Errors in Variables or Bad Leverage at Some Observations ?

Eric Blankmeyer

Department of Finance and Economics
McCoy College of Business Administration
Texas State University – San Marcos
601 University Drive
San Marcos TX 78666
eb01@txstate.edu

October 2011

Abstract. Errors-in-variables is a long-standing, difficult issue in linear regression; and progress depends in part on new identifying assumptions. I characterize measurement error as bad-leverage points and assume that fewer than half the sample observations are heavily contaminated, in which case a high-breakdown robust estimator may be able to isolate and downweight or discard the problematic data. In simulations of simple and multiple regression where eiv affected 25% of the data and R^2 was mediocre, one high-breakdown estimator had small bias, very good coverage, and precision that improved when the sample size increased.

Key words. Errors in variables, measurement error, high-breakdown estimator, minimum covariance determinant, orthogonal regression

Errors in Variables or Bad Leverage at Some Observations ?

Introduction

A venerable issue in linear-regression analysis is errors in variables (eiv, also called measurement error), when a regressor is not directly observable. Instead, a proxy is available that differs from the regressor because of random contamination. In ordinary least squares (ols) estimation, eiv produces a bias that does not vanish asymptotically. Many researchers attest that this is a pervasive and challenging problem. Theil (1971, p. 607-613) reviews some procedures for dealing with eiv, “none of which is really simple in application.” According to Malinvaud (1980, p. 416), “If in many cases structural parameters can be identified, consistent estimators appropriate to such cases are virtually useless in econometrics, since there are too few data.” Friedman (1992, p. 2131) comments that “the common practice is to regress a variable Y on a vector of variables X and then accept the regression coefficients as supposedly unbiased estimates of structural parameters, without recognizing that all variables are only proxies for the variables of real interest, if only because of measurement error, though generally also because of transitory factors that are peripheral to the subject under consideration. I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data, alleviated only occasionally by consideration of the bias introduced when ‘all variables are subject to error.’ " Dagenais and Dagenais (1997, p. 195) note that, for ols, “intended 95% confidence intervals may in practice turn out to be almost 0% intervals, even when the errors of measurement are not exceedingly large….Similarly, Student t-tests using the critical values corresponding normally to 5% type I errors may in fact correspond to tests with type I errors of size equal to almost 100%! This may have dramatic consequences since one may be induced to reject a null hypothesis when this hypothesis is true, with a probability close to 100%!” Greene (2003, p. 84) remarks that the “general assessment of the problem is not particularly optimistic. The biases introduced by measurement error can be rather severe.” And Hausman (2001, p. 58) speaks of “ the ‘Iron Law of

Econometrics' –the magnitude of the estimate is usually smaller than expected."

Eiv models assume that *all the observations* are potentially contaminated by measurement error in one or several regressors. Since in many situations this assumption will be unduly pessimistic, I explore eiv estimation when egregious measurement errors affect *only a minority of observations* –a subset that can be characterized as bad-leverage points. A robust high-breakdown estimator can then locate and downweight these bad-leverage observations. My simulations emulate data which are noisy due to eiv and also because R^2 for the correctly-measured variables is not very high. In this challenging framework, which may be fairly typical of cross-section data, the initial estimate of the minimum-covariance-determinant procedure seems to perform well whereas a robust estimator designed specifically for linear regression is less successful in terms of bias control and accurate coverage of a confidence interval.

Eiv has generated a vast literature, and a comprehensive review is beyond this essay's scope. In addition to the papers and monographs cited below, some important surveys are Klepper and Leamer (1984), Fuller (1987), Hausman (2001), Söderström (2007), and the papers edited by Van Huffel (2007). The next section is a concise review of the basic eiv model and several of the estimators proposed for it. Then follows a section in which measurement errors are interpreted as bad-leverage observations; and, subject to an identifying assumption, high-breakdown estimators are proposed to cope with eiv. Six simulations are presented and discussed, followed by a brief examination of two actual data sets and some concluding remarks.

A canonical eiv model

Although researchers have explored many variations of the measurement-error problem in linear regression, I focus on a canonical eiv model:

$$y_i = \alpha + \beta x_i + u_{yi} \quad , \qquad (1)$$

where α and β are unknown parameters, y_i is an observation on the dependent variable, and u_{yi} is an unobservable normal random variable, independently and identically distributed (iid) with zero expectation and standard deviation σ_y. The regressor x_i is also unobservable; instead a researcher observes $x_i + u_{xi}$, where u_{xi} is an iid normal error with zero expectation and standard deviation σ_x. It is assumed that x_i, u_{xi} and u_{yi} are stochastically independent of one another. The object is to obtain good estimates of α, β, and σ_y from a random sample of y_i and $x_i + u_{xi}$. The eiv problem is that x_i and u_{xi} are never observed separately but always as $x_i + u_{xi}$, a mismeasured regressor that is correlated with the regression disturbance $u_{yi} - \beta u_{xi}$. Consequently, the ols estimate of β is inconsistent: it converges to $\beta/(1 + \sigma_x^2 / \text{plim}(\Sigma x_i^2/n))$, so the bias is toward 0 –the notorious "least-squares attenuation." In multiple linear regression, which will also be examined in simulations and actual data, the eiv bias may skew all the estimated coefficients. In addition, it can happen that several regressors are mismeasured. In either case, the direction of ols bias becomes problematic in general (Greene 2003, p. 85-86).

Moreover, if the correctly-measured but unobservable regressor x_i is also normal iid, the parameters of interest are not even identifiable, hence the non-uniqueness of the maximum-likelihood estimator in the absence of additional information or assumptions. In model (1), identification is achieved if one knows (or is willing to guess) the value of σ_x^2 or $\text{plim}(\Sigma x_i^2/n)$ or their ratio. It is argued that, in some contexts, there could be extra-sample information about the size of the measurement error. In particular, if the researcher believes that $\sigma_x^2 / \sigma_y^2 \sim 1$, then the maximum-likelihood estimator is the orthogonal regression (org), the eigenvector corresponding to the smallest eigenvalue of the covariance matrix of y_i and $x_i + u_{xi}$. Org is perhaps the most widely-used eiv technique (but see Carroll and Ruppert 1996). In principle it can be extended to multiple linear regression with several mismeasured regressors although it does not seem very plausible that the required extra-sample information would often be available. Latent-variable models and factor analysis have also been used extensively to model measurement error. In that methodology, identification is of course dependent on rather subjective decisions about the number of factors to be

included and the choice of a “rotation” criterion (varimax, quartimax, and so on).

The eiv literature explores several other strategies for the identification and consistent estimation of linear models, all of which can be interpreted as instrumental variables and therefore reflect the strengths and weaknesses of that procedure. For example, it is suggested that in model (1) the sample data be split into groups according to some *a priori* criterion (the instrument) and that the regression be performed on the group means in the belief that the u_{xi} will average to zero within each group, at least asymptotically. As Malinvaud (1980, p. 416-419) explains, consistent estimation makes “two demands which are often contradictory. For it is necessary that” the groups be chosen independently of the u_{xi} but also in a way that the group means of the dependent variable do not all converge to $E(y_i)$ since there would then be little or no variation in the dependent variable. As an alternative to grouping, it is suggested that the instrument be formed from the ranking of $x_i + u_{xi}$ in the hope that the ranks will be independent of the u_{xi} but strongly correlated with the x_i.

Another instrumental-variable approach achieves identifiability by assuming that, while the u_{xi} are normally distributed, the x_i are not; then instruments can be generated from the higher-order moments of the observations $x_i + u_{xi}$. Important contributions to this literature include Dagenais and Dagenais (1997) and Erickson and Whited (2002). For model (1), the method-of-moments (mom) estimator of β is

$$\Sigma(x_i + u_{xi})(y_i)^2 / \Sigma(x_i + u_{xi})^2 y_i \qquad (2)$$

if the sample values are measured as deviations from their respective means. Unless the x_i are distinctly non-normal, exhibiting for example significant kurtosis or skewness, the instrumental variables will be weak, leading to a large standard error for the estimate of β.

Moreover, like every other eiv procedure discussed in this paper, the mom estimator raises the issue of “the whimsical character of inference, how adequately to base inferences on opinions when facts are unavailable” (Leamer 1983, p. 38). For estimator (2), the key assumption that the unobservable x_i are definitely non-normal may appear whimsical since

econometricians instinctively assume normality in many other contexts. While this brief discussion has merely skimmed the rich variety of eiv models found in the literature, I have alluded to the whimsicality or fragility of the identifying assumptions for several leading eiv estimators. Although it is futile to expect a universally acceptable strategy for identification in situations involving measurement error, progress on the eiv problem depends on the formulation of new identifying assumptions that are credible for a well-defined but reasonably wide range of real data sets.

In the sequel I offer an identifying strategy that has not, as far as I know, been proposed before. It will strike some researchers as whimsical, but I believe that it has considerable intuitive appeal. Moreover, the strategy utilizes statistical procedures that have proved effective for detecting anomalous observations in data sets from business, economics and finance (Haezendonck et al. 2001, Knox et al. 2001, Zaman et al. 2001, Boudt et al. 2008); in other social sciences (Maes et al. 1998); in technology (Mili et al. 1991, Rousseeuw and Van Aelst 1999, Rousseeuw and Van Driessen 1999, Prieto et al. 2009); and in the natural sciences, including chemistry and astronomy (Hubert et al. 2002, Rousseeuw and Van Driessen 1999).

Bad leverage and high-breakdown estimators

As previously mentioned, ols is a biased estimator of model (1) because x_i and u_{xi} are never observed separately; all the observations are potentially contaminated. While this premise is no doubt realistic in some contexts, it seems unduly pessimistic for many actual data sets, where egregious measurement error may well be limited to a minority of the observations. The high-breakdown estimators used in the sequel can in principle cope with contamination in as much as 50% of the data. The rationale for this upper bound is that, when it comes to avoiding very large biases (“breakdown”), no affine-equivariant estimator for linear regression can distinguish between valid and invalid observations if the latter are in the majority (Rousseeuw and Leroy 1987, chapters 1 and 3; Maronna et al. 2006, chapters 3, 5 and 6). One leading researcher recommends a “default coverage” of 75%, meaning that anomalous observations are assumed to

affect at most 25% of the sample (Rousseeuw, Van Aelest, Hubert 1999, p. 425). I adopt this viewpoint in order to generate the simulations in the next section, acknowledging (per Leamer) that the choice is based on a mixture of opinion and experience. More generally*, the new identifying assumption is that eiv affects less than half of the sample; in a majority of observations, u_{xi} is negligible. The estimation strategy is simply to use high-breakdown methods that can detect and downweight or eliminate the mismeasured observations, those for which u_{xi} is not negligible.*

Now Rousseeuw and Van Driessen (2006, p. 29) offer a taxonomy of outliers: a point for which y_i diverges from the linear pattern of the majority of the data but whose regressors are not outlying is called a *vertical outlier*. A point with one or more outlying regressors is a *leverage point*. A *good* leverage point lies far from the majority of observations but near to the regression plane implied by the majority. A *bad* leverage point lies far from the majority of observations and their implied regression plane. "Summarizing, a data set can contain four types of points: regular observations, vertical outliers, good leverage points, and bad leverage points. Of course, most data sets do not have all four types."

In model (1), a bad-leverage observation occurs when variation in a regressor is not matched by a corresponding variation in the dependent variable. Measurement error produces bad-leverage points because u_{xi} is uncorrelated with y_i. Figure 1 displays a pseudo sample of 2000 bivariate observations, 25% of which is contaminated with eiv. (It is in fact a simulation summarized in Table 2 and discussed in the next section.) The correctly-measured data are concentrated in the central ellipse whose principal axis has a slope of 1 approximately. The mismeasured observations mostly protrude horizontally to the left and right of the central ellipse; they are the bad-leverage points whose excess variation (the u_{xi}) flattens out or attenuates the ols slope estimate. *Accordingly, I interpret eiv as a type of bad-leverage observation.* If the proportion of mismeasured observations is not excessive, an appropriate high-breakdown estimator will focus on the data clustered in the central ellipse of Figure 1 and will therefore produce a good estimate the regression line.

The statistical properties of leading high-breakdown estimators (including the requirements for consistency and asymptotic normality) have

been detailed elsewhere, as have the algorithms for their computation (Maronna et al. 2006, chapters 5, 6 and 9; Rousseeuw and Leroy 1987, chapters 3 and 5; Rousseeuw and Van Driessen 1999, 2006). It is only necessary to remark that high-breakdown estimation proceeds in two stages. The first step is to compute an initial, very robust estimate that has low statistical efficiency; i.e., this estimate often downweights some regular observations and good-leverage points along with the vertical outliers and bad-leverage points. Starting from the initial estimate, the second stage performs one or more iterations by robustly-weighted least squares to reinstate the valid data and thereby boost the efficiency of the final estimate.

However, my simulations focus primarily on the initial estimates. This is because all the simulations except Table 1 are designed to generate challenging and realistic samples in which the "true" linear relationship between x_i and y_i is mediocre, with R-squared in the range of 0.30-0.35. Cross-section data in economics and other fields are frequently quite noisy (σ_y is relatively large), and preliminary work indicated that the second-stage (efficient) high-breakdown estimators tend to retain unacceptably large eiv biases in these difficult environments. After all, a high breakdown point guarantees that the estimator's bias is finite but not that it is small. Nowadays, moreover, data sets often have a great many observations, in which case an estimator's efficiency is less important than its ability to control bias.

Among high-breakdown estimators, it seems obvious to choose one that is designed for linear regression. I use the Robust MM method (Maronna et al. 2006, chapter 5), hereafter denoted mmr; however, the least-trimmed-squares method (Rousseeuw and Leroy 1987, Rousseeuw and Van Driessen 2006) would be equally appropriate. Nevertheless, the simulations suggest that even an initial robust estimate from a regression-based method may fail to cope with eiv when R^2 is mediocre. Accordingly, I also use the initial ("raw") estimate from the Minimum Covariance Determinant algorithm (Rousseeuw and Leroy 1987, chapter 7; Rousseeuw and Van Driessen 1999), denoted mcd. For a data set containing one or more continuous-valued regressors and a continuous-valued dependent variable, the mcd searches for the subsample containing 75% of the

observations whose covariance matrix has the smallest determinant (volume). From that subsample --hopefully uncontaminated by eiv-- the covariance matrix and its corresponding mean vector are used to compute a robustified ols regression. (In several simulations, the Mahalanobis distances from the initial mcd estimate are employed to create robust weights for one round of weighted least squares. The efficiency improvement is denoted mcdw in Tables 1, 2 and 3.)

This paper is by no means the first to juxtapose eiv and high-breakdown estimation. Previous work includes Fekri and Ruiz-Gazen 2004, 2006; Jung 2007; Maronna 2005; Rousseeuw and Leroy 1987, p. 284-285; and Zamar (1989). However, all these authors make a distinction between bad-leverage points and measurement error, whereas I see no difference in practice. These authors proceed in two stages: first they apply a high-breakdown estimator to deal with a limited number of vertical outliers and bad-leverage points; then they use some version of orthogonal regression to handle eiv, which is assumed to affect the entire sample. On the other hand, I assume that eiv seriously affects less than half the sample, where it shows up as bad-leverage points. Therefore only the first stage is required, and it also deals with vertical outliers if they are not too numerous. There is no need for the often-problematic identifying assumptions of orthogonal regression.

Some statistical software environments that currently implement high-breakdown estimators are MATLAB (Verboven and Hubert 2005), R (Konis 2011, Maechler 2011, Todorov 2011), SAS (Chen 2002), and STATA (Verardi and Croux 2009). To compute mmr, I use lmRob from Konis (2011); for mcd, I use covMcd from Maechler (2011).

Simulations

This section reports six simulations of the eiv model. Each simulation has these characteristics: 0 is the value of the intercept α; 1 is the value of the true slope coefficient(s) β; the sample is replicated 1000 times; and for each sample, the regressor(s) are contaminated with measurement error in 25 percent of the observations unless stated otherwise. For the slope coefficient(s), Tables 1 through 6 report the bias, the root mean squared

error (rmse), and the actual coverage for a nominal 90-percent confidence interval. (The interval is computed as the average value of the coefficient in the simulation ± 1.65 times the coefficient's standard deviation in the simulation. Coverage is the proportion of samples for which the computed interval contains 1, the value of β.) In the text and tables, $z \sim N(\mu,\sigma)$ denotes a normally-distributed random variable z with expectation μ and standard deviation σ; and n denotes the sample size. The estimators to be simulated are ols, mmr, mcd, mcdw, org, and mom.

For example, Table 1 reports the simulation of a bivariate eiv regression in which the correlation between y_i and x_i is rather high: R^2 = 0.800: specifically, $u_{xi} \sim N(0,4)$, $u_{yi} \sim N(0,1)$ and $x_i \sim N(0,2)$. For 200 observations, ols has a downward bias of about 50 percent with negligible sampling error; in other words, ols is efficient but has a large bias and correspondingly poor coverage. On the other hand, the downward biases of mmr and mcd are less than 5 percent; and their coverages are close to the nominal level. For mcdw, the bias is slightly larger in magnitude and the coverage is a little worse. When n = 2000, the ols results are unchanged. The biases of mmr, mcd and mcdw remain numerically small, but only mcd still has coverage near the nominal level.

For the bivariate regression in Table 2, the correlation between y_i and x_i is mediocre: R^2 = 0.310, which may be more typical of cross-section data sets. Again ols has a large bias and no coverage, but now mmr also performs poorly for both sample sizes whereas mcd and mcdw have rather small biases. When n = 200, the trade-off between mcd and mcdw is apparent: the former has less bias, the latter has less sampling variation and hence a smaller rmse; both have very good coverage. When n = 2000, the coverage of mcd remains near the nominal level but that of mcdw deteriorates notably.

The canonical eiv model (1) assumes that the expected value of u_{xi} is 0; but as Malinvaud (1980, p. 384-385) suggests, this may be unrealistic: measurement error could also change the level of the observed regressor. Table 3 repeats the simulation in Table 2 except that the mean of u_{xi} is now 10 instead of 0. The biases of ols and mmr balloon to more than 80 percent, but mcd has negligible bias and excellent coverage; also its rmse shrinks dramatically when n is 2000 instead of 200.

Table 4 summarizes a multiple regression for which R^2 would be about 0.35 in the absence of eiv. However, both regressors (x and z) are affected by eiv. Specifically, n = 1000 and each regressor has 125 measurement errors at non-overlapping observations, so total contamination is again 25 percent. Unlike ols and mmr, mcd has negligible bias and excellent coverage. When n = 200, the three estimators have similar rmse; but mcd has the largest rmse when n = 2000. This simulation indicates that, as long as total contamination is not too large, mcd can be applied in situations where it is suspected that eiv affects more than one regressor.

Table 5 reverts to bivariate regression and compares ols and mcd to org and mom. Here u_{xi} and u_{yi} have the same standard deviation, which should favor org; but u_{xi} and x_i are both normally distributed, so mom has useless instruments. When the eiv contamination is 25 percent, mcd performs best in terms of bias, rmse and coverage. The bias of org is actually positive and large while mom has a huge bias and rmse. However, org performs very well when eiv is present in all observations (100 percent contamination), while mcd is no better than ols and mom is very bad. These latter results reflect a situation that is ideal for org but fatal for mcd, which is vulnerable to breakdown whenever contamination approaches or exceeds 50 percent.

Table 6 compares the same four estimators when u_{xi} and u_{yi} have different standard deviations and x_i is a chi-squared variable with 2 degrees of freedom. This case is expected to be less favorable to org and more favorable for mom, whose instruments should be strong since x_i does not resemble a normal random variable. When eiv contamination is 25 percent, mom has the smallest bias and rmse together with excellent coverage while mcd is in second place; ols and org perform poorly. When eiv contamination is 100 percent, only mom performs well.

The preceding simulations can be extended in several directions. For example, one could explore correlations between u_{xi} and u_{yi} or between u_{xi} and x_i. Levels of eiv contamination below 25 percent and nearer to 50 percent could be examined, as could the effect of a single mismeasured regressor on the estimated coefficients of correctly-measured regressors. In addition, one could test alternative calibrations and implementations of

algorithms for the high-breakdown estimators mmr, mcd and mcdw. Of course, no set of simulations will be dispositive for the relative merits of the various eiv estimators. With this caveat in mind, I tentatively conclude that the simulations make a good case for trying mcd in situations where measurement error is suspected, especially in cross-section data sets where n is “large” and it is reasonable to assume that serious eiv contamination affects a minority of the observations.

Actual data sets

I now turn to a brief presentation of two cross-section samples for which the linear-regression estimates may be affected by eiv. In Table 7, food expenditure is regressed on income for 235 Belgian households (Koenker 2011). The slope coefficients for mmr and mcd are similar, and both are significantly larger than the ols estimate. If, as is often conjectured, income reported by households is subject to measurement error and transitory effects, then the low ols estimate probably reflects attenuation. The mmr and mcd estimates may be similar because R^2 is high.

In Table 7, the ols standard error is calculated as usual for the iid model, and Maronna et al. (2006, chapter 5) derive the asymptotic standard error for mmr. The mcd standard error is estimated using the “wild” bootstrap (Flachaire 2005). Salibian-Barrera and Zamar (2002) address the statistical and computational issues of bootstrapping high-breakdown estimators.

A simple hedonic housing-demand model is displayed in Table 8 (Bivand 2011). In a log-linear regression, the average sale price of a house is explained by the house’s age, lot size, and number of rooms. The table does not show standard errors, but the statistical significance of each regression coefficient is high, in part because of the large sample size. The eiv problem might arise because age is only a crude proxy for the house’s condition and need of repairs (indeed some older homes are *per se* more valuable), and the lot size is also a proxy that does not take into account the shape of the lot and its surface features. For these and other reasons, researchers often try to include additional regressors, especially spatially-weighted variables that capture the characteristics of neighboring houses

(LeSage and Pace 2009). In any case, Table 8 shows that the mmr estimates are closer to ols than to mcd. Does this reflect a situation like Tables 2, 3 and 4, where mmr performs poorly because R^2 is not very high?

Conclusion

Eiv is a long-standing, difficult issue in linear regression; and progress depends in part on new identifying assumptions. In this spirit, I characterize measurement error as bad-leverage points and assume that fewer than half the sample observations are heavily contaminated, in which case a high-breakdown estimator may be able to isolate and downweight or discard the problematic data. In simulations of simple and multiple regression where eiv affected 25% of the data and R^2 was mediocre, the initial (raw) mcd estimates had small bias, very good coverage of the 90% confidence interval, and precision that improved when the sample size increased (as evidenced by smaller rmse together with virtually unchanged bias). When the mcd is applied to actual data sets, the bootstrap or other resampling methods could provide standard errors and confidence intervals.

This paper has not addressed the eiv problem in non-linear and nonparametric regression (e. g., Schennach 2004a, 2004b), nor has it considered the complications that arise when the u_{xi} or the u_{yi} exhibit temporal or spatial dependence. Moreover, the presence of several dummy variables in a linear regression may raise computational and statistical issues for mmr and mcd (Blankmeyer 2006; Maronna et al. 2006, p. 361-362).

Researchers who are already making routine use of high-breakdown estimators might conclude that this paper offers them little new information, and they would be substantially correct since I am simply associating eiv with bad-leverage points. However, those researchers may want to note the poor performance of mmr when R^2 is mediocre. In Figure 1 the valid observations lie inside a compact ellipse, which suggests why mcd is preferable to mmr in this situation (compare Rousseeuw and Leroy 1987, Figure 13 on p. 70).

In view of its canonical status and pedagogical value, I have emphasized the bivariate eiv model (1). However, Tables 4 and 8 make the point that high-breakdown estimation is especially advantageous for multiple linear regression, where scatter plots are less effective in detecting bad-leverage observations and where conventional outlier diagnostics can be quite misleading (Rousseeuw and Leroy 1987, chapter 3 and 6).

References

Bivand, R. (2011). Spatial dependence: weighting schemes, statistics and models (R package 'spdep'), http://cran.r-project.org/web/packages, accessed April 30, 2011.

Blankmeyer, E. (2006). How robust is linear regression with dummy variables? http://ecommons.txstate.edu/fiaefacp/2, accessed May 31, 2011.

Boudt, K., B. Peterson, C. Croux (2008), Estimation and decomposition of downside risk for portfolios with non-normal returns. *Journal of Risk* (11): 1-25.

Carroll, R., D. Ruppert (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician* 50(1): 1-6.

Chen, C. (2002). Robust regression and outlier detection with the ROBUSTREG Procedure. Paper 265-27. Cary, NC: SAS Institute, Inc. http://www2.sas.com/proceedings/sugi27/p265-27.pdf, accessed June 1, 2011.

Dagenais, M., D. Dagenais (1997). Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics* 76: 193-221.

Erickson, T., T. Whited (2002). Two-step GMM estimation of the errors-in-variables model using higher-order moments. *Econometric Theory* 18(3): 776-799.

Fekri, M., A. Ruiz-Gazen (2004). Robust weighted orthogonal regression in the errors-in-variables model. *Journal of Multivariate Analysis* 88: 89-108.

Fekri, M., A. Ruiz-Gazen (2006). Robust estimation of the simple errors-in-variables model. *Statistics and Probability Letters* 76(16): 1741-1747.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis* 49, 361-376.

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

Friedman, M. (1992). Do old fallacies ever die ? *Journal of Economic Literature* 30(4): 2129-2132.

Greene, W. (2003). *Econometric Analysis*. Fifth edition. Upper Saddle River, NJ: Prentice-Hall.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic Perspectives* 15(4): 57-67.

Haezendonck, E., G. Pison, P. Rousseeuw, A. Struyf, A. Verbeke (2001), The core competences of the Antwerp seaport: an analysis of "port specific" advantages. *International Journal of Transport Economics* (28): 325-349.

Hubert, M., P. Rousseeuw, S. Verboven (2002), A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems* (60): 101-111.

Jung, K-M. (2007). Least trimmed squares estimator in the errors-in-variables model. *Journal of Applied Statistics* 34(3): 331-338.

Klepper,S., E. Leamer (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica* 52(1): 163-184.

Knox, K., E. Blankmeyer, J. Stutzman (2001). Organizational structure, performance, Quality, and administrative compensation in Texas nursing facilities. *Quarterly Journal of Business and Economics* (40): 45-67.

Koenker R. (2011). Quantile regression (R package 'quantreg'), http://cran.r-project.org/web/packages, accessed April 30, 2011.

Konis, K. (2011). Insightful Robust Library (R package 'robust'), http://cran.r-project.org/web/packages, accessed April 30, 2011.

LeSage, J., R. Pace (2009). *Introduction to Spatial Econometrics*. Boca Raton FL: CRC Press/Taylor & Francis.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review* 73(1): 31-43.

Maechler, M. (2011). Basic robust statistics (R package 'robusbase'), http://cran.r-project.org/web/packages, accessed April 30, 2011.

Maes, M., L. Delmeire, C. Schotte, A. Janca, T. Creten, J. Mylle, A. Struyf, G. Pison, P. Rousseeuw (1998), The two-factor symptom structure of post-traumatic stress disorder: depression-avoidance and arousal-anxiety. *Psychiatry Research* (81): 195-210.

Malinvaud, E. (1980). *Statistical Methods of Econometrics*. Amsterdam, Netherlands: North-Holland.

Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics* (47): 264-273.

Maronna, R., R. Martin, V. Yohai (2006). *Robust Statistics: Theory and Methods*. New York,NY: Wiley.

Mili L., V. Phaniraj, P. Rousseeuw (1991). Least median of squares estimation in power systems. *IEEE Transactions on Power Systems* (6)**:** 511-523.

Prieto, J., C. Croux, A. Jimenez (2009). RoPEUS: a new robust algorithm for static positioning in ultrasonic systems. *Sensors* (9): 4211-4229.

Rousseeuw, P., A. Leroy (1987). *Robust Regression and Outlier Detection*. New York, NY: Wiley.

Rousseeuw, P., S. Van Aelst (1999), Positive-breakdown robust methods in computer vision, in *Computing Science and Statistics, Vol 31*, (edited by K. Berk and M. Pourahmadi, eds.), Interface Foundation of North America, Inc., Fairfax Station, VA, 451-460.

Rousseeuw, P., S. Van Aelst, M. Hubert (1999), Rejoinder to discussion of 'Regression Depth'. *Journal of the American Statistical Association* (94): 419-433.

Rousseeuw, P., K. Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* (41): 212-223.

Rousseeuw, P., K. Van Driessen (2006). Computing LTS for large data sets. *Data Mining and Knowledge Discovery* (12): 29-45.

Salibian-Barrera, M., R. Zamar (2002). Bootstrapping robust estimates of regression. *Annals of Statistics* (30): 556-582.

Schennach, S. (2004a). Estimation of nonlinear models with measurement error. *Econometrica* (72): 33-75.

Schennach, S. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory* (20): 1046–1093.

Söderström, T. (2007). Errors-in-variables methods in system identification. *Automatica* (43): 939-958.

Theil, H. (1971). *Principles of Econometrics*. New York, NY: Wiley.

Todorov, V. (2011). Scalable robust estimators with high breakdown point (R package 'rrcov'), http://cran.r-project.org/web/packages, accessed April 30, 2011.

Van Huffel, S. (2007). Total least squares and errors-in-variables modeling. *Computational Statistics & Data Analysis* (52): 1076-1079.

Verardi, V., C. Croux (2009). Robust regression in Stata. *Stata Journal* (9): 439-453. http://www.econ.kuleuven.be/public/NDBAE06/public.htm, accessed Jun 1, 2011.

Verboven, S., M. Hubert (2005). LIBRA: a MATLAB Library for Robust Analysis, *Chemometrics and Intelligent Laboratory Systems* (75): 127-136. http://wis.kuleuven.be/stat/robust/LIBRA.html, accessed June 1, 2011.

Zaman, A., P. Rousseeuw, M. Orhan (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters* (71**):** 1-8.

Zamar, R. (1989). Robust estimation in the errors-in-variables model. *Biometrika* 76: 149-160.

Figure 1. Bivariate eiv scatter

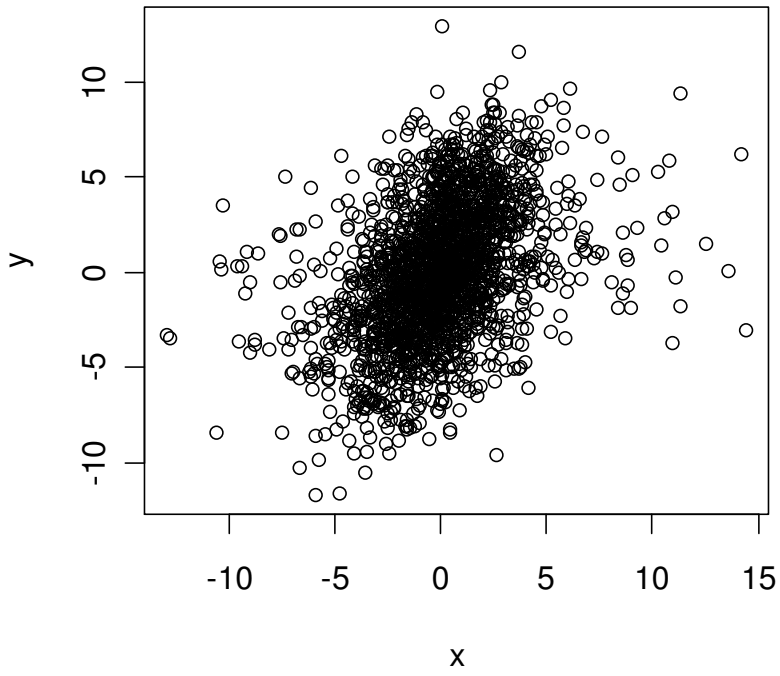

Table 1. Bivariate regression with "true" $R^2 = 0.800$

$u_{xi} \sim N(0,4)$ $u_{yi} \sim N(0,1)$ $x_i \sim N(0,2)$

	n = 200			n = 2000		
	bias	rmse	coverage	bias	rmse	coverage
ols	-0.497	0.501	0.000	-0.500	0.501	0.000
mmr	-0.039	0.088	0.876	-0.034	0.041	0.571
mcd	-0.029	0.146	0.887	-0.018	0.054	0.886
mcdw	-0.060	0.119	0.845	-0.046	0.056	0.597

Table 2. Bivariate regression with "true" $R^2 = 0.310$

$u_{xi} \sim N(0,4)$ $u_{yi} \sim N(0,3)$ $x_i \sim N(0,2)$

	n = 200			n = 2000		
	bias	rmse	coverage	bias	rmse	coverage
ols	-0.496	0.505	0.000	-0.498	0.499	0.000
mmr	-0.363	0.442	0.574	-0.374	0.385	0.006
mcd	-0.088	0.469	0.896	-0.087	0.182	0.846
mcdw	-0.146	0.345	0.861	-0.146	0.176	0.545

Table 3. Bivariate regression with "true" $R^2 = 0.310$

$u_{xi} \sim N(10,4)$ $u_{yi} \sim N(0,3)$ $x_i \sim N(0,2)$

	n = 200			n = 2000		
	bias	rmse	coverage	bias	rmse	coverage
ols	-0.851	0.852	0.000	-0.850	0.850	0.000
mmr	-0.839	0.846	0.000	-0.847	0.847	0.000
mcd	-0.011	0.384	0.902	-0.015	0.058	0.885
mcdw	0.151	0.294	0.835	-0.142	0.165	0.490

Table 4. Multiple regression with "true" $R^2 = 0.352$, n = 1000

$u_{xi} \sim N(0,4)$ $u_{zi} \sim N(0,3)$ $u_{yi} \sim N(0,4)$ $x_i, z_i \sim N(0,2)$ $r_{xz} = 0.393$

	x coefficient			z coefficient		
	bias	rmse	coverage	bias	rmse	coverage
ols	-0.249	0.258	0.021	-0.187	0.206	0.280
mmr	-0.146	0.220	0.758	-0.162	0.256	0.795
mcd	-0.023	0.248	0.896	-0.018	0.324	0.893

Table 5. Bivariate regression with "true" $R^2 = 0.310$, n = 1000

$u_{xi}, u_{yi} \sim N(0,3)$ $x_i \sim N(0,2)$

	x contamination = 25%			x contamination = 100%		
	bias	rmse	coverage	bias	rmse	coverage
ols	-0.359	0.362	0.000	-0.691	0.692	0.000
org	1.159	1.168	0.000	0.005	0.099	0.897
mom	-4.353	135.143	0.996	-29.682	948.786	0.999
mcd	-0.107	0.251	0.869	-0.688	0.708	0.008

Table 6. Bivariate regression with "true" $R^2 = 0.310$, n = 1000

$u_{xi} \sim N(0,4)$ $u_{yi} \sim N(0,3)$ $x_i \sim X^2(2 \text{ df})$

	x contamination = 25%			x contamination = 100%		
	bias	rmse	coverage	bias	rmse	coverage
ols	-0.499	0.501	0.000	-0.800	0.801	0.000
org	0.821	0.837	0.000	-0.548	0.551	0.000
mom	0.010	0.155	0.907	0.056	0.361	0.959
mcd	-0.078	0.324	0.881	-0.877	0.886	0.000

Table 7. Food Expenditures

n = 235 Belgian households
dep var = food expenditure
regressor = income

	ols	mmr	mcd
income	0.48	0.65	0.69
std err	0.01	0.02	0.03
R^2	0.83	0.88	0.83

Table 8. Hedonic Housing Demand

n = 25,357 houses sold in Lucas Co. Ohio, 1993-98
dep var = log of house price
regressors = logs of house's age, lot size and number of rooms

	ols	mmr	mcd
ln age	-0.36	-0.34	-0.51
ln lot size	0.34	0.24	0.48
ln rooms	0.71	0.87	0.63
R^2	0.49	0.65	0.59