w}_{xi} \cdot (x_i-\tilde{X}_i)^2] \quad ,$$

where

$$\check{\check{r}}_{yi} = y_i - \check{\beta}\check{X}_i \tag{23a}$$

$$\check{r}_{xi} = x_i - \check{X}_i \tag{23b}$$

are the two EV residuals associated with v and u respectively and

$$\check{\check{w}}_{yi} = \frac{\psi_y(\check{\check{r}}_{yi}/\check{\check{s}}_y)}{\check{\check{r}}_{yi}/\check{\check{s}}_y} \tag{24a}$$

$$\check{w}_{xi} = \frac{\psi_x(\check{r}_{xi}/\check{s}_x)}{\check{r}_{xi}/\check{s}_x} \tag{24b}$$

are the two weights. Notice that we now require a preliminary estimate $\check{X}_i$ for each $X_i$ as well as a $\check{\beta}$ for $\beta$.

Further on we discuss the some data-based possibilities for choosing $\check{\beta}$ and $\check{X}_i$.

Taking the (n+1) derivatives of the weighted $\min^{-1}$ condition (22) yields, after simplifying, this $6^{th}$ degree equation for BW:

$$\sum_{i=1}^{n} \{\hat{k}_i \check{w}_{yi} [\check{w}_{xi}^2 x_i y_i + (\check{w}_{xi} \check{w}_{yi} y_i^2 - \check{w}_{xi}^2 x_i^2) \cdot BW - \check{w}_{xi} \check{w}_{yi} x_i y_i (BW)^2]\} = 0 \qquad (25)$$

where

$$\hat{k}_i \equiv \frac{1}{[\check{w}_{xi} + \check{w}_{yi}(BW)^2]^2} \quad . \qquad (26)$$

### 6.3 A Simplification

Choosing $\check{X}_i$ to have the ML/LS form (15) as a function of $\check{\beta}$ we may verify easily that

$$\check{r}_x = -\check{\beta}\check{r}_y \quad . \tag{27}$$

Thus $\{\check{r}_{xi}\}$ and $\{\check{r}_{yi}\}$, for this choice of $\check{X}$, yield the same measures of scale $\check{s}_x$, $\check{s}_y$ whence (assuming $\psi_x=\psi_y$) we have

$$\check{w}_{xi} \equiv \check{w}_{yi} \; , \qquad i=1,\ldots,n \; . \tag{28}$$

Substituting this common weight ($\check{w}_i$ say) into the $\min^{-1}$ condition (22) shows that BW has the form (17) of BML except that

$$t_{12} = \frac{1}{n} \sum_{i=1}^{n} \check{w}_i x_i y_i \tag{29}$$

replaces $s_{12}$ and corresponding weighted moments $t_{11}$, $t_{22}$ replace $s_{11}$, $s_{22}$.

### 6.4 Details of the Implementation

We have used

$$\psi_x\left(\frac{r}{s}\right) = \psi_y\left(\frac{r}{s}\right) = \begin{cases} \frac{r}{s}\left(1 - \left(\frac{1}{c} \cdot \frac{r}{s}\right)^2\right)^2 & \text{if } \left(\frac{r}{s}\right) \le c \\ 0 & \text{if } \left(\frac{r}{s}\right) > c \end{cases} \tag{30}$$

with c = 5.0.

This $\psi$-function, due to Tukey, is known as the "bisquare". Our robust scale is due to Hampel:

$$s(\{r_i\}) \equiv \underset{i}{\text{median}} \quad \{(r_i - \underset{j}{\text{median}}\ \{r_j\})\} \tag{31}$$

In order to separate the issues of preliminary and final robust estimators, we set

$$\check{\beta} = \beta \tag{32}$$

for the present w-simulations. Thus BW in our simulations takes "one step away from the true $\beta$", and these results indicate the character of performances to be expected with a good data-based preliminary estimator $\check{\beta}$. See later remarks.

We also exhibit the performances of the regression w-estimator BWR, which is (19) with x in place of X.

Table 2. Same 100 samples as for Table 1

| n | $h_y$ | MSE(BWR) | MSE(BW) |
|---|---|---|---|
| 20 | 0.50 | .0274 | .0211 |
| 20 | 1.0 | .0277 | .0240 |
| 20 | 1.25 | .0277 | .0257 |
| 20 | 1.50 | .0277 | .0263 |
| 20 | 2.0 | .0273 | .0250 |
| 20 | 3.0 | .0271 | .0248 |
| 20 | 10.0 | .0317 | .0225 |
| 100 | 1.50 | .0211 | .00342 |
| 100 | 2.0 | .0213 | .00363 |
| 100 | 3.0 | .0215 | .00361 |
| 100 | 10.0 | .0223 | .00322 |

### 6.5 Comments on Table 2

Table 2 exhibits the MSE's for BW and for BWR corresponding to the situations of Table 1. Besides the artifactitious "superefficiency" induced by assuming $\check{\beta} = \beta$ (so that MSE(BWR)<MSE(BML) when $h_v$ is near 0.50), we see that MSE (BW) remains below MSE(BWR) in all cases. In other words, superposing the w-weighting allows the error-in-x correction using $\lambda$ to operate on an effectively uncontaminated sample.

### 6.6 Influence of sample size n

The ratio

$$\frac{\text{MSE(BW)}}{\text{MSE(BWR)}} \tag{33}$$

decreases as n increases for fixed $\gamma_v$, $h_v$. This accords with our expectation, for as n increases both variances decrease like 1/n but bias-squared approaches a non-zero limit in the case of BWR and zero in the case of BW because of BW's bias-correction using $\lambda$.

## 7. Data-based Preliminary Estimators

### 7.1 A preliminary slope

One straightforward method of obtaining a $\check{\beta}$ for EV might be to combine good preliminary regression estimators, $\check{\beta}_{yx}$ and $\check{\beta}_{xy}$, in the manner that BML combines BMLR and

$$\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i y_i} . \tag{34}$$

We have simulated

$$\text{BL} = \Delta + (\text{sgn(BYX)}) \cdot \sqrt{\Delta^2 + \lambda} \tag{35a}$$

where

$$\Delta = [\text{BRXY} - \lambda \cdot \frac{1}{\text{BYX}}] \tag{35b}$$

with BYX the "robust line" ("medians-of-thirds grouping" estimator) of Tukey [ 8 ], and BRXY the reciprocal of the same estimator corresponding to regression of x on y.

Table 3.

| n | $(\gamma_u,h_u)$ | $(\gamma_v,h_v)$ | MSE(BYX) | MSE(BRXY) | MSE(BL) |
|---|---|---|---|---|---|
| 20 | (.0,0) | (.0,0) | .06499 | .3777 | .1359 |
| 50 | (.0,0) | (.0,0) | .0543 | .1327 | .0260 |
| 20 | (.0,0) | (.05,3) | .0654 | .5822 | .1921 |
| 50 | (.0,0) | (.05,10) | .0599 | .6985 | .1637 |
| 20 | (.05,10) | (.0,0) | .1037 | .4244 | .1604 |
| 50 | (.05,10) | (.0,0) | .1026 | .1123 | .0417 |

### 7.2 Comments on Table 3

Either a small sample or contamination in y renders BRXY sufficiently unstable as to make MSE(BL)>MSE(BYX). But for moderately large n and contamination in x only, the estimator BL with λ-correction improves on BYX, the "regression" slope. Notice that contamination in x alleviates some of BRXY's overestimating bias at $\gamma_u$=.0!

### 7.3 LAR estimation in EV

One of the best-known proposals for a preliminary estimator in regression is the LAR (least-absolute-residuals) slope

$$\check{\beta} = \min_{\tilde{\beta}}^{-1} \sum_{i=1}^{n} | y_i - \tilde{\beta} X_i | .$$

The EV generalization requires $\check{\beta}$, $\check{X}_1,\ldots,\check{X}_n$ jointly to minimize

$$\sum_{i=1}^{n} \{| y_i - \tilde{\beta} \tilde{X}_i | + | x_i - \tilde{X}_i|\} ;$$

The solution (C. Mallows, 1973 personal communication) is $\tilde{\beta}$ to minimize the smaller of

$$\sum_{i=1}^{n} | y_i - \tilde{\beta} x_i | \quad , \quad \sum_{i=1}^{n} | x_i - \frac{y_i}{\tilde{\beta}} | \ .$$

In classical EV, we may remark that this essentially computes the LAR estimator of y on x or of x on y according as which of x or y respectively has the smaller "noise-to-signal ratio". We conjecture that this estimator would be very satisfactory in, and only in, "large" samples in contaminated EV.

BIBLIOGRAPHY

[1] Andrews, et. al., Robust Estimates of Location: Survey and Advances, Princeton University Press (Princeton, New Jersey), 1972.

[2] Beaton, A.E. and J.W. Tukey, "The Fitting of Power Series, Meaning Polynomials, Illustrated on Bandspectroscopic Data," Technometrics, Vol. 16, No. 2, May 1974, pp. 147-192.

[3] Huber, P., "The 1972 Wald Lecture. Robust Statistics: A Review," Ann. Math. Stat., Vol. 43, No. 4, 1972, pp. 1041-1067.

[4] Kendall, M.G. and Stuart, A., The Advanced Theory of Statistics, Vol. 2, 3rd ed., Griffin (London), 1973.

[5] Lindley, D.V., "Regression Lines and the Linear Functional Relationship," Jour. Royal Stat. Soc., Supp., 9, 1947, 219-244.

[6] Madansky, A., "The Fitting of Straight Lines When Both Variables are Subject to Error," Jour. Amer. Statist. Assn., 54, March 1959, pp. 173-205.

[7] Malinvaud, E., Statistical Methods of Econometrics, 2nd ed., American Elsevier Pub. Co. (New York), 1970.

[8] Tukey, J., Exploratory Data Analysis, limited preliminary edn., Vol. I, Addison-Wesley Pub. Co. (Reading, Mass.) 1972.