Calculation and Simulation in Errors-in-Variables Regression Problems

by

Leon Jay Gleser
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #78-5

May 1978

*Invited paper given at 163rd meeting of the Institute of Mathematical Statistics, New Brunswick, New Jersey, June 1, 1978. Research on this paper was sponsored by the Air Force Office of Scientific Research under Grant No. AFOSR-77-3291 at Purdue University. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

Calculation and Simulation in Errors-in-Variables Regression Problems*

by

Leon Jay Gleser
Purdue University

1. The Model. The multivariate errors-in-variables regression model can be described as follows: An unobservable vector v_i of the values of r dependent variables and an unobservable vector u_i of the values of p independent variables are generated by some natural or sociological phenomenon of interest, $i = 1,2,\ldots,n$. These unobservable vectors are related by the linear model

$$v_i = \alpha + Bu_i, \quad i = 1,2,\ldots,n, \tag{1.1}$$

where α: rx1 and B: rxp are the unknown vector of intercepts and the unknown matrix of regression slopes, respectively. We observe v_i and u_i with error. That is, what we actually observe is

$$\begin{aligned} x_{1i} &= u_i + e_{1i}, \\ x_{2i} &= v_i + e_{2i}, \qquad i = 1,2,\ldots,n, \end{aligned} \tag{1.2}$$

where

$$e_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} : (p+r)x1$$

is a vector of random errors having mean vector 0 and covariance matrix

*Invited paper given at 163rd meeting of the Institute of Mathematical Statistics, New Brunswick, New Jersey, June 1, 1978. Research on this paper was sponsored by the Air Force Office of Scientific Research under Grant No. AFOSR-77-3291 at Purdue University. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

(1.3) $\Sigma = \mathcal{E}(e_i e_i') = \sigma^2 I_{p+r}, \quad \sigma^2 > 0.$

We also assume that the e_i's are independent with a common (p+r)-variate normal distribution.

Although the assumption that the error covariance matrix is a multiple of the identity matrix is customarily adopted in classical regression analysis, the practical statistician may nonetheless be a little disturbed that such a strong restriction is placed on the error covariance matrix in the present problem. Indeed, more general models for Σ have been proposed in the literature, particularly when $r = p = 1$ (the case of simple linear regression with errors in variables). However, some restrictions on Σ must be adopted; for if Σ is allowed to be a general (p+r)x(p+r) positive definite covariance matrix, Gleser (1978) shows that the maximum of the likelihood is infinite and no maximum likelihood estimators of the unknown parameters exist. Further, a recent paper by Nussbaum (1977) shows in the case $r = p = 1$ that if Σ is general, then no consistent estimator of B can exist. Admittedly, there is some room for a model for Σ intermediate between $\Sigma = \sigma^2 I_{p+r}$ and a general Σ, but with one exception [Bhargava (1977)] such models have not been treated correctly in the literature. Of course, we can assume that $\Sigma = \sigma^2 \Sigma_0$, where Σ_0 is known, and then transform the data so that the transformed data obey an errors-in-variables model where the errors have covariance matrix $\sigma^2 I_{p+r}$ (see the Appendix for a way to do this), but this is only a minor relaxation of the restrictions on Σ. Thus, at the present "state of the art", if we wish to apply errors-in-variables regression models to data, we must be willing to make the assumption that we know the covariance matrix Σ of the measurement errors e_i, $i = 1,2,\ldots,n$, except for a scalar multiple σ^2.

2. Computation of Estimators. To compute estimators of the unknown parameters in our model, we start by computing the sample mean vector

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \tag{2.1}$$

and the sample cross-product matrix

$$W = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' : (p+r)x(p+r), \tag{2.2}$$

where

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} .$$

In order for W to be almost surely positive definite we require that

$$n-1 \geq p+r. \tag{2.3}$$

There are now two possible ways to compute our estimators. Either way, up to errors produced by roundoff, will yield the same values for the estimators.

Method 1 Find the largest p eigenvalues $d_1 \geq d_2 \geq \ldots \geq d_p$ of W and the corresponding orthonormal eigenvectors $g_1, g_2, \ldots, g_p$. Let

$$\begin{pmatrix} G_{11} \\ G_{21} \end{pmatrix} = (g_1, g_2, \ldots, g_p) : (p+r)xp, \tag{2.4}$$

so that

$$W \begin{pmatrix} G_{11} \\ G_{21} \end{pmatrix} = \begin{pmatrix} G_{11} \\ G_{21} \end{pmatrix} [\text{diag } (d_1, d_2, \ldots, d_p)]. \tag{2.5}$$

Then compute

$$\hat{B} = G_{21} G_{11}^{-1}, \tag{2.6}$$

$$\hat{\alpha} = \bar{x}_2 - \hat{B}\, \bar{x}_1, \tag{2.7}$$

$$\hat{\sigma}^2 = \frac{\operatorname{tr} W - \sum_{i=1}^{p} d_i}{nr}, \tag{2.8}$$

and

$$\hat{u}_i = \bar{x}_1 + G_{11}G_{11}'(x_{1i}-\bar{x}_1)+G_{11}G_{21}'(x_{2i}-\bar{x}_2),\ i = 1,2,\ldots,n. \tag{2.9}$$

Method 2 Find the smallest r eigenvalues $d_{p+1} \geq d_{p+2} \geq \cdots \geq d_{p+r}$ $(d_{p+1} \leq d_p)$ of W and the corresponding orthonormal eigenvectors $g_{p+1}, g_{p+2}, \ldots, g_{p+r}$. Let

$$\begin{pmatrix} G_{12} \\ G_{22} \end{pmatrix} = (g_{p+1}, g_{p+2}, \ldots, g_{p+r}): (p+r)xr, \tag{2.4'}$$

so that

$$W\begin{pmatrix} G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} G_{12} \\ G_{22} \end{pmatrix} [\operatorname{diag}(d_{p+1}, d_{p+2}, \ldots, d_{p+r})]. \tag{2.5'}$$

Then compute

$$\hat{B} = -(G_{22}')^{-1}(G_{12}'), \tag{2.6'}$$

$$\hat{\alpha} = \bar{x}_2 - \hat{B}\bar{x}_1, \tag{2.7'}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=p+1}^{p+r} d_i}{nr}, \tag{2.8'}$$

and

$$\hat{u}_i = \bar{x}_1 + (I_p - G_{12}G_{12}')(x_{1i}-\bar{x}_1) - G_{12}G_{22}'(x_{2i}-\bar{x}_2),\ i = 1,2,\ldots,n. \tag{2.9'}$$

The equivalence of these two methods of computation follows from the fact that

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \tag{2.10}$$

is a (p+r)x(p+r) orthogonal matrix.

Healy (1975) and Gleser (1978) show that $\hat{B}$, $\hat{\alpha}$, $r(p+r)^{-1}\hat{\sigma}^2$, and $\hat{u}_i$, $i = 1,2,\ldots,n$, are maximum likelihood estimators of B, α, σ^2, and u_i, $i = 1,2,\ldots,n$, respectively. They also show that if

(2.11) $$\mu = \lim_{n\to\infty} \bar{u} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} u_i \quad \text{exists}$$

and

(2.12) $$\Delta = \lim_{n\to\infty} \Delta_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (u_i-\bar{u})(u_i-\bar{u})' \quad \text{exists,}$$

and Δ is positive definite, then $\hat{B}$, $\hat{\alpha}$, and $\hat{\sigma}^2$ are strongly (a.s.) consistent estimators of B, α, and σ^2, respectively. Gleser (1978) also shows that under these same assumptions

(2.13) $$\hat{\Delta} = \frac{1}{n} G_{11}(D_{max} - \hat{\sigma}^2 I_p)G_{11}',$$

where $D_{max} = \text{diag}(d_1,d_2,\ldots,d_p)$, is a strongly consistent estimator of Δ, and that the elements of $n^{1/2}(\hat{B}-B)$ have a limiting (as $n \to \infty$) rp-variate normal distribution with zero means and a certain covariance matrix depending on B, Δ and σ^2. It follows from this last result that a $100(1-\alpha)\%$ large-sample confidence region for the elements of B is

(2.14) $$\{B: (\hat{\sigma}^2)^{-1} \text{tr}[(I_r+\hat{B}\hat{B}')^{-1}(\hat{B}-B)G_{11}D_{max}^{-1}(D_{max}-\hat{\sigma}^2 I_p)^2 G_{11}'(\hat{B}-B)'] \leq \chi^2_{rp}[1-\alpha]\},$$

where $\chi^2_{rp}[1-\alpha]$ is the $100(1-\alpha)$-th percentile of the chi-squared distribution with rp degrees of freedom.

Note that the confidence region (2.14) is an ellipse in rp-dimensional space. To my knowledge, this is the only reasonably-shaped (convex) confidence region for the elements of B published in the literature. Thus, it is of considerable interest to determine how well this confidence region performs when the sample size n is of small to moderate size. Unfortunately, the extreme distributional complexity inherent in the errors-in-variables model (see Section 3) forces us to evaluate the performance of this region by simulations.

To do simulations, it becomes necessary to generate many repetitions of the sampling situation for varying values of α, B, σ^2, and the u_i's. In consequence, it becomes very important to determine the most efficient ways to calculate the estimators and construct the confidence region (2.14), since even minor inefficiencies in calculation mean a major cost in computation time when totaled over hundreds of simulations.

Which method should we use to calculate the estimators - Method 1 or Method 2? Looking at the formulas for the estimators, we see that $\hat{\alpha}$ is computed in the same way under either method. Further, since the $\hat{u}_i$'s are not consistent estimators and are not needed to construct the confidence region (2.14) [and are also rarely of much interest in practice], we need not consider differences in computation efficiency between Methods 1 and 2 for these estimators. Thus, we can compare the computational efficiency of Methods 1 and 2 in terms of the number of elementary arithmetic operations needed to compute $\hat{B}$, $\hat{\sigma}^2$, and the confidence region (2.14).

At first glance, the choice of method seems to be obvious. To compute $\hat{B}$ under Method 1, we must invert a pxp matrix and then multiply an rxp matrix times a pxp matrix, while to compute $\hat{\sigma}^2$ we must average p quantities. To compute $\hat{B}$ under Method 2, we must invert an rxr matrix and then multiply an rxr matrix times an rxp matrix, while to compute $\hat{\sigma}^2$ we must average r quantities. Thus, it appears that we should choose Method 1 if $p < r$, and choose Method 2 if $p > r$, with either method being used if $p = r$. If the choice were this easy, I would have been embarrassed to discuss it. What has been forgotten in this comparison is that we must use an eigenvalue-eigenvector routine to obtain the matrices and scalars used in computing $\hat{B}$ and $\hat{\sigma}^2$.

There are basically two types of eigenvalue-eigenvector routines available to the statistician. One type, which is called the iterative type, was once the only available method, and is the method described in most textbooks for hand computation. In this method, the largest eigenvalue d_1 of W is obtained first, along with its associated eigenvector g_1. Then the influence of this largest eigenvalue and associated eigenvector is subtracted from W, and the largest eigenvalue and associated eigenvector of the remaining matrix are obtained. (These are, of course, actually the second largest eigenvalue d_2 of W, and the associated eigenvector g_2.) This process is repeated over and over, so that d_3 and g_3 are obtained next, then d_4 and g_4, and so on.

The second type of eigenvalue-eigenvector routine became feasible only with the coming of the big fast computer. All variants of this type first rotate W to a special form, using symmetric orthogonal matrices known as Householder matrices. Once W is rotated to a special form W*, the characteristic polynomial of W* is easily calculated, and the so-called Sturm polynomials are used to locate the roots of this polynomial within intervals. Numerical techniques which are basically variants of Newton's method then are used to find the eigenvalues of W* (which are also the eigenvalues of W). Once the eigenvalues of W* are found, the eigenvector g_i^* corresponding to the eigenvalue d_i is found by solving the linear system of equations

$$W^* g_i^* = d_i g_i^*,$$

and then the eigenvector g_i of W corresponding to d_i is found from g_i^* by reversing the rotations that rotated W to W*. Such a computational routine is most efficient when the eigenvalues of W* (and W) are found in a certain order (smallest to largest, or largest to smallest) and the routine is stopped

before all eigenvalues are obtained. Routines of this kind are often called QR (or QL or QRL) routines. They tend to be both more efficient and more accurate than the iterative routines, and thus are to be preferred in practice.

Suppose, however, that we use an iterative eigenvalue-eigenvector routine to calculate the d_i's and corresponding g_i's. Such a routine will use an order of magnitude greater computing time than the simple matrix operations used to compute $\hat{B}$, even if the procedure is stopped before all eigenvalues are obtained. If we use Method 2 to compute $\hat{B}$ and $\hat{\sigma}^2$, we must carry out the iterative routine until the last eigenvalue and eigenvector are obtained; whereas if we use Method 1, we can stop the routine once d_p and g_p have been obtained. The extra calculations required to run the iterative eigenvalue-eigenvector routine to the end under Method 2 will lose the saving in computations for $\hat{B}$ and $\hat{\sigma}^2$ over Method 1 noted earlier, and much more. Further, it is well known that roundoff errors in the early eigenvalues and eigenvectors propogate as the iteration continues, so that $d_{p+1},\ldots d_{p+r}$ and $g_{p+1},\ldots,g_{p+r}$ are calculated less accurately than $d_1,\ldots,d_p$ and $g_1,\ldots,g_p$. For these reasons, statisticians using the iterative method to compute the eigenvalues and eigenvectors of W should use Method 1 to compute the estimators almost regardless of the relative magnitudes of p and r. The only exception to this rule is when r is 1 or 2 and p is very large, and even then the convenience of being able to stop the iterative eigenvalue-eigenvector routine early would argue in favor of Method 1.

As long as we are not planning to construct the confidence region (2.14) and provided that we are able to use an eigenvalue-eigenvector routine of the noniterative (QL, QR, QRL) type which is flexible enough to permit us to obtain eigenvalues either from smallest to largest, or from largest to smallest, stopping at any time we have found all the eigenvalues we want, then our earlier

analysis of the relative merits of Method 1 and Method 2 remains unchanged. That is, we use Method 1 if $p < r$ and Method 2 if $p > r$. However, this comparison is based on computational efficiency alone. It is well known that accuracy of computation of eigenvalues and eigenvectors increases with increase in the spread of the eigenvalues. In large samples, Gleser (1978) shows that

$$\lim_{n\to\infty} \frac{1}{n} W = \sigma^2 I_{p+r} + \begin{pmatrix} I_p \\ B \end{pmatrix} \Delta \begin{pmatrix} I_p \\ B \end{pmatrix}',$$

so that

$$\lim_{n\to\infty} \frac{1}{n} d_{p+i} = \sigma^2, \quad i = 1,2,\ldots,r,$$

while

$$\lim_{n\to\infty} \frac{1}{n} d_i = \sigma^2 + \tau_i, \quad i = 1,2,\ldots,p,$$

where $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_p > 0$ are the positive (non-zero) eigenvalues of $(I_p, B')'\Delta(I_p, B')$. Thus, assuming that the τ_i's are unequal, the p largest eigenvalues of W will be more widely dispersed than the r smallest eigenvalues, and thus Method 1 is likely (assuming p and r are not greatly different) to give greater accuracy than Method 2.

If we plan to calculate the confidence region (2.14), however, Method 1 is always preferable to Method 2. The reason for this assertion is that regardless of how we compute $\hat{B}$ and $\hat{\sigma}^2$, to construct the region (2.14) we must know the value of the matrix G_{11}. Using only G_{12} and G_{22} and the orthogonality of the matrix G in (2.10), we can determine G_{11} up to an orthogonal rotation, but a look at (2.14) shows us that this is not enough. Thus, our choice becomes one of using Method 2, in which case we must obtain all eigenvalues and eigenvectors of W in order to obtain G_{12}, G_{22}, and G_{11}, or Method 1, in which case we need compute only the largest p eigenvalues and their associated eigenvectors.

The choice now is clear. We should use Method 1!

In this section, we have seen how one can come to erroneous conclusions in statistical computation by not considering the entire process of data manipulation. Theoreticians are in particular prone to this kind of error—treating the symbols in their formulas as if they were given pieces of data, rather than quantities that must be computed from data, often in very expensive ways. Although I am only sometimes a theoretician, I must admit that I was trapped in the error of preferring Method 2 until I was forced by the need to do simulations on the region (2.14) into considering the entire process by which the statistical computations must be done. Thus, this section is partly in expiation for my earlier sin.

3. A Canonical Distributional Form. Where theoreticians have an advantage over data analysts and other practical people is that they are not as prone to becoming wedded to a particular coordinate system. I have seen many complicated (and poorly designed) simulation studies which could have been greatly simplified (and thereby improved) by a little use of invariance theory. Thus, for the present problem, I rotated the variables in various ways trying to achieve distributional simplification.

Let us start this exercise by counting the number of parameters that would have to be specified if we started simulation of our model directly from (1.1), (1.2), (1.3). There are r components in α, rp components in B, pn components in

$$U = (u_1, u_2, \ldots, u_n): \ p \times n, \tag{3.1}$$

and 1 component in σ^2, making a total of (n+r)p+r+1 parameters to be specified. Even when r = p = 1 and n = 3, this means that our simulation

design must be crossed over the values of 6 parameters!

However, we are interested only in simulating to study how accurately our estimators of α, B, and σ^2 estimate these parameters, and how close to $100(1-\alpha)\%$ the true probability of coverage of the large-sample confidence region (2.14) is in moderate samples. We note that $\hat{\alpha}$, $\hat{B}$, and $\hat{\sigma}^2$ are functions of the data only through $\bar{x}_1$, $\bar{x}_2$, and W. It can be shown that

$$\text{(3.2)} \qquad \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \bar{u} \\ \alpha+B\bar{u} \end{pmatrix}, \frac{\sigma^2}{n} I_{p+r} \right),$$

where $\bar{u} = n^{-1} \sum_{i=1}^{n} u_i$; that W has a noncentral Wishart distribution with $(n-1)$ degrees of freedom, covariance matrix parameter $\sigma^2 I_{p+r}$ and noncentrality matrix parameter

$$\text{(3.3)} \qquad \begin{aligned} &(\sigma^2)^{-1} \begin{pmatrix} I_p \\ B \end{pmatrix} \sum_{i=1}^{n} (u_i-\bar{u})(u_i-\bar{u})' \begin{pmatrix} I_p \\ B \end{pmatrix}' \\ &= n(\sigma^2)^{-1} \begin{pmatrix} I_p \\ B \end{pmatrix} \Delta_n \begin{pmatrix} I_p \\ B \end{pmatrix}', \end{aligned}$$

and that $(\bar{x}_1', \bar{x}_2')'$ and W are statistically independent. Thus, we can perform our simulations by specifying the parameters $\bar{u}$, α, B, σ^2, and Δ_n, and then simulating $(\bar{x}_1', \bar{x}_2')'$ and W. Noting that (3.3) is a $(p+r)\times(p+r)$ matrix of rank p, an algorithm developed by the present author [Gleser (1976)] for simulating the noncentral Wishart distribution with singular noncentrality matrix will be useful.

Counting parameters once again, we now have p components in $\bar{u}$, r components in α, rp components in B, 1 component in σ^2, and p^2 components in Δ_n for a total of $(r+p)(p+1)+1$ parameter values that must be specified. Notice that the sample size n no longer enters into our counting of design parameters, although n must be specified in order to do the simulations. If $r = p = 1$, we

need to specify $(1+1)(1+1)+1 = 5$ design parameters.

In making the reduction in parameterization above, we are in essence utilizing an invariance argument. What we have shown is that the joint distribution of $(\bar{x}_1', \bar{x}_2')$ and W, and thus the joint distribution of our estimators, is the same for all

$$(3.4) \qquad U = (u_1, u_2, \ldots, u_n) = \bar{u} 1_n + n^{1/2} \begin{pmatrix} I_p \\ B \end{pmatrix} \Delta_n^{1/2} \Gamma_n$$

where $1_n = (1,1,\ldots,1)'$: nx1, $\Delta_n^{1/2}$ is the symmetric square root of Δ_n, and Γ_n is any (p+r)xn row orthogonal matrix for which $\Gamma_n 1_n = 0$.

Let

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1n} \\ x_{21} & x_{22} \cdots x_{2n} \end{pmatrix} = (x_1, x_2, \ldots, x_n)$$

and

$$E = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = \begin{pmatrix} e_{11} & e_{12} \cdots e_{1n} \\ e_{21} & e_{22} \cdots e_{2n} \end{pmatrix} = (e_1, e_2, \ldots, e_n),$$

where both X and E are (p+r)xn. In terms of X, U, E, our model (1.1), (1.2), (1.3) is

$$(3.5) \qquad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} U \\ \alpha 1_n + BU \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}$$

where the columns of $E = (E_1', E_2')'$ are i.i.d. with a common $MVN(0, \sigma^2 I_{p+r})$ distribution. Also, letting

$$C_n = I_n - n^{-1} 1_n 1_n',$$

we have

$$(3.6) \qquad \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = n^{-1} X 1_n = n^{-1} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} 1_n,$$

and

(3.7) $$W = X\, C_n\, X'.$$

Consider the linear transformation

(3.8) $$X^* = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} = k \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} X = \begin{pmatrix} k\Gamma_1 & X_1 \\ k\Gamma_2 & X_2 \end{pmatrix},$$

where $k \neq 0$, Γ_1 is a pxp orthogonal matrix, and Γ_2 is an rxr orthogonal matrix. Then from (3.5) and (3.8),

(3.9) $$X^* = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} = \begin{pmatrix} k\Gamma_1 U \\ k\Gamma_2 \alpha 1_n + \Gamma_2 B \Gamma_1' (k\Gamma_1 U) \end{pmatrix} + \begin{pmatrix} k\Gamma_1 E_1 \\ k\Gamma_2 E_2 \end{pmatrix}$$

$$= \begin{pmatrix} U^* \\ \alpha^* 1_n + B^* U^* \end{pmatrix} + \begin{pmatrix} E_1^* \\ E_2^* \end{pmatrix},$$

where

(3.10) $$U^* = k\Gamma_1 U,\ \alpha^* = k\Gamma_2 \alpha,\ B^* = \Gamma_2 B \Gamma_1'$$

and the columns of

$$E^* = \begin{pmatrix} k\Gamma_1 E_1 \\ k\Gamma_2 E_2 \end{pmatrix}$$

are i.i.d. with common distribution $MVN(0, k^2\sigma^2 I_{p+r})$. Also

(3.11) $$\bar{x}^* = \begin{pmatrix} \bar{x}_1^* \\ \bar{x}_2^* \end{pmatrix} = k \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} \begin{pmatrix} \frac{1}{n} X 1_n \end{pmatrix} = \begin{pmatrix} k\Gamma_1 \bar{x}_1 \\ k\Gamma_2 \bar{x}_2 \end{pmatrix},$$

and

$$W^* = (X^*)(C_n)(X^*)' = k^2 \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} X\, C_n\, X' \begin{pmatrix} \Gamma_1' & 0 \\ 0 & \Gamma_2' \end{pmatrix}$$

$$= k^2 \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} W \begin{pmatrix} \Gamma_1' & 0 \\ 0 & \Gamma_2' \end{pmatrix}.$$

Let $d_1^* \geq d_2^* \geq \ldots \geq d_{p+r}^*$ be the (ordered) eigenvalues of W* and let

$$G^* = \begin{pmatrix} G_{11}^* & G_{12}^* \\ G_{21}^* & G_{22}^* \end{pmatrix} \tag{3.12}$$

be the (p+r)x(p+r) orthogonal matrix whose ith column is the eigenvector corresponding to d_i^*, i = 1,2,...,p+r. Then

$$d_i^* = k^2 d_i, \qquad i = 1,2,\ldots,p+r, \tag{3.13}$$

and

$$G^* = \begin{pmatrix} G_{11}^* & G_{12}^* \\ G_{21}^* & G_{22}^* \end{pmatrix} = \begin{pmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{pmatrix} G = \begin{pmatrix} \Gamma_1 G_{11} & \Gamma_1 G_{12} \\ \Gamma_2 G_{21} & \Gamma_2 G_{22} \end{pmatrix}, \tag{3.14}$$

from which we conclude that

$$\begin{aligned} \hat{B}^* &= G_{21}^* (G_{11}^*)^{-1} = \Gamma_2 G_{21} (\Gamma_1 G_{11})^{-1} = \Gamma_2 G_{21} G_{11}^{-1} \Gamma_1' \\ &= \Gamma_2 \hat{B} \Gamma_1', \end{aligned} \tag{3.15}$$

$$\hat{\alpha}^* = \bar{x}_2^* - \hat{B}^* \bar{x}_1^* = k\Gamma_2 \bar{x}_2 - \Gamma_2 \hat{B} \Gamma_1' (k\Gamma_1 \bar{x}_1) \tag{3.16}$$

$$= k\, \Gamma_2\, \hat{\alpha},$$

and

$$(\hat{\sigma}^*)^2 = \frac{\sum_{i=p+1}^{p+r} d_i^*}{rp} = \frac{k^2 \sum_{i=p+1}^{p+r} d_i}{rp} = k^2 \hat{\sigma}^2. \tag{3.17}$$

Also

$$\begin{aligned} \Delta_n^* &= n^{-1} \sum_{i=1}^{n} (u_i^* - \bar{u}^*)(u_i^* - \bar{u}^*)' = n^{-1} U^* C_n (U^*)' \\ &= k^2 \Gamma_1 U C_n U' \Gamma_1' = k^2 \Gamma_1 \Delta_n \Gamma_1', \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} \hat{\Delta}^* &= n^{-1} (G_{11}^*)(D_{max}^* - (\hat{\sigma}^*)^2 I_p)(G_{11}^*)' \\ &= k^2 \Gamma_1 \hat{\Delta} \Gamma_1'. \end{aligned} \tag{3.19}$$

Finally, note that the confidence region (2.14) for B has the form

$$\{B: F(B; \hat{B}, G_{11}, D_{max}, \hat{\sigma}^2) \leq \chi^2_{rp}[1-\alpha]\}$$

where

$$F(B; \hat{B}, G_{11}, D_{max}, \hat{\sigma}^2)$$

(3.20)

$$= (\hat{\sigma}^2)^{-1} \operatorname{tr}[(I_r + \hat{B}\hat{B}')^{-1}(\hat{B}-B)G_{11}D_{max}^{-1}(D_{max}-\hat{\sigma}^2 I_p)^2 G_{11}'(\hat{B}-B)'].$$

Using (3.10), (3.14), (3.15), and (3.17), it is easily shown that

$$(3.21) \qquad F(B^*; \hat{B}^*, G_{11}^*, D_{max}^*, (\hat{\sigma}^*)^2) = F(B; \hat{B}, G_{11}, D_{max}, \hat{\sigma}^2),$$

so that the probability of coverage of the region (2.14) is invariant under transformations of the form (3.8).

What this argument has shown is that the properties of the estimators $\hat{\alpha}$, $\hat{B}$, $\hat{\sigma}^2$, $\hat{\Delta}$, and the confidence region (2.14) when $\bar{u} = \bar{u}_0$, $\alpha = \alpha_0$, $B = B_0$, $\sigma^2 = \sigma_0^2$, $\Delta_n = \Delta_n^{(0)}$ can easily be inferred from the properties of these estimators and this confidence region when

$$\bar{u} = k\Gamma_1\bar{u}_0, \ \alpha = k\Gamma_2\alpha_0, \ B = \Gamma_2 B_0 \Gamma_1', \ \sigma^2 = k^2\sigma_0^2, \ \Delta_n = k^2\Gamma_1\Delta_n^{(0)}\Gamma_1',$$

and k, Γ_1, Γ_2 are chosen at our convenience. In particular, we can let $k = \sigma_0^{-1/2}$, Γ_1 be the pxp orthogonal matrix whose rows are eigenvectors of $\Delta_n^{(0)}$, so that

$$\Gamma_1\Delta_n^{(0)}\Gamma_1' = D_\delta = \operatorname{diag}(\delta_1, \delta_2, \ldots, \delta_p),$$

where $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_p$ are the eigenvalues of $\Delta_n^{(0)}$, and Γ_2 be the rxr orthogonal matrix for which (when $r \leq p$):

$$\Gamma_2(B_0\Gamma_1') = (H_1, H_2),$$

where H_1: rxr is upper triangular, H_2: rx(p-r) is arbitrary. When r > p, Γ_2 is chosen so that

$$\Gamma_2(B_0\Gamma_1') = \begin{pmatrix} H_1 \\ 0 \end{pmatrix},$$

where H_1 is pxp upper triangular, 0 is a (r-p)xp matrix of zeros. That is, our parameters are now $\sigma^2 = 1$, and:

Parameter	Number of Components
α = no special form	r
$\bar{u}$ = no special form	p
$B = (H_1, H_2)$ or $\begin{pmatrix} H_1 \\ 0 \end{pmatrix}$	$\frac{(2p-r+1)r}{2}$ or $\frac{p(p+1)}{2}$
$\Delta = \text{diag}(\frac{\delta_1}{\sigma_0^2}, \frac{\delta_2}{\sigma_0^2}, \ldots, \frac{\delta_p}{\sigma_0^2})$	p

for a total of either (1/2) [2p(r+2)-r(r-3)] parameters when $r \leq p$, or $(1/2)(p^2+5p+2r)$ parameters when $p < r$. Although, we have dramatically reduced the number of parameters that must be specified in order to do our simulations, the number of parameters that must be specified in our simulation design (apart from n, p, and r) is still large, as illustrated by the following table:

p \ r	1	2	3	4	5
1	4	5	6	7	8
2	7	9	10	11	12
3	10	13	15	16	17
4	13	17	20	22	23
5	16	21	25	28	30

If we are willing to consider simulating only in order to evaluate the coverage of the confidence region (2.14), then since (2.14) depends on the data only through W, and the distribution of W depends only on σ^2, B and Δ_n

and not on α or $\bar{u}$, we need not specify the values of α and $\bar{u}$. This reduces the number of parameters we need by p+r, so that we now need to specify only $\Delta = \text{diag}(\delta_1/\sigma_0^2, \delta_2/\sigma_0^2, \ldots, \delta_p/\sigma_0^2)$ and B. Thus we need to specify the values of $(1/2)[2p(r+1)-r(r-1)]$ parameters when $r \leq p$, and $(1/2)(p^2+3p)$ parameters when $p < r$. Our table of the number of parameters which we need to specify in our simulation design now becomes:

p \ r	1	2	3	4	5
1	2	2	2	2	2
2	4	5	5	5	5
3	6	8	9	9	9
4	8	11	13	14	14
5	10	14	17	19	20

Well designed simulations in this case appear feasible when $p = 1 \leq r \leq 5$, and possibly even for $p = 2$, $r = 1$. For other values of r and p, we might consider concentrating on specific choices of B and Δ for our simulations.

One final comment about the simulations needed to evaluate the confidence region (2.14) should be noted. In performing the simulations, it is helpful and not very expensive, to print out a histogram of values of $F(B; \hat{B}, G_{11}, D_{max}, \hat{\sigma}^2)$. By looking at this histogram, we can see whether $\chi^2_{rp}[1-\alpha]$ exceeds the $100(1-\alpha)$th percentile of the histogram distribution (in which case the confidence of our region exceeds $100(1-\alpha)\%$). What we gain is information that may later allow us to change the cutoff point $\chi^2_{rp}[1-\alpha]$ so as to obtain greater control of the coverage of the region in small samples.

REFERENCES

[1] Bhargava, A. K. (1977). Maximum likelihood estimation in a multivariate 'errors in variables' regression model with unknown error covariance matrix. Comm. Statist.-Theor. Meth. A6, 587-601.

[2] Gleser, Leon Jay (1976). A canonical representation for the noncentral Wishart distribution useful for simulation. Journ. Amer. Statist. Assoc. 71, 690-695.

[3] Gleser, Leon Jay (1977). Multivariate regression with errors in variables: Linear functional relationships. Invited paper presented at 157th meeting of IMS, Madison, Wisconsin, May 17, 1977. Available from author.

[4] Gleser, Leon Jay (1978). Estimation of a linear transformation: Large sample results. Revision of Purdue University Department of Statistics, Mimeo Series No. 453. To appear in Ann. Statist.

[5] Healy, John Douglas (1975). Estimation and tests for unknown restrictions in multivariate linear models. Purdue University Department of Statistics, Mimeo Series No. 471.

[6] Nussbaum, N. (1977). Asymptotic optimality of estimators of a linear functional relation if the ratio of the error variances is known. Math. Op. Statist. Ser. Statist. 8, 173-198.

Appendix. We here show how to transform an errors-in-variables model in which the error covariance matrix $\Sigma = \sigma^2\Sigma_0$, where

$$\text{(A.1)} \qquad \Sigma_0 = \begin{pmatrix} \Sigma_{11}^{(0)} & \Sigma_{12}^{(0)} \\ \Sigma_{21}^{(0)} & \Sigma_{22}^{(0)} \end{pmatrix}$$

is known, into an errors-in-variables model in which the error covariance matrix is $\sigma^2 I_{p+r}$. Suppose that

$$\text{(A.2)} \qquad \begin{aligned} z_{1i} &= w_i + f_{1i}, \\ z_{2i} &= \alpha^* + B^* w_i + f_{2i}, \end{aligned}$$

where z_{1i}, w_i and f_{1i} are px1; z_{2i}, α^* and f_{2i} are rx1, and B^* is rxp. We assume that

$$f_i = \begin{pmatrix} f_{1i} \\ f_{2i} \end{pmatrix}, \quad i = 1,2,\ldots,n,$$

are i.i.d. with a common (p+r)-variate normal distribution with mean vector 0 and covariance matrix $\sigma^2\Sigma_0$.

To define the matrix T: (p+r)x(p+r) which will transform the data, first let

$$\text{(A.3)} \qquad \Sigma_{22.1}^{(0)} = \Sigma_{22}^{(0)} - (\Sigma_{21}^{(0)})(\Sigma_{11}^{(0)})^{-1}(\Sigma_{12}^{(0)}),$$

and then define

$$\text{(A.4)} \qquad T_{11}^{-1} = (\Sigma_{11}^{(0)})^{1/2}, \quad T_{22}^{-1} = (\Sigma_{22.1}^{(0)})^{1/2},$$

where $(\Sigma_{11}^{(0)})^{1/2}$ and $(\Sigma_{22.1}^{(0)})^{1/2}$ are any square roots of $\Sigma_{11}^{(0)}$ and $\Sigma_{22.1}^{(0)}$, respectively. [For computational purposes, lower-triangular square roots are recommended.] Also let

(A.5) $$T_{21} = -T_{22}\Sigma_{21}^{(0)}T_{11}$$

and define

(A.6) $$T = \begin{pmatrix} T_{11} & 0 \\ T_{21} & T_{22} \end{pmatrix} : (p+r)x(p+r).$$

Using block matrix multiplication and the definitions of the elements of T, it can be shown that

(A.7) $$T\Sigma_0 T' = I_{p+r}.$$

We can now use T tc transform the old observations $z_i = (z_{1i}', z_{2i}')'$ to new observations $x_i = (x_{1i}', x_{2i}')'$, i = 1,2,...,n:

(A.8) $$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = Tz_i = \begin{pmatrix} T_{11}z_{1i} \\ T_{21}z_{1i}+T_{22}z_{2i} \end{pmatrix} .$$

It follows from (A.2), (A.6), and (A.8) that

(A.9)
$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} T_{11}z_{1i} \\ T_{21}z_{1i}+T_{22}z_{2i} \end{pmatrix}$$
$$= \begin{pmatrix} T_{11}w_i \\ T_{22}\alpha^*+(T_{21}+T_{22}B^*)T_{11}^{-1}T_{11}w_i \end{pmatrix} + Tf_i$$
$$= \begin{pmatrix} u_i \\ \alpha + Bu_i \end{pmatrix} + e_i, \quad i = 1,2,...,n,$$

where

$$u_i = T_{11}w_i, \quad i = 1,2,...,n,$$

(A.10) $$\alpha = T_{22}\alpha^*,$$
$$B = (T_{21}+T_{22}B^*)T_{11}^{-1},$$

and the new error vectors

$$e_i = Tf_i, \quad i = 1,2,\ldots,n,$$

are i.i.d. with a common multivariate normal distribution with mean vector 0 and covariance matrix

$$T(\sigma^2\Sigma_0)T' = \sigma^2 T\Sigma_0 T' = \sigma^2 I_{p+r}.$$

If we now form estimators of α, B, σ^2, and u_i, $i = 1,2,\ldots,n$ based on the transformed data x_i, $i = 1,2,\ldots,n$, then we can use these estimators to construct estimators of the parameters α^*, B*, σ^2, and w_i, $i = 1,2,\ldots,n$, of our original model as follows:

$$\hat{\alpha}^* = T_{22}^{-1}\hat{\alpha} = (\Sigma_{22.1}^{(0)})^{1/2}\hat{\alpha},$$

$$\hat{B}^* = T_{22}^{-1}(\hat{B}T_{11} - T_{21}) = [(\Sigma_{22.1}^{(0)})^{1/2}\hat{B} + \Sigma_{21}^{(0)}](\Sigma_{11}^{(0)})^{-1/2}, \tag{A.11}$$

$$\hat{w}_i = T_{11}^{-1}\hat{u}_i = (\Sigma_{11}^{(0)})^{1/2}\hat{u}_i, \quad i = 1,2,\ldots,n.$$

Of course, $\hat{\sigma}^2$ is the same under both models.