Structural equation models for regression with measurement error: Just enough theory (Draft One)

Jerry Brunner
Department of Statistics, University of Toronto
100 St. George St.
Toronto M5G 3G3, Canada

September 23, 2007

Abstract

This article shows how structural equation modelling methods may be used to carry out a valid regression analysis when independent variables are measured with error. An essential part of the process is to overcome the problem of model identification. A general solution, called the "double measurement design," is described; this involves measuring each independent variable twice. When data are collected according to the double measurement design, model identification is guaranteed, and the data analyst need not struggle with mathematical details.

Keywords: Errors in variables, Measurement error, Regression, Structural equation models.

Introduction

In a survey, suppose that a respondent's annual income is "measured" by simply asking how much he or she earned last year. Will this measurement be completely accurate? Of course not. Some people will lie, some will forget and give a reasonable guess, and still others will suffer from legitimate confusion about what constitutes income. Even physical variables like height, weight and blood pressure are subject to some inexactness of measurement, no matter how skilled the personnel conducting the measurement. In fact, very few of the variables in the typical data set are measured completely without error. An exception might be something like the amount of drug administered in a clinical trial. Here, laboratory procedures guarantee that for all practical purposes, the amount of drug a subject receives is exactly what you think it is. But in general, if a variable is simply measured rather than being experimentally manipulated, there is usually at least a little bit of measurement error.

Random variables that cannot be directly observed are called *latent variables*. The ones we can observe are sometimes called "manifest," but in this paper they will be called "observed" or "observable," which is also a common usage. Upon reflection, it is clear that most of the time, we are interested in relationships among latent variables, but at best our data consist only of their imperfect, observable counterparts. One is reminded of the allegory of the cave in Plato's *Republic*, where human beings are compared to prisoners in a cave, with their heads chained so that they can only look at a wall. Behind them is a fire, which casts flickering shadows on the wall. They cannot observe reality directly; all they can see are the shadows.

In ordinary least-squares regression, the only latent variable is the error term. Measurement error in the dependent variable can perhaps be absorbed into the error term, but there is no provision for measurement error in the independent variables. Unfortunately, when independent variables are measured with error, the results can be disastrous. Estimated regression coefficients are biased even as the sample size approaches infinity, and Type I error rates can be seriously inflated.

This has been known for a long time. The alarm about biased regression coefficients was sounded by Stouffer (1936), and by the seventh edition of *Statistical methods for research workers*, Fisher (1938) was warning scientists about the problem. For a modern and readable discussion of what happens to ordinary least-squares regression when measurement error is ignored, the classic article by Cochran (1968) is an excellent source. Fuller (1987) provides an authoritative treatment of regression models that incorporate measurement error; also see Cheng and Van Ness (1999). And the classical structural equation models (for example Goldberger and Duncan, 1973; Jöreskog, 1978; McArdale, 1980; McDonald, 1978; Bentler and Weeks, 1980; Bollen, 1989) include regression with and without measurement error as special cases.

Nevertheless, few regression texts outside Econometrics provide guidance about what to do when the independent variables are measured with error. The present article attempts to fill this gap. It uses language and notation associated with the LISREL structural equation model (Jöreskog, 1978; Bollen, 1989) rather than the arguably more sophisticated approach of Fuller (1987), in order to be accessible to advanced undergraduates in Statistics. Another advantage of the structural equation modelling approach is that high-quality commercial software is available. SAS `proc calis` (SAS Institute, 1999) is available in many academic environments, and LISREL (Jöreskog and Sörbom, 1996) and AMOS (Arbuckle, 2006) are excellent programs with free student versions. There is also a structural equation modelling package for R (Fox, 2006).

Here is the plan of the paper. Section 1 presents almost the simplest possible regression model with measurement error. There is one independent variable, no intercept, additive measurement error, and everything is normally distributed. We will see that even in this case, the model parameters cannot be successfully estimated from the data. The problem is *model identification*. When a statistical model is not identified, it is impossible to recover the parameters even from an infinite amount of data.

Section 2 discusses model identification, and arrives at a well-known principle that

applies to all structural equation models, including models of regression with measurement error. The principle is this. The mean and covariance matrix of the observable variables are always functions of the model parameters. If the model parameters are also functions of the mean and covariance matrix, then those parameters are identified.

Section 3 describes a general solution of the identification problem for regression with measurement error: the *double measurement design*. This consists of measuring all the independent variables twice, preferably on two different occasions, with different measurement procedures. If this can be done in such a way that the errors of measurement on the two occasions are independent, then model identification is taken care of automatically, and the analysis can proceed in a routine manner.

The double measurement design is similar to the idea of "tau-equivalent measures" (for example Bollen, 1987, p. 208, or cite Lord and Novick?), except that all measurement errors need not be independent. In fact, a very desirable feature of the double measurement design is that while errors of measurement from different measurement procedures must be independent, errors of measurement from the same measurement procedure are allowed to be correlated. For example, one should always expect correlated measurement errors for self-report data; these would arise from consistent individual differences in style of responding to questionnaires and in desire to make a favorable impression. And when measurement errors are correlated, adopting a model where they are uncorrelated can have effects that are just as bad as ignoring measurement error altogether.

The double measurement model of Section 3 employs the classical Structural Equation Modelling trick of "centering" all the variables by subtracting off the means, and then conducting the analysis under the assumption that all expected values are zero. In Section 4, the model is expanded to include intercepts. But in most cases this just makes the model parameters harder to identify, and does not providing any additional information about the relationship between the independent and dependent variables. The final conclusion is that most of the time, including intercepts is not worth the extra trouble.

In the development of this theory, assuming multivariate normality simplifies the exposition but is not really necessary. In Section 5, the normal assumption is relaxed. For independent variables that are measured without error (for example, the dummy variables for factors that are experimentally manipulated), the distribution does not matter at all. For independent variables that are measured with error, the double measurement design guarantees identification of a necessary *function* of the parameters of a distribution-free model.

The double measurement design also points to estimators of the regression coefficient that are consistent and asymptotically normal, by a straightforward application of the Central Limit Theorem. This would provide the basis for a full set of large-sample tests and confidence intervals, but it is unnecessary to go there. In fact, the estimators and tests based on a multivariate normal assumption enjoy robustness properties that make them superior to one method (the weighted least-squares approach of Browne, 1984) that was specifically designed to avoid the assumption of normality. They are probably also superior to the methods suggested by the double measurement design, which are very

similar to Browne's.

The moral of the story is comforting in its simplicity. For data that are collected according to the double measurement recipe, just fit a classical structural model with no intercepts and everything normally distributed; this is close to the default settings of most available software.

1 Regression through the origin with one independent variable

Even in the simplest case, when we try to incorporate measurement error into a regression model, we immediately encounter a technical difficulty: model identification. In a simple regression, suppose the dependent variable is related to an independent variable. We can observe the dependent variable, but not the actual value of the independent variable. All we can see is the independent variable plus a piece of random noise.

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \gamma\xi_i + \zeta_i \\ X_i &= \xi_i + \delta_i, \end{aligned} \tag{1}$$

where ξ_i, ζ_i and δ_i and are independent normal random variables with expected value zero, $Var(\xi_i) = \phi$, $Var(\zeta_i) = \psi$, and $Var(\delta_i) = \theta_\delta$. The regression coefficient γ is a fixed constant. The notation here is taken from the LISREL structural equation model (Jöreskog, 1978; Bollen, 1989) for compatibility with later parts of this paper, and because familiarity with this notation will make it easier for students to use structural equation modelling software.

Data from Model (1) are just the pairs (X_i, Y_i) for $i = 1, \ldots, n$. The true independent variable ξ_i is a latent variable whose value cannot be known exactly. The model implies that the (X_i, Y_i) are independent bivariate normal with mean zero and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \phi + \theta_\delta & \gamma\phi \\ \gamma\phi & \gamma^2\phi + \psi \end{bmatrix}. \tag{2}$$

A multivariate normal distribution with mean zero is completely characterized by its covariance matrix, so even an infinite amount of data can only tell us the three unique values in the matrix $\mathbf{\Sigma}$. But there are four parameters in the model: γ, ϕ, ψ and θ_δ. Recovering all four parameters from the unique elements of $\mathbf{\Sigma}$ amounts to solving three equations in four unknowns — an impossibility. Maximum likelihood estimation will fail, with a non-unique maximum at an infinite number of points along a curve in four dimensions.

The problem is that Model (1) is not uniquely identified in the model parameters. The concept of model identification is unfamiliar to most students, because typically (except in the case of exploratory factor analysis) we present them with statistical models that

are nicely identified, and the issue does not arise. Thus, a general discussion of model identification may be helpful.

2 Model identification

Suppose we have a vector of observable data $\mathbf{D} = (D_1, \ldots, D_n)$, and a statistical model (a set of assertions implying a probability distribution) for $\mathbf{D}$. The model depends on a parameter θ, which is usually a vector. If the probability distribution of $\mathbf{D}$ corresponds uniquely to θ, then we say that the model is *identified*. But if any two different parameter values yield the same probability distribution, then the model is not identified. In this case, the data cannot be used to decide between the two parameter values, and standard methods of parameter estimation will fail.

In Model (1), $\theta = (\gamma, \phi, \psi, \theta_\delta)$, $D_i = (X_i, Y_i)$, and the probability distribution of $\mathbf{D}$ is completely determined by $\mathbf{\Sigma}$. The two variances and one covariance in $\mathbf{\Sigma}$ cannot correspond uniquely to the four elements of θ, so the model is not identified. To really nail it down, the two distinct parameter values $\theta_1 = (2, 4, 9, 1)$ and $\theta_2 = (2, \frac{8}{3}, 1, \frac{7}{3})$ both yield

$$\mathbf{\Sigma} = \begin{bmatrix} 5 & 8 \\ 8 & 25 \end{bmatrix}.$$

The clearest way to prove a model is non-identified is with a simple numerical example like this, but frequently other arguments are more convenient.

When a model is not identified, consistent estimation for all the points in the parameter space is an impossibility. Recall that an estimator is said to be *consistent* if, for any arbitrarily small neighborhood of the true parameter value, the probability of the estimator being in that neighborhood approaches one as a limit, as the sample size tends to infinity. Consistency is about the least one can ask of an estimator — basically that for a large enough sample, it will probably be close to the right answer.

So, let $\widehat{\theta}$ be an estimator of the parameter θ, and suppose the model is not identified. Then there exist two different parameters values θ_1 and θ_2 that generate exactly the same distribution of the sample data, and hence of $\widehat{\theta}$. Take neighborhoods around θ_1 and θ_2 small enough so they do not overlap. Suppose that the estimator $\widehat{\theta}$ is consistent, regardless of whether θ equals θ_1 or θ_2. Since the probability distribution of $\widehat{\theta}$ is identical for both parameters, it must become concentrated in *both* the neighborhood around θ_1 and the neighborhood around θ_2. This cannot be, since the neighborhoods are disjoint. Hence, the supposition that "the estimator $\widehat{\theta}$ is consistent, regardless of whether θ equals θ_1 or θ_2" has to be wrong; this is what we wanted to show.

In practical terms, if two parameter values yield the same probability distribution of the data, then the data cannot be used to distinguish between them. In the typical case of a non-identified model, infinitely many parameter values yield the same distribution, for each point in the parameter space. The parameter has a true value, but you cannot know it, even with an infinite amount of data.

It is possible for certain *functions* of the parameter vector to be identified, even when the entire model is not. If full knowledge of the probability distribution of $\mathbf{D}$ implies knowledge of some function of θ, then that function is said to be identified, and consistent estimation of it is a possibility. For example, let $D_1, \ldots, D_n$ be i.i.d. Poisson random variables with mean $\lambda_1 + \lambda_2$, where $\lambda_1 > 0$ and $\lambda_1 > 0$. The parameter is the pair $\theta = (\lambda_1, \lambda_2)$. The model is not identified because any pair of λ values satisfying $\lambda_1 + \lambda_2 = c$ will produce exactly the same probability distribution. Notice also how maximum likelihood estimation will fail in this case; the likelihood function will have a ridge, a non-unique maximum along the line $\lambda_1 + \lambda_2 = \overline{D}$, where $\overline{D}$ is the sample mean. The function $\lambda = \lambda_1 + \lambda_2$, of course, is identified.

The Normal distribution Suppose we have a random sample $X_1, \ldots, X_n$ from a normal distribution with parameters μ and σ^2. Of course this model must be identified because we use it all the time, but how can one show it? Think of the cumulative distribution function of X_1 not as a formula involving μ and σ^2, but as a curve, a set of (x, y) points. If we can produce the values of μ and σ^2 as functions of the curve, then the model will be identified, because function values are unique (this is the difference between a function and a relation). But an integral with respect to a distribution is a function of that distribution, so when we compute $E(X_1)$ and $E(X_1^2)$ and then solve for μ and σ^2, we have proved identification. Identification of all the common probability models (including the multivariate normal) follows in this way.

Back to regression with measurement error Classical structural equation models, including models for regression with measurement error, are based on systems of simultaneous linear equations. Assuming simple random sampling from a large population, the observable data are independent and identically distributed, with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ that may be written as functions of the model parameters in a straightforward way. If it is possible to solve uniquely for a given model parameter in terms of the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then that parameter is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which in turn are functions of the probability distribution of the data. A function of a function is a function, and so the parameter is a function of the probability distribution of the data. Hence, it is identified.

To summarize, we have arrived a simple way to check model identification for any linear simultaneous equation model, not just measurement error regression. *First, calculate the expected value and covariance matrix of the observable data, as a function of the model parameters. If it is possible to solve uniquely for the model parameters in terms of the means, variances and covariances of the observable data, then the model parameters are identified.* If all the random vectors in the model are multivariate normal, this condition is necessary as well as sufficient.

Example: Instrumental variables In a model like (1), suppose that we have access to data for another two variables that depend on the latent independent variable ξ. Our main interest is still in Y; the other two are are called *instrumental* variables because they are just tools for obtaining an identified model.

Here is the expanded version of Model (1). The original dependent variable Y is now called Y_1. Following the usual convention in structural equation modelling, the subscript i has been omitted to reduce notational clutter. The model is presented for a single observation, and implicitly everything is independent and identically distributed, for $i = 1, \ldots, n$.

$$\begin{aligned} Y_1 &= \gamma_1\xi + \zeta_1 \\ Y_2 &= \gamma_2\xi + \zeta_2 \\ Y_3 &= \gamma_3\xi + \zeta_3 \\ X &= \xi + \delta, \end{aligned} \tag{3}$$

where δ, ξ, ζ_1, ζ_2 and ζ_3 are all independent, $Var(\xi) = \phi$, $Var(\zeta_1) = \psi_1$, $Var(\zeta_2) = \psi_2$, $Var(\zeta_3) = \psi_3$, $Var(\delta) = \theta_\delta$, all expected values are zero, and the regression coefficients γ_1, γ_2 and γ_3 are fixed constants.

Writing the vector of observable data (for subject i) as $\mathbf{D} = (X, Y_1, Y_2, Y_3)'$, elements of the covariance matrix $\mathbf{\Sigma}$ may be obtained by elementary one-variable calculations, like $Var(X) = Var(\xi + \delta) = Var(\xi) + Var(\delta) = \phi + \theta_\delta$, and

$$\begin{aligned} Cov(X, Y_1) &= E(X, Y_1) = E([\xi + \delta][\gamma_1\xi + \zeta_1]) = E(\gamma_1\xi^2 + \xi\zeta_1 + \gamma_1\delta\xi + \delta\zeta_1) \\ &= \gamma_1 E(\xi^2) + E(\xi\zeta_1) + \gamma_1 E(\delta\xi) + E(\delta\zeta_1)) \\ &= \gamma_1 Var(\xi) + E(\xi)E(\zeta_1) + \gamma_1 E(\delta)E(\xi) + E(\delta)E(\zeta_1) \\ &= \gamma_1\phi \end{aligned}$$

In this way, we obtain

$$\mathbf{\Sigma} = \begin{bmatrix} \phi + \theta_\delta & \gamma_1\phi & \gamma_2\phi & \gamma_3\phi \\ & \gamma_1^2\phi + \psi_1 & \gamma_1\gamma_2\phi & \gamma_1\gamma_3\phi \\ & & \gamma_2^2\phi + \psi_2 & \gamma_2\gamma_3\phi \\ & & & \gamma_3^2\phi + \psi_3 \end{bmatrix}. \tag{4}$$

To prove model identification, we need to solve for the model parameters in terms of $\mathbf{\Sigma}$. Denote the i, j element of $\mathbf{\Sigma}$ by σ_{ij}. The task is to solve the following ten equations in eight unknowns

$$\begin{aligned} \sigma_{11} &= \phi + \theta_\delta \\ \sigma_{12} &= \gamma_1\phi \\ \sigma_{13} &= \gamma_2\phi \\ \sigma_{14} &= \gamma_3\phi \end{aligned} \tag{5}$$

$$
\begin{aligned}
\sigma_{22} &= \gamma_1^2\phi + \psi_1 \\
\sigma_{23} &= \gamma_1\gamma_2\phi \\
\sigma_{24} &= \gamma_1\gamma_3\phi \\
\sigma_{33} &= \gamma_2^2\phi + \psi_2 \\
\sigma_{34} &= \gamma_2\gamma_3\phi \\
\sigma_{44} &= \gamma_3^2\phi + \psi_3
\end{aligned}
$$

for ϕ, θ_δ, γ_1, γ_2, γ_3, ψ_1, ψ_2, and ψ_3.

The fact that there are more equations than unknowns does not guarantee the existence of a unique solution; it merely tells us that a unique solution is possible. Suppose that γ_2 and γ_3 are both non-zero. This is reasonable, because to be useful, the instrumental dependent variables must have some relationship to the independent variable. In this case,

$$
\frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} = \frac{\gamma_2\gamma_3\phi^2}{\gamma_2\gamma_3\phi} = \phi. \tag{6}
$$

Then, simple substitutions allow us to solve for the rest of the parameters, yielding the complete solution

$$
\begin{aligned}
\phi &= \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} \\
\theta_\delta &= \sigma_{11} - \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} \\
\gamma_1 &= \frac{\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{14}} \\
\gamma_2 &= \frac{\sigma_{34}}{\sigma_{14}} \\
\gamma_3 &= \frac{\sigma_{34}}{\sigma_{13}} \\
\psi_1 &= \sigma_{22} - \frac{\sigma_{12}^2\sigma_{34}}{\sigma_{13}\sigma_{14}} \\
\psi_2 &= \sigma_{33} - \frac{\sigma_{13}\sigma_{34}}{\sigma_{14}} \\
\psi_3 &= \sigma_{44} - \frac{\sigma_{14}\sigma_{34}}{\sigma_{13}}
\end{aligned} \tag{7}
$$

This proves model identification. The solution is thorough but somewhat tedious, even for this simple example. The student may wonder how much work really needs to be shown. I would suggest showing the calculations leading to the covariance matrix (4), saying "Denote the i, j element of $\mathbf{\Sigma}$ by σ_{ij}," skipping the system of equations (5) because

they are present in (4), and showing the solution for ϕ in (6), *including* the stipulation that γ_2 and γ_3 are both non-zero. Then, instead of the explicit solution (7), write something like

$$\begin{aligned}
\theta_\delta &= \sigma_{11} - \phi \\
\gamma_1 &= \frac{\sigma_{12}}{\phi} \\
\gamma_2 &= \frac{\sigma_{13}}{\phi} \\
\gamma_3 &= \frac{\sigma_{14}}{\phi} \\
\psi_1 &= \sigma_{22} - \gamma_1^2\phi \\
\psi_2 &= \sigma_{33} - \gamma_2^2\phi \\
\psi_3 &= \sigma_{44} - \gamma_3^2\phi
\end{aligned} \tag{8}$$

Notice how once I have solved for a model parameter, I use it to solve for other parameters without explicitly substituting in terms of σ_{ij}. The objective is to prove that a unique solution exists by showing how to get it. An exact statement of the solution is not necessary.

Two additional comments are in order. First, this model had no intercepts, and the random variables all had expected value zero. This is typical of the classical structural equation models, in which inference is based solely on the sample covariance matrix and not the means. One speaks of "centering" all the variables by subtracting off the sample means (for example Bollen, 1989). For large samples, this is almost the same as subtracting off the population means. Since all the confidence intervals and tests are based on large-sample theory anyway, no harm is done. Later, we shall consider models with intercepts.

A second comment is that even for the most complex models, proving model identification as in the preceding example involves only elementary mathematics. But it can be long and messy, especially for models with lots of independent variables — and almost all real-life regressions have lots of independent variables. Furthermore, for a given data set, it is not always possible to come up with a realistic model that is identified. A sensible alternative is to plan the statistical analysis in advance, and to ensure model identification by collecting the right kind of data. The next section describes a way to do this. The key is to measure the independent variables twice, preferably using different methods or measuring instruments.

3 The double measurement design

For regression with measurement error, the model identification problem is solved if we measure all the independent variables on more than one occasion, in such a way that

errors of measurement on different occasions are independent. We begin with a classical structural equation model in which all random variables have expected value zero and there no intercepts. In Section 4, the model is extended to include intercepts and non-zero expected values, but ultimately Model (9) below is recommended for most purposes.

For each of n independent observations, assume the following simultaneous equation model. Implicitly, all the random quantities involved have a subscript i, $i = 1, \ldots, n$.

$$\begin{aligned} \mathbf{X}_1 &= \boldsymbol{\xi} + \boldsymbol{\delta}_1 \\ \mathbf{X}_2 &= \boldsymbol{\xi} + \boldsymbol{\delta}_2, \\ \mathbf{Y} &= \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \end{aligned} \tag{9}$$

where

$\mathbf{Y}$ is an $m \times 1$ random vector of observable dependent variables, so the regression can be multivariate.

$\boldsymbol{\Gamma}$ is an $m \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each dependent variable and one column for each independent variable.

$\boldsymbol{\xi}$ is a $p \times 1$ random vector of latent independent variables, with expected value zero and variance-covariance matrix $\boldsymbol{\Phi}$, an $m \times m$ symmetric and positive definite matrix of unknown constants.

$\boldsymbol{\zeta}$ is the error term of the latent regression. It is an $m \times 1$ random vector with expected value zero and variance-covariance matrix $\boldsymbol{\Psi}$, an $m \times m$ symmetric and positive definite matrix of unknown constants.

$\mathbf{X}_1$ and $\mathbf{X}_2$ are $p \times 1$ observable random vectors, each representing $\boldsymbol{\xi}$ plus a different piece of random error.

$\boldsymbol{\delta}_1$ is the measurement error in $\mathbf{X}_1$. It is a $p \times 1$ random vector of error terms, with expected value zero and variance-covariance matrix $\boldsymbol{\Theta}_1$, a $p \times p$ symmetric and positive definite matrix of unknown constants.

$\boldsymbol{\delta}_2$ is the measurement error in $\mathbf{X}_2$. It is a $p \times 1$ random vector of error terms, with expected value zero and variance-covariance matrix $\boldsymbol{\Theta}_2$, a $p \times p$ symmetric and positive definite matrix of unknown constants.

$\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are all independent.

Notice that in this model, measurement errors in the independent variables can be correlated in one sense, but not in another. Because the variance-covariance matrices of the error terms ($\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$) need not be diagonal, the model allows, for example, farmers

who overestimate their number of pigs to also overestimate their number of cows. On the other hand, if one thinks of $\mathbf{X}_1$ and $\mathbf{X}_2$ as measurements of the independent variables by two different methods, then the errors of measurement by different methods must *not* be correlated. For example, if the number of pigs were counted once by the farm manager at feeding time (an element of $\mathbf{X}_1$) and on another occasion by a research assistant from an areal photograph (the corresponding element of $\mathbf{X}_2$), then the requirement of uncorrelated measurement errors would surely be satisfied.

To emphasize an important practical point, the matrices $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ must be of the same size, but none of their corresponding elements need be equal. This means that if measurements of the independent variables are obtained by two different methods, the methods need not be equally precise.

Proof of model identification The following proof illustrates how model identification is established for structural equation models in general. Collecting $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Y}$ into a single long data vector $\mathbf{D}$, we write its variance-covariance matrix as a partitioned matrix:

$$\mathbf{\Sigma} = \left[\begin{array}{c|c|c} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{13} \\ \hline \mathbf{\Sigma}_{12}' & \mathbf{\Sigma}_{22} & \mathbf{\Sigma}_{23} \\ \hline \mathbf{\Sigma}_{13}' & \mathbf{\Sigma}_{23}' & \mathbf{\Sigma}_{33} \end{array}\right], \tag{10}$$

where the covariance matrix of $\mathbf{X}_1$ is $\mathbf{\Sigma}_{11}$, the covariance matrix of $\mathbf{X}_2$ is $\mathbf{\Sigma}_{22}$, the matrix of covariances between $\mathbf{X}_1$ and $\mathbf{Y}$ is $\mathbf{\Sigma}_{13}$, and so on.

The parameters of the model consist of the non-redundant elements of the matrices $\mathbf{\Gamma}$, $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$. Assuming multivariate normality, the probability distribution of the observable random variables corresponds uniquely to $\mathbf{\Sigma}$. Thus, to prove model identification, we need to show we can express the model parameters in terms of the $\mathbf{\Sigma}_{ij}$ quantities. First, we use Model (9) to write the $\mathbf{\Sigma}_{ij}$ matrices in terms of the parameter matrices.

$$\begin{aligned} \mathbf{\Sigma}_{11} &= \mathbf{\Phi} + \mathbf{\Theta}_1 \\ \mathbf{\Sigma}_{12} &= \mathbf{\Phi} \\ \mathbf{\Sigma}_{13} &= \mathbf{\Phi\Gamma}' \\ \mathbf{\Sigma}_{22} &= \mathbf{\Phi} + \mathbf{\Theta}_2 \\ \mathbf{\Sigma}_{23} &= \mathbf{\Phi\Gamma}' \\ \mathbf{\Sigma}_{33} &= \mathbf{\Gamma\Phi\Gamma}' + \mathbf{\Psi} \end{aligned} \tag{11}$$

This system of matrix equations is readily solved for the parameter matrices to yield

$$\begin{aligned} \mathbf{\Phi} &= \mathbf{\Sigma}_{12} \\ \mathbf{\Theta}_1 &= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12} \\ \mathbf{\Theta}_2 &= \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12} \end{aligned} \tag{12}$$

$$\begin{aligned}\boldsymbol{\Gamma} &= \boldsymbol{\Sigma}'_{13}\boldsymbol{\Sigma}_{12}^{-1} = \boldsymbol{\Sigma}'_{23}\boldsymbol{\Sigma}_{12}^{-1} \\ \boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{33} - \boldsymbol{\Sigma}'_{13}\boldsymbol{\Sigma}_{12}^{-1}\boldsymbol{\Sigma}_{13}.\end{aligned}$$

This shows that Model (9) is identified, so that if data are collected following the test-retest recipe, then the data analyst may proceed without giving further thought to model identification.

4 Intercepts

We now expand Model (9) to include intercepts and non-zero expected values. However, we will see that this leads to complications that are seldom worth the trouble, and the classical models with zero expected value and no intercepts are usually preferable. Let

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ \mathbf{X}_1 &= \boldsymbol{\nu_1} + \boldsymbol{\xi} + \boldsymbol{\delta}_1 \\ \mathbf{X}_2 &= \boldsymbol{\nu_2} + \boldsymbol{\xi} + \boldsymbol{\delta}_2,\end{aligned} \tag{13}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\nu_1}$ and $\boldsymbol{\nu_2}$ are vectors of constants, and $E(\boldsymbol{\xi}) = \boldsymbol{\kappa}$. Everything else is as in Model (9).

Again, the observable data $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Y}$ are collected into a data vector $\mathbf{D}$, with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a function of the probability distribution of $\mathbf{D}$. If the parameter matrices of Model (13) are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then they are also functions of the distribution of $\mathbf{D}$, and they will be identified.

Since the addition of constants has no effect on variances or covariances, the contents of $\boldsymbol{\Sigma}$ are given by (10) and (11), as before. The expected value $\boldsymbol{\mu}$ is the partitioned vector

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu_1} \\ \hline \boldsymbol{\mu_2} \\ \hline \boldsymbol{\mu_3} \end{array}\right] = \left[\begin{array}{c} E(\mathbf{X}_1) \\ \hline E(\mathbf{X}_2) \\ \hline E(\mathbf{Y}) \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\nu_1} + \boldsymbol{\kappa} \\ \hline \boldsymbol{\nu_2} + \boldsymbol{\kappa} \\ \hline \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa} \end{array}\right]. \tag{14}$$

To demonstrate the identification of Model (13), one would need to solve the equations in (14) uniquely for $\boldsymbol{\nu_1}$, $\boldsymbol{\nu_2}$, $\boldsymbol{\kappa}$ and $\boldsymbol{\alpha}$. Even with $\boldsymbol{\Gamma}$ considered known and fixed because it is identified in (12), this is impossible, because there are still more unknowns than equations.

If either $\boldsymbol{\nu_1}$ or $\boldsymbol{\nu_2}$ can be assumed zero (or if $\boldsymbol{\kappa} = \mathbf{0}$) then the system can be solved uniquely and the model is identified, but we doubt that such an assumption could be justified very often in practice. Most of the time, all we can do is identify the parameter matrices that appear in the covariance matrix, and also the *functions* $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ and $\boldsymbol{\mu_3}$ of the parameter vector. This can be viewed as a re-parameterization of the model.

It is instructive to see how this works in the multivariate normal case, where the parameters would be estimated by maximum likelihood. For $i = 1, \ldots, n$, we collect the

observed data $\mathbf{x}_{i,1}$, $\mathbf{x}_{i,2}$ and $\mathbf{y}_i$ into a vector $\mathbf{d}_i$, of length $m+2p$. We then write -2 times the log likelihood as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simplifying, we obtain

$$\begin{aligned} -2\log L(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= n[\,(m+2p)(\log|\boldsymbol{\Sigma}|+\log 2\pi)+tr(\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}}) \\ &\quad +(\overline{\mathbf{d}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{d}}-\boldsymbol{\mu})\,]. \end{aligned} \tag{15}$$

The goal, of course, is to minimize (15) over all the parameters making up $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Now for any value of $\boldsymbol{\Sigma}$ (so long as it is non-singular), the quadratic form in the second line of (15) is zero and the entire function is minimized when $\boldsymbol{\mu}$ equals $\overline{\mathbf{d}}$. This means that "centering the data" by subtracting off sample means and then pretending that all variables have expected value zero is equivalent to starting with a model like (13) that contains intercepts, re-parameterizing the components of $\boldsymbol{\mu}$ in (14) as $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$ and $\boldsymbol{\mu_3}$, and then estimating those functions by the corresponding sample means (yielding the MLE of $\boldsymbol{\mu}$).

Notice that this minimization works for any value of the matrix of regression slopes $\boldsymbol{\Gamma}$, so that the even though $\boldsymbol{\Gamma}$ appears in the expression for $\boldsymbol{\mu}$, its MLE is determined entirely by the first line of (15). In this sense, the mean vector contains no information about the relationships between independent and dependent variables. We believe that except in special circumstances, this makes it reasonable to employ the classical no-intercept structural equation models to do regression with latent variables.

5 Normality

The discussion of model identification mentions multivariate normality, but this is not necessary. Suppose that the no-intercept Model (9) holds, and that the distributions of of the latent independent variables and error terms are unknown, except for possessing covariance matrices. In this case the parameter of the model could be expressed as $\theta = (\boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Theta}_1, \boldsymbol{\xi}_2, F_{\boldsymbol{\Phi}}, F_{\boldsymbol{\zeta}}, F_{\boldsymbol{\delta_1}}, F_{\boldsymbol{\delta_2}})$, where $F_{\boldsymbol{\xi}}$, $F_{\boldsymbol{\zeta}}$, $F_{\boldsymbol{\delta_1}}$ and $F_{\boldsymbol{\delta_2}}$ are the (joint) cumulative distribution functions of $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\delta_1}$ and $\boldsymbol{\delta_2}$ respectively.

Note that the parameter in this "non-parametric" problem is of infinite dimension, but this presents no conceptual difficulty. The probability distribution of the observed data is still a function of the parameter, and to show model identification, we would have to be able to recover the parameter from the probability distribution of the data. While in general we cannot recover the entire parameter vector, we certainly can recover a useful *function* of it, namely $\boldsymbol{\Gamma}$. In fact, $\boldsymbol{\Gamma}$ is the only quantity of interest; the remainder of the parameter vector consists only of nuisance parameters, whether the model is normal or not.

Again using $\boldsymbol{\Sigma}$ to denote the covariance matrix of the observed data, we see that $\boldsymbol{\Sigma}$ is a function of the probability distribution of the observed data. The calculations leading to (12) still hold, showing that $\boldsymbol{\Gamma}$ is a function of $\boldsymbol{\Sigma}$, and hence of the probability distribution of the data. This means that $\boldsymbol{\Gamma}$ is identified.

This is all very well, but can we actually *do* anything without knowing what the distributions are? Certainly! For example, a reasonable though non-standard estimator is

$$\widehat{\mathbf{\Gamma}} = \frac{1}{2}(\widehat{\mathbf{\Sigma}}'_{13}\widehat{\mathbf{\Sigma}}^{-1}_{12} + \widehat{\mathbf{\Sigma}}'_{23}\widehat{\mathbf{\Sigma}}^{-1}_{12}), \tag{16}$$

where $\widehat{\mathbf{\Sigma}}$ is the sample variance-covariance matrix. Consistency follows from the Law of Large Numbers and a continuity argument. All this assumes the existence only of second moments and cross-moments. With the assumption of fourth moments, the multivariate Central Limit Theorem would provide a routine basis for large-sample interval estimation and testing.

However, there is no need to bother. Research on the robustness of the normal model for structural equation model (Amemiya, Fuller and Pantula, 1987; Anderson and Rubin, 1956; Anderson and Amemiya, 1988; Anderson, 1989; Anderson and Amemiya, 1990; Browne, 1988; Browne and Shapiro, 1988; Satorra and Bentler, 1990) shows that procedures for (such as likelihood ratio and Wald tests) based on a multivariate normal model are asymptotically valid even when the normal assumption is false. And Satorra and Bentler (1990) describe Monte Carlo work suggesting that normal-theory methods generally perform better than at least one method (Browne, 1984) that is specifically designed to be distribution-free. Since the methods suggested by the estimator (16) are similar to Browne's weighted least squares approach, they are also likely to be inferior to the standard normal-theory tools.

It is important to note that while the normal-theory tests and confidence intervals for $\mathbf{\Gamma}$ can be trusted when the data are not normal, this does not extend to the other model parameters. For example, if the vector of latent variables $\boldsymbol{\xi}$ is not normal, then normal-theory inference about its covariance matrix will be flawed.

In any event, the method of choice is maximum likelihood, with interpretive focus on the regression coefficients in $\mathbf{\Gamma}$ rather than on the other model parameters.

6 Discussion

In general, data collection should be planned with the statistical analysis in mind. In keeping with this idea, the double measurement design is both a statistical model – specifically, Model (9) – and a set of guidelines for data collection. It assumes that measurement error is present, and that when data are collected by a common method or in a common setting, the errors of measurement will naturally be correlated with one another. It also assumes that each independent variable can be measured more than once, ideally on different occasions and in ways that are different enough so that errors of measurement are independent between occasions.

A great deal of effort can be saved by following this recipe. The data are tailored to satisfy the technical requirements of the model, while the model allows for the inevitable correlations among measurement errors within occasions and is automatically identified,

allowing clear conclusions to be drawn from the data. The only remaining issue is choosing good software and making sure that one knows what it is actually doing.

Unfortunately, most observational data sets are assembled without any awareness of measurement error as a statistical issue. Variables tend to be measured in only one way, and often at more or less the same time by the same personnel. Only after the data are collected do the investigators possibly start to think about fitting a model with measurement error. Many times, it is only at this point that a statistician enters the picture.

This is a difficult situation, but not necessarily hopeless. The most plausible model that includes measurement error is unlikely to be identified, but the instrumental variables example of Section 2 tells us that model identification can sometimes be purchased by adding more dependent variables. (Watch out, though! Dependent variables are usually measured with error too, and one needs specific reason to believe that those measurement errors are unrelated to measurement errors in the independent variables.) Sometimes, a model can be simplified or constrained, perhaps by assuming that certain covariances are zero, and the simplified model will be identified and still fairly realistic.

Fixing up a non-identified model after the data are already collected requires the quantitative sophistication to check model identification (repeatedly), and the subject-matter sophistication to tell whether the model is still scientifically meaningful when a given technical constraint is imposed. Either one person has to know a lot, or statistician and scientist must work closely together for an extended period, without any guarantee of ultimate success. It's a lot easier to plan the study properly in the first place.

One final comment is that from the statistician's viewpoint, a non-identified model is a "bad" model because it does not allow us to find out about the model parameters, and will probably generate a pile of warnings and error messages if we try to run the software anyway. But it's not the model's fault! Think of the very first example, the simple regression through the origin of Model (1). A model like this could be reasonable and even approximately correct, but the *data* we have will not allow us to estimate the parameters.

Now consider what happens when a fairly complicated initial model turns out not to be identified. The typical approach is to start imposing constraints that will make it identified. But this makes the model better only in a formal, statistical sense. Actually, the initial model was probably the most natural and believable one, and what we are doing is to chop pieces off for purely technical reasons. The best we can hope is that this does not cripple the model too much.

It's not the model's fault; it's the data's fault. Or, to put it delicately, there is an opportunity here for scientists to make their research even better by collecting data that allow reasonable models to be estimated.

References

Amemiya, Y., Fuller, W.A. and Pantula, S.C. (1987). "The asymptotic distributions of some estimators for a factor analysis model." *Journal of Multivariate Analysis*, 22, 51-64.

Anderson, T.W. (1989). "Linear latent variable models and covariance structures." *Journal of Econometrics*, 41, 91-119.

Anderson, T. W. and Amemiya, Y. (1988). "The asymptotic normal distribution of estimators in factor analysis under general conditions." *Annals of Statistics*, 16, 759-771.

Anderson, T. W. and Amemiya, Y. (1990). "Asymptotic chi-square tests for a large class of factor analysis models." *Annals of Statistics*, 18, 1453-1463.

Anderson, T. W. and Rubin, H. (1956). "Statistical inference in factor analysis." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5, 111-150.

Arbuckle, J. L. (2006). *Amos 7.0 User's Guide.* Chicago: SPSS Inc.

Bentler, P. M. and Weeks, D. G. (1980). "Linear structural equations with latent variables." *Psychometrika*, 45, 289-308.

Bentler, P. M. and Woodward, J. A. (1978). "A Head Start re-evaluation: Positive effects are not yet demonstrable." *Evaluation Quarterly*, 2, 493-510.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Browne, M.W. (1984). "Asymptotically distribution-free methods for the analysis of covariance structures." *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.

Browne, M.W. (1987). "Robustness of statistical inference in factor analysis and related models." *Biometrika*, 74, 375-384.

Browne, M.W. and Shapiro, A. (1988). "Robustness of normal theory methods in the analysis of linear latent variable models." *British Journal of Mathematical and Statistical Psychology*, 41, 193-208.

Cheng, C. L. and Van Ness, J. W. (1999). *Statistical regression with measurement error.* London: Chapman & Hall.

Cochran, W. G. (1968). "Errors of measurement in statistics." *Technometrics*, 10, 637-666.

Fisher, R. A. F. (1938). *Statistical methods for research workers (7th ed.).* London: Oliver and Boyd.

Fox, J. (2006). "Structural equation modeling with the `sem` package in R." *Structural equation modelling*, 13, 465-486.

Fuller, W. A. (1987). *Measurement error models.* New York: Wiley.

Goldberger, A. S. and Duncan, O. D. (1973). *Structural equation models in the social sciences.* New York: Academic Press.

Jöreskog, K. G. (1978). "Structural analysis of covariance and correlation matrices." *Psychometrika*, 43, 443-477.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8: Structural equation modelling with the SIMPLIS command language.* London: Scientific Software International.

McArdale, J. J. (1980). "Causal modelling applied to psychonomic systems simulation." *Behavior research methods & Instrumentation.* 12, 193-209.

McDonald, R. P. (1978). "A simple comprehensive model for the analysis of covariance structures." *British Journal of Mathematical and Statistical Psychology.* 31, 59-72.

SAS Institute, Inc. (1999). "SAS/STAT User's guide, Version 8." Cary, N. C.: SAS Institute, Inc. 3884 pp.

Satorra, A. and Bentler, P. M. (1990). "Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations." *Computational Statistics and Data Analysis.* 10, 235-249.

Stouffer, S. A. (1936). "Evaluating the effect of inadequately measured variables in partial correlation analysis." *J. Am. Statist. Assoc.*, 31, 348-360.