CHAPTER TWO

Model Notation, Covariances, and Path Analysis

Readers of this book are likely to have diverse backgrounds in statistics. There is a need to establish some common knowledge. I assume that readers have prior exposure to matrix algebra. Appendix A at the end of the book provides a summary of basic matrix algebra for those wishing to review it. Appendix B gives an overview of asymptotic distribution theory which I use in several chapters. This chapter discusses three basic tools essential to understanding structural equation models. They are model notation, covariances, and path analysis.

MODEL NOTATION

Jöreskog (1973, 1977), Wiley (1973), and Keesling (1972) developed the notation on which I rely. Jöreskog and Sörbom's LISREL (LInear Structural RELationships) computer program popularized it, and many refer to it as the LISREL notation. I introduce the basic notation in this section and save the more specialized symbols for the later chapters where they are needed.

The full model consists of a *system of structural equations*. The equations contain random variables, structural parameters, and sometimes, nonrandom variables. The three types of random variables are latent, observed, and disturbance/error variables. The nonrandom variables are explanatory variables whose values remain the same in repeated random sampling (fixed or nonstochastic variables). These are less common than random explanatory variables.

The links between the variables are summarized in the *structural parameters*. The structural parameters are invariant constants that provide the "causal" relation between variables. The structural parameters may describe the causal link between unobserved variables, between observed variables, or between unobserved and observed variables. I further discuss the meaning of causality and the structural parameters in Chapter 3. The system of structural equations has two major subsystems: the latent variable model and the measurement model.

Latent Variable Model

Latent random variables represent unidimensional concepts in their purest form. Other terms for these are *unobserved* or *unmeasured variables* and *factors*. The observed variables or indicators of a latent variable contain random or systematic measurement errors, but the latent variable is free of these. Since all latent variables correspond to concepts, they are hypothetical variables. Concepts and latent variables, however, vary in their degree of abstractness. Intelligence, social class, power, and expectations are highly abstract latent variables that are central to many social science theories. Also important, but less abstract, are variables such as income, education, population size, and age. The latter type of latent variables are directly measurable, whereas the former are capable of being only indirectly measured. An example containing both types of latent variables is Emile Durkheim's hypothesis of the inverse relationship between social cohesion and suicide. Social cohesion refers to group solidarity, a fairly abstract latent variable. Suicide is directly observable. But this direct-indirect demarcation becomes blurred when one considers that some suicides are disguised or misclassified as some other form of death. Thus the measurement of suicide may not be as direct as it initially appears. I make no distinction between directly and indirectly observable latent variables for the latent variable models. Analytically they may be treated the same. Chapter 6 provides further discussion of the nature of latent variables, their measurement, and their scales.

The latent variable model encompasses the structural equations that summarize the relationships between latent variables. Sometimes this part of the model is called the "structural equation" or "causal model." I depart from this practice because it can be misleading. All equations in the model, both those for the latent variables and those for the measurement model, describe structural relationships. To apply structural to only the latent variable part of the full model suggests that the measurement model is not structural.

I use the relationship of political democracy to industrialization in developing countries to introduce the notation for latent variable models. International development researchers disagree about whether industrialization is positively associated with political democracy in Third World countries. The alternation between dictatorships and electoral regimes in some of these societies makes it difficult to discern whether any general association exists. Political democracy refers to the extent of political rights (e.g., fairness of elections) and political liberties (e.g., freedom of the press) in a country. Industrialization is the degree to which a society's economy is characterized by mechanized manufacturing processes. It is some of the consequences of industrialization (e.g., societal wealth, an educated population, advances in living standards) that are thought to enhance the chances of democracy. However, to keep the model simple, I do not include these intervening variables. Suppose that I have three latent random variables: political democracy in 1965 and 1960, and industrialization in 1960. I assume that political democracy in 1965 is a function of 1960 political democracy and industrialization. The 1960 industrialization level also affects the 1960 political democracy level. Nothing is said about the determinants of industrialization that lie outside of the model. Industrialization is an *exogenous* ("independent") latent variable and is symbolized as ξ_1 (xi). It is exogenous because its causes lie outside the model. The latent political democracy variables are *endogenous*; they are determined by variables within the model. Each endogenous latent variable is represented by η_i (eta). Political democracy in 1960 is represented as η_1 and 1965 democracy by η_2. The latent endogenous variables are only partially explained by the model. The unexplained component is represented by ζ_i (zeta) which is the random disturbance in the equation.

The terms exogenous and endogenous are model specific. It may be that an exogenous variable in one model is endogenous in another. Or, a variable shown as exogenous, in reality, may be influenced by a variable in the model. Regardless of these possibilities, the convention is to refer to variables as exogenous or endogenous based on their representation in a particular model.

The latent variable model for the current example is

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1 \tag{2.1}$$

$$\eta_2 = \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2 \tag{2.2}$$

The equations are *linear in the variables and linear in the parameters*. We can sometimes estimate equations that are nonlinear in the variables and linear in the parameters as in regression analysis. To date, however, practical and general means of implementing this for nonlinear functions of latent variables measured with error do not exist (see Chapter 9).

The random errors, ζ_1 and ζ_2, have expected values (means) of zero and are uncorrelated with the exogenous variable, industrialization (ξ_1). A constant is absent from the equations because the variables are deviated from their means.[1] This deviation form will simplify algebraic manipulations but does not affect the generality of the analysis. The β_{21} (beta) coefficient is the structural parameter that indicates the change in the expected value of η_2 after a one-unit increase in η_1 holding ξ_1 constant. The γ_{11} (gamma) and γ_{21} regression coefficients have analogous interpretations. The β_{21} coefficient is associated with the latent endogenous variable, whereas γ_{11} and γ_{21} are associated with the exogenous latent variable.

Equations (2.1) and (2.2) may be rewritten in matrix notation:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix} [\xi_1] + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \tag{2.3}$$

which is more compactly written as

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2.4}$$

The equation (2.4) is the general matrix representation of the structural equations for the latent variable model. Table 2.1 summarizes the notation for models relating latent variables, including each symbol's name, phonetic spelling, dimension, and definition.

Starting with the first variable, $\boldsymbol{\eta}$ is an $m \times 1$ vector of the latent endogenous random variables. In the industrialization–political democracy example m is equal to 2. The $\boldsymbol{\xi}$ vector is $n \times 1$, and it represents the n exogenous latent variables. Here, as in most cases, $\boldsymbol{\xi}$ is a vector of random variables. Occasionally, one or more of the ξ's are nonrandom. For the current example n is one since only industrialization (ξ_1) is exogenous. The errors in the equations or disturbances are represented by $\boldsymbol{\zeta}$, an $m \times 1$ vector. A ζ_i is associated with each η_i, with i running from 1 to m. The example has two ζ_i variables. The $\boldsymbol{\zeta}$ vector generally contains random variables.[2] As in a regression analysis the disturbance ζ_i includes those variables that influence η_i but are excluded from the η_i equation. We assume that these numerous omitted factors fused into ζ_i have $E(\zeta_i) = 0$

[1] In Chapters 4, 7, and 8 I will explain how to estimate models with a constant.

[2] One exception is if an equation is an identity (e.g., $\eta_1 = \eta_2 + \eta_3$) so that ζ_1 is zero and therefore constant.

Table 2.1 Notation for Latent (Unobserved) Variable Model

Structural Equation for Latent Variable Model

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

Assumptions

$E(\boldsymbol{\eta}) = 0$
$E(\boldsymbol{\xi}) = 0$
$E(\boldsymbol{\zeta}) = 0$
$\boldsymbol{\zeta}$ uncorrelated with $\boldsymbol{\xi}$
$(\mathbf{I} - \mathbf{B})$ nonsingular

Symbol	Name	Phonetic Spelling	Dimension	Definition
			Variables	
$\boldsymbol{\eta}$	eta	ā′ t ə (or ē′t ə)	$m \times 1$	latent endogenous variables
$\boldsymbol{\xi}$	xi	zī (or ksē)	$n \times 1$	latent exogenous variables
$\boldsymbol{\zeta}$	zeta	zā′t ə (or zē′t ə)	$m \times 1$	latent errors in equations
			Coefficients	
$\mathbf{B}$	beta	bā′t ə (or bē′t ə)	$m \times m$	coefficient matrix for latent endogenous variables
$\boldsymbol{\Gamma}$	gamma	gam′ə	$m \times n$	coefficient matrix for latent exogenous variables
			Covariance Matrices	
$\boldsymbol{\Phi}$	phi	fī (or fē)	$n \times n$	$E(\boldsymbol{\xi}\boldsymbol{\xi}')$ (covariance matrix of $\boldsymbol{\xi}$)
$\boldsymbol{\Psi}$	psi	sī (or psē)	$m \times m$	$E(\boldsymbol{\zeta}\boldsymbol{\zeta}')$ (covariance matrix of $\boldsymbol{\zeta}$)

and are uncorrelated with the exogenous variables in $\boldsymbol{\xi}$. Otherwise, inconsistent coefficient estimators are likely.

We also assume that ζ_i is homoscedastic and nonautocorrelated. To clarify this assumption, suppose that I add an observation index to ζ_i so that ζ_{ik} refers to the value of ζ_i for the kth observation and ζ_{il} is ζ_i for the lth observation. The homoscedasticity assumption is that the VAR(ζ_i) is constant across cases [i.e., $E(\zeta_{ik}^2) = \mathrm{VAR}(\zeta_i)$ for all k]. The no autocorrelation assumption means that ζ_{ik} is uncorrelated with ζ_{il} for all k and l, where $k \neq l$ (i.e., $\mathrm{COV}(\zeta_{ik}, \zeta_{il}) = 0$ for $k \neq l$). Corrections for heteroscedastic or autocorrelated disturbances are well known for econometric-type models but hardly studied for the general structural equation model with latent variables. The homoscedasticity and no autocorrelation assumptions do *not* mean that the disturbances from two *different* equations need be uncorrelated nor that they need have the same variance. That is, $E(\zeta_{ik}^2) = \mathrm{VAR}(\zeta_i)$ is not the same as $E(\zeta_i^2) = E(\zeta_j^2)$, nor does $\mathrm{COV}(\zeta_{ik}, \zeta_{il}) = 0$ mean that $\mathrm{COV}(\zeta_i, \zeta_j) = 0$, where ζ_i and ζ_j are from separate equations.

The coefficient matrices are $\mathbf{B}$ and Γ. The $\mathbf{B}$ matrix is an $m \times m$ coefficient matrix for the latent endogenous variables. Its typical element is β_{ij} where i and j refer to row and column positions. The model assumes that $(\mathbf{I} - \mathbf{B})$ is nonsingular so that $(\mathbf{I} - \mathbf{B})^{-1}$ exists. This assumption enables (2.4) to be written in reduced form. Solving (2.4) algebraically so that only $\boldsymbol{\eta}$ appears on the left-hand side leads to the reduced form that I discuss in Chapter 4. The Γ matrix is the $m \times n$ coefficient matrix for the latent exogenous variables. Its elements are symbolized as γ_{ij}. For the industrialization–political democracy example,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \quad \boldsymbol{\zeta} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix}, \quad \boldsymbol{\xi} = [\xi_1] \tag{2.5}$$

The main diagonal of $\mathbf{B}$ is always zero. This serves to remove η_i from the right-hand side of the ith equation for which it is the dependent variable. That is, we assume that a variable is not an immediate and instantaneous cause of itself. A zero in $\mathbf{B}$ also indicates the absence of an effect of one latent endogenous variable on another. That a zero appears in the (1, 2) position of $\mathbf{B}$ in (2.5) indicates that η_2 does not affect η_1. The Γ matrix in equation (2.5) is 2×1 since there are two endogenous latent variables and one exogenous latent variable. Since ξ_1 affects both η_1 and η_2, Γ contains no zero elements.

Two covariance matrices are part of the latent variable model in Table 2.1. A covariance matrix is an "unstandardized correlation matrix" with the variances of a variable down the main diagonal and the covariance (the product of the correlation between two variables times their standard deviations) of all pairs of variables in the off-diagonal.[3] The $n \times n$ covari-

[3] I treat covariances more fully later in this chapter.

ance matrix of the latent exogenous variables (or the ξ's) is $\boldsymbol{\Phi}$ (phi) with elements ϕ_{ij}. Like all covariance matrices, it is symmetric. If the variances of the ξ variables are equal to one, then $\boldsymbol{\Phi}$ is a correlation matrix. In the industrialization–political democracy example only one variable appears in $\boldsymbol{\xi}$, so $\boldsymbol{\Phi}$ is a scalar (i.e., ϕ_{11}) that equals the variance of ξ_1.

The $m \times m$ covariance matrix of the errors in the equations is $\boldsymbol{\Psi}$ (psi) with elements ψ_{ij}. Each element of the main diagonal of $\boldsymbol{\Psi}$ (ψ_{ii}) is the variance of the corresponding η_i variable that is unexplained by the explanatory variables included in the ith equation. In the current example $\boldsymbol{\Psi}$ is 2×2. The $(1,1)$ element is the variance of ζ_1, the $(2,2)$ element is the variance of ζ_2, and the off-diagonal elements—$(1,2)$ and $(2,1)$—are both equal to the covariance of ζ_1 with ζ_2. In this example I assume that the off-diagonal elements are zero.[4] As I show in Chapters 4 and 8, the covariance matrix for $\boldsymbol{\eta}$ is a function of $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. It does not have a special symbol.

Readers familiar with econometric texts will note the similarity between the structural equations for the latent variable model of Table 2.1 and the general representation of simultaneous equation systems (e.g., $\mathbf{By} + \boldsymbol{\Gamma}\mathbf{x} = \mathbf{u}$, in Johnston 1984, 450). One difference is that both the endogenous and exogenous variables may be written on the left-hand side, leaving only $\boldsymbol{\zeta}$ on the right-hand side: $\mathbf{B}^*\boldsymbol{\eta} + \boldsymbol{\Gamma}^*\boldsymbol{\xi} = \boldsymbol{\zeta}$, with $\mathbf{B}^* = (\mathbf{I} - \mathbf{B})$ and $\boldsymbol{\Gamma}^* = -\boldsymbol{\Gamma}$. Most of the time some symbol other than $\boldsymbol{\zeta}$ represents the error in the equation (e.g., $\mathbf{u}$). These alternative representations matter little. Other differences are that in most econometric presentations $\mathbf{y}$ replaces $\boldsymbol{\eta}$ and $\mathbf{x}$ replaces $\boldsymbol{\xi}$. This difference is more than just a change in symbols. The classical econometric treatment assumes that the observed y and x variables are perfect measures of the latent η and ξ variables. Structural equations with latent variable models no longer have this assumption. In fact, the second major part of these models consists of structural equations linking the latent variables (the η's and ξ's) to the measured variables (the y's and x's). This part of the system is the measurement model.

Measurement Model

Like the latent variables, the observed variables have a variety of names, including manifest variables, measures, indicators, and proxies. I use these terms interchangeably. The latent variable model of industrialization and political democracy as described so far is exclusively in terms of unobserved variables. A test of this theory is only possible if I collect observable

[4]The ζ_1 and ζ_2 could be positively correlated since they influence the same latent variable separated only by five years. In other more elaborate models ψ_{12} might be a free parameter, but in this case it is not identified if freed. I return to the issue of identification in Chapter 4.

measures of these latent variables. One strategy is to use single indicators or proxy variables of political democracy and industrialization. Another option is to construct an index with two or more indicator variables for each of the concepts. The empirical analysis is of these observed indicators or indices, and researchers treat the results as tests of the relationships between the latent variables.

The underlying assumption of the preceding strategies is that the observed variables are perfectly correlated (or at least nearly so) with the latent variables that they measure. In most cases this is not true. Nearly all measures of abstract factors such as political democracy have far from perfect associations with the factor. The measurement model has structural equations that represent the link between the latent and observed variables, as an imperfect rather than a deterministic one. For this example I select three indicators of industrialization in 1960: gross national product (GNP) per capita (x_1), inanimate energy consumption per capita (x_2), and the percentage of the labor force in industry (x_3). For political democracy I have the same four indicators for 1960 and 1965: expert ratings of the freedom of the press (y_1 in 1960, y_5 in 1965), the freedom of political opposition (y_2 and y_6), the fairness of elections (y_3 and y_7), and the effectiveness of the elected legislature (y_4 and y_8). Thus each latent variable is measured with several observed variables.

Equations (2.6)–(2.7) provide a measurement model for these variables:

$$\begin{aligned} x_1 &= \lambda_1 \xi_1 + \delta_1 \\ x_2 &= \lambda_2 \xi_1 + \delta_2 \\ x_3 &= \lambda_3 \xi_1 + \delta_3 \end{aligned} \tag{2.6}$$

$$\begin{aligned} y_1 &= \lambda_4 \eta_1 + \epsilon_1, & y_5 &= \lambda_8 \eta_2 + \epsilon_5 \\ y_2 &= \lambda_5 \eta_1 + \epsilon_2, & y_6 &= \lambda_9 \eta_2 + \epsilon_6 \\ y_3 &= \lambda_6 \eta_1 + \epsilon_3, & y_7 &= \lambda_{10} \eta_2 + \epsilon_7 \\ y_4 &= \lambda_7 \eta_1 + \epsilon_4, & y_8 &= \lambda_{11} \eta_2 + \epsilon_8 \end{aligned} \tag{2.7}$$

As in the latent variable model, the variables in the measurement model are deviated from their means. The x_i variables ($i = 1, 2, 3$) stand for the three measures of ξ_1, industrialization, the y_1 to y_4 variables are measures of η_1, 1960 political democracy, and y_5 to y_8 measure η_2, 1965 democracy. Note that all the manifest variables depend on the latent variables. In some cases indicators may cause latent variables. This situation is discussed more fully in Chapters 3, 6, and 7.

The λ_i (lambda) coefficients are the magnitude of the expected change in the observed variable for a one unit change in the latent variable. These coefficients are regression coefficients for the effects of the latent variables on the observed variables. We must assign a scale to the latent variable to fully interpret the coefficients. Typically, analysts set a latent variable's scale equal to one of its indicators or standardize the variance of the latent variable to one. I will discuss this issue in more detail in chapters six and seven.

The δ_i (delta) and ϵ_i (epsilon) variables are the *errors of measurement* for x_i and y_i, respectively. They are disturbances that disrupt the relation between the latent and observed variables. The assumptions are that the errors of measurement have an expected value of zero, that they are uncorrelated with all ξ's, η's, and ζ's, and that δ_i and ϵ_j are uncorrelated for all i and j.

A correlation of δ_i and ϵ_j with any ξ's or η's can lead to inconsistent parameter estimators in a fashion analogous to a disturbance correlated with an explanatory variable in regression analysis. Sometimes in factor analysis δ_i and ϵ_j are called unique factors, and each δ_i and ϵ_j is divided into specific and nonspecific components. I will say more about this in Chapters 6 and 7, but till then I refer to the δ's and ϵ's as errors of measurement. Finally, we assume that each δ_i or ϵ_i is homoscedastic and nonautocorrelated across observations. This assumption parallels that made for the ζ_i's, the disturbances for the latent variable model.

Equations (2.6) and (2.7) may be more compactly written in matrix form as:

$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \tag{2.8}$$

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{2.9}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \boldsymbol{\Lambda}_x = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}, \quad \boldsymbol{\xi} = [\xi_1], \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} \tag{2.10a}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix}, \quad \boldsymbol{\Lambda}_y = \begin{bmatrix} \lambda_4 & 0 \\ \lambda_5 & 0 \\ \lambda_6 & 0 \\ \lambda_7 & 0 \\ 0 & \lambda_8 \\ 0 & \lambda_9 \\ 0 & \lambda_{10} \\ 0 & \lambda_{11} \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \end{bmatrix} \tag{2.10b}$$

Equations (2.8) and (2.9) also are shown at the top of Table 2.2, which provides the notation for the measurement model. The random variables in $\mathbf{x}$ are indicators of the latent exogenous variables (the ξ's).[5] The random variables in $\mathbf{y}$ are indicators of the latent endogenous variables (the η's). In general, $\mathbf{x}$ is $q \times 1$ (where q is the number of indicators of $\boldsymbol{\xi}$) and $\mathbf{y}$ is $p \times 1$ (where p is the number of indicators of $\boldsymbol{\eta}$).

The Λ_y and Λ_x matrices contain the λ_i parameters, which are the structural coefficients linking the latent and manifest variables. The Λ_x matrix is $q \times n$ (where n is the number of ξ's) and Λ_y is $p \times m$ (where m is the number of η's). In confirmatory factor analysis or in models where $\mathbf{y} = \boldsymbol{\eta}$ or $\mathbf{x} = \boldsymbol{\xi}$, I double subscript λ (λ_{ij}). The i refers to x_i (y_i) and the j refers to the ξ_j (η_j) that influences x_i (y_i). When $\mathbf{x}$, $\boldsymbol{\xi}$, $\mathbf{y}$, and $\boldsymbol{\eta}$ are used in a model such as (2.10), the double subscripts can be confusing since each λ_{ij} can refer to two different parameters. Therefore I use a single subscript to consecutively number the λ's in these cases. The errors of measurement vector for $\mathbf{x}$ is $\boldsymbol{\delta}$, and $\boldsymbol{\delta}$ is $q \times 1$. The error vector for $\mathbf{y}$, $\boldsymbol{\epsilon}$, is $p \times 1$. Generally, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ contain vectors of random variables.[6]

The last two matrices, Θ_δ and Θ_ϵ, are covariance matrices of the errors of measurement. The main diagonals contain the error variances associated with the indicators. The off-diagonal elements are the covariances of the errors of measurement for the different indicators. The Θ_δ matrix is $q \times q$ and has the error variances and their covariances for the x variables, and Θ_ϵ is a $p \times p$ matrix that contains the error variances and their covariances for the y variables. In this example I assume that the errors of measurement for the indicators of industrialization (x_1 to x_3) are uncorrelated so that Θ_δ is a diagonal matrix. This assumption is less defensible for Θ_ϵ because I have the same set of indicators at two points in time. It is likely that the error in measuring an indicator in 1960 is correlated with the error in measuring the same indicator in 1965. In addition the 1960 measures of y_2 and y_4 and the 1965 ones of y_6 and y_8 are from the same data source. The corresponding errors of measurement might be positively correlated due to systematic biases present in the source. Therefore the (4, 2), (5, 1), (6, 2), (7, 3), (8, 4), and (8, 6) off-diagonal elements of Θ_ϵ may be nonzero.

This example reveals some of the major features of structural equations with latent variables that are distinct from the standard regression approach. The models are more realistic in their allowance for measurement error in the observed variables. They allow random measurement error in $\boldsymbol{\epsilon}$

[5]The observed variables in $\mathbf{x}$ are not random when $\mathbf{x} = \boldsymbol{\xi}$ and $\mathbf{x}$ is fixed. See Chapter 4 for a discussion of fixed $\mathbf{x}$ variables.

[6]When any $\mathbf{y}$ or $\mathbf{x}$ has no measurement error the corresponding element in $\boldsymbol{\epsilon}$ or $\boldsymbol{\delta}$ is zero, a constant.

Table 2.2 Notation for Measurement Model

Structural Equations for Measurement Model

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta}$$
$$\mathbf{y} = \Lambda_y \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

Assumptions

$E(\boldsymbol{\eta}) = 0$, $E(\boldsymbol{\xi}) = 0$, $E(\boldsymbol{\epsilon}) = 0$, and $E(\boldsymbol{\delta}) = 0$
$\boldsymbol{\epsilon}$ uncorrelated with $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\delta}$
$\boldsymbol{\delta}$ uncorrelated with $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and $\boldsymbol{\epsilon}$

Symbol	Name	Phonetic Spelling	Dimension	Definition
			Variables	
$\mathbf{y}$	—	—	$p \times 1$	observed indicators of $\boldsymbol{\eta}$
$\mathbf{x}$	—	—	$q \times 1$	observed indicators of $\boldsymbol{\xi}$
$\boldsymbol{\epsilon}$	epsilon	e*p*′ sə lo*n*′ (e*p*′ səlen)	$p \times 1$	measurement errors for **y**
$\boldsymbol{\delta}$	delta	de*l*′ tə	$q \times 1$	measurement errors for **x**
			Coefficients	
Λ_y	lambda *y*	la*m*′ də*y*	$p \times m$	coefficients relating **y** to $\boldsymbol{\eta}$
Λ_x	lambda *x*	lam′ də*x*	$q \times n$	coefficients relating **x** to $\boldsymbol{\xi}$
			Covariance Matrices	
$\boldsymbol{\Theta}_\epsilon$	theta-epsilon	th*ā*′ tə (th*ē*′ tə)-e*p*′ sə lon′	$p \times p$	$E(\boldsymbol{\epsilon\epsilon}')$ (covariance matrix of ϵ)
$\boldsymbol{\Theta}_\delta$	theta-delta	th*ā*′ tə (th*ē*′ tə)-de*l*′ tə	$q \times q$	$E(\boldsymbol{\delta\delta}')$ (covariance matrix of δ)

and $\boldsymbol{\delta}$, and systematic differences in scale are introduced with the λ coefficients. The error in measuring one variable can correlate with that of another. Multiple indicators can measure one latent variable. Furthermore researchers can analyze the relation between latent variables unobscured by measurement error. All of these features bring us closer to testing the hypotheses set forth in theories.

COVARIANCE

Covariance is a central concept for the models. In fact, another name for the general structural equation techniques is *analysis of covariance structures*. I review two aspects of covariances. One is covariance algebra which helps in deriving properties of the latent variable and measurement models. The other includes the factors that influence sample covariances which in turn can affect parameter estimates. I consider the covariance algebra first.

Covariance Algebra

Table 2.3 provides a summary of the definitions and common rules of covariance algebra. In Table 2.3 the $E(\cdot)$ refers to the expected value of the expression within the parentheses. The top half of Table 2.3 defines both the covariance and the variance. The capital X_1, X_2, and X_3 signify the original random variables rather than the mean deviation forms. When X_1 and X_2 have a positive linear association, the $\text{COV}(X_1, X_2)$ is positive. If they are inversely related the $\text{COV}(X_1, X_2)$ is negative, but it is zero if there is no linear association. Note that in the definition of covariance, I employ capital letters "COV" to signify *population* covariances. The population covariance matrix of the observed variables is Σ. I will discuss sample covariances shortly. The variance is the covariance of a variable with itself. Capital $\text{VAR}(X_1)$ represents the population variance of X_1. The main diagonal of Σ contains the variances of the observed variables.

Table 2.3 Definitions and Common Rules of Covariances

Definitions

$$\begin{aligned}\text{COV}(X_1, X_2) &= E[(X_1 - E(X_1))(X_2 - E(X_2))]\\ &= E(X_1X_2) - E(X_1)E(X_2)\\ \text{VAR}(X_1) &= \text{COV}(X_1, X_1)\\ &= E[(X_1 - E(X_1))^2]\end{aligned}$$

Rules

c is a constant

X_1, X_2, X_3 are random variables

(1) $\text{COV}(c, X_1) = 0$

(2) $\text{COV}(cX_1, X_2) = c\,\text{COV}(X_1, X_2)$

(3) $\text{COV}(X_1 + X_2, X_3) = \text{COV}(X_1, X_3) + \text{COV}(X_2, X_3)$

Several examples help to illustrate the covariance rules. In the examples I assume that all disturbances have expected values of zero and that all random variables are deviated from their means. Unless stated otherwise, the lowercase x and y represent deviation forms of the original random variables X and Y. For the first example, suppose that you know the covariance of a latent variable, ξ_1, with an observed variable x_1. If you add a constant to x_1, the covariance is

$$\begin{aligned}\mathrm{COV}(\xi_1, x_1 + c) &= \mathrm{COV}(\xi_1, x_1) + \mathrm{COV}(\xi_1, c)\\ &= \mathrm{COV}(\xi_1, x_1)\end{aligned}$$

Thus the covariance of a latent variable and an observed variable is unchanged if a constant is added to the observed variable. This result is important given that we rarely have widely agreed-upon baselines (zero points) for measures of social science concepts. It shows that if the measure is changed by some constant value, this will not influence the covariance it has with the latent variable. The same example illustrates another point. Suppose that c is the mean of X_1 in its original form. Then $x_1 + c$ leads to X_1. The preceding example shows that the covariance of any random variable with another is the same regardless of whether the variables are in deviation or original form. If, however, the scale is changed to cx_1, this does change the covariance to $c\,\mathrm{COV}(\xi_1, x_1)$.

A second example addresses an issue in measurement. In psychometrics and other social science presentations it is often argued that two indicators, each positively related to the same concept, should have a positive covariance. Suppose that we have two indicators each related to the same ξ_1 such that $x_1 = \lambda_1\xi_1 + \delta_1$, $x_2 = \lambda_2\xi_1 + \delta_2$, $\mathrm{COV}(\xi_1, \delta_1) = \mathrm{COV}(\xi_1, \delta_2) = 0$, and λ_1 and $\lambda_2 > 0$. Must the $\mathrm{COV}(x_1, x_2)$ be positive?

$$\begin{aligned}\mathrm{COV}(x_1, x_2) &= \mathrm{COV}(\lambda_1\xi_1 + \delta_1, \lambda_2\xi_1 + \delta_2)\\ &= \lambda_1\lambda_2\phi_{11}\end{aligned}$$

For nonzero ϕ_{11}, the indicators x_1 and x_2 must have a positive covariance. (Since ϕ_{11} is a variance, it is positive for all nonconstant ξ_1.) As a second part, consider x_1 and x_2 as indicators of ξ_1 such that $\xi_1 = \lambda_1 x_1 + \lambda_2 x_2 + \delta_1$, $\mathrm{COV}(x_1, \delta_1) = \mathrm{COV}(x_2, \delta_1) = 0$, and λ_1 and $\lambda_2 > 0$. Must the $\mathrm{COV}(x_1, x_2) > 0$?

$$\mathrm{COV}(x_1, x_2) = ?$$

The x_1 and x_2 variables are exogenous. Their covariance is not determined within the model so that it can be positive, zero, or negative even if both x_1 and x_2 are positively related to ξ_1. An example is if the latent variable is exposure to discrimination (ξ_1) and the two indicators are race (x_1) and sex (x_2). Although x_1 and x_2 indicate exposure to discrimination, we would not expect them to have a positive correlation. Or, the latent variable could be social interaction (ξ_1) and the indicators time spent with friends (x_1) and time spent with family (x_2). These indicators might even have a negative covariance. This simple example shows that general statements about the necessity of indicators of the same concept to be positively associated require qualification.[7]

The covariance definitions and rules also apply to matrices and vectors. For instance, if $\mathbf{c}'$ is a vector of constants that conforms in multiplication with $\mathbf{x}$, then $\mathrm{COV}(\mathbf{x}, \mathbf{c}') = \mathbf{0}$. The $\mathrm{VAR}(\mathbf{x}) = \mathrm{COV}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\Sigma}$, where all three symbols represent the population covariance matrix of $\mathbf{x}$. Typically, I use $\boldsymbol{\Sigma}$, sometimes subscripting it to refer to specific variables (e.g., $\boldsymbol{\Sigma}_{xy}$ = covariance matrix of $\mathbf{x}$ with $\mathbf{y}$).

Sample Covariances

So far I have limited the discussion to *population* covariances and variances. In practice, sample estimates of the variance and covariances are all that are available. The unbiased sample estimator of covariance is

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1} \tag{2.11}$$

where $\mathrm{cov}(X, Y)$ represents the sample estimator of the covariance between the N random sample values of X and Y. The X_i and Y_i represent the values of X and Y for the ith observation. The $\bar{X}$ and $\bar{Y}$ are the sample means. In Appendix A I show that the sample covariance matrix is computed as

$$\mathbf{S} = \left(\frac{1}{N - 1}\right)\mathbf{Z}'\mathbf{Z} \tag{2.12}$$

where $\mathbf{Z}$ is a $N \times (p + q)$ matrix of deviation (from the means) scores for

[7]The implications of this for the internal consistency perspective on measurement is treated in Bollen (1984).

the $p + q$ observed variables. The **S** matrix is square and symmetric with the sample variances of the observed variables down its main diagonal and the sample covariances off the diagonal.

The sample covariance matrix is crucial to estimates of structural equation models. In LISREL, EQS, and other computer programs for the analysis of covariance structures, the sample covariance (or correlation) matrix often is the only data in the analysis. The parameter estimates depend on functions of the variances and covariances. For instance, in simple regression analysis the ordinary least squares (OLS) estimator for a regression coefficient is the ratio of the sample covariance of y and x to the variance of x [i.e., $\text{cov}(x, y)/\text{var}(x)$]. In more general structural equation systems the estimates result from far more complicated functions. For both the simple and the complex situation, factors that affect elements of the sample covariance matrix, **S**, have the potential to affect the parameter estimates.

Although, on average, the sample covariance matrix **S** equals the population covariance matrix $\mathbf{\Sigma}$, some samples lead to an **S** closer to $\mathbf{\Sigma}$ than others. In addition to sampling fluctuations of **S**, other factors can affect its elements. One is a nonlinear relation between variables. The sample covariance or correlation are measures of *linear* association. Covariances like correlations can give a misleading impression of the association between two variables that have a curvilinear relation. In some substantive areas research is sufficiently developed to alert researchers to possible nonlinear links, and this knowledge can be incorporated into a model. For instance, demographic transition theory suggests that as mortality rates in countries decline, they follow a typical curvilinear pattern. Another example is that earnings initially increase with age, then eventually stabilize or even decline as individuals grow older. The more common situation is when the exact form of the relation is unknown. Scatterplots, partial plots (Belsley, Kuh, and Welsch 1980), comparisons of equation fits upon variable transformations, or related devices help to detect nonlinearity in observed variable equations such as regression models. Detection of nonlinearities in models with latent variables is a relatively underdeveloped area. McDonald (1967a, 1967b) suggests some procedures for exploring nonlinearities in factor analysis, but much more needs to be done to have procedures applicable to all types of structural equation models.

A second factor that can affect sample covariances and correlations are *outliers*. Outliers are observations with values that are distinct or distant from the bulk of the data. Outliers can have large or small effects on an analysis. Outliers that lead to substantial changes are *influential* observations. When influential cases are present the covariances provide a misleading summary of the association between most of the cases.

Table 2.4 U.S. Disposable Income Per Capita and Consumers' Expenditures Per Capita in Constant Dollars, 1922 to 1941

	Disposable Income (income)	Consumers' Expenditures (consum)
1922	433	394
1923	483	423
1924	479	437
1925	486	434
1926	494	447
1927	498	447
1928	511	466
1929	534	474
1930	478	439
1931	440	399
1932	372	350
1933	381	364
1934	419	392
1935	449	416
1936	511	463
1937	520	469
1938	477	444
1939	517	471
1940	548	494
1941	629	529

Source: Haavelmo (1953).

Detection of outliers can start with examining the univariate distributions of the observed variables. To illustrate this, I take data from Haavelmo (1953). His goal is to analyze the marginal propensity to consume as a function of income with aggregate U.S. data. That is, he seeks to estimate the proportion of U.S. disposable income spent for consumers' expenditures rather than "spent" for investment expenditures. Two variables in his analysis are U.S. disposable income per capita (income) and U.S. consumers' expenditures per capita (consum), both in constant dollars and for each year from 1922 to 1941. Table 2.4 lists these data.

A convenient means of summarizing sample univariate distributions are *stem-and-leaf* diagrams. The stem-and-leaf displays for consum and income are in Figure 2.1. To construct the display, each value of a variable is rounded to two digits. For instance, the 1930 income value of 478 is rounded to 48. Then all two-digit values are placed in ranked order for each variable. The first digit of each value is the stem, and it is indicated in the

Consumption

```
Stem-Leaf          #
  5 3              1
  4 55677779       8
  4 0223444        7
  3 5699           4
  3
   ----+----+----+----+
Multiply Stem-Leaf by 100
```

Income

```
Stem-Leaf          #
  6 3              1
  5 5              1
  5 011223         6
  4 5888899        7
  4 234            3
  3 78             2
   ----+----+----+----+
Multiply Stem-Leaf by 100
```

Figure 2.1 Stem-and-Leaf Displays for Haavelmo's (1953) U.S. Consumption and Income Data, 1922 to 1941

first column of the stem-and-leaf. The second digit is the leaf, and this is placed in the row that corresponds to its stem. At the base of the stem-and-leaf is the 10^p power that a value should be multiplied by to recover the proper decimal place for the original variable. Consider the top row for income in Figure 2.1. The stem is 6 and the leaf is 3, while the multiple is 10^2 which leads to 630 (the rounded value of 629 for 1941 income). The next to the bottom row represents 420, 430, and 440 which are the original values of 419 (in 1934), 433 (in 1922), and 440 (in 1931). The stem-and-leaf diagram is like a histogram turned on its side, but unlike the histogram, it allows the recovery of the original values with a two-digit degree of accuracy.[8]

The stem-and-leaf diagrams reveal that most income values are in the 450 to 500 range and most consumption values fall between 400 and 490.

[8]A brief introduction to stem-and-leaf displays and other tools of exploratory data analysis is in Leinhardt and Wasserman (1978). Also, if the range of values is large enough, the stem may have two or more digits so that the degree of accuracy would be greater than two digits.

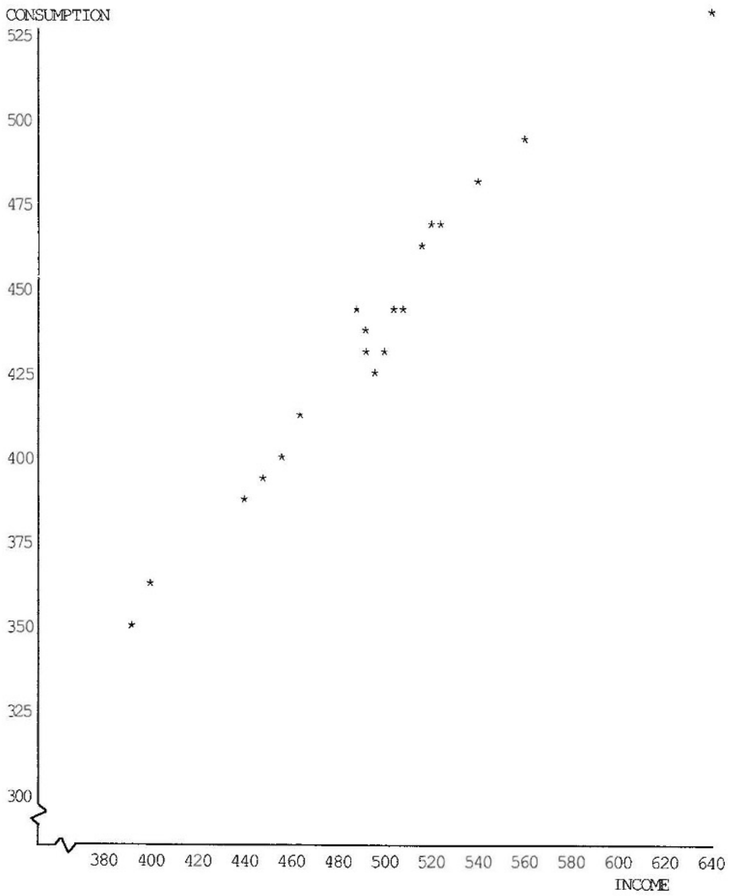


Figure 2.2 Scatterplot of Consumption (cons) and Income (inc) U.S. 1922 to 1941

The income value of 630 is the highest, and it is the most distant from the other values. For consumption the highest value, 530, is less distant from the other sample values. Both of these highest values are for 1941, which suggests that this observation bears special attention in subsequent analyses. A transformation of the variables might change the outlier status of this year, but to conform to Haavelmo's analysis, I stay with the original data.

With only two variables a scatterplot can aid the identification of outliers. Figure 2.2 is the scatterplot of consumption (consum) by income. This shows a very close linear association between the variables. The observation most distant from the bulk of the points in the upper-right quadrant is the one for 1941. If we imagine a straight line drawn through the points, this last case would fall slightly below the line. Whether this point would seriously affect parameter estimates can be assessed by doing analyses with and without the observation and by comparing estimates.

The sample covariance matrices of consumption and income with and without the 1941 data are

$$\mathbf{S} = \begin{bmatrix} 1889 & \\ 2504 & 3421 \end{bmatrix}, \qquad \mathbf{S}_{(i)} = \begin{bmatrix} 1505 & \\ 1863 & 2363 \end{bmatrix}$$

where $\mathbf{S}_{(i)}$ is the sample covariance matrix removing the ith observation (1941 in this case). When 1941 is dropped, there are substantial drops in the variances and covariance. Does this mean that the 1941 observation has a large effect on all estimates that are functions of the covariance matrix? Not necessarily, since it depends on how these elements combine to form an estimate. For instance, the correlation coefficient is the covariance of two variables divided by the product of their standard deviations. Dropping the 1941 case for consumption and income changes the correlation hardly at all ($r = 0.985$ vs. $r_{(i)} = 0.988$). If the object of study were the correlation, we would conclude that this outlier was not influential. But we cannot generalize from the effect on correlations to the effects on other estimates that are different functions of the elements of $\mathbf{S}$. In Chapter 4 I will examine the influence of this case on other more complex functions of $\mathbf{S}$.

The consumption and income example is atypical in that it contains only two variables. The standard application involves many more variables. Bivariate scattergrams often can identify deviant cases, but they do not always reveal multidimensional outliers (Daniel and Wood 1980, 50–53). Detection of multidimensional outliers is not a fully solved problem. In certain areas, such as single-equation regression analysis, several procedures are available.[9] In factor analysis or general systems of structural equations, the techniques are less developed. One general screening device, however, is to form a $N \times (p + q)$ matrix $\mathbf{Z}$ that contains all of the observed variables written as deviations from their mean values. Then define an $N \times N$ matrix $\mathbf{A} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. The main diagonal of $\mathbf{A}$, called a_{ii}, has several useful interpretations. First, a_{ii} gives the "distance" of the ith case from the means for all of the variables. It has a range between zero and one such that the closer to one the more distant it is, while the closer to zero the nearer is the observation to the means. The $\Sigma_i^N a_{ii}$ equals $p + q$, the number of observed variables. This means that the average size of a_{ii} is $(p + q)/N$ so each a_{ii} can be compared to the average value as a means of judging its

[9]See, e.g., Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), or Bollen and Jackman (1985).

magnitude. The relative size of a_{ii} also can be assessed by examining the univariate distribution of a_{ii} and noting any values that are much larger than the others.

I illustrate this procedure with the data in Table 2.5. The data come from a study that assesses the reliability and validity of human perceptions versus physical measures of cloud cover (Cermak and Bollen 1984). The first three columns of Table 2.5 contain the perception estimates of three judges of the percent of the visible sky containing clouds in each of 60 slides. The slides showed identical views of the sky at different times and days. The judgments range from 0% to 100% cloud cover.

I formed deviation scores for each variable and calculated $\mathbf{A}$ as $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. A stem-and-leaf display of the a_{ii}, the main diagonal of $\mathbf{A}$, is given in Figure 2.3. The sum of the a_{ii}'s is three, which corresponds to the three observed variables. The average value is 3/60, or 0.05, and all a_{ii} values lie in the 0 to 1 range. Two or three outliers are evident in the stem-and-leaf. The highest two a_{ii} have values of 0.301 and 0.307 and correspond to the 40th and 52nd observation. Both of these are over six times larger than the average a_{ii} value of 0.05. The third largest a_{ii} is 0.183, which is over three and a half times larger than the average and is observation 51. A weaker case could be made that the fourth highest value, 0.141, also is an outlier, but I concentrate on the first three.

I begin to assess the influence of the three outliers by calculating the covariance matrix with and without these three cases:

$$\mathbf{S} = \begin{bmatrix} 1301 & & \\ 1020 & 1463 & \\ 1237 & 1200 & 1404 \end{bmatrix}, \qquad \mathbf{S}_{(i)} = \begin{bmatrix} 1129 & & \\ 1170 & 1494 & \\ 1149 & 1313 & 1347 \end{bmatrix}$$

where $\mathbf{S}_{(i)}$ refers to the covariance matrix with the outliers removed and where i now refers to the set of observations 40, 51, and 52. In the consumption and income example all elements of the covariance matrix decreased when the outlier was removed. In this example some elements increase while others drop. For instance, the variance for the first cloud cover variable drops from 1301 to 1129, whereas the covariance for the first and second cloud cover variables increases from 1020 to 1170. The largest change in the correlation matrix for these variables is for the second cloud cover variable. The correlation of the first and second cloud cover variables shifts from 0.74 to 0.90 and that for the third and second variables increases from 0.84 to 0.93 when the outliers are removed. Thus these outliers influence the covariance and correlation matrices. I show in Chapter 7 that the outliers also affect a confirmatory factor analysis of these data.

Table 2.5 Three Estimates of Percent Cloud Cover for 60 Slides

OBS	COVER1	COVER2	COVER3
1	0	5	0
2	20	20	20
3	80	85	90
4	50	50	70
5	5	2	5
6	1	1	2
7	5	5	2
8	0	0	0
9	10	15	5
10	0	0	0
11	0	0	0
12	10	30	10
13	0	2	2
14	10	10	5
15	0	0	0
16	0	0	0
17	5	0	20
18	10	20	20
19	20	45	15
20	35	75	60
21	90	99	100
22	50	90	80
23	35	85	70
24	25	15	40
25	0	0	0
26	0	0	0
27	10	10	20
28	40	75	30
29	35	70	20
30	55	90	90
31	35	95	80
32	0	0	0
33	0	0	0
34	5	1	2
35	20	60	50
36	0	0	0
37	0	0	0
38	0	0	0
39	15	55	50
40	95	0	40
41	40	35	30
42	40	50	40
43	15	60	5

Table 2.5 *(Continued)*

OBS	COVER1	COVER2	COVER3
44	30	30	15
45	75	85	75
46	100	100	100
47	100	90	85
48	100	95	100
49	100	95	100
50	100	99	100
51	100	30	95
52	100	5	95
53	0	0	0
54	5	5	5
55	80	90	85
56	80	95	80
57	80	90	70
58	40	55	50
59	20	40	5
60	1	0	0

With influential cases like these, it is worthwhile determining why they are outliers. The cloud cover estimates for observations 40, 51, and 52 in Table 2.5 are drastically different from one another. For example, observation 52 has estimates of 100%, 5%, and 95%. The slides that correspond to the outliers reveal that the pictures were taken when the scene was obstructed by considerable haze. Under these conditions the judges could not easily estimate the cloud cover, and this led to the differences in estimates. The outliers indicate the inadequacy of this means of assessing cloud cover when hazy conditions are present and suggest that instructions for judging slides under these conditions should be developed.[10]

In general, outliers can increase, decrease, or have no effect on covariances and on the estimates based on these covariances. To determine which of these possibilities holds, researchers need to screen their data for maverick observations and to perform their analysis with and without these cases. A detailed examination of the outliers can suggest omitted variables, incorrect functional forms, clerical errors, or other neglected aspects of a study. In this way outliers can be the most valuable cases in an analysis.

[10] It also could be that the error variance for the cloud cover measures are associated with the degree of haziness, a possibility suggested to me by Ron Schoenberg.

```
Stem-Leaf                                #
  30 17                                  2
  28
  26
  24
  22
  20
  18 3                                   1
  16
  14 1                                   1
  12
  10 56                                  2
   8 2590                                4
   6 123588                              6
   4 445682799                           9
   2 002456                              6
   0 44613333456667778888888888888      29
     ----+----+----+----+----+
Multiply Stem-Leaf by -0.01
```

Figure 2.3 Stem-and-Leaf Display for a_{ii} of Cloud Cover Data

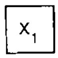
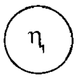
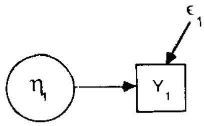
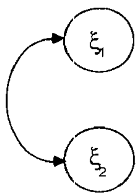
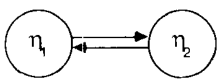
PATH ANALYSIS

As mentioned in Chapter 1, Sewall Wright's (1918, 1921) *path analysis* is a methodology for analyzing systems of structural equations. Contemporary applications emphasize three components of path analysis: (1) the path diagram, (2) decomposing covariances and correlations in terms of model parameters, and (3) the distinctions between direct, indirect, and total effects of one variable on another. I treat each of these in turn.

Path Diagrams

A *path diagram* is a pictorial representation of a system of simultaneous equations. One of the main advantages of a path diagram is that it presents a picture of the relationships that are assumed to hold. For many researchers this picture may represent the relationships more clearly than the equations. To understand path diagrams, it is necessary to define the symbols involved. Table 2.6 provides the primary symbols. The observed variables are enclosed in boxes. The unobserved or latent variables are circled, with the exception of the disturbance terms which are not enclosed. Straight single-headed arrows represent causal relations between the vari-

Table 2.6 Primary Symbols Used in Path Analysis

	rectangular or square box signifies an observed or manifest variable
	circle or ellipse signifies an unobserved or latent variable
	unenclosed variable signifies a disturbance term (error in either equation or measurement) straight arrow signifies assumption that variable at base of arrow "causes" variable at head of arrow
	curved two-headed arrow signifies unanalyzed association between two variables
	two straight single-headed arrows connecting two variables signifies feedback relation or reciprocal causation

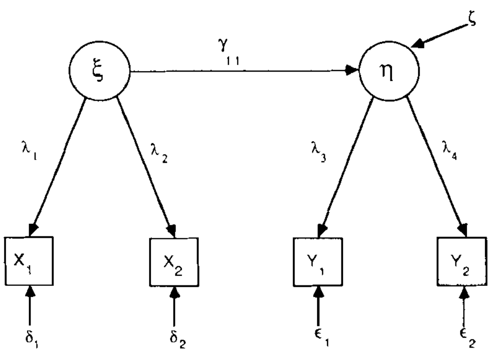


Figure 2.4 An Example of a Path Diagram

ables connected by the arrows. A curved two-headed arrow indicates an association between two variables. The variables may be associated for any of a number of reasons. The association may be due to both variables depending on some third variable(s), or the variables may have a causal relationship but this remains unspecified.

The path diagram in Figure 2.4 is equivalent to the following simultaneous system of equations:

$$\eta = \gamma_{11}\xi + \zeta$$

$$x_1 = \lambda_1\xi + \delta_1, \qquad x_2 = \lambda_2\xi + \delta_2$$

$$y_1 = \lambda_3\eta + \epsilon_1, \qquad y_2 = \lambda_4\eta + \epsilon_2$$

Assuming

$$\begin{array}{cccc}
\mathrm{COV}(\xi,\delta_1) = 0 & \mathrm{COV}(\xi,\delta_2) = 0 & \mathrm{COV}(\xi,\epsilon_1) = 0 & \mathrm{COV}(\xi,\epsilon_2) = 0 \\
\mathrm{COV}(\xi,\zeta) = 0 & \mathrm{COV}(\eta,\epsilon_1) = 0 & \mathrm{COV}(\eta,\epsilon_2) = 0 & \mathrm{COV}(\delta_1,\delta_2) = 0 \\
\mathrm{COV}(\delta_1,\epsilon_1) = 0 & \mathrm{COV}(\delta_1,\epsilon_2) = 0 & \mathrm{COV}(\delta_2,\epsilon_1) = 0 & \mathrm{COV}(\delta_2,\epsilon_2) = 0 \\
\mathrm{COV}(\epsilon_1,\epsilon_2) = 0 & \mathrm{COV}(\delta_1,\zeta) = 0 & \mathrm{COV}(\delta_2,\zeta) = 0 & \mathrm{COV}(\epsilon_1,\zeta) = 0 \\
\mathrm{COV}(\epsilon_2,\zeta) = 0 & & &
\end{array}$$

I purposely have written all of the assumptions for disturbance terms since the same information is explicitly shown in the path diagram. In fact, all of the relationships are represented in the path diagram. For example, the fact that there is no arrow connecting δ_1 and δ_2 or ξ and ϵ_1 is equivalent to the assumption of a zero covariance between these variables. Thus path diagrams are another means of representing systems of equations.

Decomposition of Covariances and Correlations

Path analysis allows one to write the covariance or correlation between two variables as functions of the parameters of the model. One means of doing this is with covariance algebra.[11] To illustrate this, consider the simple model in Figure 2.5. It represents a single latent variable (ξ_1) that has four indicators (x_1 to x_4). All the errors of measurement (δ_i's) are uncorrelated, except for δ_2 and δ_3. The errors of measurement (δ_i's) are assumed to be uncorrelated with ξ_1 and $E(\delta_i) = 0$ for all i.

[11]An alternative more complicated means to decomposition is with the "first law of path analysis." See Kenney (1979) for a discussion of this.

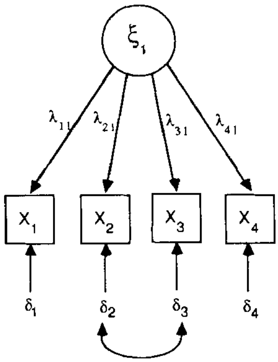


Figure 2.5 Path Diagram of a Single Latent Variable with Four Indicators

The decomposition of the $\text{COV}(x_1, x_4)$ is

$$\begin{aligned}\text{COV}(x_1, x_4) &= \text{COV}(\lambda_{11}\xi_1 + \delta_1, \lambda_{41}\xi_1 + \delta_4)\\ &= \lambda_{11}\lambda_{41}\phi_{11}\end{aligned}$$

The right-hand side of the top equation follows from the equations for x_1 and x_4 defined in the path diagram. This shows that the $\text{COV}(x_1, x_4)$ is a function of the effect of ξ_1 on x_1 and x_4 (i.e., λ_{11} and λ_{41}) and the variance of the latent variable ξ_1.

For complicated models this covariance algebra can be tedious. An alternative is to use matrix algebra to decompose covariances (or correlations) into the model parameters. As an example, consider the covariance matrix for the **x** variables or $\boldsymbol{\Sigma}$. The covariance matrix for **x** is the expected value of **xx**′, where $\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$:

$$\begin{aligned}\mathbf{xx}' &= (\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})'\\ &= (\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\xi}'\boldsymbol{\Lambda}_x' + \boldsymbol{\delta}')\\ &= \boldsymbol{\Lambda}_x\boldsymbol{\xi}\boldsymbol{\xi}'\boldsymbol{\Lambda}_x' + \boldsymbol{\Lambda}_x\boldsymbol{\xi}\boldsymbol{\delta}' + \boldsymbol{\delta}\boldsymbol{\xi}'\boldsymbol{\Lambda}_x' + \boldsymbol{\delta}\boldsymbol{\delta}'\\ E(\mathbf{xx}') &= \boldsymbol{\Lambda}_x E(\boldsymbol{\xi}\boldsymbol{\xi}')\boldsymbol{\Lambda}_x' + \boldsymbol{\Lambda}_x E(\boldsymbol{\xi}\boldsymbol{\delta}') + E(\boldsymbol{\delta}\boldsymbol{\xi}')\boldsymbol{\Lambda}_x' + E(\boldsymbol{\delta}\boldsymbol{\delta}')\\ \boldsymbol{\Sigma} &= \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta\end{aligned}$$

In this case $\boldsymbol{\Sigma}$, the covariance matrix of **x**, is decomposed in terms of the elements in $\boldsymbol{\Lambda}_x$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Theta}_\delta$. As I show in Chapters 4, 7, and 8 the covariances for all observed variables can be decomposed into the model parameters in a similar fashion. These decompositions are important because they show that the parameters are related to the covariances and different parameter values lead to different covariances.

Total, Direct, and Indirect Effects

Path analysis distinguishes three types of effects: direct, indirect, and total effects. The direct effect is that influence of one variable on another that is unmediated by any other variables in a path model. The indirect effects of a variable are mediated by at least one intervening variable. The sum of the direct and indirect effects is the total effects:

$$\text{Total effects} = \text{Direct effect} + \text{Indirect effects}$$

The decomposition of effects always is with respect to a specific model. If the system of equations is altered by including or excluding variables, the estimates of total, direct, and indirect effects may change.

To make these types of effects more concrete, consider the model introduced in the model notation section relating 1960 industrialization to 1960 and 1965 political democracy in developing countries. The equations for this model are in (2.3), (2.10), (2.11), and in the assumptions in the discussion surrounding these equations.

The path model for these equations and assumptions is represented in Figure 2.6. Here ξ_1 is industrialization with indicators x_1 to x_3. The η_1 variable is 1960 political democracy with its four measures, y_1 to y_4, and η_2 is 1965 democracy measured by y_5 to y_8. An example of a direct effect is the effect of η_1 on η_2, that is, β_{21}. A one-unit change in η_1 leads to an expected direct change of β_{21} in η_2 net of ξ_1. There are no mediating variables between η_1 and η_2. The direct effect of ξ_1 on η_2 is γ_{21}, while λ_8 is the direct effect of η_2 on y_5.

To illustrate indirect effects, consider the influence of ξ_1 on η_2. The intervening variable in this case is η_1. A one-unit change in ξ_1 leads to an expected γ_{11} change in η_1. This γ_{11} change in η_1 leads to an expected β_{21} change in η_2. Thus the indirect effect of ξ_1 on η_2 is $\gamma_{11}\beta_{21}$. Following a similar procedure the indirect effect of η_1 on y_7 is $\beta_{21}\lambda_{10}$.

One variable's total effect on another is the sum of its direct effect and indirect effects. For instance, the total effect of ξ_1 on η_2 is

$$\begin{aligned}\text{Total effect} &= \text{Direct effect} + \text{Indirect effects}\\ &= \quad \gamma_{21} \quad + \quad \gamma_{11}\beta_{21}\end{aligned}$$

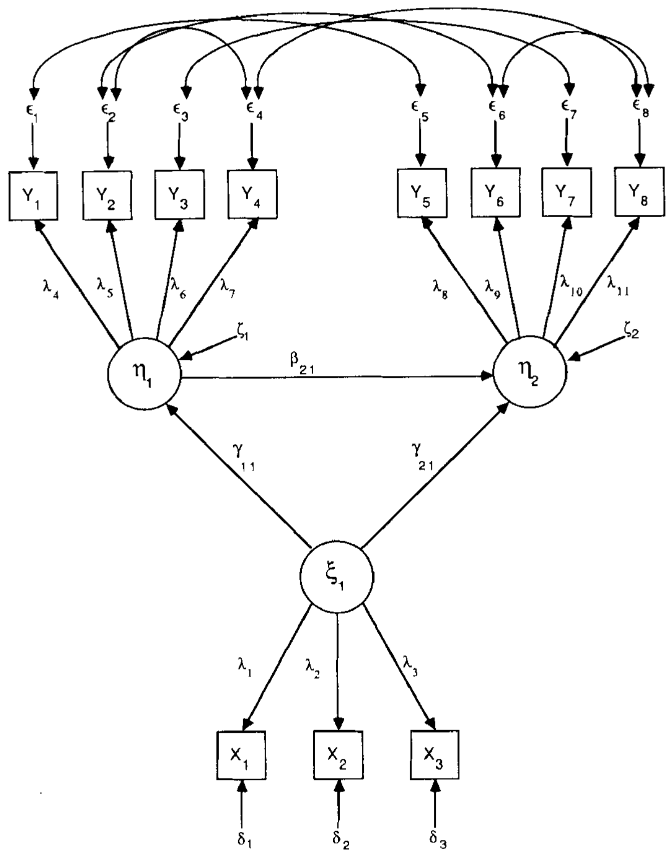


Figure 2.6 Path Diagram of Industrialization and Political Democracy Model

The total effect of ξ_1 on y_8 is

$$
\begin{aligned}
\text{Total effect} &= \text{Direct effect} + \quad \text{Indirect effects} \\
&= \qquad 0 \qquad\quad + (\gamma_{21}\lambda_{11} + \gamma_{11}\beta_{21}\lambda_{11}).
\end{aligned}
$$

Since ξ_1 has no direct effect on y_8, the total effect comprises only indirect effects.

Considering each type of effect leads to a more complete understanding of the relation between variables than if these distinctions are not made. In the typical regression analysis the regression coefficient is an estimate of the direct effect of a variable. If we ignore the indirect effects that a variable may have through other variables, we may be grossly off in the assessment of its overall effect. For example, if in Figure 2.6 we claim that industrialization's (ξ_1) effect on 1965 political democracy (η_2) is γ_{21}, we would overlook a possibly large indirect effect of $\gamma_{11}\beta_{21}$. Similarly, if we claim that industrialization has no effect on 1965 political democracy based on a nonsignificant estimate of γ_{21}, we again would be in error if the indirect effect $\hat{\gamma}_{11}\hat{\beta}_{21}$ is significant. As I show in Chapter 8, matrix algebra provides an easier means of deriving these effects in all types of structural equation models.

To conclude this section, I should mention several misunderstandings about path analysis. One is the belief that path analysis is a methodology only appropriate when reciprocal or feedback relations are absent. This belief seems to stem from many of the early applications of path analysis in the social sciences which were restricted to one-way causation without any feedback. However, path diagrams, decompositions of covariances, or determination of total, direct, and indirect effects are not limited in this way. Indeed, as I mentioned in Chapter 1, Wright pioneered the estimation of nonrecursive models using path analysis.

A second belief is that path analysis deals exclusively with *standardized* regression coefficients. The standardized coefficients are defined as the usual unstandardized coefficient multiplied by the ratio of the standard deviation of the explanatory variable to the standard deviation of the variable it affects (see Chapter 4). To date, many analyses have reported standardized coefficients in path analyses. Path analysis, however, is not restricted to standardized coefficients. In fact, most of the examples I present are with unstandardized coefficients. A third incorrect belief about path analysis is that curvilinear relationships are not possible in path analysis. This too is not true. Just as with linear regression techniques, transformations of variables to capture curvilinear relationships may be used in path analysis. However, as I mentioned earlier, complications arise if latent variables have curvilinear relations with one another (see Chapter 9).

A final misunderstanding is the idea that path analysts are assuming that a path model that fits well proves causation. As I discuss in the next chapter, we can never prove causation with any technique. Rather, the purpose of path analysis is to determine if the *causal inferences* of a researcher are *consistent* with the data. If the path model does not fit the data, then revisions are needed, since one or more of the model assumptions are in error. If the path model is consistent with the data, this does not prove causation. Rather, it shows that our assumptions are not contradicted and may be valid. We only can say "may be valid" because other models and assumptions also may fit the data.

SUMMARY

This chapter provided three tools essential to subsequent chapters of the book. I repeatedly will apply the notation, covariances, and path analysis throughout the book. The next chapter examines the idea of causality that implicitly underlies the various structural equation techniques.