ADJUSTING DATA FOR MEASUREMENT ERROR TO REDUCE BIAS WHEN ESTIMATING COEFFICIENTS OF A QUADRATIC MODEL

Jeff Bay, Maryville College and L. A. Stefanski, North Carolina State University
Jeff Bay, Division of Mathematics and Computer Science, Maryville College, Maryville, Tennessee, 37804

Key Words: Measurement Error, Moments, Regression

Introduction

When sample values are measured with error, parameter estimates that are nonlinear functions of the data are biased due to measurement error (Fuller, 1987; Carroll, Ruppert and Stefanski, 1995). Typically, adjustments are made to the estimates to account for measurement error. Instead, we create an alternative data set that is adjusted for measurement error. We might think of this as a measurement error transformation of the data. Parameter estimates calculated from the transformed data are unbiased or have reduced bias and, therefore, the method provides a simple way to adjust statistics for measurement error when it otherwise may not be clear how to do so.

One case of particular interest is estimating the coefficients in the regression model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$, where the independent variable is measured with error. It is well known that the least squares estimates of the β's will be biased due to the measurement error. Though it is straightforward to correct the least squares estimates in simple linear regression, it is not so for the quadratic model. We will show that adjusting the data prior to finding the least squares estimates leads to estimators with reduced bias and mean squared error.

The Model

The true data consist of n observations (X_i, Y_i), i=1,..., n, where X_i is subject to measurement error and Y_i is a measurement error-free variable. Because X_i is measured with error, we observe (W_i, Y_i), i=1,..., n, where

$$W_i = X_i + U_i,$$

$U_1, \ldots, U_n$ are independent $N(0, \sigma_u^2)$ and are independent of $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$. For the purpose of illustrating the method, we assume that σ_u^2 is known. When σ_u^2 is not known, it would be estimated using either replicate measurements or validation data, see Carroll, Ruppert, and Stefanski (1995).

The Measurement Error Transformation

Under the regression model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$, where X is a random variable with mean μ_x and variance σ_x^2, and ε is independent of X, has mean zero and variance σ_ε^2, the naïve least squares estimators follow from solving the least squares equations,

$$\begin{aligned}
\beta_0 n + \beta_1 \Sigma W_i + \beta_2 \Sigma W_i^2 &= \Sigma Y_i \\
\beta_0 \Sigma W_i + \beta_1 \Sigma W_i^2 + \beta_2 \Sigma W_i^3 &= \Sigma W_i Y_i \\
\beta_0 \Sigma W_i^2 + \beta_1 \Sigma W_i^3 + \beta_2 \Sigma W_i^4 &= \Sigma W_i^2 Y_i.
\end{aligned}$$

Though $(1/n)\Sigma W_i$ is an unbiased estimator of $E(X)$ and $(1/n)\Sigma W_i Y_i$ is an unbiased estimator of $E(XY)$, the remaining estimators are all biased. Specifically,

$n^{-1}\Sigma W_i^2$ unbiasedly estimates $E(X^2)+\sigma_u^2$,
$n^{-1}\Sigma W_i^3$ unbiasedly estimates $E(X^3)+3\sigma_u^2 E(X)$,
$n^{-1}\Sigma W_i^4$ unbiasedly estimates $E(X^4)+6\sigma_u^2 E(X^2)+3\sigma_u^4$,
$n^{-1}\Sigma W_i^2 Y_i$ unbiasedly estimates $E(X^2 Y)+\sigma_u^2 E(Y)$.

Thus, the estimated parameters will also be biased.

To reduce this bias, the observed data $W_1, \ldots, W_n$ are transformed to $X_1^*, \ldots, X_n^*$ such that

$$E(n^{-1}\Sigma X_i^{*r}) = E(X^r), \quad r = 1, \ldots, 4,$$

where $n^{-1}\Sigma X_i^{*r}$ converges in probability to a constant for r = 1, ..., 4, and, therefore,

$$n^{-1}\Sigma X_i^{*r} \rightarrow E(X^r), \quad r = 1, \ldots, 4.$$

Also, the adjusted data $X_1^*, \ldots, X_n^*$ meet the constraint that

$$E(n^{-1}\Sigma X_i^{*r} Y_i) = E(X^r Y), \quad r = 1, 2,$$

where $n^{-1}\Sigma X_i^{*r} Y_i$ converges in probability to a constant for r = 1, 2, and, therefore,

$$n^{-1}\Sigma X_i^{*r} Y_i \rightarrow E(X^r Y), \quad \mathrm{r} = 1, 2.$$

Details of the method are given in Bay (1997), but a brief overview is given here. The first step is to find estimators, m_{rs}, with the property that

$$E(m_{rs}) = E(n^{-1}\Sigma X_i^r Y_i^s), \quad r=1, \ldots, 4, \quad s=0, 1.$$

The estimators, m_{rs}, are found using Hermite Polynomials (Stefanski, 1990). These estimators define the constraints under which we maximize the likelihood for the unknown $X_1, \ldots X_n$.

Given that $U \sim N(0, \sigma_u^2)$ and only $(W_1, \ldots, W_n)$ are observed, the likelihood for the unknown $X_1, \ldots X_n$ is

$$f_{W|X} = \prod (2\pi\sigma_u^2)^{-1/2} exp[-(W_i - X_i)^2/(2\sigma_u^2)],$$

and the negative log likelihood is proportional to $\sum(W_i - X_i)^2$. We minimize this sum of squares, subject to the moment and cross-product constraints, thus ensuring that the estimated $X_1, \ldots, X_n$, have the desired moments and cross-products asymptotically.

For matching four moments and the xy and x^2y cross-product terms, the objective function is

$$q = \textstyle\sum_{i=1 \text{ to } n}(1/2)(W_i - X_i)^2 + \sum_{r=1 \text{ to } 4}(n/r)\lambda_r(n^{-1}\sum_{i=1 \text{ to } n}X_i^r - m_{r0}) + \sum_{r=1 \text{ to } 2}(n/r)\lambda_{4+r}(n^{-1}\sum_{i=1 \text{ to } n} X_i^r Y_i - m_{r1}),$$

where $\lambda_1, \ldots, \lambda_6$ are Lagrange multipliers and the constants are included to simplify derivatives. Differentiating q with respect to X_i gives us

$$\partial q/\partial X_i = X_i - W_i + \textstyle\sum_{r=1 \text{ to } 4} \lambda_r X_i^{r-1} + \sum_{r=1 \text{ to } 2} \lambda_{4+r} X_i^{r-1} Y_i.$$

Setting this partial derivative equal to zero implicitly defines $X_i^* = h(W_i, Y_i, \Lambda^*)$, where $\Lambda = (\lambda_1, \ldots, \lambda_6)^T$. Because $h(W_i, Y_i, \Lambda^*)$ is only implicitly defined, we must solve for $X_1^*, \ldots, X_n^*$ and $\lambda_1^*, \ldots, \lambda_6^*$ numerically. We have used the Newton-Raphson algorithm successfully. Although n may be very large, the Newton-Raphson algorithm can be solved while only needing to invert a diagonal matrix and a 6x6 matrix.

There are times when the Newton-Raphson algorithm does not converge (Bay, 1997). If the problem lies in matching four moments, the number of moments is reduced to two. In the very few cases when the problem is due to the xy cross-product, the estimated cross-product is multiplied by .98, and the algorithm retried. In general, these adjustments work well in allowing the algorithm to converge to create an adjusted data set that leads to reduced bias and mean squared error when estimating the coefficients in a quadratic model.

In a certain percentage of cases, the estimated x^2y cross-product is not able to be matched, and for this study a new data set is simulated when this is the case. Typically, we only run into problems when the measurement error variance, σ_u^2, is as large or larger than the random error variance, σ_ε^2.

Results

Table 1 compares the observed and adjusted data in terms of the bias and mean squared error (MSE) for the least squares estimates of the coefficients of the regression model $Y = 1 + 1X - .5X^2 + \varepsilon$, with σ_ε^2 either 0.25 or 1.0. The true data $X_1, \ldots, X_{300}$, were generated from a normal distribution with mean zero and variance one. The true data were contaminated with one of two levels of measurement error, σ_u^2=0.25 and σ_u^2=1.0, to form $W_1, \ldots, W_{300}$, the observed values. The table gives the mean bias and MSE for 1000 simulations, where the MSE is calculated as the Monte Carlo variance plus the square of the mean bias.

When possible, the first four moments were matched along with the two cross-product terms that appear in the least squares equations. When four moments could not be matched, the first two moments were matched. When the xy cross-product could not be matched, it was multiplied by .98 (this occurred in less than one percent of the simulation runs). The results in the table only include cases when the x^2y cross-product term could be matched. Approximately 25 percent of the time the x^2y term could not be matched when σ_u^2=1.0 and σ_ε^2=0.25. In other combinations of σ_u^2 and σ_ε^2, this problem occurred less than ten percent of the time. When the x^2y term could not be matched, it is possible to "match" it by reducing its value slightly, for example, multiplying $(1/n)(\sum W_i^2 Y_i - \sigma_u^2 \sum Y_i)$ by 0.98 and then finding adjusted values to match the new value.

The algorithm was less successful matching the x^2y term for the model $Y = 1 - 3X + X^2 + \varepsilon$; it was possible to match the x^2y term in only approximately 60 to 65 percent of the runs when σ_u^2=1.0 and σ_ε^2=0.25 (data not shown). Of the four combinations of $\sigma_u^2$and σ_ε^2, it would seem that this combination is the least likely to reflect a situation that may arise in practice.

σ_u^2	σ_ε^2	Parameter	Bias			MSE		
			W	X_1	X_2	*W*	X_1	X_2
.25	.25	β_0	-0.1026	-0.0974	0.0129	0.0134	0.0129	0.0067
		β_1	-0.2002	-0.1029	0.0073	0.0424	0.0140	0.0063
		β_2	0.1808	0.0945	-0.0179	0.0339	0.0115	0.0071
.25	1.0	β_0	-0.0988	-0.0945	0.0148	0.0166	0.0164	0.0120
		β_1	-0.1981	-0.1012	0.0083	0.0438	0.0164	0.0097
		β_2	0.1800	0.0947	-0.0163	0.0347	0.0131	0.0093
1.0	.25	β_0	-0.2657	-0.2739	-0.0831	0.0757	0.0813	0.0350
		β_1	-0.5117	-0.3094	-0.0428	0.2640	0.1014	0.0241
		β_2	0.3848	0.2766	0.0872	0.1487	0.0804	0.0318
1.0	1.0	β_0	-0.2608	-0.2588	-0.0104	0.0772	0.0784	0.0665
		β_1	-0.5031	-0.2910	0.0084	0.2567	0.0937	0.0438
		β_2	0.3802	0.2555	0.0023	0.1457	0.0722	0.0660

Table 1: *The table compares bias and mean squared error (MSE) for estimating the coefficients of $Y = 1 + 1X - .5X^2 + \varepsilon$ with σ_ε^2 either 0.25 or 1.0. W refers to the observed data, measured with error, X_1 refers to the adjusted data when only the first-order cross-product is matched, X_2 refers to the adjusted data created by matching both the xy and x^2y cross-products.*

Using the transformed data leads to large reductions in the bias and MSE for the least squares estimates. For example, Table 1 shows that when σ_u^2=0.25 and σ_ε^2=1.0, β_2 is overestimated by 0.1800, on average, when using the observed data. It is still overestimated, though the bias is reduced by about half, when using transformed data that matches moments and the xy cross-product. However, when the x^2y cross-product term is also matched, the bias is reduced to only –0.0163, on average. This is accompanied by a reduction in the mean squared error from 0.0347 to 0.0093. Similar results are obtained for the other coefficients and for other combinations of σ_u^2 and σ_ε^2.

Approximating Variance

It is important to account for the variability introduced when transforming the sample values; we cannot simply estimate variance by treating the transformed values as though they were the observed values. Table 2 compares the results from the Monte Carlo standard errors of the 1000 simulated data sets with the means of the standard errors estimated using M-estimation.

σ_u^2	σ_ε^2	Parameter	Monte Carlo	M-estimation
.25	.25			
		β_0	0.0809	0.0767
		β_1	0.0793	0.0757
		β_2	0.0822	0.0718
.25	1.0			
		β_0	0.1085	0.1026
		β_1	0.0982	0.0945
		β_2	0.0952	0.0852
1.0	.25			
		β_0	0.1676	0.2117
		β_1	0.1491	0.1841
		β_2	0.1557	0.2091
1.0	1.0			
		β_0	0.2577	0.2990
		β_1	0.2091	0.2416
		β_2	0.2569	0.3022

Table 2: *The table compares the Monte Carlo standard errors with the mean approximate standard errors from M-estimation. The Monte Carlo standard errors are the square roots of the Monte Carlo variances for 1000 simulated data sets and the mean approximate standard errors are the means of the square roots of the approximate variances.*

While the M-estimation standard errors appear to slightly overestimate the standard error when measurement error is relatively large, this overestimation decreases as the sample size increases (data not shown).

Summary

When data are measured with error, the measurement error transformation leads to least squares estimators for a quadratic model that have reduced bias and mean squared error compared to those found using the observed data. The primary benefit of this method may be when conducting exploratory analyses. In exploratory analyses it would be tedious and time-consuming to constantly correct parameter estimates for measurement error. Yet ignoring measurement error may lead to qualitatively different conclusions about the relationships between variables. Our method allows for the correction to be made prior to carrying out the exploratory analyses and estimated parameters calculated from the transformed data will be corrected for measurement error. It is also possible, using M-estimation, to approximate standard errors for estimates calculated from the transformed data.

Bay (1997) shows that this method also works well for estimating a cumulative distribution function, the coefficients in a linear regression, and the coefficients in a logistic regression.

References

Bay, J.M. (1997). "Adjusting data for measurement error," Ph.D. Thesis, North Carolina State University.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). Measurement Error in Nonlinear Models. Chapman & Hall, London.

Fuller, W.A. (1987). Measurement Error Models. John Wiley and Sons, New York.

Stefanski, L.A. (1990). "Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models," *Comm. Statist.*, 18, 4335-4358.