Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models

Raymond J. CARROLL, Helmut KÜCHENHOFF, F. LOMBARD, and Leonard A. STEFANSKI

Cook and Stefanski have described a computer-intensive method, the SIMEX method, for approximately consistent estimation in regression problems with additive measurement error. In this article we derive the asymptotic distribution of their estimators and show how to compute estimated standard errors. These standard error estimators can either be used alone or as pivoting devices in a bootstrap analysis. We also give theoretical justification to some of the phenomena observed by Cook and Stefanski in their simulations.

KEY WORDS: Asymptotics; Bootstrap; Computationally intensive methods; Measurement error models.

1. INTRODUCTION

We consider regression problems where some of the predictors are measured with additive error. The response is denoted by \( Y \) and the predictors by \( (Z, X) \), but \( X \) cannot be observed. Instead, we can observe \( W = X + \varepsilon U \), where \( U \) has mean zero and variance 1. We will consider the structural case, so that \( (Y_i, Z_i, X_i, W_i) \) for \( i = 1, \ldots, n \) are independent and identically distributed. For linear regression, this is the classical additive structural measurement error model described in detail by Fuller (1987).

In linear regression, as well as nonlinear models such as the generalized linear models, it is well known that the naive estimator that ignores measurement error leads to inconsistent regression parameter estimates. Correction for this bias in linear regression has a long history, but the analysis of nonlinear measurement error models is of more recent vintage. One of the more useful general methods is what we call regression calibration, wherein \( X \) is replaced by an estimate of \( E(X|Z, W) \) and the standard analysis then performed (see Carroll, Ruppert, and Stefanski 1995, Carroll and Stefanski 1990, Gleser 1990, Rosner, Spiegelman, and Willett 1990, and Rosner, Willett, and Spiegelman 1989). Regression calibration, which is simply the usual correction for attenuation in linear regression, is consistent for the slope in models with mean that is an exponential function of a linear combination of the predictors and is very nearly consistent for slopes in most applications of generalized linear models. In general, for small measurement error, the bias is of the order \( O(\sigma^4) \). The regression calibration estimates are usually easy to compute, have straightforward standard error estimates (Carroll and Stefanski 1990, 1994) and are amenable to bootstrap analysis.

An alternative general method has recently been proposed by Cook and Stefanski (1994), hereafter denoted by CS. Their idea, called SIMEX for simulation and extrapolation, relies on computer simulation to generate parameter estimates. SIMEX has essentially all the same properties outlined earlier for regression calibration, with two exceptions: (a) it is as easy to program but far more computationally intensive, and (b) the small measurement error bias of SIMEX is of order \( O(\sigma^6) \), suggesting that it might prove superior in some highly nonlinear models.

The idea behind the SIMEX method is most clearly seen in simple linear regression when the independent variable is subject to measurement error. Suppose that the regression model is \( E(Y|X) = \alpha + \beta X \) and that \( W = X + \varepsilon U \), rather than \( X \), is observed where \( U \) has mean zero and variance 1 and the measurement error variance \( \sigma^2 \) is known. It is well known that the ordinary least squares estimate of the slope from regressing \( Y \) on \( W \) converges to \( \beta \sigma_x^2 (\sigma_x^2 + \sigma^2)^{-1} \), where \( \sigma_x^2 \) denotes the variance of \( X \).

For any fixed \( \lambda > 0 \), suppose that one repeatedly "adds on," via simulation, additional error with mean zero and variance \( \sigma^2 \lambda \) to \( W \); computes the ordinary least squares slope each time and then takes the average. This simulation estimator consistently estimates

\[
g(\lambda) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma^2 (1 + \lambda)} \beta.
\]

Since, formally at least, \( g(-1) = \beta \), the idea is to plot \( g(\lambda) \) against \( \lambda \geq 0 \), fit a model to this plot, and then extrapolate back to \( \lambda = -1 \). CS showed that this procedure will yield a consistent estimate of \( \beta \) if one fits the model \( g(\lambda) = \gamma_0 + \gamma_1 (\gamma_2 + \lambda)^{-1} \).

One thing is missing from the SIMEX methodology: an asymptotic distribution theory, including verification of asymptotic normality and implementation of standard error estimates. Our purpose is to provide this theory and the requisite standard error estimates. CS suggested that inference be done by means of the bootstrap, but of course asymptotic standard error estimates still have a role to play. Quite apart from verifying that SIMEX estimators are asymptotically normally distributed, standard error estimates have two im-

© 1996 American Statistical Association
Journal of the American Statistical Association
March 1996, Vol. 91, No. 433, Theory and Methods

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
portant roles. First, the SIMEX estimator is itself computationally intensive, usually a few hundred times slower to compute than the regression calibration estimator. Because the bootstrap itself is computationally intensive, this means that naive bootstrap analyses may well be too time-consuming in the initial stages of model fitting, when having standard error estimates available can serve as a rough guide. In addition, standard error estimates can be used to construct pivots to improve the precision of bootstrap confidence intervals and inferences (see Hall 1988).

The primary aims of this article are to provide an asymptotic theory for the SIMEX method and to construct standard error estimates. Our approach is simple and transparent. We consider problems in which the regression parameter estimates are the solution to unbiased estimating equations. Then, using asymptotic linearization results, we develop representations for each of the building blocks in the SIMEX atmosphere. Having studied the component parts of the algorithm, it is then easy to combine them into an overall linearization result for the final SIMEX estimator. Consistent standard error estimates are easily constructed; they require only the first two derivatives of the basic estimating equation and are simple by-products of the SIMEX algorithm.

The article is arranged as follows. In Section 2 we give a precise description of the SIMEX algorithm as used by CS. The basic idea in the case of known measurement error variance $\sigma^2$ is described in Section 3. In Section 4 we show how one can extend these results in the common case that the measurement error variance is estimated. In Section 5 we describe further asymptotic theory relating to choices of the number of simulations performed by SIMEX and the type of location functional used. CS name the method described in Section 2 the IID method. They also describe a refinement of their method, which they call the NON-IID case; in Section 6 we argue that in practice, the asymptotic standard errors that we have computed will be applicable even in the NON-IID case. A small simulation study for linear regression is described in Section 7. An analysis of data from the Framingham Heart Study is given in Section 8. Concluding remarks are given in Section 9.

2. REVIEW OF THE SIMEX METHOD

Here we provide a brief description of the SIMEX algorithm. CS should be consulted for more details and motivation. The algorithm consists of a simulation step, followed by an extrapolation step. We review each step in turn.

2.1 Simulation Step

Suppose that one has an unknown vector parameter $\Theta_0$. If one could observe the $X$'s, then we suppose that one could estimate $\Theta_0$ by solving an estimating equation,

$$0 = \sum_{i=1}^{n} \psi(Y_i, Z_i, X_i, \Theta).$$  (1)

Estimating equations and their theory have a long history in statistics and include as special cases most of the forms of regression currently used in practice. As used by CS, SIMEX works as follows. Fix $B > 0$ (they used $B = 50, 100$) and for $b = 1, \ldots, B$, generate via computer independent standard normal random variables $\{e_{ib}\}$. Then the variance of $W_i + \sigma \lambda^{1/2} e_{ib}$ given $X_i$ is $\sigma^2(1 + \lambda)$. For each $b$, define $\hat{\Theta}_{b, \sigma^2(1 + \lambda)}$ as the solution to

$$0 = \sum_{i=1}^{n} \psi(Y_i, Z_i, W_i + \sigma \lambda^{1/2} e_{ib}, \Theta).$$  (2)

Now form the average of the $\hat{\Theta}_{b, \sigma^2(1 + \lambda)}$'s, namely

$$\hat{\Theta}_{S, \sigma^2(1 + \lambda)} = B^{-1} \sum_{b=1}^{B} \hat{\Theta}_{b, \sigma^2(1 + \lambda)},$$  (3)

the subscript $S$ emphasizing the simulation nature of the estimator. Of course, when $\lambda = 0$ the simulation process is vacuous and $\hat{\Theta}_{S, 0}$ denotes the naive estimator of $\Theta_0$. By standard estimating equation theory, under sufficient regularity conditions, $\hat{\Theta}_{S, \sigma^2(1 + \lambda)}$ converges in probability to $\Theta_{\sigma^2(1 + \lambda)}$, the solution in $\Theta$ of

$$0 = E\{\psi(Y, Z, W + \sigma \lambda^{1/2} e, \Theta)\}.$$  (4)

CS also considered the idealized limit for infinite $B$,

$$\hat{\Theta}_{\sigma^2(1 + \lambda)} = E\{\hat{\Theta}_{b, \sigma^2(1 + \lambda)}(Y_i, Z_i, W_i), i = 1, \ldots, n\}.$$  (5)

2.2 Extrapolation Step

CS suggested that one compute $\hat{\Theta}_{S, \sigma^2(1 + \lambda)}$ on a fixed grid $\Lambda = (\lambda_1, \ldots, \lambda_M)$, thus yielding an understanding of the behavior of the estimators for different amounts of measurement error. Then, as motivated in section 1, they suggested that one fit a parametric model $G(\Gamma, \lambda)$ in a vector parameter $\Gamma$ to the $\hat{\Theta}_{S, \sigma^2(1 + \lambda)}$'s as a function of the $\lambda$'s, thus resulting in estimates $\hat{\Gamma}$. Finally, the SIMEX estimator of $\Theta_0$ is

$$\hat{\Theta}_{SIMEX} = G(\hat{\Gamma}, -1).$$  (6)

Various parametric models are possible; CS suggested linear models, quadratic models, and, with $\Gamma = (\gamma_0, \gamma_1, \gamma_2)$, the nonlinear model $G(\lambda, \Gamma) = \gamma_0 + \gamma_1(\lambda + \gamma_2)^{-1}$. They fit the models by ordinary unweighted least squares.

3. ASYMPTOTIC DISTRIBUTION THEORY

Like the algorithm, the asymptotic distribution theory of the SIMEX estimator splits into an analysis of the simulation and extrapolation steps. Both analyses are based on estimating equation theory. In accord with the practice used by CS, we assume that $B$ is fixed (but in practice, fairly large). The case that $B = \infty$ is discussed in Section 5.

3.1 Simulation Step

For any fixed $b$, standard asymptotic theory yields the expansion

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
\( n^{1/2} \{ \hat{\Theta}_{\sigma^2(1+\lambda)} - \Theta_{\sigma^2(1+\lambda)} \} = -A^{-1}(\sigma^2, \lambda, \Theta_{\sigma^2(1+\lambda)}) \)
\[ \times n^{-1/2} \sum_{i=1}^{n} \psi(Y_i, Z_i, W_i + \sigma^{1/2} \varepsilon_{ib}, \Theta_{\sigma^2(1+\lambda)}) + o_p(1), \]  

(7)

\[ \mathcal{A}(\sigma^2, \lambda, \Theta_{\sigma^2(1+\lambda)}) = E \{(\partial/\partial \Theta^T) \psi(Y, Z, W + \sigma^{1/2} \varepsilon, \Theta_{\sigma^2(1+\lambda)}). \} \]  

(8)

Although the results are phrased here in terms of estimating equations, they rely only on expansions of the form (7). The expansion is actually better than indicated, as the remainder term in (7) is typically of order \( O(n^{-1/2+\epsilon}) \) almost surely for any \( \epsilon > 0 \). Because \( B \) is fixed, the expansion (7) is uniform in \( b = 1, \ldots, B \) and \( \lambda \in \Lambda \); that is, the remainder converges uniformly to zero in probability, as long as \( \Lambda \) has a finite number of elements.

Define
\[ \chi_{B,i}(\sigma^2, \lambda, \Theta_{\sigma^2(1+\lambda)}) = B^{-1} \sum_{b=1}^{B} \psi(Y_i, Z_i, W_i + \sigma^{1/2} \varepsilon_{ib}, \Theta_{\sigma^2(1+\lambda)}). \]  

(9)

Setting \( A^{-1}(\cdot) = A^{-1}(\sigma^2, \lambda, \Theta_{\sigma^2(1+\lambda)}) \), by (7) we find that
\[ n^{1/2} \{ \hat{\Theta}_{S,\sigma^2(1+\lambda)} - \Theta_{\sigma^2(1+\lambda)} \} \]
\[ = -A^{-1}(\cdot)n^{-1/2} \sum_{i=1}^{n} \chi_{B,i}(\sigma^2, \lambda, \Theta_{\sigma^2(1+\lambda)}) + o_p(1). \]  

(10)

The terms \( \chi_{B,i}(\cdot) \) are independent and identically distributed with mean zero, so that (10) is simply a standard asymptotic linearization result for \( \hat{\Theta}_{S,\sigma^2(1+\lambda)}. \)

### 3.2 Combining the Estimators

The SIMEX algorithm computes \( \hat{\Theta}_{S,\sigma^2(1+\lambda)}(\cdot) \) on a grid of values \( \Lambda = (\lambda_1, \ldots, \lambda_M) \). Denote the resulting vector of estimators vec\( \{ \hat{\Theta}_{S,\sigma^2(1+\lambda)}, \lambda \in \Lambda \} \) by \( \hat{\Theta}_{S}(\cdot) \) and the corresponding vector of estimands \( \Theta_{S,\sigma^2(1+\lambda)}(\lambda) \) by \( \Theta_{S}(\cdot) \). Define \( \Psi_{B,i}(\cdot) = \{ \sigma^2, \lambda, \Theta_{S}(\cdot) \} = \psi(Y_i, Z_i, W_i + \sigma^{1/2} \varepsilon_{ib}, \Theta_{S,\sigma^2(1+\lambda)}), \lambda \in \Lambda \}, \) and \( \mathcal{A}_{11}(\cdot) = \{ \sigma^2, \lambda, \Theta_{S}(\cdot) \} = \text{diag} \{ A_{11}(\cdot) \} \). Then, using (10), the joint limit distribution of \( n^{1/2} \{ \hat{\Theta}_{S}(\cdot) - \Theta_{S}(\cdot) \} \) is seen to be multivariate normal \( (0, \Sigma) \) with
\[ \Sigma = \mathcal{A}_{11}^{-1}(\cdot) \mathcal{C}_{11}(\cdot) \mathcal{A}_{11}^{-1}(\cdot) \]  

(11)

and
\[ \mathcal{C}_{11}(\cdot) = \text{cov} \{ \Psi_{B,i}(\cdot) \sigma^2, \lambda, \Theta_{S}(\cdot) \}. \]  

(12)

### 3.3 Extrapolation Step

Having computed the vector of estimates \( \hat{\Theta}_{S}(\cdot) \) on the grid \( \Lambda = (\lambda_1, \ldots, \lambda_M) \), one now sets \( \hat{\Theta}_{S,\sigma^2(1+\lambda)} = \hat{G}(\Gamma, \lambda \mid \lambda) \) for a vector of parameters \( \Gamma \) and fits \( \hat{G}(\Gamma, \lambda), \lambda \in \Lambda \), to the elements of \( \hat{\Theta}_{S}(\cdot) \). Define \( \hat{G}_T(\Gamma, \lambda) = (\partial/\partial \Gamma)\hat{G}(\Gamma, \lambda), s(\Gamma) = \{ \hat{G}_T(\Gamma, \lambda_1), \ldots, \hat{G}_T(\Gamma, \lambda_M) \}, \) and
\[ R(\Gamma) = \hat{G}_s(\Gamma) - \{ \hat{G}(\Gamma, \lambda_1), \ldots, \hat{G}(\Gamma, \lambda_M) \}^T. \]  

For any symmetric positive definite matrix \( C \) of weights, an estimate \( \hat{\Gamma} \) of \( \Gamma \) is obtained by minimizing \( R(\Gamma)C^{-1}R(\Gamma) \), which has estimating equation \( 0 = s(\Gamma)C^{-1}R(\Gamma) \). One possibility is to use weighted least squares and set \( C = \Sigma \) defined in (11), but our numerical experience with this method was not particularly encouraging because of the near singularity in the estimate of \( \Sigma \) (Sec. 7). This is likely to be due in part to the fact that the same values \( \varepsilon_{ib} \) are used for all values of \( \lambda \). CS took \( C \) to be the identity matrix and used the following three models \( \hat{G}(\Gamma, \lambda): \gamma_0 + \gamma_1 + \lambda \) (linear), \( \gamma_0 + \gamma_1 + \lambda + \gamma_2 \lambda^2 \) (quadratic), and \( \gamma_0 + \gamma_1 + \gamma_2 + \lambda^{-1} \) (nonlinear).

Define \( D(\Gamma) = s(\Gamma)C^{-1}s(\Gamma) \). Then a standard asymptotic analysis shows that
\[ \hat{\Gamma} - \Gamma \approx \text{Normal} \{ \hat{0}, \Sigma(\Gamma) = D^{-1}(\Gamma)s(\Gamma)C^{-1} \times \Sigma C^{-1}s(\Gamma)D^{-1}(\Gamma) \}. \]  

(13)

Using the delta method, the SIMEX estimator \( \hat{G}_{S,\sigma^2(1+\lambda)}(\cdot) \) has asymptotic covariance matrix \( \hat{G}_{s}(\Gamma, -1) = \Sigma(\Gamma) \hat{G}_{s}(\Gamma, -1) \).

Estimating the asymptotic covariance matrix requires estimates of \( \mathcal{A}_{11}(\cdot) \) and \( \mathcal{C}_{11}(\cdot) \); see (8) and (12). We use the sandwich method (Drum and McCullough 1993, and references therein), which gives consistent estimates without making any distributional assumptions. Take \( \hat{C}_{11}(\cdot) \) to be the sample covariance matrix of the terms \( \hat{G}_{B,i}(\cdot) = \{ \sigma^2, \lambda, \hat{G}(\Gamma, \lambda) \} \) and \( \mathcal{A}_{11}(\cdot) = \text{diag} \{ \hat{A}_{m}(\cdot) \} \) for \( m = 1, \ldots, M \), where
\[ \hat{A}_{m}(\cdot) = (nB)^{-1} \sum_{i=1}^{n} \sum_{b=1}^{B} (\partial/\partial \Theta^T) \psi(Y_i, Z_i, W_i + \sigma^{1/2} \varepsilon_{ib}, \hat{G}_{S,\sigma^2(1+\lambda)}). \]

The estimated matrices \( \hat{A}_{11}(\cdot) \) and \( \hat{C}_{11}(\cdot) \) are often easily calculated, the only real complication arising from the fact that the former requires the derivative of \( \psi(\cdot) \) in (2), but this is not usually a major problem. The latter matrix requires no more than the estimating equation itself.

### 4. ESTIMATED MEASUREMENT ERROR VARIANCE

The development thus far has assumed a known measurement error variance. But our estimating equation approach easily accommodates unknown measurement error variance. For specificity, consider the important special case where two independent replicate measurements are available at each \( X_i \). Suppose, for instance, that
\[ W_{i(j)} = X_i + 2^{1/2} \sigma U_{i(j)} \text{ for } j = 1, 2 \text{ and } i = 1, \ldots, n. \]

Set
\[ W_i = (W_{i(1)} + W_{i(2)})/2, U_i = (U_{i(1)} + U_{i(2)})/2^{1/2}, \]
and
\[ D_i = (W_{i(1)} - W_{i(2)})^2/4. \]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Then $W_i = X_i + \sigma U_i$, where $U_i$ has zero mean and unit variance, and an unbiased estimating equation for $\sigma^2$ is

$$
0 = \sum_{i=1}^{n} (D_i - \sigma^2) = \sum_{i=1}^{n} \Psi_{i(2)}(D_i, \sigma^2),
$$

(14)
say. Let $A_11$ and $C_{11}$ be as defined in Section 3.2. By combining $\Psi_{B,i(1)}$ and $\Psi_{i(2)}$ into a single estimating equation and applying standard theory, the covariance matrix of the joint distribution of $\tilde{\Theta}_{S^2}(\Lambda)$ and $\tilde{\sigma}$ is

$$
n^{-1} \begin{bmatrix} A_{11}(\cdot) & A_{12}(\cdot) \\ C_{11}(\cdot) & C_{12}(\cdot) \end{bmatrix}^{-1} \begin{bmatrix} A_{11}(\cdot) & A_{12}(\cdot) \\ C_{11}(\cdot) & C_{12}(\cdot) \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}(\cdot) & A_{12}(\cdot) \\ C_{11}(\cdot) & C_{12}(\cdot) \end{bmatrix}^{-1},
$$

(15)

where

$$
A_{12}(\sigma^2, \Lambda, \Theta_*(\Lambda)) = n^{-1} \sum_{i=1}^{n} E \left[ \frac{\partial}{\partial \sigma^2} \Psi_{B,i(1)}(\sigma^2, \Lambda, \Theta_*(\Lambda)) \right],
$$

$$
A_{22}(\sigma^2) = n^{-1} \sum_{i=1}^{n} E \left[ \frac{\partial}{\partial \sigma^2} \Psi_{i(2)}(\sigma^2) \right] = -1,
$$

and

$$
\begin{bmatrix} C_{11}(\cdot) & C_{12}(\cdot) \\ C_{12}(\cdot) & C_{22}(\cdot) \end{bmatrix} = C_*(\cdot) = \text{cov} \left[ \Psi_{B,i(1)}(\sigma^2, \Lambda, \Theta_*(\Lambda)), \Psi_{i(2)}(\sigma^2) \right].
$$

Estimating these quantities via the sandwich method is easy. For $A_{12}$ and $A_{22}$, remove the expectation symbol and replace $\{\sigma^2, \Theta_*(\Lambda)\}$ by $\{\hat{\sigma}^2, \tilde{\Theta}_S(\Lambda)\}$. The covariance matrix $C_*(\cdot)$ can be estimated by the sample covariance matrix of the vectors,

$$
\begin{bmatrix} \Psi_{B,i(1)}(\hat{\sigma}^2, \Lambda, \tilde{\Theta}_S(\Lambda)) \\ \Psi_{i(2)}(\hat{\sigma}^2) \end{bmatrix}.
$$

Having made the substitutions to estimate the joint covariance matrix (15), we obtain an estimate of the covariance matrix of $\tilde{\Theta}_S(\Lambda)$. Estimated standard errors for the SIMEX estimate now follow exactly as in Section 3.

As a second example, consider a problem in which $\sigma^2$ has been estimated based on replicates in an entirely independent data set of size $m$, so that $\sigma^2$ is asymptotically normally distributed with variance $m^{-1} \xi$. Then, as in the preceding argument, merely set $A_{22}(\cdot) = -m/n, C_{12}(\cdot) = 0$ and $C_{22}(\cdot) = \xi(m/n)$.

5. ROBUST SUMMARIES, AND PARTIAL ASYMPTOTICS FOR LARGE $B$

This section briefly investigates what happens if instead of sample means in (3), we use other location functionals such as the median. CS stated that “our experience suggests that using ... the sample mean as the location estimator works as well as any other ... location functional studied, at least in large samples” (p. 1327). In this section we briefly indicate that this statement, based on extensive simulations and experimentation with real data, has some theoretical foundation.

Three types of asymptotics can be investigated: (a) $n \to \infty$ for $B$ fixed, (b) $n \to \infty$ and $B \to \infty$ simultaneously, and (c) $B = \infty$ and $n \to \infty$. We have already described the SIMEX estimator in case (a). For case (b), the same results apply as long as $B n^{-1/2+a} \to 0$ for some $a > 0$. The reason for this is that the error term in (7) is typically of order $O_p(n^{-1/2+\varepsilon})$, so that under the stated condition on $B$, the expansions in (7) are uniform in $b$. For case (c), or in general if $B$ is large relative to $n$, one can replace the terms $\chi_{B,i}$ in (10) by the term (21) defined later.

We first investigate the use of other location functionals in cases (a)–(b). This is the main result of the section, and it indicates that all location functionals should have approximately the same behavior for large sample sizes. The argument is simple and transparent. The second part of this section briefly summarizes what is known in case (c). The arguments are neither simple nor transparent, even in simple linear regression when using the mean functional, but in special cases, under strong regularity conditions, the conclusions agree with the results in cases (a)–(b).

5.1 Other Location Functions when $B$ is Large

We have proposed the estimator (3), based on taking the sample mean of the terms $\tilde{\Theta}_{b,\sigma^2(i+1)}$ for $b = 1, \ldots, B$. Other location functions can be used; for example, the sample median and trimmed mean. We now show informally that for large $B$, using other location functions should result in the same asymptotic distribution as obtained by (3). In this section we consider case (a), namely, that $n \to \infty$ and then $B \to \infty$; case (b) has the same analysis.

We distinguish between a sample location functional, such as the sample mean or median, and a population location functional. Let $T$ be any population location functional of a random variable $V$, such as the population mean or median, satisfying $T(a+bV) = a+bT(V)$. The population location functional also must satisfy the condition that if $V$ is normally distributed with mean zero, then $T(V) = 0$.

It is easiest to understand the behavior of the SIMEX estimator for large $B$ if one first fixes $B$, lets $n \to \infty$, and then lets $B \to \infty$. For fixed $B$, we have shown that the random variables $\{\tilde{\Theta}_{b,\sigma^2(i+1)}\}_{b=1}^{B}$ form a set of correlated normal random variables as $n \to \infty$, with common variance and correlation. They thus satisfy the one-way random effects model

$$
n^{1/2}(\tilde{\Theta}_{b,\sigma^2(i+1)} - \Theta_{\sigma^2(i+1)}) \approx \xi + \nu_b,
$$

(16)
say, where $\xi$ is normally distributed and independent of $\{\eta_1, \ldots, \eta_B\}$, which are themselves independent and identically distributed normal random variables with mean zero. As $B \to \infty$, under our stated conditions and subject to further regularity conditions, all sample location functionals of the left side of (16), including the mean and the mean.
5.2 $B = \infty$

A more involved analysis occurs if asymptotics are done for $B = \infty$. We conjecture that the same point holds—namely that for large $n$, there should be little effect due to the choice of functional location. Although the heuristic argument presented here in support of this conjecture is fairly straightforward, the actual technical verification of the argument seems extraordinarily difficult. We have verified the steps in special cases under strong regularity conditions.

As indicated by CS and in (5), if $B = \infty$, then for any given $\lambda$, the SIMEX building blocks are the terms

$$ T\{\hat{\theta}_{b,\sigma^2(1+\lambda)}(Y_i, Z_i, W_i), i = 1, \ldots, n\}. \quad (17) $$

It is important to emphasize that $b$ in (17) is any single, fixed $b$. Equation (5) arises when $T$ is the expectation functional, whereas when using the median, $T$ is the median of the indicated distribution.

At the risk of repetition, (17) is a functional of the conditional distribution of a random variable. If we can understand the behavior of the random variable for large sample sizes as a function of the data, then under sufficient regularity conditions, we can compute its conditional distribution.

Our analysis then requires two steps:

**Step 1.** For any fixed $b$, find an expansion for

$$ n^{1/2}(\hat{\theta}_{b,\sigma^2(1+\lambda)} - \Theta_{\sigma^2(1+\lambda)}) $$

This describes the unconditional distribution of $\hat{\theta}_{b,\sigma^2(1+\lambda)}$ as a function of the data $(Y_i, Z_i, W_i)$ for $i = 1, \ldots, n$.

**Step 2.** Compute the conditional distribution of the expansion.

Here is Step 1. For any fixed $b$ (remember, $B = \infty$, and we are not letting $B \to \infty$), from (7), as $n \to \infty$,

$$ n^{1/2}(\hat{\theta}_{b,\sigma^2(1+\lambda)} - \Theta_{\sigma^2(1+\lambda)}) $$

$$ = -n^{-1/2} \sum_{i=1}^{n} A^{-1}(\cdot)\psi(Y_i, Z_i, W_i + \sigma \lambda^{1/2} \varepsilon_{ib}, $$

$$ \Theta_{\sigma^2(1+\lambda)}) + o_p(1), $$

where the $(\varepsilon_{ib})_{i=1}^{n}$ are independent and identically distributed standard normal random variables. Let $\mathcal{G}(Y, Z, W; \Theta)$ be the conditional mean of $-A^{-1}(\cdot)\psi(Y, Z, W + \sigma \lambda^{1/2} \varepsilon, \Theta)$ given $(Y, Z, W)$; note that it has unconditional mean zero. It then follows that

$$ n^{1/2}(\hat{\theta}_{b,\sigma^2(1+\lambda)} - \Theta_{\sigma^2(1+\lambda)}) $$

$$ = n^{-1/2} \sum_{i=1}^{n} \mathcal{G}(Y_i, Z_i, W_i, \Theta_{\sigma^2(1+\lambda)}) $$

$$ + n^{-1/2} \sum_{i=1}^{n} \begin{bmatrix} A^{-1}(\cdot)\psi(Y_i, Z_i, W_i + \lambda^{1/2} \varepsilon_{ib}, $$

$$ \Theta_{\sigma^2(1+\lambda)}) $$

$$ - \mathcal{G}(Y_i, Z_i, W_i, \Theta_{\sigma^2(1+\lambda)}) \end{bmatrix} + o_p(1). $$

(18)

The two terms on the right side of (18) are uncorrelated. Further, given $(Y_i, Z_i, W_i)_{i=1}^{n}$, the last term in (18), which we write as $\mathcal{G}_n$, converges to a normal random variable, say $\mathcal{G}$, which has a mean zero and variance $\Omega = \Omega(\{(Y_i, Z_i, W_i)_{i=1}^{n}\})$. Thus

$$ n^{1/2}(\hat{\theta}_{b,\sigma^2(1+\lambda)} - \Theta_{\sigma^2(1+\lambda)}) $$

$$ = \mathcal{G}_n + n^{-1/2} \sum_{i=1}^{n} \mathcal{G}(Y_i, Z_i, W_i, \Theta_{\sigma^2(1+\lambda)}) $$

$$ + o_p(1). $$

(19)

The two terms on the right side of (19) are uncorrelated and asymptotically normally distributed and hence asymptotically independent.

Here is Step 2. Condition on the terms $(Y_i, Z_i, W_i)_{i=1}^{n}$. Under sufficient regularity conditions, it follows that

$$ n^{1/2}(T(\hat{\theta}_{b,\sigma^2(1+\lambda)}) - T(\Theta_{\sigma^2(1+\lambda)})) $$

$$ = T\{\mathcal{G}_n + n^{-1/2} \sum_{i=1}^{n} \mathcal{G}(Y_i, Z_i, W_i, \Theta_{\sigma^2(1+\lambda)}) + o_p(1)\} $$

$$ = n^{-1/2} \sum_{i=1}^{n} \mathcal{G}(Y_i, Z_i, W_i, \Theta_{\sigma^2(1+\lambda)}) + T\{\mathcal{G}_n + o_p(1)\} $$

$$ = n^{-1/2} \sum_{i=1}^{n} \mathcal{G}(Y_i, Z_i, W_i, \Theta_{\sigma^2(1+\lambda)}) + o_p(1), $$

(20)

the last step following from the fact that $\mathcal{G}_n$ is conditionally normally distributed with mean zero and $T(V) = 0$ when $V$ is normally distributed with mean zero.

We have thus shown heuristically that, subject to regularity conditions in going from Step 1 to Step 2, if $B = \infty$, then as $n \to \infty$, all SIMEX estimators are asymptotically equivalent. Because in practice $B$ must be fixed anyway, to develop standard errors, one can thus use the theory developed in Sections 3 and 4. Alternatively, if $B$ is large, then the results for $B = \infty$ hold approximately, and the expansion in (20) can be used in place of (10), by making the substitution

$$ X_{B,i}(\cdot) = E(\psi(Y_i, Z_i, W_i + \sigma \lambda^{1/2} \varepsilon, \Theta_{\sigma^2(1+\lambda)})|Y_i, Z_i, W_i), $$

(21)

a function that can be computed numerically.

Finally, a remark on regularity conditions. The $o_p(1)$ term in Equation (19) is an unconditional one, whereas the $o_p(1)$ term in (20) is conditional. Justifying this passage can be extraordinarily difficult and technical and is far beyond the scope of this article. Even in simple linear regression, with $T$ the mean functional, the argument is very involved. We have developed explicit arguments in special cases under very strong regularity conditions. The arguments can be obtained as a postscript file through anonymous ftp to picard.tamu.edu, in the directory /pub/rjcarroll, and in the file simex.theory.appendix.ps.
Table 1. Results of a Simulation Study for Linear Regression

<table>
<thead>
<tr>
<th></th>
<th>Linear extrapolation</th>
<th>Quadratic extrapolation</th>
<th>Nonlinear extrapolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic value</td>
<td>.837</td>
<td>.903</td>
<td>1.000</td>
</tr>
<tr>
<td>$\sigma^2$ known</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulated mean ($B = 100$)</td>
<td>.843</td>
<td>.912</td>
<td>1.024</td>
</tr>
<tr>
<td>Simulated std. dev. ($B = 100$)</td>
<td>.081</td>
<td>.091</td>
<td>.126</td>
</tr>
<tr>
<td>Average std. err. ($B = 100$)</td>
<td>.081</td>
<td>.091</td>
<td>.126</td>
</tr>
<tr>
<td>90% coverage: $\beta$ ($B = 100$)</td>
<td>.382</td>
<td>.734</td>
<td>.912</td>
</tr>
<tr>
<td>95% coverage: $\beta$ ($B = 100$)</td>
<td>.514</td>
<td>.802</td>
<td>.958</td>
</tr>
<tr>
<td>90% coverage: $\beta_2$ ($B = 100$)</td>
<td>.898</td>
<td>.898</td>
<td>.912</td>
</tr>
<tr>
<td>95% coverage: $\beta_2$ ($B = 100$)</td>
<td>.942</td>
<td>.944</td>
<td>.958</td>
</tr>
<tr>
<td>$\sigma^2$ estimated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulated mean ($B = 100$)</td>
<td>.843</td>
<td>.912</td>
<td>1.036</td>
</tr>
<tr>
<td>Simulated std. dev. ($B = 100$)</td>
<td>.085</td>
<td>.099</td>
<td>.158</td>
</tr>
<tr>
<td>Average std. err. ($B = 100$)</td>
<td>.083</td>
<td>.096</td>
<td>.156</td>
</tr>
<tr>
<td>90% coverage: $\beta$ ($B = 100$)</td>
<td>.402</td>
<td>.716</td>
<td>.918</td>
</tr>
<tr>
<td>95% coverage: $\beta$ ($B = 100$)</td>
<td>.502</td>
<td>.800</td>
<td>.950</td>
</tr>
<tr>
<td>90% coverage: $\beta_2$ ($B = 100$)</td>
<td>.888</td>
<td>.894</td>
<td>.918</td>
</tr>
<tr>
<td>95% coverage: $\beta_2$ ($B = 100$)</td>
<td>.946</td>
<td>.934</td>
<td>.950</td>
</tr>
</tbody>
</table>

NOTE: “Asymptotic value” refers to the asymptotic limit of each of the estimates. Coverage for $\beta$ refers to the coverage probability for the actual slope; coverage for $\beta_2$ refers to the coverage probability for the asymptotic value of the slope estimate.

6. NON-IID ERRORS

In defining the SIMEX estimator through the building-block solutions to (2), we have assumed that the $\epsilon_{ib}$ are independent and identically distributed. CS proposed instead that for each $b$, one modifies the $\epsilon_{ib}$’s to have mean zero, and variance 1 and to be uncorrelated with the observed data; they called these NON-IID errors. It is possible to modify the results derived in Section 3 to allow for NON-IID errors, but in practice we believe that such modifications (and additional complications!) will generally be unnecessary.

In their numerical work, CS found that these modifications had a negligible effect on the performance of the estimators. As we will show, the reason for this is that although $\hat{\Theta}_{SIMEX}$ differs from its limit by order $O_p((n^{-1/2})$, using NON-IID errors changes the estimator only by the order $O_p((nB)^{-1/2})$. Because $B$ is usually quite large, this suggests that in practice one may use the standard error estimates derived in Section 3.

It is easiest to see the result if one merely centers the $\epsilon_{ib}$’s, using instead $\epsilon_{ib*} = (\epsilon_{ib} - \bar{\epsilon}_b)$, where $\bar{\epsilon}_b$ is the sample mean of the $\epsilon_{ib}$ for $i = 1, \ldots, n$. The other modifications only increase the algebra without changing the idea. If we center, then (7) still holds but with $\epsilon_{ib}$ replaced by $\epsilon_{ib*}$.

Let $\hat{\Theta}_{\sigma^2}^{(1+\lambda)}$ and $\hat{\chi}_{B*,\sigma^2}^{(1+\lambda)}$ be defined as in (3) and (9) but with $\epsilon_{ib}$ replaced by $\epsilon_{ib*}$, and let $\xi$ be the derivative of $\psi$ with respect to its third element. Let $A_\xi = E\xi(Y, \mathbf{Z}_i, \mathbf{W}_i + \sigma^{1/2}\epsilon_{ib}, \Theta_{\sigma^2}^{(1+\lambda)})$. Using the same analysis that led to (9) and (10) and suppressing arguments, we find

$$\hat{\chi}_{B*,\sigma^2}^{(1+\lambda)}(\cdot) - \chi_{B*,\sigma^2}^{(1+\lambda)}(\cdot) = \sigma^{1/2}B^{-1}\sum_{b=1}^{B} \xi_{\cdot,\bar{\epsilon}_b}(\cdot) + O_p(n^{-1}),$$

which yields

$$A_\xi(\hat{\Theta}_{\sigma^2}^{(1+\lambda)} - \hat{\Theta}_{\sigma^2}^{(1+\lambda)})$$

$$= n^{-1}\sum_{i=1}^{n} (\chi_{B*,\sigma^2}^{(1+\lambda)}(\cdot) - \chi_{B*}^{(1+\lambda)}(\cdot)) + o_p(n^{-1/2})$$

$$= \sigma^{1/2}B^{-1}\sum_{b=1}^{B} \bar{\epsilon}_b n^{-1}\sum_{i=1}^{n} \xi_{\cdot,\bar{\epsilon}_b}(\cdot) + o_p(n^{-1/2}).$$

For any $b$, the inner sum over $n$ equals $A_\xi + o_p(n^{-1/2})$, and hence, because $B$ is fixed,

$$A_\xi(\hat{\Theta}_{\sigma^2}^{(1+\lambda)} - \hat{\Theta}_{\sigma^2}^{(1+\lambda)})$$

$$= \sigma^{1/2}A_\xi B^{-1}\sum_{b=1}^{B} \bar{\epsilon}_b + o_p(n^{-1/2}).$$

The last term is clearly of order $O_p((nB)^{-1/2})$, completing the argument.

7. NUMERICAL EXAMPLES

To evaluate the standard error estimates in a simple case, we simulated linear regression with $Y = \alpha + \beta X + \epsilon$, where $(\alpha, \beta) = (0, 1)$, $\epsilon$ was normally distributed with mean zero and variance 25, $X$ was normally distributed with mean zero and variance 100, and $U$ was normally distributed with mean zero and variance $\sigma^2 = 49$. The sample size was $n = 100$, and there were 500 iterations to the simulations. We worked with $B = 100$. As expected, there were no major differences in SIMEX between simulating IID or NON-IID errors, and we report results only for the latter. The SIMEX estimator was computed for a two-point linear extrapolation at $\lambda = 0, 1$; a three-point quadratic extrapolation at $\lambda = 0, 1, 2$; and a five-point fit to the nonlinear model with $\lambda = 0, 5, 1.0, 1.5, 2.0$. The results are displayed in Table 1.

Of course, as CS emphasized, different extrapolant functions lead to different results, and consistency of estimation requires that the “right” function be chosen. They discuss this issue in some detail, and we will not repeat their arguments here.

Two simulations were done. In the first case, we assumed that the measurement error variance $\sigma^2$ was known. In the second case we assumed that it was estimated by replication with standard errors computed as in Section 4.

The main result from the simulation is that in this particular problem, the estimated standard errors from our theory perform very well. The average standard error is nearly the same as the standard deviation of the estimates, and the coverage probabilities are very nearly the nominal.

We did some experimentation with a weighted five-point fit to the nonlinear model. In our numerical work, the estimated covariance matrix of the slopes was usually very nearly singular; the cases $\lambda = 1.5, 2.0$ were almost perfectly explained by the others. This near singularity made the weighted method fairly unstable and no more efficient than the unweighted method, at least in the sample sizes studied here.

8. FRAMINGHAM HEART STUDY

We illustrate the methods using data from the Framing-
ham Heart Study, correcting for bias due to measurement error in systolic blood pressure measurements. The Framingham study consists of a series of exams taken 2 years apart. We use Exam #3 as the baseline. There are 1,615 men age 31–65 in this data set, with the outcome Y indicating the occurrence of coronary heart disease (CHD) within an 8-year period following Exam #2: there were 128 such cases of CHD. Predictors used in this example are the patient’s age at Exam #2, smoking status at Exam #1, serum cholesterol level at Exam #3, and systolic blood pressure (SBP) at Exam #3. The latter is the average of two measurements taken by different examiners during the same visit. In addition to the measurement error in SBP measurements, there is also measurement error in the cholesterol measurements. But for this example we ignore the latter source of measurement error and illustrate the methods under the assumption that only SBP is measured with error.

The covariates measured without error (Z) are age, smoking status, and serum cholesterol level, along with an indicator for the intercept. For W, we use a modified version of a transformation originally due to Cornfield and discussed by Carroll, Spiegelman, Lan, Bailey, and Abbott (1984), setting $W = \log(\text{SBP-50})$. Implicitly, we are defining X as the long-term average of W.

In addition to the aforementioned variables, we also have SBP measured at Exam #2. The mean transformed SBP at Exams #2 and #3 are 4.37 and 4.35. Their difference has mean .02 and standard error .0040, so that the large-sample test of equality of means has p value < .0001. Thus in fact the measurement at Exam #2 is not exactly a replicate, but the difference in means from Exam #2 to Exam #3 is negligible for all practical purposes.

We present two sets of analyses. The first analysis uses the full complement of replicate measurements from Exam #2. In the second analysis we illustrate the procedures for the case when only a single measurement is used and $\sigma^2$ is estimated by a small validation data set obtained in this example by randomly selecting a subset of the Exam #2 SBP measurements.

### 8.1 Full Replication

This analysis uses the replicate SBP measurements from Exams #2 and #3 for all study participants. The transformed data are $W_{i,j}$, where $i$ denotes the individual and $j = 1, 2$ refers to the transformed SBP at Exams #2 and #3. The overall surrogate is $W_i$, the sample mean for each individual. The model is

$$W_{i,j} = X_i + U_{i,j}^*,$$

where the $U_{i,j}^*$ have mean zero and variance $\sigma_u^2$. The components of variance estimator

$$\hat{\sigma}_u^2 = n^{-1} \sum_{i=1}^n \sum_{j=1}^2 (W_{ij} - \bar{W}_i)^2$$

is .01259. In the notation of Section 4, $U_{i,j}^* = \sigma_u U_{i,j}$ and $\sigma_u = 2^{1/2}.\sigma$.

We use SIMEX with $W_i^* = W_i$, and $U_{i,j}^* = \bar{U}_i$. The sample variance of the terms $(W_i^*)^4$ is $\delta_{w,*} = .04543$, and the estimated measurement error variance is $\hat{\sigma}_u^2 = \hat{\delta}_{u,*}^2 = .00630$. Thus the linear model correction for attenuation (i.e., the inverse of the reliability ratio) for these data is 1.16093. There are 1,614 degrees of freedom for estimating $\hat{\sigma}_u^2$, and thus for practical purposes the measurement error variance is known.

In Table 2 we list the results of three analyses:

- the naive analysis that ignores measurement error
- the regression calibration analysis, in which X is replaced by the best linear estimate of $E(X|Z, W)$ (Carroll and Stefanski 1990; Gleser 1990; see Carroll et al. 1995, chap. 3, for detailed formulas)
- the SIMEX analysis.

For the naive analysis, sandwich and information refer to the sandwich and information standard errors; the latter is the output from standard statistical packages.

For the regression calibration analysis, the first sandwich and information standard errors are those obtained from a standard logistic regression analysis after substituting the estimates of $E(X|Z, W)$ for X and ignoring the fact that the equation is estimated. The second set of sandwich standard errors takes into account the estimation of $\sigma^2$ and uses the delta method (see Carroll et al. 1995, chap. 3, for details).
For the SIMEX estimator, M estimator refers to estimates derived from our theory (Sec. 4) for the case where \( \sigma^2 \) is estimated from the replicate measurements. Sandwich and Information refer to estimates derived by Stefanski and Cook (1996) for the case of known measurement error variance \( \sigma^2 \).

Figure 1 contains plots of the logistic regression coefficients \( \hat{\theta}(\lambda) \) for eight equally spaced values of \( \lambda \) spanning \( [0, 2] \) (solid circles). For this example, \( B = 2,000 \). The points plotted at \( \lambda = 0 \) are the naive estimates \( \hat{\theta}_{\text{naive}} \).

The nonlinear least squares fits of \( G_{\text{NL}}(\lambda, \Gamma) \) to the components of \( \{\hat{\theta}(\lambda_j), \lambda_j\}_{j=1}^g \) (solid curves) are extrapolated to \( \lambda = -1 \) (dashed curves), resulting in the SIMEX estimators (crosses). The open circles are the SIMEX estimators that result from fitting quadratic extrapolants. To preserve clarity, the quadratic extrapolants were not plotted. Note that the quadratic extrapolant estimates are conservative relative to the nonlinear extrapolant estimates in the sense that they fall between the nonlinear extrapolant estimates and the naive estimates.

A careful examination of Figure 1 reveals that the effect of measurement error is of practical significance only on the intercept and the coefficient of log(SBP-50).

The parameter estimates from SIMEX and regression calibration are strikingly similar in this example, which is to be expected given the relatively small amount of measurement error, combined with the fact that the observed data \( W \) are at least approximately normally distributed. All the methods give approximately the same standard errors, which can be explained here both by the relatively large sample size and by the fact that \( \sigma^2 \) is estimated extremely well.

8.2 Partial Replication

We now illustrate the analyses for the case where \( \sigma^2 \) is estimated from a small validation data set. The measured predictor, \( W \), is the single SBP measurement from Exam #3. A randomly selected subset of 30 replicate measurements...
from Exams #2 and #3 were used to estimate $\sigma^2$. For these data, the sample variance of $W$ is .05252 and the estimate of $\sigma^2$ is .01306. The estimated linear model correction for attenuation, or inverse of the reliability ratio, is 1.33121.

There are two major differences between this set of analyses and those from the previous section: (a) the measurement error variance is twice as large resulting in greater attenuation in the naive estimate; and (b) the measurement error variance is estimated with far less precision—29 degrees of freedom versus 1,614—resulting in less precise corrected estimates.

Both of these differences are reflected in the results reported in Table 3. The standard errors are calculated as in Table 2, with the exception that bootstrap standard errors were not calculated for the regression calibration estimates. One notes that the SIMEX standard errors are smaller than the M estimation standard errors proposed in this article, because $\sigma^2$ now is estimated with only 29 degrees of freedom.

9. DISCUSSION

This article has investigated the asymptotic distribution of the SIMEX estimator. We have shown that the estimator (3) is typically asymptotically normally distributed. We have developed methods for consistent estimation of the asymptotic variance of the SIMEX estimators, allowing the measurement error variance $\sigma^2$ to be known or estimated.

Although our description of the SIMEX algorithm is based on generating standard normal random variables, our asymptotic theory (and the method itself) applies to any distribution for the $\varepsilon_{ip}$'s.

Our definition of the SIMEX estimator (3) uses the average of the $\hat{\theta}_b(\lambda)$'s. CS discussed the possibility of using other measures of location such as the median. We have focused on the means because, as CS noted in their numerical examples, “using the sample mean as the location estimator works as well as any other . . . location functional studied” (p. 1327)—at least for the most common examples. Their numerical experience is explained by the results in Section 5, where we indicated that for large values of $B$, and subject to regularity conditions, the standard location functionals can be expected to have the same asymptotic behavior.

One should note, however, that for fixed and small $B$, location functionals other than (3) typically result in estimators that are not asymptotically normally distributed. To see this, remember that for fixed $\lambda$, the terms $\hat{\theta}_b(\lambda)$ for $b = 1, \ldots, B$ are jointly asymptotically normally distributed with common mean. To understand the basic point, consider the median as the location functional and the extreme case that $B = 3$. Because the median of three correlated normal random variables is not normally distributed, none of the terms corresponding to our $\hat{\theta}(\lambda)$ are asymptotically normally distributed, and hence neither would the median SIMEX estimator be asymptotically normally distributed.

[Received August 1993. Revised April 1995.]

REFERENCES