3

# ESTIMATION OF LINEAR MODELS WITH INCOMPLETE DATA

*Paul D. Allison*

UNIVERSITY OF PENNSYLVANIA

*In estimating linear models, it is often necessary to combine data from two or more samples or subsamples, each with a somewhat different set of variables. Conventional methods for doing this are statistically inefficient or have unknown statistical properties. This paper describes a maximum likelihood method that is both consistent and efficient and that can be implemented with LISREL. Unlike other maximum likelihood algorithms, this method can estimate overidentified models, and it produces consistent estimates of standard errors. Using this approach, one can*

For helpful suggestions, I am indebted to Scott Long, Donald Rubin, Robert Hauser, Jeffrey Liker, Peter Mossel, Kenneth Spenner, William Bielby, Arne Kalleberg, Peter Marsden, and several anonymous reviewers. Work on this paper was completed while I was supported by a fellowship from the John Simon Guggenheim Foundation.

*estimate linear regression models, path models, confirmatory factor models, errors-in-variables models, and nonrecursive models with incomplete data.*

In estimating linear models, researchers often need to combine data from two or more samples or subsamples, each with a somewhat different set of variables. Consider the following examples:

1. *Remeasurement studies.* Sometimes additional data is collected for some fraction of the original sample to evaluate the quality of the data collection. For example, Bielby, Hauser, and Featherman (1977*a*, 1977*b*) had data on the status attainment of 25,223 nonblack males in the U.S. To estimate the degree of measurement error, they reinterviewed a random subsample of 578 males to obtain repeated measurements of the variables of interest. After estimating a measurement model for this subsample, they used the resulting reliability estimates to correct for measurement error in the entire sample—an *ad hoc* procedure with unknown statistical properties.

2. *Sibling studies.* A currently popular way to estimate the effects of family background on various outcomes in adulthood is to collect data on siblings (Taubman 1977). The covariation in outcomes among siblings can then be used to estimate models in which family characteristics are treated as latent variables. However, a universal problem in such studies is that some persons have no siblings, and it is not obvious whether and how the data for only-children can be combined with the data for sibling pairs. In studies that include *all* siblings, variation in sibship size further compounds the problem.[1]

3. *Multiple data sources.* It is increasingly common for investigators to construct data sets that combine results from questionnaire surveys with government data files and other public records. In attempting to combine data from different sources, one often finds that a substantial fraction of the records cannot be matched. In such cases, the total sample can be divided into subsamples according to which records are present and which are absent.

4. *Attrition in panel studies.* In multiwave panel studies, there is often substantial attrition from one wave to the next. As a result, there

[1] The same problems occur in the estimation of neighborhood effects (Bielby 1981) when the number of neighbors in the sample differs across neighborhoods.

may be one subsample with data from the first wave only, a second subsample with data from the first two waves only, etc.

5. *General missing data problems*. Examples 3 and 4 are usually treated by conventional methods for handling missing data. More generally, in any missing data problem, the sample can be decomposed into subsamples, each having a distinct set of variables present and absent. In many cases, however, these subsamples will be numerous and small.

In each of these examples, the statistical problem is to combine the data from the multiple subsamples in such a way that the resulting parameter estimates are both consistent (i.e., converge to the true values as the sample gets large) and efficient (i.e., have standard errors that are as small as possible). Of nearly equal importance are good estimates of the standard errors, which enable us to construct valid hypothesis tests and confidence intervals.

Previous approaches to the problem are deficient in several respects. For example, listwise and pairwise deletion are known to be consistent (if the data are missing completely at random), but both are inefficient (Glasser 1964; Afifi and Elashoff 1966). Listwise deletion, in particular, can discard an enormous amount of potentially useful data. Pairwise deletion may be more efficient than listwise deletion in many cases, but for some data structures it is known to be less efficient (Donner and Rosner 1982; Brown 1983). Moreover, the standard error estimates produced by most pairwise deletion algorithms are inconsistent estimates of the true standard errors.

In this paper I describe a maximum likelihood (ML) method that produces direct estimates of the parameters of a large class of linear models using data from subsamples with different sets of observed variables. Under appropriate assumptions, these estimates are consistent, asymptotically efficient, and asymptotically normally distributed. The method also produces consistent estimates of the standard errors of the parameter estimates. In essence, the method treats missing variables as latent variables, possibly without indicators. The model is then estimated simultaneously for all subsamples while appropriate equality constraints are imposed across subsamples.

An attractive feature of this method is that it can be implemented with recent versions of LISREL (Jöreskog and Sörbom 1981, 1983), which is widely available and frequently used in social science research. Although this is currently the only known program that can

implement the method, the recently developed EQS program (Bentler 1983) may eventually have this capability. The method is primarily useful when the number of subsamples with distinct sets of variables present is relatively small and when the number of cases in each subsample is large. Although the technique can be used for general missing data problems with numerous subsamples, it is likely to be tedious and expensive in those applications.

This method is closely related to other recent uses of ML to handle missing data problems, and I will later point out some of those connections. A similar approach was suggested by Werts, Rock, and Grandy (1979), but their method did not yield true ML estimates and applied only to confirmatory factor models. In contrast, the method considered here applies to any linear structural equation model subsumed under the general model proposed by Jöreskog (1977), which includes multiple regression, path analysis, confirmatory factor analysis, seemingly unrelated regressions, and nonrecursive structural equation models. Detailed examples will be presented.

## *ML ESTIMATION WITH INCOMPLETE DATA*

In this section I discuss the requirements for ML estimation of linear models with interval-level variables when the data are incomplete. I also comment on previous uses of ML estimation with incomplete data.

*Models.* The class of models considered here is defined by Jöreskog's (1977) general linear structural relations model. This is specified in part by

$$\eta = \mathbf{B}\eta + \Gamma\xi + \zeta, \tag{1}$$

where $\eta$ is a vector of endogenous variables, $\xi$ is a vector of exogenous variables, $\zeta$ is a vector of unobserved, exogenous disturbances, and $\mathbf{B}$ and $\Gamma$ are matrices of coefficients. $\zeta$ and $\xi$ are assumed to be uncorrelated. For convenience, all variables are assumed to have means of zero, thus eliminating the need for intercept terms in the model. To achieve identification, one must usually impose restrictions on $\mathbf{B}$, $\Gamma$, and var($\zeta$). Important special cases of equation (1) include multivariate regression ($\mathbf{B} = \mathbf{0}$), multiple regression ($\mathbf{B} = \mathbf{0}$, and $\eta$ is a scalar), and recursive systems ($\mathbf{B}$ is subdiagonal and var($\zeta$) is diagonal).

To facilitate its estimation with missing data, this model can be equivalently expressed without the $\xi$'s:

$$\eta^* = \mathbf{B}^*\eta^* + \zeta^*, \tag{2}$$

where $\eta^* = \begin{pmatrix} \eta \\ \xi \end{pmatrix}$, $\zeta^* = \begin{pmatrix} \zeta \\ \xi \end{pmatrix}$, and $\mathbf{B}^* = \begin{pmatrix} \mathbf{B} & \Gamma \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. In essence, this reformulation forces some of the endogenous $\eta$'s to be identically equal to some of the exogenous $\zeta$'s.[2] Since equation (2) is the focus of the remainder of the paper, the asterisks will be dropped for notational simplicity.

While equation (2) (or the equivalent equation [1]) is quite standard in the econometric literature, the class of models is greatly expanded by allowing for the possibility that $\eta$ may not be directly measured. This is accomplished by adding equations that specify the effects of the latent variables on a set of observed indicators:

$$\mathbf{y} = \Lambda_y\eta + \varepsilon. \tag{3}$$

Here, $\mathbf{y}$ is a vector of observed variables, and $\varepsilon$ is a vector representing random measurement error. $\Lambda_y$ is a matrix of coefficients that must usually be restricted in some way to achieve identification. Although $\varepsilon$ is assumed to be uncorrelated with $\eta$ and $\zeta$, correlations are allowed among the elements of $\varepsilon$ so long as the model is still identified. Equation (3) is essentially a factor model, and the elements of $\Lambda_y$ can be interpreted as factor loadings.

If we let $\Sigma$ be the population covariance matrix for $\mathbf{y}$, equations (2) and (3) imply that $\Sigma = \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\Psi(\mathbf{I} - \mathbf{B}')^{-1}\Lambda'_y + \theta_\varepsilon$, where $\Psi = \mathrm{var}(\zeta)$ and $\theta_\varepsilon = \mathrm{var}(\varepsilon)$. The model can also be generalized to allow for multiple populations simply by adding a subscript to each of the parameter matrices.

*Data.* The data are presumed to consist of two or more subsamples, each having measurements on a somewhat different set of variables. One of these subsamples may be complete; that is, it may have observations on all the variables of interest. This is not essential, however. What *is* essential depends, in part, on the identification status of the model to be estimated. Within the class of models that are identified, some are "more identified" than others:

[2] The $\eta$'s that were previously $\xi$'s are still exogenous because they are identically equal to the $\zeta$'s, which are defined to be exogenous.

1. In just-identified models, the number of parameters to be estimated equals the number of population moments (usually variances and covariances). Ordinary multiple regression models, for example, are just-identified. In just-identified models, every possible pairing of variables must occur in at least one of the subsamples, making it possible to estimate the covariance (or correlation) for each pair. In an extreme case, each subsample would have measurements on only two variables; if there were $K$ variables in the model, this would require $K(K-1)/2$ subsamples.

2. In overidentified models, the number of population moments exceeds the number of parameters. Hence, some of the moments are redundant and need not be directly estimated in any of the subsamples. Unfortunately, it is often difficult to determine which moments are redundant; it depends entirely on the model of interest. Confirmatory factor models and simultaneous equation models are commonly overidentified.

For application of LISREL, the number of cases in each subsample must exceed the number of variables measured for that subsample. Otherwise, the resulting correlation or covariance matrix would not be positive definite.

All the conditions described so far are sample properties that can be directly verified. It is also necessary to make some assumptions about how the observations come to be in one or another of the subsamples. Conventional approaches to missing data usually rest on the assumption that the data are, in some sense, missing at random, but there has been much confusion about what this actually means. In a definitive treatment, Rubin (1976) showed that there are actually two different conditions that determine whether or not the mechanism generating the missing data can be ignored.

1. Data are missing at random if the probability of obtaining the particular pattern of missing data found in the sample does not depend on the values of the data that are missing. It may, however, depend on the values of the data that are observed.

2. Data are observed at random if the probability of obtaining the missing data pattern found in the sample does not depend on the data that are observed; however, it may depend on the data that are missing.

To clarify these distinctions, consider the following simple example. Suppose that for $i = 1, \ldots, n$, we have a joint variable $(x_i, y_i)$ that

is independent and identically distributed (across cases). For concreteness, imagine that $x$ is years of schooling and $y$ is income. Now suppose that income ($y$) is not observed for the first $m$ cases but that education ($x$) is observed for all. Let $D_i$ be a dummy variable with a value of 1 if $y_i$ is missing, 0 otherwise, and define $g(x_i, y_i) = \text{pr}(D_i = 0 | x_i, y_i)$. The data are missing at random if $g$ does not depend on the value of $y_i$ for $i = 1, \ldots, m$ when $x_i$ is fixed at its observed value. The data are observed at random if $g$ does not depend on the value of $x_i$ for $i = 1, \ldots, m$ and if $g$ does not depend on the values of $x_i$ or $y_i$ for $i = m + 1, \ldots, n$.

In other words, the data are *not* missing at random if persons with high income are less likely to report their income. On the other hand, suppose that persons with high education are less likely to report their income but that among those with the same years of schooling, income is unrelated to the probability that they will report. Then, the data are missing at random but are not observed at random.

Most missing data techniques rest on the rather strong assumption that the data are both missing at random and observed at random. If both these conditions are satisfied, the data are said to be missing *completely* at random. One of the virtues of the ML method, by contrast, is that it retains its desirable properties when the data are missing at random but *not* observed at random. Thus, even if people with high education are less likely to report their income, it is still appropriate to use ML to estimate the regression of income on education and other explanatory variables. For a more detailed discussion of these points, see Marini, Olsen, and Rubin (1979).

Even the weaker assumption that the data are missing at random is likely to be violated for many of the applications envisioned here. In panel studies, for example, attrition may well depend on values of the variables that would have been observed in later waves. In kinship studies, a couple's decision to have only one child may be based on anticipated undesirable characteristics of later children. As with other statistical assumptions, however, the missing-at-random assumption may be a useful approximation even when it is believed to be false. Although models that do not make this assumption are possible, they must be specially constructed for each application (Little 1982, 1983). Such models are usually complex, difficult to estimate, and untestable with the data at hand (Rubin 1977; Greenlees, Reece, and Zieschang 1982; Heckman 1979).

Finally, the ML method described here rests on the assumption that the data are drawn from a multivariate normal distribution, which implies that the sufficient statistics are the sample means, variances, and covariances. This assumption is fairly common for multivariate problems and will be familiar to users of LISREL. While it is unlikely that this assumption will be exactly satisfied in practice, there are several reasons to believe that violations may not seriously compromise the estimates:

1. Regardless of the true distribution, the proposed estimators are consistent.

2. For multiple regression models, normality is not required for independent variables that have no missing data (Rubin 1974).

3. The same estimators can be justified by arguments that do not require multivariate normality. If attention is restricted to estimators that are functions only of sample means, variances, and covariances, the method considered here yields estimators of the population means, variances, and covariances that are approximately minimum-variance unbiased (Beale and Little 1975; Hocking and Smith 1968).

Thus, even with nonnormal data, ML estimators should have reasonably good properties relative to competing estimators. On the other hand, the standard error estimates may be more sensitive to departures from normality. It should also be noted that the approach described here can be used with estimation methods that rest on somewhat weaker distributional assumptions—e.g., generalized least squares and unweighted least squares methods.

*ML estimation.* We have known for many years that ML has several advantages over other methods of handling missing or incomplete data (Wilks 1932). But that knowledge had little impact on applied data analysis because ML estimation with missing data required enormous computational resources. In recent years, however, improved algorithms and reduced costs of computation have made ML a feasible option.

ML estimation methods have been developed for interval-level data, for categorical data (Fuchs 1982), and for a combination of the two (Little and Schluchter 1985). I consider only interval-level data here. There are presently three widely known computational methods for getting ML estimates of an unrestricted covariance (or correlation) matrix when data are missing or incomplete: factoring the likelihood, the Newton-Raphson algorithm, and the expectation maximization

(EM) algorithm. All three methods produce identical estimates of the covariance matrix, but they differ in cost, ease of implementation, and range of applications.

1. Factoring the likelihood (Anderson 1957; Rubin 1974; Marini et al. 1979) is computationally simple but is only applicable if the subsamples follow a monotone or "nested" pattern. This occurs when the variables with missing data can be arranged in a sequence $x_1, x_2, \ldots, x_n$ such that, for a given respondent and for $i > j$, $x_i$ is missing whenever $x_j$ is missing.

2. Newton-Raphson (Hartley and Hocking 1971; Hocking and Marx 1979) is an iterative algorithm that handles general missing data problems, but it is quite expensive computationally. It also produces standard error estimates for the means, variances, and covariances.

3. The EM algorithm (see Orchard and Woodbury 1972; Dempster, Laird, and Rubin 1977), an iterative method available in the BMDP package (Dixon 1981), handles general missing data problems but is less expensive than the Newton-Raphson procedure. Like the factorization method, however, it does not produce estimates of standard errors.

In their generally available implementations, none of these algorithms is ideally suited to the estimation of linear models. For multiple regression models, which are just-identified, the estimated covariance matrix produced by any of these algorithms could be input to standard regression programs to get true ML estimates of the coefficients. But the resulting standard errors would not be consistent estimates of the true standard errors. (Such estimates can often be obtained by additional, nonstandard calculations, however.) For overidentified models, simply inputting the ML covariance matrix will not yield efficient estimates of the model parameters. Efficient estimation requires that the overidentifying restrictions be incorporated into the ML estimation procedure. Again, it may be possible to modify or extend the algorithms to accomplish this.

## *ML ESTIMATION WITH LISREL*

I now describe how to use LISREL to get ML estimates of linear models with incomplete data. The method capitalizes on the ability of LISREL (versions IV through VI) to estimate simultaneously the same model for two or more samples. Individual parameters can be

either constrained equal across samples or allowed to vary. For incomplete data problems, the sample is divided into subsamples, each having a different (although possibly overlapping) set of variables present. The model is then estimated simultaneously for all subsamples, constraining corresponding parameters to be equal across subsamples. For this to work, however, a number of special techniques are necessary.

Before describing these techniques, I will briefly indicate why the method works and why the special techniques are needed. Suppose that the sample is divided into $G$ subsamples ($g = 1, \ldots, G$) in such a way that each subsample has a distinct set of variables present and missing. Hartley and Hocking (1971) showed that for an unrestricted multivariate normal distribution with data missing at random, the log-likelihood function is given by

$$-\frac{1}{2}\sum_{g=1}^{G} n_g\left[\log|\boldsymbol{\Sigma}_g| + \operatorname{tr}\left(\mathbf{S}_g\boldsymbol{\Sigma}_g^{-1}\right) + \operatorname{tr}\left(\mathbf{H}_g\boldsymbol{\Sigma}_g^{-1}\right) + C_g\right]. \tag{4}$$

For subsample $g$, $n_g$ is the number of cases, $\boldsymbol{\Sigma}_g$ is the true covariance matrix for the variables present, $\mathbf{S}_g$ is the observed covariance matrix for the variables present, $C_g$ is a term that depends on the data but not on the parameters, and $\mathbf{H}_g = (\hat{\boldsymbol{\mu}}_g - \boldsymbol{\mu}_g)(\hat{\boldsymbol{\mu}}_g - \boldsymbol{\mu}_g)'$, where $\boldsymbol{\mu}_g$ is the vector of true means for variables present and $\hat{\boldsymbol{\mu}}_g$ is the corresponding vector of sample means. Compare this with the fitting function that is maximized under the multiple group option in LISREL (Jöreskog and Sörbom 1981):

$$-\frac{1}{2}\sum_{g=1}^{G} n_g\left[\log|\boldsymbol{\Sigma}_g| + \operatorname{tr}\left(\mathbf{S}_g\boldsymbol{\Sigma}_g^{-1}\right) + C_g'\right]. \tag{5}$$

Although expressions (4) and (5) are clearly similar, they are different in three key respects: (a) unlike (4), the LISREL function in (5) does not include a term for the differences between sample and population means; (b) the LISREL function in (5) is defined for situations in which the same set of variables appears in all groups, but (4) allows for a different subset of variables in each group; (c) expression (4) is maximized with respect to the elements of $\boldsymbol{\Sigma}$, but (5) is maximized with respect to the set of parameter matrices that determine $\boldsymbol{\Sigma}$.

These differences can be reconciled by applying some special techniques to both data input and model specification in LISREL. One

set of techniques, documented in the LISREL manual (Jöreskog and Sörbom 1981), allows for the inclusion of structured means in the LISREL model. This is accomplished by (a) specifying an additional "variable" with a constant value of 1.0 and (b) analyzing a matrix of mean sums of squares and cross-products rather than a covariance matrix. A second set of techniques makes it possible to have a different set of observed variables in each of the multiple groups. This is accomplished by (a) inputting pseudo-values for the missing sample moments and (b) specifying fixed values for certain factor loadings and error variances so that these pseudo-values are fitted exactly. Following are detailed instructions for implementing these two techniques. A proof that these techniques do indeed produce an equivalence between expressions (4) and (5) is provided in Appendix A.

*Data input.* To incorporate means in LISREL VI, do the following:

1. In the DATA statement, specify MATRIX = AM. This indicates that the matrix to be analyzed is an "augmented" moment matrix. This is a matrix of moments about zero, whose last row (and column) consists of sample means followed by the element 1.0.
2. Unless raw data are read in, input both the sample covariance matrix and the sample means. Read in the latter using the ME statement.

To incorporate means in LISREL V, do the following:

1. In the DATA statement, specify MATRIX = MM. This indicates that the matrix to be analyzed is a matrix of moments about zero.
2. Set NINPUT equal to one more than the actual number of variables.
3. Add a row (and implicitly a column) of zeros to the sample covariance matrix. This corresponds to an $x$ variable with a constant value 1.0.
4. Read in the sample means using the ME statement. The last mean, corresponding to the additional $x$ variable, should be 1.0.

To allow for missing variables, for each subsample, set any missing covariances to 0.0, missing means to 0.0, and missing variances to 1.0.

*Model specification.* To incorporate means, formulate the model of interest as in equation (2), where all observed variables are $y$'s and all latent variables are $\eta$'s. In addition, include the following specifications on the MODEL statement:

1. Set NY equal to the number of observed variables, and set NETA equal to one more than the number of unobserved variables.
2. Set NX = 1 and specify FIXEDX.

In declaring fixed and free parameters, make the following specifications:

3. In the PSI matrix (the covariance matrix for $\zeta$), fix at 0.0 all the elements of the last row (and column).
4. In the GAMMA matrix (which here is a vector), fix at 0.0 all elements except the last, which should be fixed at 1.0.

The LY matrix will have one more column than usual, the last column consisting of free parameters that correspond to the population means of the observed variables.

The aim of this specification is to force $x = \xi = \eta = 1.0$ and to let this "variable" directly affect all the $y$'s. The model actually being estimated is

$$\begin{aligned} \mathbf{y} &= \Lambda_y^* \boldsymbol{\eta}^* + \boldsymbol{\varepsilon}, \\ \boldsymbol{\eta}^* &= \mathbf{B}^* \boldsymbol{\eta}^* + \Gamma\xi + \boldsymbol{\zeta}^*, \\ x &= \xi, \end{aligned} \tag{6}$$

where

$$\begin{gathered} \boldsymbol{\eta}^* = \begin{pmatrix} \boldsymbol{\eta} \\ 1 \end{pmatrix}, \qquad \xi \equiv 1, \qquad \boldsymbol{\zeta}^* = \begin{pmatrix} \boldsymbol{\zeta} \\ 0 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}, \\ \Lambda_y^* = \begin{pmatrix} & \mu_1 \\ \Lambda_y & \mu_2 \\ & \vdots \end{pmatrix}, \qquad \mathbf{B}^* = \begin{pmatrix} \mathbf{B} & & 0 \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}. \end{gathered} \tag{7}$$

The unstarred parameters in equation (7) are the same as the starred parameters in equation (2).

To allow for missing variables, any variable that is missing from any of the subsamples must be an indicator of a latent variable. If this is not already the case, a new latent variable must be added to the model. The coefficients in the $\Lambda_y$ matrix are then used to "switch" the variable "on" or "off," depending on whether it is present or absent in a particular subsample. Specifically, in subsamples with data missing for a particular variable $y_i$, all elements in row $i$ of $\Lambda_y$ must be fixed at 0.0. One must also fix $\text{var}(\varepsilon_i) = 1.0$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0.0$ for $i \neq j$. These constraints ensure that the pseudo-values of 0.0 and 1.0 in the sample covariance matrix for that subsample will be fitted exactly.

In subsamples with data present for that variable $y_i$, the treatment depends on whether or not the model allows for random error in the measurement of that variable. If the model does not allow for random error, one of the $\lambda_{ij}$ coefficients should be fixed at 1.0 and $\text{var}(\varepsilon_i)$ should be fixed at 0.0. If random error is allowed, no special constraints are needed; the $\lambda$ coefficient and the error variance should be left as free parameters to be estimated. Finally, all other parameters are constrained to be equal across subsamples.

I now consider two examples. Both are deliberately much simpler than typical applications so that the mechanics of the technique are not obscured by the complexity of the illustrations. In each case, the model to be estimated is relatively simple, and the sample is divided into only two subsamples. In the first example, the data are incomplete by design rather than by accident, and they are missing completely at random. In the second example, the data are missing at random but are not observed at random.

## *A CONFIRMATORY FACTOR MODEL*

Suppose the aim is to estimate the correlation between father's occupational status (*FAOC*) and father's educational attainment (*FAED*) for black men in the U.S. Using a sample of 2,020, Bielby et al. (1977*b*) estimated that correlation to be 0.433. They recognized, however, that this correlation may be attenuated by random measurement error. To estimate and possibly correct for this error, they took a random subsample of 348 black males from the original sample of 2,020 and reinterviewed them approximately three weeks later. Consequently, their original sample can be divided into two groups: a small

TABLE 1
Covariance Matrices for Measures of Father's Occupation and Father's Education

| | Father's Occupation | | Father's Education | |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| Complete-data subsample ($N = 348$) | | | | |
| $y_1$ | 180.90 | | | |
| $y_2$ | 126.77 | 217.56 | | |
| $y_3$ | 23.96 | 30.20 | 16.24 | |
| $y_4$ | 22.86 | 30.47 | 14.36 | 15.13 |
| Mean | 16.62 | 17.39 | 6.65 | 6.75 |
| Incomplete-data subsample ($N = 1{,}672$) | | | | |
| $y_1$ | 217.27 | | | |
| $y_2$[a] | 0.0 | 1.0 | | |
| $y_3$ | 25.57 | 0.0 | 16.16 | |
| $y_4$[a] | 0.0 | 0.0 | 0.0 | 1.0 |
| Mean | 16.98 | 0.0 | 6.83 | 0.0 |

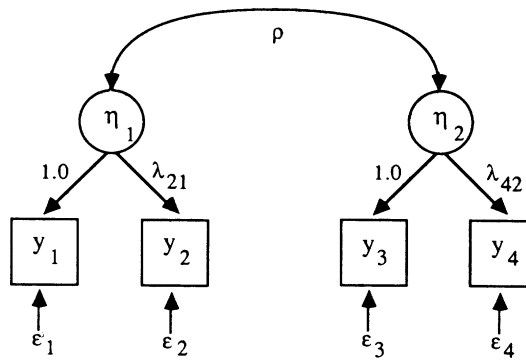
[a] Data for these variables are missing.

subsample of 348 with complete data and a larger subsample of 1,672 with incomplete data. The small subsample had two indicators of *FAOC* (denoted by $y_1$ and $y_2$) and two indicators of *FAED* (denoted by $y_3$ and $y_4$). The large subsample had only $y_1$ and $y_3$. This design virtually guarantees that the missing data are missing completely at random.

Table 1 gives sample variances and covariances for these two groups.[3] For the missing variables in the large subsample, pseudo-values

[3] The covariance matrix for the complete-data subsample was obtained directly from the correlation matrix and standard deviations reported by Bielby et al. (1977*b*). The variances and covariances for the incomplete subsample were more difficult to obtain because what was actually reported was the correlation matrix and standard deviations for the *entire* sample, a combination of the incomplete and complete subsamples. The calculations were performed by (a) converting the reported correlations and standard deviations into sums of squares and cross-products, (b) subtracting the sums for the remeasurement sample from those for the full sample, and (c) using the result to reconstruct the covariance matrix for the incomplete subsample.

FIGURE 1. Path diagram of confirmatory factor model.




of 1.0 have been entered for the variances, and pseudo-values of 0.0 have been entered for the covariances with the other variables. As described above, this substitution is a necessary part of the estimation procedure.

Let us assume (as did Bielby et al.) that the data were generated by the simple confirmatory factor model diagrammed in Figure 1 and represented algebraically by the equations

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \equiv \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{8}$$

and

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1.0 & 0 \\ \lambda_{21} & 0 \\ 0 & 1.0 \\ 0 & \lambda_{42} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}, \tag{9}$$

which are just special cases of equations (2) and (3). The latent factors $\eta_1$ and $\eta_2$ represent the true values of *FAOC* and *FAED*, respectively, and the $\varepsilon$'s represent random measurement error for the four observed variables. The model says that $y_1$ and $y_2$ load exclusively on $\eta_1$ and that $y_3$ and $y_4$ load exclusively on $\eta_2$. The nonzero $\lambda$ coefficients for $y_1$ and $y_3$ are fixed at 1.0 to define metrics for the latent variables (otherwise the model would be underidentified). There are nine parameters to be estimated: the two unconstrained $\lambda$'s, the variances of the

four $\varepsilon$'s, the variances of $\eta_1$ and $\eta_2$, and the covariance of $\eta_1$ and $\eta_2$.[4] Since there are a total of ten population variances and covariances for the four observed variables, the model is overidentified with a single overidentifying restriction: $\sigma_{13}\sigma_{24} = \sigma_{14}\sigma_{23}$, where $\sigma_{ij} = \text{cov}(y_i, y_j)$.

For the subsample with complete data, this model can be readily estimated with LISREL following standard procedures. The estimate of $\text{cov}(\eta_1, \eta_2)$ is 23.31 with an estimated standard error of 3.13. The corresponding correlation is 0.623, which is substantially higher than the uncorrected correlation of 0.433.

The problem with this approach (which is equivalent to listwise deletion) is that it discards the data on $y_1$ and $y_3$ for the large subsample with incomplete data. Incorporating this additional data might substantially increase the precision of the estimates.

Now consider the ML method for incomplete data. To incorporate the means, the model must be reformulated by adding one $\eta$, one $\xi$, and one $x$ variable, all of which are forced to be identically equal to 1.0. Thus we have

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \xi + \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{pmatrix} \tag{10}$$

and

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1.0 & 0 & \lambda_{13} \\ \lambda_{21} & 0 & \lambda_{23} \\ 0 & 1.0 & \lambda_{33} \\ 0 & \lambda_{42} & \lambda_{43} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}, \tag{11}$$

where $\xi = x = \eta_3 = 1.0$ and $\text{var}(\zeta_3) = 0.0$. Note that $\lambda_{i3}$ $(i = 1, \ldots, 4)$ correspond to the population means of the four $y$'s.

LISREL VI control statements for estimating this model are displayed in Appendix B. The data, including the sample means, are read in as described above. For the subsample with no observations on $y_2$ and $y_4$, we set $\lambda_{21}$, $\lambda_{23}$, $\lambda_{42}$, and $\lambda_{43}$ equal to 0.0 and $\text{var}(\varepsilon_2)$ and $\text{var}(\varepsilon_4)$ equal to 1.0. All the free parameters are constrained to be equal across subsamples.

LISREL VI estimates are given in Table 2. The derived ML estimate of the correlation between $\eta_1$ and $\eta_2$ is 0.616, slightly lower

[4] The covariance between the two $\eta$'s is equivalent to the covariance between the two $\zeta$'s, which is what is actually reported in the LISREL runs.

TABLE 2
Parameter Estimates for Confirmatory Factor Model

| Parameter | Estimate | SE | Standardized Estimate |
|---|---|---|---|
| $\lambda_{11}$ | 1.0[a] | — | 0.74 |
| $\lambda_{21}$ | 1.25 | 0.09 | 0.89 |
| $\lambda_{32}$ | 1.0[a] | — | 0.94 |
| $\lambda_{42}$ | 1.003 | 0.04 | 0.97 |
| $\text{cov}(\eta_1, \eta_2)$ | 25.17 | 1.41 | 0.62 |
| $\text{var}(\eta_1)$ | 116.63 | 10.20 | — |
| $\text{var}(\eta_2)$ | 14.29 | 0.70 | — |
| $\text{var}(\varepsilon_1)$ | 94.20 | 8.81 | — |
| $\text{var}(\varepsilon_2)$ | 47.10[b] | 12.45 | — |
| $\text{var}(\varepsilon_3)$ | 1.88 | 0.49 | — |
| $\text{var}(\varepsilon_4)$ | 0.77[b] | 0.48 | — |

[a] The parameters are fixed at this value.
[b] These estimates are for the complete-data subsample only.

than that obtained using the complete-data subsample alone. More importantly, the standard error for the estimate of $\text{cov}(\eta_1, \eta_2)$ is cut in half, from 3.13 to 1.41. Thus, there is a major gain in using all the available data.

LISREL also routinely reports a likelihood ratio $\chi^2$ statistic for evaluating the goodness of fit of the model to the data. For this example, the $\chi^2$ was 7.80 with 15 degrees of freedom (*df*), yielding a *p* value of 0.93. This would ordinarily be considered an excellent fit. Unfortunately, the reported *df* is artifactually high and must be corrected. LISREL calculates the *df* by subtracting the number of estimated parameters from the total number of moments that are read in as data. In this example, there are 13 parameters and a total of 28 elements in the two moment matrices for the two subsamples, yielding the 15 *df*. Nevertheless, 9 of the elements in the moment matrix for the incomplete subsample are pseudo-values of 1.0 and 0.0, which are perfectly fitted as part of the estimation procedure. We must therefore subtract 9 from 15 to get the correct number, 6. A $\chi^2$ of 7.80 with 6 *df* has a *p* value of 0.25, which is still acceptable but not nearly as good as the first impression.

There is still more to be said about this $\chi^2$ statistic. The 6 *df* (as well as the statistic itself) can be decomposed into two parts: 1 *df* pertains to the fit of the model to the complete-data subsample; the

other 5 *df* pertain to the equality constraints across the two subsamples. To get the $\chi^2$ for the single *df*, we can fit the model to the complete-data subsample by itself (which has already been done). The resulting $\chi^2$ is 1.96 with a $p$ value of 0.16. To test the equality constraints, we calculate $\chi^2$ as $7.80 - 1.96 = 5.84$, with 5 *df* and a $p$ value of 0.32. Thus, no matter how one looks at the $\chi^2$, the fit is still acceptable.

What if the $\chi^2$ for the equality constraints had been large and statistically significant? That would suggest that at least some of the parameters were not really the same for the two subsamples or, equivalently, that the data were not really observed at random. There would be no cause for concern, however, because the procedure still produces valid ML estimates when the data are not observed at random as long as they are missing at random. Unfortunately, there is no general test for the missing-at-random assumption because the observed data are always consistent with some missing-at-random model (Rubin 1976).

Before concluding this example, two additional comments are in order. First, the example shows that the model does not have to be identified for every subsample. The confirmatory factor model was identified for the complete-data subsample, but it was grossly underidentified for the incomplete-data subsample. In a more extreme situation, the model may be underidentified in each subgroup and yet identified for all the groups taken together. The equality constraints across the subsamples are what make it possible to estimate the model under such conditions. Second, the estimation procedure applied here is very natural for confirmatory factor models that contain latent variables. When a single latent variable has multiple indicators, these indicators are substitutable for one another, and the loss of any one of them is not crucial to the estimation of the model. Moreover, the technique of switching variables off by fixing the appropriate $\lambda$ parameters to 0.0 is straightforward, since these parameters are already part of the original model. I turn now to an example in which this technique is not so obvious or straightforward.

## *A MULTIPLE REGRESSION MODEL*

In this example, the aim is to estimate a linear regression model in which the dependent variable is years of schooling (*ED*) and the independent variables are father's years of schooling (*FAED*) and father's occupational status (*FAOC*). The purpose of this example is

(a) to illustrate the use of LISREL when none of the variables in the model is a latent variable, (b) to demonstrate the superiority of ML estimation when data are missing at random but not observed at random, and (c) to show how to correct for certain kinds of sample selection bias.

Suppose that the sample is restricted, for reasons that need not concern us, to persons who have at least some college education. That is, $ED > 12$. It is well known that when a sample is truncated by values of the dependent variable, OLS regression may suffer from severe sample selection bias (e.g., Heckman 1979; Berk 1983). Let us also suppose, however, that we have an auxiliary sample in which we observe *ED* (but not *FAED* or *FAOC*) for persons who have 12 or fewer years of education. Thus, we have one sample with complete data and a second sample with data missing on *FAED* and *FAOC*. As we shall see, when the information in these two samples is combined using the LISREL method, the sample selection bias is eliminated. Note that the only information required from the auxiliary sample is the mean and the variance of the dependent variable. This information can often be obtained from published or publicly available sources.

Let us pursue this example with the simulated data in Table 3, which gives means, variances, and covariances for the two samples. By using simulated data, we can control the process generating the missing

TABLE 3
Covariance Matrices for Education, Father's Education and Father's Occupation

| | *ED* | *FAED* | *FAOC* |
|---|---|---|---|
| Complete-data subsample ($N = 492$) | | | |
| *ED* | 3.043 | | |
| *FAED* | 1.987 | 13.07 | |
| *FAOC* | 10.38 | 37.94 | 496.43 |
| Mean | 14.25 | 10.59 | 37.65 |
| Incomplete-data subsample ($N = 508$) | | | |
| *ED* | 3.240 | | |
| *FAED*[a] | 0.0 | 1.0 | |
| *FAOC*[a] | 0.0 | 0.0 | 1.0 |
| Mean | 9.614 | 0.0 | 0.0 |

[a] Data for these variables are missing.

TABLE 4
Coefficient Estimates for Regression of *ED* on *FAED* and *FAOC*

| | *FAED* | | *FAOC* | |
|---|---|---|---|---|
| Estimation Method | Estimate | SE | Estimate | SE |
| OLS (no missing data) | 0.272 | 0.0230 | 0.0283 | 0.00400 |
| OLS (listwise deletion) | 0.106 | 0.0222 | 0.0173 | 0.00386 |
| ML | 0.271 | 0.0328 | 0.0276 | 0.00557 |

data and compare results of missing data methods with results obtained when no data are missing. To get the data in Table 3, a random number generator was used to produce 1,000 multivariate normal observations based on the moment matrix for *ED*, *FAED*, and *FAOC* reported by Bielby et al. (1977*a*). Observations in which $ED > 12$ were assigned to the complete-data subsample ($n = 492$), and observations in which $ED \leq 12$ were assigned to the incomplete subsample ($n = 508$), in which values of *FAED* and *FAOC* were suppressed. (As in the previous example, missing means and covariances are given a value of 0.0 in the table, and missing variances are given a value of 1.0.) These data are missing at random, since the probability that an observation is assigned to the incomplete subsample does not depend on the variables that have missing data—*FAED* and *FAOC*. The data are *not* observed at random, however, because the value of *ED* completely determines whether or not an observation has missing data.

The effect of sample selection bias can be seen in the first two rows of Table 4. The first row, which may be taken as a standard of comparison, gives coefficients and standard errors for OLS applied to the complete data for all 1,000 observations. The second row gives coefficients and standard errors for OLS with listwise deletion of missing data (i.e., the subsample in which $ED \leq 12$ is discarded). The listwise coefficient for *FAED* is less than half the estimate obtained from the complete-data sample, and the listwise coefficient for *FAOC* is about 60 percent of the estimate obtained from the complete sample.

Now consider the ML method using LISREL. Although the model of interest is simpler than in the previous example, it is less straightforward because there are no latent variables. To apply the method, the model must be reformulated by postulating latent vari-

ables corresponding to each of the three observed variables. The observed variables are assumed to be measured without error when no data are missing. This is, in fact, the approach usually taken in LISREL to estimate models in which all variables are directly observed. The difference here is that in the subsample with missing data, the latent variables corresponding to missing variables are treated as true unobservables with no observed indicators. To accomplish this and to incorporate means, the model may be specified as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 & 0 & \lambda_{14} \\ 0 & \lambda_{22} & 0 & \lambda_{24} \\ 0 & 0 & 1 & \lambda_{34} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \tag{12}$$

and

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \xi + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{bmatrix}, \tag{13}$$

where $y_1$ is *FAED*, $y_2$ is *FAOC*, $y_3$ is *ED*, and $x = \xi = 1.0$.

Equation (12), which is a special case of equation (3), says that the observed variables $y_1$, $y_2$, and $y_3$ depend on the latent variables. The model forces $\eta_4$ to be identically equal to 1.0, which implies that $\lambda_{14}$, $\lambda_{24}$, and $\lambda_{34}$ are the population means of $y_1$, $y_2$, and $y_3$. In the complete-data subsample, we set $\lambda_{11} = \lambda_{22} = 1.0$ and $\text{var}(\varepsilon_1) = \text{var}(\varepsilon_2) = \text{var}(\varepsilon_3) = 0.0$. For the subsample with data missing on $y_1$ and $y_2$, we set $\lambda_{11} = \lambda_{22} = \lambda_{14} = \lambda_{24} = 0.0$, $\text{var}(\varepsilon_1) = \text{var}(\varepsilon_2) = 1.0$, and $\text{var}(\varepsilon_3) = 0.0$.

Equation (13), which is a special case of equation (2), is the same for both subsamples. The regression coefficients of interest are $\beta_{31}$ and $\beta_{32}$. Although the $\beta_{21}$ coefficient is not of direct interest, it is included in the model as a convenient way to allow for a nonzero covariance between *FAED* and *FAOC*. In estimating the model, the matrix of $\beta$'s, the covariance matrix for $\zeta$, and the $\lambda_{34}$ coefficients are constrained to be equal across the two groups. We must also fix $\text{var}(\zeta_4) = 0$.

LISREL VI control statements for estimating this model are given in Appendix B. These statements should be relatively straightforward for anyone experienced with LISREL, with one exception. By

default, LISREL V and VI calculate starting values for the iterative algorithm using a least squares and instrumental variables method. This method requires that each latent variable in each subsample have at least one indicator with a fixed, nonzero $\lambda$ coefficient. For the subsample with missing data, however, this condition is not satisfied, since $\eta_1$ and $\eta_2$ have no indicators with nonzero coefficients. Consequently, the automatic starting values must be suppressed (by coding NS on the OUTPUT card), and initial estimates must be supplied. These can be obtained from the listwise-deletion OLS.

ML coefficients and their standard errors are given in the last row of Table 4. The coefficients are very close to those obtained by applying least squares to the original data with no truncation and no missing values.

The $\chi^2$ statistic for this model was 996.37 with 2 *df* (after correction), suggesting correctly that the data are *not* observed at random. Virtually all of this statistic comes from constraining the population mean of $y_3$ for the incomplete subsample to be equal to the population mean of $y_3$ for the complete subsample. Of course, the missing data mechanism ensures that this constraint must be false, because observations are allocated to the complete and incomplete subsamples according to whether they are high or low on $y_3$. This does not mean that the constraint should be relaxed, however. Violation of the observed-at-random condition does not vitiate ML estimation, and failure to impose the constraint yields estimates that are not true ML estimates. In this example, relaxing the constraint gives estimates that are quite close to those obtained with listwise deletion.

It should also be noted that in this example, the missing data follow a monotone pattern, implying that ML estimates can be obtained by factoring the likelihood (Marini et al. 1979). This is easily accomplished by hand calculations performed on the sample moments. The resulting estimates are identical to those obtained with the LISREL method. Nevertheless, as previously observed, this method of obtaining ML estimates does not yield standard errors of those estimates.

## *DISCUSSION*

Although the models and missing data patterns in the two examples were quite simple, extensions to more complicated situations should be straightforward. On the other hand, the LISREL method is

obviously cumbersome when there are many missing data patterns.[5] And if the number of variables is at all large, the number of possible missing data patterns is enormous.[6]

As we have seen, there are many kinds of applications that tend to produce a small number of missing data patterns containing most of the observed cases. Even in those situations, however, there are often a number of minor missing data patterns that each contain only a handful of cases. In practice, several *ad hoc* approaches may be used to eliminate these minor patterns. One is to delete all observations that do not fall into one of the major patterns. A second is to use pairwise deletion or imputation *within* each of the minor patterns to make it conform to one of the major patterns. Finally, there are some situations in which it may be possible to make a minor pattern conform to a major pattern by suppressing some of the observed variables (see Marini et al. 1979 for details). While none of these methods is ideal, any one of them seems preferable to abandoning the ML approach altogether, since that would require the application of *ad hoc* methods to the entire sample.

Another bothersome characteristic of the proposed method is the inclusion of sample means, which requires that the model be reformulated in a cumbersome and confusing way.[7] Unfortunately, the means are essential for getting true ML estimates. Since some authors have proposed similar methods that do not require the means (e.g., Lee 1986), it is worth considering the possible consequences of ignoring them. My own experience in applying the method to several examples suggests that the consequences are slight when the data are missing *completely* at random. Specifically, the estimates and their standard errors differed only slightly when the method was applied without the means.

On the other hand, the information contributed by the means appears to be crucial when the data are missing at random but not

[5] This limitation is not inherent in the proposed method, since it could be overcome by efficient computer programming. The entire process could be automated to the point at which specifying the model could be no more difficult than specifying a LISREL model for a single sample. This could be done either by modifying LISREL itself or by creating a preprocessor to generate the control statements for LISREL.

[6] If there are $k$ variables, the number of possible missing data patterns is $2^k - 1$.

[7] The forthcoming LISREL VII will make it possible to include means in a much more direct, straightforward fashion.

observed at random (as in the second example above). When the means are omitted, the estimates may change drastically. As previously noted, while most conventional missing data techniques presume that the data are missing completely at random, ML estimators are consistent and asymptotically efficient under the weaker assumption that the data are missing at random but not observed at random. Thus, one of the principal advantages of ML estimation may be lost without the inclusion of sample means.

Even the weaker assumption that the data are missing at random but not observed at random will be dubious in most applications, but there are some applications in which one can with total confidence invoke the assumption that the data are missing *completely* at random. These are studies in which the data are incomplete by design, as in the confirmatory factor example. The availability of efficient statistical methods should now make such designs much more attractive. I will briefly mention two possible applications here; a more thorough treatment will appear in a later paper.

1. Suppose the aim is to estimate a linear regression model in which one of the explanatory variables is extremely expensive to measure. Instead of measuring all the variables for all the cases, a more cost-effective approach might be to measure the expensive variable for only a random subsample. Using the LISREL method, one could combine the complete data from this subsample with the incomplete data from the remaining respondents to get consistent estimates of the regression coefficients.

2. Since long questionnaires can lead to respondent fatigue and consequent response errors, one solution might be to assign respondents randomly to different shortened versions of the questionnaire. Then one could use the LISREL method to combine the results from the incomplete subsamples.

A final comment concerns the relationship between missing data and unobservable variables. As we have seen, the way to get LISREL to estimate models with missing data is to treat missing data as unobservable variables. In effect, this reverses an approach that is now popular among statisticians. The EM algorithm, discussed earlier, is a general approach to missing data problems. Early in the development of this algorithm, it was realized that unobservables could be treated as missing data and that a variety of latent variable models could be estimated in this way (Dempster et al. 1977; Rubin and Thayer 1982;

Bentler and Tanaka 1983). The method proposed here demonstrates that the process can be turned around: Methods for estimating latent variable models can be used to estimate models with missing data. This should not be too surprising, since unobservables are merely missing data carried to an extreme.

## *APPENDIX A: PROOF THAT THE MODIFIED LISREL ALGORITHM MAXIMIZES THE LIKELIHOOD WHEN DATA ARE MISSING*

The proof is presented in two parts. In part 1, I show that the proposed techniques for handling missing variables allow expression (5) to be applied when there are different variables measured for each subgroup. In part 2, I show that the proposed techniques for incorporating means into the LISREL algorithm produce an equivalence between expressions (4) and (5).

*Part 1.* Consider a $p \times 1$ random vector $\mathbf{y}$ of potentially observable variables that depend on a set of $q$ latent variables by equation (3). Now suppose that we have $G$ subsamples with observations on $\mathbf{y}$, such that each subsample has a distinct set of variables present and missing. We focus on a single, arbitrary subsample with $r$ variables observed and $s = p - r$ variables not observed. Without loss of generality, let us assume that the observed variables are the first $r$ elements of $\mathbf{y}$, and let $\tilde{\mathbf{y}}$ be an $r \times 1$ vector of those elements. Thus, we can write

$$\tilde{\mathbf{y}} = \tilde{\boldsymbol{\Lambda}}_y \boldsymbol{\eta} + \tilde{\boldsymbol{\varepsilon}},$$

where $\tilde{\boldsymbol{\Lambda}}_y$ is an $r \times q$ matrix consisting of the first $r$ rows of $\boldsymbol{\Lambda}_y$ and $\tilde{\boldsymbol{\varepsilon}}$ is a vector consisting of the first $r$ elements of $\boldsymbol{\varepsilon}$. Define $\tilde{\boldsymbol{\theta}}_\varepsilon = \mathrm{var}(\tilde{\boldsymbol{\varepsilon}})$.

Let $\mathbf{S}$ be the $r \times r$ sample covariance matrix for $\mathbf{y}$, and define the $p \times p$ matrix

$$\mathbf{S}^* = \begin{pmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is an $s \times s$ identity matrix. The techniques described in the main text specify that $\mathbf{S}^*$ is the sample covariance matrix that is read into LISREL for the subsample under consideration. Those techniques also require that two of the parameter matrices take the following

form:

$$\Lambda_y^* = \begin{pmatrix} \tilde{\Lambda}_y \\ 0 \end{pmatrix}, \qquad \theta_\varepsilon^* = \begin{pmatrix} \tilde{\theta}_\varepsilon & 0 \\ 0 & \mathbf{I} \end{pmatrix},$$

where $\Lambda_y^*$ has $p$ rows and $\theta_\varepsilon^*$ is $p \times p$.

We define $\Sigma^* = \Lambda_y^* \Omega \Lambda_y^{*\prime} + \theta_\varepsilon^*$, where $\Omega$ is the $q \times q$ population covariance matrix for $\eta$. We therefore have

$$\Sigma^* = \begin{pmatrix} \tilde{\Lambda}_y \Omega \tilde{\Lambda}_y' + \tilde{\theta}_\varepsilon & 0 \\ 0 & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \Sigma & 0 \\ 0 & \mathbf{I} \end{pmatrix},$$

where $\Sigma$ is the population covariance matrix for the $r$ observed variables.

For a single group, the function that is maximized by LISREL is

$$n\left[\log|\Sigma^*| + \operatorname{tr}(\mathbf{S}^*\Sigma^{*-1}) + C\right].$$

It then suffices to show that $|\Sigma^*| = |\Sigma|$ and $\operatorname{tr}(\mathbf{S}^*\Sigma^{*-1}) = \operatorname{tr}(\mathbf{S}\Sigma^{-1}) + s$. The equality of the determinants is easily shown by repeated expansion by cofactors. The second equality follows from

$$\mathbf{S}^*\Sigma^{*-1} = \begin{pmatrix} \mathbf{S} & 0 \\ 0 & \mathbf{I} \end{pmatrix}\begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{S}\Sigma^{-1} & 0 \\ 0 & \mathbf{I} \end{pmatrix}.$$

Therefore, $\operatorname{tr}(\mathbf{S}^*\Sigma^{*-1}) = \operatorname{tr}(\mathbf{S}\Sigma^{-1}) + \operatorname{tr}(\mathbf{I}) = \operatorname{tr}(\mathbf{S}\Sigma^{-1}) + s$. We have thus shown that maximization of expression (5) with the proposed modifications to $\mathbf{S}$, $\Lambda_y$, and $\theta_\varepsilon$ allows the set of observed variables to be different in each subsample.

*Part 2.* For simplicity, we assume that there is no missing data, but the proof works equally well if there is missing data and the techniques in part 1 have already been applied. If there is no missing data, then expressions (4) and (5) apply to a single sample with no summation involved.

*Notation.* We observe $\mathbf{y}$ distributed as $N_p(\mu, \Sigma)$. Let $\mathbf{S}$ be the sample covariance matrix and $\hat{\mu}$ be the sample mean vector for $\mathbf{y}$. Define $\mathbf{M} = \mathbf{S} + \hat{\mu}\hat{\mu}'$, the matrix of sample moments about zero, and define $\Omega = \Sigma + \mu\mu'$, the matrix of population moments about zero. We also define three augmented moment matrices:

$$\mu^* = [\mu \quad 1.0]', \quad \mathbf{M}^* = \begin{pmatrix} \mathbf{M} & \hat{\mu} \\ \hat{\mu}' & 1.0 \end{pmatrix}, \quad \text{and} \quad \Omega^* = \begin{pmatrix} \Omega & \mu \\ \mu' & 1.0 \end{pmatrix}.$$

$\mathbf{M}^*$ is the matrix that is analyzed by LISREL under the proposed techniques.

*Lemma 1:* The modified model in equations (6) and (7) implies that $\Omega^*$ is the population covariance matrix fitted by *LISREL*.

*Proof:* By equation (1.4) in Jöreskog and Sörbom (1981), the population covariance matrix expressed in terms of the parameters is

$$\Sigma^* = \begin{pmatrix} \Lambda_y^*(\mathbf{I}-\mathbf{B}^*)^{-1}(\Gamma\Phi\Gamma'+\Psi^*)(\mathbf{I}-\mathbf{B}^{*\prime})^{-1}\Lambda_y^{*\prime}+\theta_\varepsilon & \Lambda_y^*(\mathbf{I}-\mathbf{B}^*)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'(\mathbf{I}-\mathbf{B}^{*\prime})^{-1}\Lambda_y^{*\prime} & \Lambda_x\Phi\Lambda_x'+\theta_\delta \end{pmatrix},$$

where $\Psi^* = \mathrm{var}(\zeta^*)$, $\theta_\varepsilon = \mathrm{var}(\varepsilon)$, $\Phi = \mathrm{var}(\xi)$, and $\theta_\delta = \mathrm{var}(\mathbf{x} - \Lambda_x\xi)$. The other matrices are defined in equations (6) and (7). Under the proposed techniques, the FIXEDX specification in LISREL sets $\theta_\delta = 0.0$, $\Lambda_x = 1.0$, and $\Phi = 1.0$. We thus have

$$\Sigma^* = \begin{pmatrix} \Lambda_y^*(\mathbf{I}-\mathbf{B}^*)^{-1}(\Gamma\Gamma'+\Psi^*)(\mathbf{I}-\mathbf{B}^*)^{-1}\Lambda_y^{*\prime}+\theta_\varepsilon & \Lambda_y^*(\mathbf{I}-\mathbf{B}^*)^{-1}\Gamma \\ \Gamma'(\mathbf{I}-\mathbf{B}^{*\prime})^{-1}\Lambda_y^{*\prime} & 1.0 \end{pmatrix}.$$

Next, we note that

$$\Lambda_y^*(\mathbf{I}-\mathbf{B}^*)^{-1}\Gamma = \begin{bmatrix}\Lambda_y & \mu\end{bmatrix}\begin{pmatrix} (\mathbf{I}-\mathbf{B})^{-1} & \mathbf{0} \\ \mathbf{0}' & 1.0 \end{pmatrix}\begin{pmatrix} \mathbf{0} \\ 1.0 \end{pmatrix} = \mu.$$

It follows that

$$\Sigma^* = \begin{pmatrix} \Lambda_y^*(\mathbf{I}-\mathbf{B}^*)^{-1}\Psi^*(\mathbf{I}-\mathbf{B}^{*\prime})^{-1}\Lambda_y^{*\prime}+\theta_\varepsilon+\mu\mu' & \mu \\ \mu' & 1.0 \end{pmatrix}.$$

But since

$$\Psi^* = \begin{pmatrix} \Psi & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix},$$

it follows that

$$\Sigma^* = \begin{pmatrix} \Lambda_y(\mathbf{I}-\mathbf{B})^{-1}\Psi(\mathbf{I}-\mathbf{B}')^{-1}\Lambda_y'+\theta_\varepsilon+\mu\mu' & \mu \\ \mu' & 1.0 \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma+\mu\mu' & \mu \\ \mu' & 1.0 \end{pmatrix} = \Omega^*.$$

*Lemma 2:* $\operatorname{tr}(\mathbf{M}^*\Omega^{*-1}) = \operatorname{tr}(\mathbf{S}\Sigma^{-1}) + \operatorname{tr}[(\hat{\mu}-\mu)(\hat{\mu}-\mu)'\Sigma^{-1}] + 1.$

*Proof:* By a standard inversion formula for partitioned symmetric matrices (Theil 1971, p. 18),

$$\Omega^{*-1} = \begin{pmatrix} (\Omega - \mu\mu')^{-1} & -(\Omega-\mu\mu')^{-1}\mu \\ -\mu'(\Omega-\mu\mu')^{-1} & 1+\mu'(\Omega-\mu\mu')^{-1}\mu \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1}\mu \\ -\mu'\Sigma^{-1} & 1+\mu'\Sigma^{-1}\mu \end{pmatrix}.$$

Then,

$$\mathbf{M}^*\Omega^{*-1} = \begin{pmatrix} \mathbf{M}\Sigma^{-1} - \hat{\mu}\mu'\Sigma^{-1} & -\mathbf{M}\Sigma^{-1}\mu + \hat{\mu}(1+\mu'\Sigma^{-1}\mu) \\ \hat{\mu}'\Sigma^{-1} - \mu'\Sigma^{-1} & -\hat{\mu}'\Sigma^{-1}\mu + 1 + \mu'\Sigma^{-1}\mu \end{pmatrix}.$$

Therefore,

$$\begin{aligned} \operatorname{tr}(\mathbf{M}^*\Omega^{*-1}) &= \operatorname{tr}(\mathbf{M}\Sigma^{-1} - \hat{\mu}\mu'\Sigma^{-1}) + \operatorname{tr}(-\hat{\mu}'\Sigma^{-1}\mu + 1 + \mu'\Sigma^{-1}\mu) \\ &= 1 + \operatorname{tr}\left[(\mathbf{S}+\hat{\mu}\hat{\mu}')\Sigma^{-1} - \hat{\mu}\mu'\Sigma^{-1}\right] \\ &\quad + \operatorname{tr}(\mu\mu'\Sigma^{-1} - \mu\hat{\mu}'\Sigma^{-1}) \\ &= 1 + \operatorname{tr}\left[(\mathbf{S}+\hat{\mu}\hat{\mu}' + \mu\mu' - \hat{\mu}\mu' - \mu\hat{\mu}')\Sigma^{-1}\right] \\ &= 1 + \operatorname{tr}(\mathbf{S}\Sigma^{-1}) + \operatorname{tr}\left[(\hat{\mu}-\mu)(\hat{\mu}-\mu)'\Sigma^{-1}\right]. \end{aligned}$$

*Lemma 3:* $|\Omega^*| = |\Sigma|$.

*Proof:* $\Omega^*$ can be factored as follows:

$$\Omega^* = \begin{pmatrix} \mathbf{I} & \mu \\ 0' & 1.0 \end{pmatrix}\begin{pmatrix} \Sigma & 0 \\ \mu' & 1.0 \end{pmatrix},$$

where both matrices are square and nonsingular. In this situation, the determinant of the product is the product of the determinants; therefore,

$$|\Omega^*| = \begin{vmatrix} \mathbf{I} & \mu \\ 0' & 1.0 \end{vmatrix}\begin{vmatrix} \Sigma & 0 \\ \mu' & 1.0 \end{vmatrix} = |\Sigma|.$$

*Theorem:* Under the techniques described in the main text, the LISREL algorithm maximizes

$$n\left[\log|\Omega^*| + \operatorname{tr}(\mathbf{M}^*\Omega^{*-1}) + C\right]$$
$$= n\left[\log|\Sigma| + \operatorname{tr}(\mathbf{S}\Sigma^{-1})\right.$$
$$\left. + \operatorname{tr}\left((\hat{\mu} - \mu)(\hat{\mu} - \mu)'\Sigma^{-1}\right) + C'\right],$$

which is equivalent to expression (4).

*Proof:* Apply lemmas 1, 2, and 3.

## *APPENDIX B: LISREL VI CONTROL STATEMENTS FOR EXAMPLES*

*Confirmatory factor model*

```
FATHERS SES—COMPLETE DATA
DATA NI=4 NOBS=348 MA=AM NG=2
CM
180.90
126.77 217.56
23.96 30.20 16.24
22.86 30.47 14.36 15.13
ME
16.62 17.39 6.65 6.75
MODEL NY=4 NX=1 NE=3 FIXEDX PS=FI GA=FI
FREE LY 2 1 LY 4 2 LY 1 3 LY 2 3 LY 3 3
FREE LY 4 3 PS 1 1 PS 2 2 PS 1 2
VALUE 1.0 GA 3
ST 0.5 ALL
MA LY
1 0 16
1 0 16
0 1 6
0 1 6
OUTPUT SE TO SS NS
FATHERS SES—INCOMPLETE DATA
DATA NOBS=1672
```

```
CM
217.27
 0.0  1.0
25.57 0.0 16.16
 0.0  0.0  0.0 1.0
ME
16.98 0.0 6.83 0.0
MODEL PS=IN GA=IN LY=FI
FIX TE 2 TE 4
VALUE 1.0 TE 2 TE 4
FREE LY 1 3 LY 3 3
ST 0.5 ALL
MA LY
1 0 16
0 0 0
0 1 6
0 0 0
EQ TE 1 1 1 TE 1 1
EQ TE 1 3 3 TE 3 3
EQ LY 1 1 3 LY 1 3
EQ LY 1 3 3 LY 3 3
OU
```

*Multiple regression model*

```
EDUCATION REGRESSION—COMPLETE DATA
DATA NI=3 NOBS=492 MA=AM NG=2
CM
13.07
37.94 496.4
1.987 10.38 3.043
ME
10.59 37.65 14.25
MODEL NY=3 NX=1 NE=4 FIXEDX TE=FI BE=SD,FR
TD=FI GA=FI PS=DI,FR
FREE LY 1 4 LY 2 4 LY 3 4
FIX PS 4 4
VALUE 1.0 LY 1 1 LY 2 2 LY 3 3 GA 4
ST 11 LY 1 4
ST 40 LY 2 4
```

```
ST 14 LY 3 4
ST 3 BE 2 1
ST 0.3 BE 3 1
ST 0.03 BE 3 2
ST 300 PS 1
ST 7 PS 2
ST 3 PS 3
OU TO NS SE SS
EDUCATION REGRESSION—INCOMPLETE DATA
DATA NOBS=508
CM
1
0 0 1
0 0 0 3.24
ME
0 0 9.614
MODEL GA=IN BE=IN PS=IN LY=FU,FI
FREE LY 3 4
VALUE 1.0 LY 3 3 TE 1 TE 2 2
EQ LY 1 3 4 LY 3 4
OU
```

## *REFERENCES*


AFIFI, A. A., AND R. M. ELASHOFF. 1966. "Missing Observations in Multivariate Statistics. 1. Review of the Literature." *Journal of the American Statistical Association* 61:595–604.

ANDERSON, T.W. 1957. "Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations Are Missing." *Journal of the American Statistical Association* 52:200–203.

BEALE, E. M. L., AND RODERICK J. A. LITTLE. 1975. "Missing Values in Multivariate Analysis." *Journal of the Royal Statistical Society*, ser. B, 37:129–45.

BENTLER, PETER. 1983. "Some Contributions to Efficient Statistics in Structural Models: Specification and Estimation of Moment Structures." *Psychometrika* 48:493–517.

BENTLER, PETER, AND JEFFREY S. TANAKA. 1983. "Problems with EM Algorithms for ML Factor Analysis." *Psychometrika* 48:247–51.

BERK, RICHARD A. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Sociological Review* 48:386–98.

BIELBY, WILLIAM T. 1981. "Neighborhood Effects: A LISREL Model for Clustered Samples." *Sociological Methods and Research* 10:82–111.

BIELBY, WILLIAM T., ROBERT M. HAUSER, AND DAVID L. FEATHERMAN. 1977*a*. "Response Errors of Nonblack Males in Models of the Stratification Process." *Journal of the American Statistical Association* 72:723–35.

_____. 1977*b*. "Response Errors of Black and Nonblack Males in Models of the Intergenerational Transmission of Socioeconomic Status." *American Journal of Sociology* 82:1242–88.

BROWN, C. HENDRICKS. 1983. "Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings." *Psychometrika* 48:269–91.

DEMPSTER, A. P., N. M. LAIRD, AND DONALD B. RUBIN. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistical Society*, ser. B, 39:1–22.

DIXON, W. J., ed. 1981. *BMDP Statistical Software 1981*. Berkeley: University of California Press.

DONNER, ALLAN, AND BERNARD ROSNER. 1982. "Missing Values in Multiple Linear Regression with Two Independent Variables." *Communications in Statistics—Theory and Methods* 11:127–40.

FUCHS, CAMIL. 1982. "Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data." *Journal of the American Statistical Association* 77:270–78.

GLASSER, M. 1964. "Linear Regression Analysis with Missing Observations Among the Independent Variables." *Journal of the American Statistical Association* 59:834–44.

GREENLEES, JOHN S., WILLIAM S. REECE, AND KIMBERLY D. ZIESCHANG. 1982. "Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed." *Journal of the American Statistical Association* 77:251–61.

HARTLEY, H. O., AND R. R. HOCKING. 1971. "The Analysis of Incomplete Data." *Biometrics* 27:783–823.

HECKMAN, JAMES J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 45:153–61.

HOCKING, R. R., AND D. L. MARX. 1979. "Estimation with Incomplete Data: An Improved Computational Method and the Analysis of Nested Data." *Communications in Statistics—Theory and Methods* A8:1155–81.

HOCKING, R. R., AND W. B. SMITH. 1968. "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations." *Journal of the American Statistical Association* 63:159–73.

JÖRESKOG, KARL G. 1977. "Structural Equation Models in the Social Sciences: Specification, Estimation and Testing." Pp. 265–87 in *Applications of Statistics*, edited by P. R. Krishnaiah. Amsterdam: North-Holland.

JÖRESKOG, KARL G., AND DAG SÖRBOM. 1981. *LISREL V: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Chicago: National Educational Resources.

_____. 1983. *LISREL VI: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Mooresville, IN: Scientific Software Inc.

LEE, SIK-YUM. 1986. "Estimation for Structural Equation Models with Missing Data." *Psychometrika* 51:93–99.

LITTLE, RODERICK J. A. 1982. "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 77:237–50.

_____. 1983. "Superpopulation Models for Nonresponse: The Non-ignorable Case." Pp. 383–416 in *Incomplete Data in Sample Surveys*, vol. 2, edited by W. G. Madow, Ingram Olkin, and D. B. Rubin. New York: Academic Press.

LITTLE, RODERICK J. A., AND MARK D. SCHLUCHTER. 1985. "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values." *Biometrika* 72:497–512.

MARINI, MARGARET M., ANTHONY R. OLSEN, AND DONALD B. RUBIN. 1979. "Maximum Likelihood Estimation in Panel Studies with Missing Data." Pp. 314–57 in *Sociological Methodology 1980*, edited by Karl F. Schuessler. San Francisco: Jossey-Bass.

ORCHARD, TERENCE, AND MAX A. WOODBURY. 1972. "A Missing Information Principle: Theory and Applications." *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics* 1:697–715.

RUBIN, DONALD B. 1974. "Characterizing the Estimation of Parameters in Incomplete Data Problems." *Journal of the American Statistical Association* 69:467–74.

_____. 1976. "Inference and Missing Data." *Biometrika* 63:581–92.

_____. 1977. "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72:538–43.

RUBIN, DONALD B., AND DOROTHY T. THAYER. 1982. "EM Algorithms for ML Factor Analysis." *Psychometrika* 47:69–76.

TAUBMAN, PAUL, ed. 1977. *Kinometrics: Determinants of Socioeconomic Success Within and Between Families*. Amsterdam: North-Holland.

THEIL, H. 1971. *Principles of Econometrics*. New York: Wiley.

WERTS, C. E., D. A. ROCK, AND J. GRANDY. 1979. "Confirmatory Factor Analysis Applications: Missing Data Problems and Comparison of Path Models Between Populations." *Multivariate Behavioral Research* 14:199–213.

WILKS, S. S. 1932. "Moments and Distribution of Estimates of Population Parameters from Fragmentary Samples." *Annals of Mathematical Statistics* 3:163–95.