11

Path Modeling

Structural Equation Modeling With Measured Variables

In this chapter, we continue our journey beyond multiple regression and begin discussing structural equation modeling (SEM). This chapter focuses on the technique of path analysis, which can be considered the simplest form of SEM. Because we used path-type models as a way of displaying and understanding regression models throughout Part 1 of this text, this transition to a formal presentation of path modeling should be a natural extension of our work so far. As you will see, many path analyses can be solved using multiple regression analysis, although we will soon begin using specialized structural equation modeling software for both simple and complex path models.

In the final chapter of Part 1, we reviewed one of the difficulties with multiple regression analysis, the fact that we can come to different conclusions about the effects of one variable on another depending on which type of multiple regression we use and which statistics from the analysis we interpret. (If you are beginning the book here, I recommend that you read Chapter 9 as a review of multiple regression.) As you will see, this difficulty is obviated in path analysis and structural equation modeling, where it is natural to focus not only on direct effects but also on indirect and total effects (total effects are the sum of direct and indirect effects). We will use both simultaneous and sequential MR in path analysis, an exercise that will clarify the relation between these two methods. In the process, we will focus more explicitly on explanation, and on the issues of cause and effect. I think that path analysis makes many aspects of multiple regression more understandable, and it is often a better choice for the explanatory analysis of nonexperimental data.

Before we begin, let's deal with a little jargon. The general type of analysis discussed in this part of the book, SEM, is also referred to as analysis of covariance structures, or causal analysis. Path analysis, one form of SEM, is the subject of this and the next two chapters; it may also be considered as a component of SEM. Confirmatory factor analysis (CFA) is another component. More complex forms of SEM are often referred to as latent variable SEM, or simply as SEM. SEM is also sometimes referred to as LISREL analysis, which is actually the first computer program for conducting latent variable SEM and stands for *l*inear structural *rel*ations. We will discuss these and other topics in subsequent chapters, including this and other SEM computer programs. Now we introduce path analysis.

INTRODUCTION TO PATH ANALYSIS

A Simple Model

Let's return to the example we used in Chapter 9, in which we were interested in the effects of Family Background, Ability, Academic Motivation, and Academic Coursework on high school Achievement. For the sake of simplicity, we will focus on only three of the variables: Ability, Motivation, and Achievement. Suppose, then, we are interested in the effects of Motivation on Achievement. Although presumably motivation is manipulable, it is not a variable that you can assign at random, and thus you will probably need to conduct a nonexperimental analysis, as was done in Chapter 9. Intellectual Ability is included in the model to control for this variable. More specifically, we believe that Ability may affect both Motivation and Achievement, and we know that it is important to control for such *common causes* if we are to estimate accurately the effects of one variable on another.

Figure 11.1 illustrates the data we collected. Motivation is a composite of items reflecting academic motivation (student ratings of their interest in school, willingness to work hard in school, and plans for post-high school education); Achievement is a composite of achievement tests in reading, math, science, civics, and writing. We also collected data on Intellectual Ability (a composite of two verbal ability tests), with the notion that ability should be controlled because it may affect both Motivation and Achievement. The curved lines in the figure represent correlations among the three variables. The figure essentially presents the correlation matrix in graphic form. The correlation between Ability and Motivation, for example, is .205. (The data are from the correlation matrix used in Chapter 9.)

Unfortunately, the data as presented in Figure 11.1 do little to inform our question of interest: understanding the effects of Motivation on Achievement. The correlations are statistically

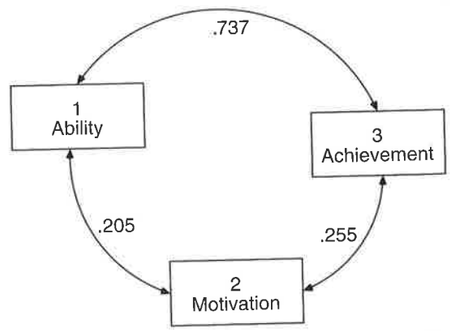

Figure 11.1 Correlations among Ability, Motivation, and Achievement. An "agnostic" model.

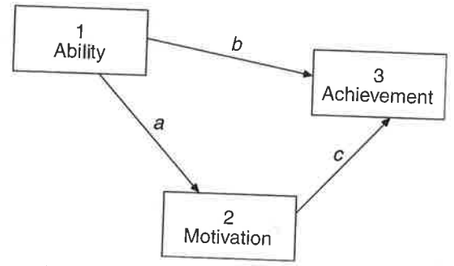

Figure 11.2 Presumed causal structure of the three variables. Note that the assumptions about causal direction were not based on the correlations.

significant, but we have no information on the effects of one on the other. We can think of this figure, then, as an "agnostic" model. In Figure 11.2 we take the first bold step in solving this dilemma by drawing arrows or paths from presumed causes to presumed effects. The purpose of this research was to determine the *effect* of Motivation on Achievement, so it certainly makes sense to draw a path from Motivation to Achievement. Ability was included in the research because we worried that it might *affect* both Motivation and Achievement; therefore, paths drawn from Ability to Motivation and Achievement are the embodiment of this supposition. Our drawing of the paths asserting presumed cause and effect was not so bold after all; it simply made obvious the reasoning underlying our study and the data we collected.

What exactly do these paths mean? They assert what is called a *weak causal ordering*, meaning that the path from Motivation to Achievement does not assert that Motivation directly causes Achievement, but rather that *if* Motivation and Achievement are causally related the cause is in the direction of the arrow, rather than the reverse. Note that we did not use the correlations or the data to make these inferences about causality; instead, our informal causal thinking guided the data we collected and used! Figure 11.2 formalizes our notions of how these three variables are related and thus represents our model of the nature of the relations among these three variables.

The data shown in Figure 11.1 may be used to solve for the paths in the model shown in Figure 11.2. The easiest way to do so is to use the tracing rule: "the correlation between two variables X and Z is equal to the sum of the product of all paths from each possible tracing between X and Z [in Figure 11.2]. These tracings include all possible routes between X and Z, with the exceptions that (1) the same variable is not entered twice per tracing and (2) a variable is not both entered and exited through an arrowhead" (Keith, 1999, p. 82; cf. Kenny, 1979, p. 30). Thus, the correlation between Ability and Achievement (r_{13}) would be equal to path b plus the product of path a times path c: $r_{13} = b + ac$. Two other formulas (for the other two correlations) may be derived: $r_{23} = c + ab$ and $r_{12} = a$. You may wonder why the third equation does not include the tracing bc. The reason is that this tracing would violate the second exception (the same variable was entered and exited through an arrowhead).

We now have three equations and three unknowns (the three paths). If you recall high school algebra, you can use it to solve for the three unknowns:[1]

$$a = r_{12}$$

$$b = \frac{r_{13} - r_{12}r_{13}}{1 - r_{12}^2}$$

$$c = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}$$

(If you don't recall high school algebra, note 1 shows how these three equations were generated.) Substituting the actual correlations in these equations, we calculate

$$a = .205$$

$$b = \frac{.737 - .205 \times .255}{1 - .205^2} = .715$$

$$c = \frac{.255 - .205 \times .737}{1 - .205^2} = .108$$

The solved paths are included in the model in Figure 11.3. The model may be interpreted as demonstrating the effects of Ability and Motivation on Achievement, along with the effects of Ability on Motivation (given several assumptions). The paths shown are the standardized path coefficients and are interpreted in standard deviation units. Thus, the path from Motivation to Achievement of .108 suggests that, given the adequacy of our model, each *SD* increase in Motivation will result in a .108 increase in Achievement.[2]

If this sounds familiar, it should. This type of interpretation is the same as that for standardized regression coefficients. A closer inspection of the formulas above will show striking similarity to those in Chapter 2 for regression coefficients. In fact, these formulas *are* the formulas for standardized regression coefficients. We don't need to use algebra to solve for the paths; we can use good old multiple regression analysis!

To solve for the paths using multiple regression, regress Achievement on Ability and Motivation. The β's from this regression are equal to the standardized paths, calculated previously, from Ability and Motivation to Achievement. The path from Ability to Motivation is estimated through the regression of Motivation on Ability. Relevant portions of the output are shown in Figure 11.4. The first table of coefficients is from the first regression and estimates the paths to Achievement; the second table of coefficients is from the second regression and shows the path to Motivation. Compare the results to those shown in Figure 11.3.

We can use and interpret that printout and model in the same fashion as we previously did with multiple regression. The model thus suggests that Motivation has a moderate effect (using the rules of thumb from Chapter 4) on Achievement, after taking students' Ability into account.[3] Ability, in turn, has a moderate effect on Motivation and a very large effect on Achievement. We can use the rest of the regression output as we have previously. Just as in other forms of MR, the unstandardized regression coefficients—used as estimates of the unstandardized paths—may be more appropriate for interpretation, for example, when the variables are in a meaningful metric. In the present example, the standardized coefficients

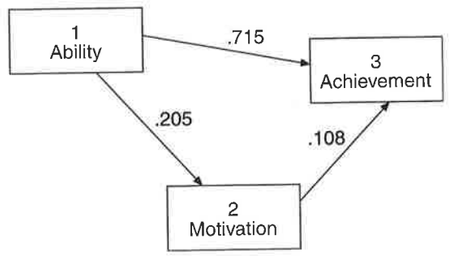

Figure 11.3 We used the data from Figure 11.1 to solve for the paths from Figure 11.2. The paths represent the standardized effect of one variable on another, given the adequacy of the model.

Model Summary

Model	R	R Square	F	df1	df2	Sig. F
1	.745[a]	.554	620.319	2	997	.000

a. Predictors: (Constant), MOTIVATE, ABILITY

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3.075	1.627		-1.890	.059	-6.267	.118
	ABILITY	.477	.014	.715	33.093	.000	.448	.505
	MOTIVATE	.108	.022	.108	5.022	.000	.066	.151

a. Dependent Variable: ACHIEVE

Model Summary

Model	R	R Square	F	df1	df2	Sig. F
1	.205[a]	.042	43.781	1	998	.000

a. Predictors: (Constant), ABILITY

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	36.333	2.089		17.396	.000	32.235	40.432
	ABILITY	.137	.021	.205	6.617	.000	.096	.177

a. Dependent Variable: MOTIVATE

Figure 11.4 Using simultaneous multiple regression to solve the paths.

are probably more interpretable. (You may wonder why the unstandardized and standardized paths from Motivation to Achievement are the same. The reason is because the *SD*s for the two variables are the same.) In addition, we can use the *t*'s and standard errors from the output to determine the statistical significance of the path coefficients, as well as confidence intervals around the paths. The 95% confidence interval around the (unstandardized) path from Motivation to Achievement was .066 to .151.

The model shown in Figure 11.3 is not entirely complete. Conceptually and statistically, it should be clear that the model does not include all influences on Achievement or Motivation. You can no doubt think of many other variables that should affect high school achievement: family background, coursework, homework, and others. And what about effects on Motivation; if Ability only affects Motivation at a level of .205, obviously many influences are unaccounted for. The model shown in Figure 11.5 rectifies these deficiencies by including "disturbances" in the model, symbolized as d1 and d2. Disturbances represent *all other* influences on the outcome variables other than those shown in the model. Thus, the circled variable d2 represents all influences on Achievement other than Ability and Motivation. The disturbances are enclosed in circles or ellipses to signify that they are *unmeasured* variables. We obviously don't measure all variables that affect Achievement and include them in the model; the disturbances, then, are unmeasured, rather than measured variables.

When I say that the disturbances represent all other influences on the outcomes besides the variables in the model, this explanation may ring a bell, as well. You might think that the

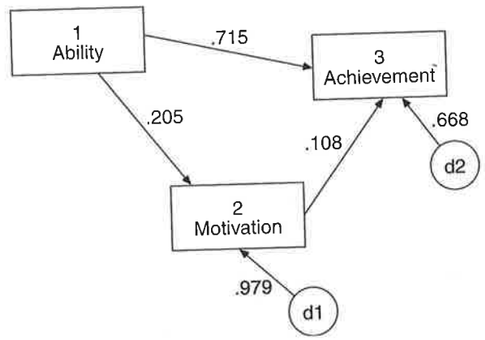

Figure 11.5 The full, standardized, solved model, including disturbances of the presumed effects. Disturbances represent all other, unmeasured variables that affect a variable other than the variables already pointing to it.

disturbances should somehow be related to the residuals, which we at one point described as what was left over or unexplained by the variables in the model. If you had this sense, then reward yourself with a break or a chocolate, because the disturbances are basically the same as the residuals from MR. You have probably encountered instances in research and statistics where two different names are used to describe the same concept; this is another instance of this practice. Although many sources use the term disturbances to describe these other influences (e.g., Bollen, 1989; Kenny, 1979), others continue to use the term *residual,* and others simply refer to these outside influences as *errors.* The paths associated with the disturbances are calculated as the square root of $1-R^2$ ($\sqrt{1-R^2}$) from each regression equation. Focus again on Figure 11.4. For the first equation, the regression of Achievement on Ability and Motivation, R^2 was equal to .554, and thus $\sqrt{1-R^2} = .668$, the value shown for the path from d2 to Achievement. Take a moment to calculate the disturbance for Motivation.

Cautions

With all this talk of cause and effect, you may feel a little queasy. After all, aren't we here breaking the one cardinal rule of elementary statistics: Don't infer causation from correlations? If you are having such misgivings, I first urge you to revisit the short quiz on this same topic in Chapter 1. Second, I point out that, no, we did not infer causality from the correlations. Yes, we had the correlations, but recall that they did not lead to or even enter into our causal inferences. We made the inference of causality when we drew paths from one variable to another, and we drew these paths *without* reference to the correlations. Neither the magnitude nor the sign (positive or negative) of the correlations entered our consideration of cause and effect.

How did we, and how could we, make these inferences of cause and effect? Several lines of evidence can be used to make such inferences and thus to draw the paths. First is *theory.* School learning theories generally include both motivation and ability (or some similar construct) as influences on academic achievement and thus justify the paths from Ability and Motivation to Achievement (Walberg, 1986). And even when formal theory is not available, informal theory can often inform such decisions. Talk to an observant teacher and he or she will tell you that if you can increase a child's level of motivation his or her achievement will likely increase.

Second, we should attend to *time precedence.* As far as we know, causality cannot operate backward in time and so, if we can establish that one variable occurs prior to another in time, it makes it easier to draw the path. This is one reason that longitudinal data are so valued in research; we can feel more confident about inferring cause and effect when our "effect" is measured after our "cause." Yet even with cross-sectional data it is often possible to determine logical time precedence. In the current example, it is well known that ability is a relatively stable characteristic, for most people, from about the time children start school. Logically, then, Ability, stable from an early age, occurs prior to high school motivation and achievement, and thus it makes sense to draw a path from Ability to both Motivation and Achievement. For an even more striking example, consider if we had the variable Sex in our model. For almost everyone (excepting those who have sex change operations!), Sex is stable from conception on. Thus, no matter when Sex is measured, we can feel confident placing it prior to variables that logically occur after conception.

Third, you should have a competent understanding of the relevant research. Previous research may well highlight the proper causal inference. Even if it doesn't—even if you find that other researchers have had to make these same inferences—previous research may help you understand the logic by which others have decided that A affected B rather than B affecting A.

Our fourth and final line of evidence we'll call logic, although it is probably a combination of logic, observation, understanding, and common sense. Go back to the illustration of what I termed informal theory. Teachers observe children every day in their classes; they are keen observers of the process of learning. If you were to ask a teacher, "Which is more likely, an increase in students' levels of motivation affecting their learning or an increase in their learning affecting their motivation?" most would pick the former possibility. You can use the same sort of process to make such inferences. Imagine the ways in which A could affect B, and then imagine the ways in which B could affect A. If you are familiar with the phenomena you are considering, if you have observed them carefully, you will often find it easy to imagine the cause going in one direction but may require mental gyrations to imagine it going in the other. This logical exercise, then, will often suggest that one direction of causation is much more plausible than the other.

Again, these lines of evidence are how we make such inferences of cause and effect. Once we have made those inferences, the correlations merely provide fuel for our calculations.

More formally, three conditions are necessary before we can make a valid inference of causality (see Kenny, 1979, or Kline, 2011, for additional discussion of these conditions; for a considerably expanded discussion of the concept of causality, see Pearl, 2009; 2011). First, there must be a relation between the variables being considered. If two variables are unrelated, then they are also *causally* unrelated. This condition is generally satisfied by the presence of a correlation between the variables (although there are exceptions). Second, and as already discussed, the presumed cause must have time precedence over the presumed effect. Causality does not operate backward in time. Third, the relation between the variables must be a true, rather than a spurious, relation. This is the hardest condition to satisfy and gets to the heart of what we have been calling the problem of omitted common causes. We will delve into this problem more deeply in the next chapter, but for now simply note that this condition means that all common causes are taken into account. Given that these three conditions are satisfied, it is perfectly reasonable to make an inference of cause and effect. What makes nonexperimental research so interesting and challenging is that we can often be very confident that we have satisfied these three conditions but never completely sure. (As it turns out, however, we can never be sure in experimental research either.)[4]

Just to make sure we are all on the same page, let's be completely clear as to what we mean by cause. When we say one variable "causes" another, we do *not* mean that one variable

directly and immediately results in change in another. When we say, for example, that smoking causes lung cancer, we do not mean that every person who smokes will necessarily and quickly develop lung cancer. What we mean is that if you smoke you will, as a result of smoking, increase your probability of developing lung cancer. The term cause is thus a probabilistic statement.

Jargon and Notation

I've been sneaking some of the jargon of SEM into the chapter as we introduce path analysis. Before we move to an expanded example, let's spend a little time going over such jargon so that it will be familiar. I have already noted that the variables representing other influences from outside the model are often called disturbances in path analysis, although many researchers use the term with which you are already familiar, residuals. In addition, I have noted that variables that we wish to symbolize but which we have not measured (unmeasured variables) are generally enclosed in circles or ovals. In contrast, measured variables, variables that we have measured in our data, are generally enclosed in rectangles. Paths or arrows represent influences from presumed cause to presumed effect, whereas curved, double-headed arrows represent correlations without an inference of causality.

Recursive and Nonrecursive Models

The models shown in Figures 11.2 and 11.3 are called *recursive* models, meaning that paths, and presumed causes, go in one direction only. It is also possible to have feedback loops in a model, to specify that two variables affect each other in a reciprocal fashion. Such models are termed *nonrecursive;* an example is shown in Figure 11.6, where Variable 2 is assumed to both affect (path *c*) and be affected by Variable 3 (path *d*). You cannot solve for the equations for nonrecursive models using the tracing rule, although you can generate the correct equations using the first law (see Note 2). Likewise, nonrecursive models cannot be estimated through multiple regression (you can estimate such models with MR, but the results will be incorrect). It is possible to estimate nonrecurive models using specialized SEM software or through a method called two-stage least squares regression, although such estimation is often tedious (and, as we will see momentarily, *this* model could not be estimated). It is tempting, especially for those new to SEM, to solve difficult questions of presumed cause and effect by deciding that such effects are reciprocal. Can't decide whether Motivation affects Achievement or Achievement affects Motivation? Draw paths in both directions! Generally, however, this is equivocation rather than decision. Nonrecursive models may require additional

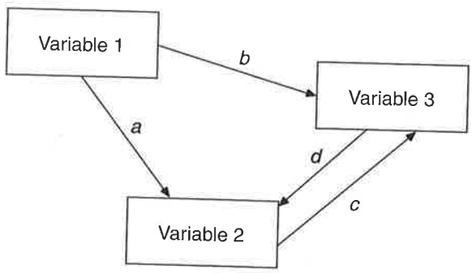

Figure 11.6 Nonrecursive model. The model is also underidentified and cannot be solved without additional assumptions.

constraints to avoid underidentification (see below) and, in my experience, often end up suggesting that the effect is indeed in the direction we would have guessed had we done the difficult work of making such decisions. I am not suggesting that you develop a cavalier attitude toward making decisions about the correct direction of causality; it often requires tough work and deep thought. Instead, I am arguing that you should not try to avoid this work by defaulting to nonrecursive models. Save such models for those instances when you have real, substantive questions about causal direction or when effects really appear to go in both directions. Some authors (e.g., Kenny, 1979) refer to recursive models as hierarchical models and nonrecursive models as nonhierarchical, but such usage may be confusing because sequential regression is also often termed hierarchical regression.

Identification

The model shown in Figure 11.3 is also a *just-identified* model. In a simplistic sense, what this means is that we had just enough information to estimate the model. Focus again on the Figures 11.1 through 11.3. We had three unknowns (the three paths in Figure 11.2), and we solved for these three paths using the three correlations from Figure 11.1. We had just enough information to solve for the paths. In addition to being a nonrecursive model, the model shown in Figure 11.6 is an *underidentified* model. For this model, we still have three correlations, but we now have four paths that we need to estimate. Unless we make some additional assumptions (e.g., assuming that paths *d* and *c* are equal), we cannot solve for the paths in this model.

The model shown in Figure 11.7, in contrast, is *overidentified.* For this model, we have more correlations than paths. The result is that we could, in fact, develop two separate sets of equations to solve for paths *a* and *b.* Consider the three equations generated from the tracing rule:

$$r_{13} = b \quad r_{12} = a \quad r_{23} = ab.$$

Using these equations to solve for a (and substituting for b), for example, we could generate the equations $a = r_{12}$ and $a = r_{23}/r_{13}$. And for b, $b = r_{13}$ and $a = r_{23}/r_{12}$ At first blush, the possibility of calculating two different estimates of each path might seem a problem. But consider for a minute what it would mean if our two estimates of the same path were very close to one another versus considerably divergent? Wouldn't you be more likely to believe a model in which you could estimate a path several different ways and always get the same result? We won't explore this topic in any greater depth right now but will return to it later. In the meantime, simply recognize that overidentified models are not problematic, but, rather, overidentification may help us evaluate the quality of our models.

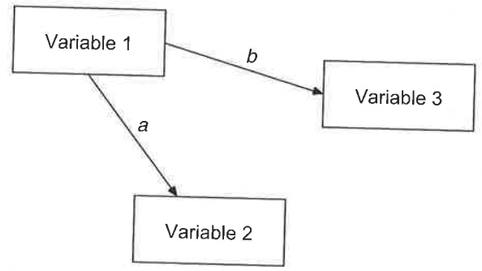

Figure 11.7 Overidentified model. The paths can be estimated more than one way.

This discussion has been a necessary simplification of the topic of identification, which can be much more complex than has been presented here. For example, it is possible for portions of a model to be overidentified and other portions to be underidentified. The primary rule presented for determining identification—comparing the number of correlations to the number of unknown paths—is really more of a necessary but insufficient condition for identification. Nevertheless, this rule generally works well for the simple path models of the type presented in this and the next chapter. For a more detailed discussion of the topic of identification with simple or complex models, see Bollen (1989).

Exogenous and Endogenous Variables

In SEM, the presumed causes (e.g., Ability in Figure 11.3) in a model are often referred to as *exogenous* variables. In medicine or biology, exogenous means "having a cause external to the body" (Morris, 1969, p. 461). An exogenous variable has causes outside the model or not considered by the model. Or, more simply, exogenous variables are ones that have no arrows pointing toward them. In contrast, variables that are affected by other variables in the model, variables that have arrows pointed toward them, are termed *endogenous* variables (meaning, loosely, from within). In Figure 11.3, Motivation and Achievement are endogenous variables.

Measured and Unmeasured Variables

In the discussion of disturbances, I noted that we generally symbolize unmeasured variables in path models by enclosing them in circles or ellipses. Unmeasured variables are variables that we wish to include in a path model, but we have no measures of these variables in our data. For now, the only unmeasured variables we will deal with are disturbances, but in later chapters we will focus on other types of unmeasured variables. Unmeasured variables are also known as *latent* variables or *factors*.

Variables enclosed in rectangles are measured variables for which we have actual measures in our data. These include all sorts of items, scales, and composites. Indeed, all the variables we have discussed so far in this book, with the exception of disturbances and residuals, are measured variables. Measured variables are also known as *manifest* or *observed* variables.

A MORE COMPLEX EXAMPLE

Now that you have a handle on the basics of path analysis, let's expand our example to a more realistic level. We will now focus on the effects of Family Background characteristics, Ability, Motivation, and Academic Coursework on High School Achievement. These are, then, the same data and the same example from Chapter 9, but in path analytic form. The comparison of the results of the path analysis to the results for the different forms of multiple regression will be instructive and help illustrate important concepts about both methods.

Steps for Conducting Path Analysis

Here are the steps involved in conducting a path analysis (Kenny, 1979; Kline, 2011).

Develop the Model

The first step in path analysis is to develop and draw the model based on formal and informal theory, previous research, time precedence, and logic. Figure 11.8 shows my model, or theory, of how these variables are related to one another. School learning theories consistently include variables reflecting Ability (e.g., ability, aptitude, previous achievement), Motivation

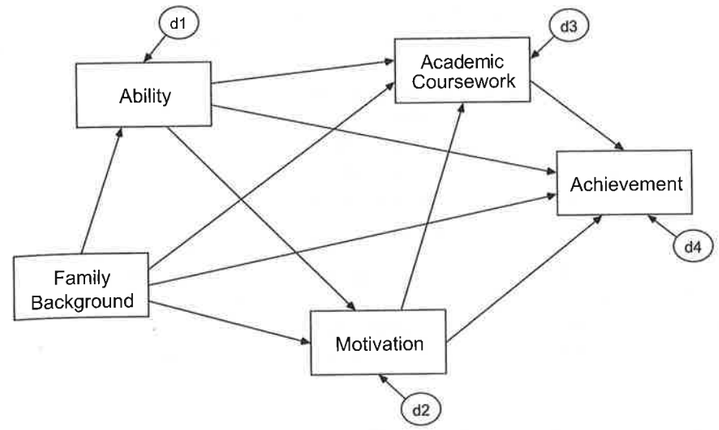

Figure 11.8 Model of the effects of Family Background, Ability, Motivation, and Academic Coursework on Achievement.

(internal motivation, perseverance), and Coursework (quantity of instruction, time spent learning, opportunity to learn) as influences on learning and achievement (e.g., Walberg, 1986). School learning theory, therefore, supports our drawing paths from Ability, Motivation, and Coursework to Achievement. You can probably easily justify these paths in other ways, as well.

Family Background is basically a background variable. By this I mean that it is included in the model because it seems needed to make the model valid (i.e., I think it may be a common cause of some of the variables and Achievement), but I'm not really interested in its effects on any of the other variables in the model. The fact that I consider this a background variable is not, however, justification for placing it first in the model. The likelihood that Family Background occurs before the other variables in time can be used to draw such paths, however, and you may find that the notion of *background variable* often is related to time. In the present case, Family Background is a parent variable, and most of its components—parents' level of education, occupational status—were likely in place, for many families, before children were even born. Even in cases in which parents were still in school or not yet employed when their children were born, time precedence would seem to flow from Family Background to the other variables in the model. Think about it: is it more likely that parents' SES will affect their child's ability (or motivation, etc.) or that a child's ability will affect his or her parents' SES? I suppose the second option is possible (children's ability affecting parents' SES), but it requires some mental gyrations to come up with plausible scenarios. Such reasoning may be used to draw paths from Family Background to each of the other variables in the model.

Time precedence, along with previous research, may also be used to justify the paths from Ability to each subsequent variable in the model. Ability, intelligence, or academic aptitude is relatively stable from an early elementary level on, and there is ample evidence that Ability affects many aspects of life and schooling, from Motivation to Achievement (Jensen, 1980, 1998).

This leaves the path from Motivation to Coursework. Imagine two high school students of equal ability and background. It is not hard to imagine one student taking a relatively easy mix of courses in high school and the other taking courses like pre-calculus, physics,

and advanced English. Academic Motivation—the desire to work hard and persevere in school, the expectation that schooling and what is learned in school will be important for the future—is likely a key difference between these students. Many of you can probably think of such examples in your own family, siblings or children who were highly motivated taking tough courses versus others just getting by. In essence, it makes a great deal of sense to posit that students with high levels of academic motivation will, other things being equal, take a tougher mix of academic courses than will students with lower levels of motivation. (Keith and Cool, 1992, further bolstered this time precedence by measuring Motivation 2 years prior to Coursework.)

This reasoning justifies the directions of the paths in the model, but what about the *variables* in the model? In particular, are there variables that should be included in the model that have not been included? That is, have I neglected an important common cause? Are there variables in the model that are unnecessary? I will postpone in-depth discussion of these issues until the next chapter. For now, I simply note that theory and previous research can help answer these questions, as well.

Check the Identification Status of the Model

Make sure that the model is either just-identified or overidentified so that the model may be estimated. The model shown in Figure 11.8 is just-identified. The correlation matrix includes 10 correlations, and there are 10 paths to be solved for. The model appears to be just-identified and can probably be estimated.

Measure the Variables in the Model

We next need to decide how to measure the variables in the model. This may mean selecting tests and items designed to measure the constructs of interest and then administering these measures to a sample of participants. When using existing data, such as the NELS data, this may mean seeing if items that measure the variables of interest have already been administered to a sample of participants. In the present case, the variables in the model were already measured in the High School and Beyond data set; the authors selected items and composites to measure these constructs.

Estimate the Model

Our next step is to estimate the model. We are currently discussing how to estimate such models using multiple regression analysis; in subsequent chapters we will learn how to estimate such models using SEM software. To estimate the paths to Achievement using MR, we regress Achievement on Family Background, Ability, Motivation, and Academic Coursework. Partial results of this regression are shown in Figure 11.9. The *b*'s and β's from the regression are the estimates of the unstandardized and standardized path coefficients, respectively, from each variable to Achievement. The R^2 is used to calculate the path from the disturbance (d4) to Achievement: $\sqrt{1-R^2} = \sqrt{1-.629} = .609$.

The paths to Academic Coursework are estimated by regressing Courses on Family Background, Ability, and Motivation, and the path from d3 to Coursework is estimated from the R^2 from that regression ($R^2 = .348$). Results from this regression are shown in Figure 11.10. The paths to Motivation are estimated from the regression of Motivation on Family Background and Ability, and the path from Family Background to Ability is estimated by the regression of Ability on Family Background. The relevant regression results are shown in Figure 11.11.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.793[a]	.629	.627	6.103451

a. Predictors: (Constant), COURSES, FAM_BACK, MOTIVATE, ABILITY

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.434	1.692		3.803	.000	3.114	9.753
	FAM_BACK	.695	.218	.069	3.194	.001	.268	1.122
	ABILITY	.367	.016	.551	23.698	.000	.337	.398
	MOTIVATE	1.26E-02	.021	.013	.603	.547	-.028	.054
	COURSES	1.550	.120	.310	12.963	.000	1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 11.9 Using simultaneous regression to estimate the paths to Achievement.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.590[a]	.348	.346	1.617391

a. Predictors: (Constant), MOTIVATE, FAM_BACK, ABILITY

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3.661	.433		-8.454	.000	-4.511	-2.811
	FAM_BACK	.330	.057	.165	5.827	.000	.219	.442
	ABILITY	4.99E-02	.004	.374	13.168	.000	.042	.057
	MOTIVATE	5.34E-02	.005	.267	10.138	.000	.043	.064

a. Dependent Variable: COURSES

Figure 11.10 Estimating the paths to Academic Coursework through simultaneous multiple regression.

Figure 11.12 shows the path model with all the standardized path coefficients added. You should compare the model to the regression results to help you understand where each path came from, including those from the disturbances.

Interpretation: Direct Effects

So, what do these findings tell us? If you focus first on the paths to Achievement, you will see these findings and their interpretation are the same as those from the simultaneous multiple regression of Achievement on these four variables in Chapter 9. Ability and Academic Coursework each had a strong effect on Achievement (.551 and .310, respectively), whereas Family Background had a small, but statistically significant effect (.069). As in the simultaneous

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.235[a]	.055	.053	9.729581

a. Predictors: (Constant), ABILITY, FAM_BACK

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	39.850	2.279		17.488	.000	35.379	44.322
	FAM_BACK	1.265	.339	.127	3.735	.000	.601	1.930
	ABILITY	.101	.023	.152	4.495	.000	.057	.146

a. Dependent Variable: MOTIVATE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.417[a]	.174	.173	13.640426

a. Predictors: (Constant), FAM_BACK

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	100.000	.431		231.831	.000	99.154	100.846
	FAM_BACK	6.255	.432	.417	14.494	.000	5.408	7.102

a. Dependent Variable: ABILITY

Figure 11.11 Estimating the paths to Motivation and Ability.

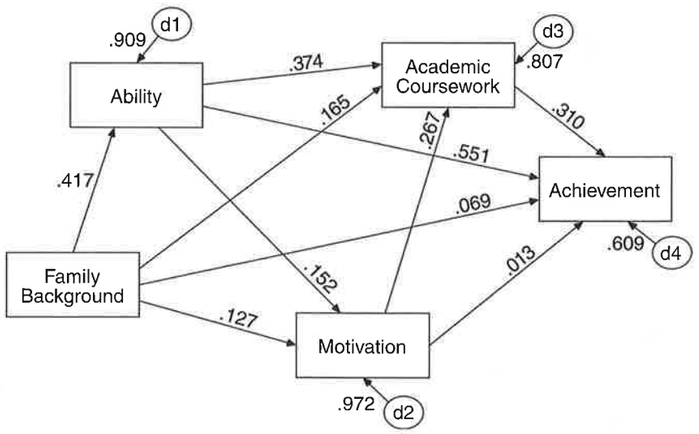

Figure 11.12 Solved model explaining Achievement, showing all standardized paths and disturbances.

regression of these same data in Chapter 9, the effect of Motivation on Achievement was small and not statistically significant.

The path model includes much more than this single simultaneous regression, however, because it also includes information about the effects *on* Coursework, Motivation, and Ability. Which of these variables affect the courses students take in high school? As hypothesized (and given the adequacy of the model), students' level of Academic Motivation had a strong effect on Coursework (.267); students who are more motivated take a more academic mix of courses than do students with lower levels of motivation. The largest effect on Coursework was from Ability (.374); more able students also take more academic courses in high school. Finally, Family Background also had a moderate effect on Coursework (.165), meaning that students from more advantaged backgrounds are more likely to take academic courses in high school than are students from less advantaged backgrounds.

The solved model also speaks to the extent to which Family Background and Ability affect Motivation; higher levels of both Ability and Family Background lead to higher levels of Academic Motivation. In addition, students from more advantaged backgrounds also show higher levels of Ability.

As an aside, notice the paths from the disturbances to each of the endogenous variables. As a general rule, these get smaller the farther to the left in the model. Don't read too much into this phenomenon. Achievement has four paths pointing toward it, four variables explaining it in the model, whereas Ability has only one explanatory variable (Family Background) pointing toward it. Other things being equal, it is natural that our model should explain more of the variance of Achievement than Ability, and thus the paths from the disturbances from Achievement should be smaller.

Indirect and Total Effects

The model (Figure 11.12) includes other information, beyond what we would get in the usual MR, as well (e.g., Chapter 9). The results of this analysis suggest that Motivation affects Coursework, which in turn affects Achievement. This makes sense: more motivated students take more academic courses in high school, and this coursework, in turn, improves their achievement. Thus, although Motivation has little direct effect on Achievement, it does have an indirect effect, through Coursework. In fact, we can easily calculate this indirect effect: multiply the path from Motivation to Coursework times the path from Coursework to Achievement ($.267 \times .310 = .083$), which is the indirect effect of Motivation on Achievement through Coursework. We can also add the direct and indirect effects to determine the *total* effect of Motivation on Achievement ($.083 + .013 = .096$).[5]

It is slightly more complex to calculate the indirect and total effects of Ability or Family Background, because the farther back you go in the model, the more possible indirect effects there are. To calculate the indirect effect of Ability on Achievement, for example, you would need to calculate the indirect effect through Coursework ($.374 \times .310 = .116$), Motivation ($.152 \times .013 = .002$), and both Motivation and Coursework ($.152 \times .267 \times .310 = .013$). These indirect effects are then summed to calculate the total indirect effect, .131, and added to the direct effect (.551) to calculate the total effect, .682. Table 11.1 shows the standardized direct, indirect, and total effects for each variable on Achievement. Calculate the indirect and total effects of Family Background on Achievement to see if your results match mine. Note also that there are no indirect effects for Coursework on Achievement. This is, of course, because our model includes no intervening variables between Coursework and Achievement. If it did, there would be indirect effects for Coursework as well.

Table 11.1 Standardized Direct, Indirect, and Total Effects of School Learning Variables on High School Achievement

Variable	*Direct Effect*	*Indirect Effect*	*Total Effect*
Academic Coursework	.310	—	.310
Motivation	.013	.083	.096
Ability	.551	.131	.682
Family Background	.069	.348	.417

Using Sequential Regression to Estimate Total and Indirect Effects

Recall that in Part 1 of this book we focused on differences in findings from simultaneous (or forced entry) and sequential (hierarchical) regression. I noted at the time that the reason for this difference is that simultaneous regression focuses on direct effects, whereas sequential regression focuses on total effects. We have seen in this chapter that the *b*'s and β's from simultaneous regression may be used as estimates of the direct effects in path analysis. Figure 11.13 shows some of the output for the sequential regression of Achievement on the variables in the school learning model, reproduced from Chapter 9. The figure shows the table of coefficients, with the variables entered into the equation based on their order of appearance in the model; that is, the first (exogenous) variable (Family Background) was entered first, followed by Ability, and so on. Focus on the standardized coefficients, β's, as each variable is added to the model; these coefficients are in italic boldface in the figure. Compare these coefficients to the total effects shown in Table 11.1 and you will see that they are the same, within errors of rounding. Thus, sequential regression may be used to estimate the *total effects* of each variable on the outcome for a path model. To do so, regress the endogenous variable of interest on each presumed cause in the order of their appearance in the model. The β for the variable entered at each step is the estimate of the variable's *total standardized effect* on the endogenous variable. The *b* for the variable entered at each step is the estimate of the variable's total unstandardized effect. If you are interested in the statistical significance of the total effects, however, you need to correct the degrees of freedom, using the value with all variables in the model. That is, look up the statistical significance of the *t*'s using 995 *df* (total *N*—*k*—1), rather than the *df* from each equation. Using this method, we can calculate the indirect effects via simple subtraction: we subtract the direct effect from the total effect to estimate the total indirect effects of each variable on the outcome. Try this subtractive method to calculate the indirect effects in Table 11.1. (To calculate the standard errors, confidence intervals, and statistical significance of indirect effects you will need to do a little hand calculation [Baron & Kenny, 1986]. See the discussion of mediation in Chapter 8. See also Kris Preacher's Web page on mediation mentioned in that chapter: www.quantspy.org/sobel. Alternatively, you can estimate the model with a SEM program, which will calculate standard errors of direct, indirect, and total effects.)

Note we could also calculate the total effects of each variable on each of the other endogenous variables in the model (in addition to their effects on Achievement). To estimate the total effects of each variable on Coursework, for example, we sequentially regress Coursework on Family Background, followed by Ability, and followed by Motivation. The coefficient for the variable entered at each step equals its total effect on Coursework. The coefficients for the final step in the multiple regression equal the direct effects for each variable on Coursework. We can calculate the indirect effects by subtracting the direct from the total effect for each variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.000	.288		173.873	.000	49.436	50.564
	FAM_BACK	***4.170***	.288	***.417***	14.494	.000	3.605	4.735
2	(Constant)	4.557	1.559		2.923	.004	1.498	7.617
	FAM_BACK	1.328	.232	.133	5.729	.000	.873	1.782
	ABILITY	***.454***	.015	***.682***	29.416	.000	.424	.485
3	(Constant)	.759	1.766		.430	.667	-2.706	4.224
	FAM_BACK	1.207	.231	.121	5.221	.000	.753	1.661
	ABILITY	.445	.015	.667	28.768	.000	.414	.475
	MOTIVATE	***9.53E-02***	.021	***.095***	4.439	.000	.053	.137
4	(Constant)	6.434	1.692		3.803	.000	3.114	9.753
	FAM_BACK	.695	.218	.069	3.194	.001	.268	1.122
	ABILITY	.367	.016	.551	23.698	.000	.337	.398
	MOTIVATE	1.26E-02	.021	.013	.603	.547	-.028	.054
	COURSES	***1.550***	.120	***.310***	12.963	.000	1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 11.13 Using sequential multiple regression to estimate the total effects of each variable on Achievement. The indirect effects are then calculated through subtraction (total–direct).

Even with only five variables in the path model, it soon becomes tedious to solve for indirect and total effects directly, that is, by multiplying and summing paths. There are several possible shortcuts for doing such calculations. The one we have illustrated here—using sequential regression to estimate total effects and then calculating indirect effects by subtraction—is one of the easiest and has the advantage of illuminating the previously puzzling relation between sequential and simultaneous regression. The reason simultaneous and sequential regression tell different stories is because they focus on different questions; simultaneous regression focuses on direct effects, whereas sequential regression focuses on total effects. I hope the method also illustrates the importance of proper order of entry in sequential regression. If you wish to interpret sequential regression results in a causal fashion, you must enter the variables in their proper causal order.

It should be clear that this method of estimating total and indirect effects *does* work, but it may not be clear *why* it works. Recall that for the next to last variable in the causal chain (Coursework) the direct effects were equal to the total effects. The reason, of course, is there are no intervening or mediating variables between Coursework and Achievement and thus no possible indirect effect. The total and direct effects for Coursework on Achievement are the same. All the effect of one variable on another, then, is a direct effect *when there are no intervening variables.* It then stands to reason that one way of calculating total effects is to remove intervening variables.

In essence, what we have done with our sequential regression is to temporarily remove the intervening variables. Focus on Figures 11.14 through 11.16. The first step in the sequential regression, in which Achievement was regressed on Family Background, estimates the model shown in Figure 11.14. In this model, all intervening variables between Family Background and Achievement are removed. The total effect of Family Background remains the same whether there are no intervening variables or whether there are three, or even 30, intervening variables; the total effects *are always the same.* Therefore, when we estimated *this* model, with the intervening variables removed, the direct effects and total effects are the same. The regression coefficient from this regression (.417) can then be used as an estimate of the total effect for the full model with all intervening variables. Figure 11.15 removes the intervening

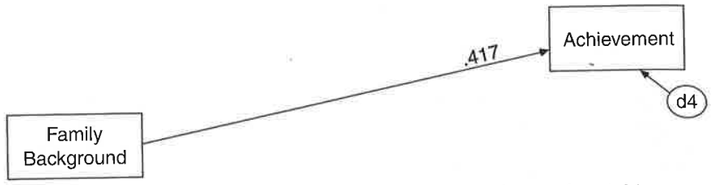

Figure 11.14 "Model" used to estimate the total effect of Family Background on Achievement.

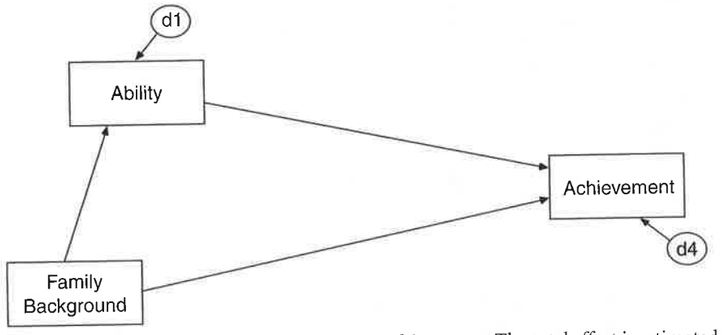

Figure 11.15 Estimating the total effect of Ability on Achievement. The total effect is estimated by the β (or *b*) for the variable added at this stage of the sequential regression.

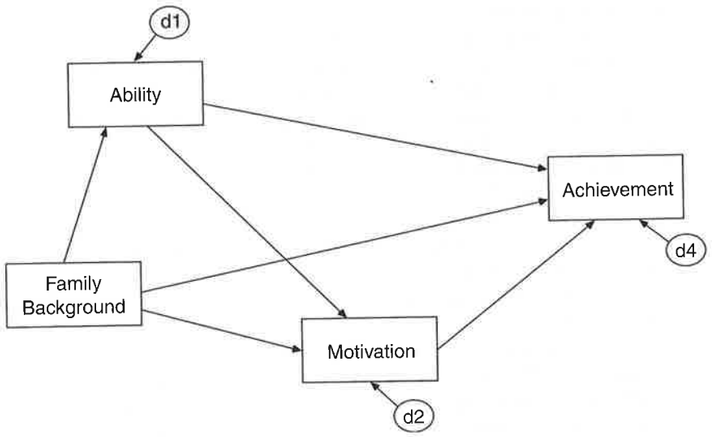

Figure 11.16 Estimating the total effect of Motivation on Achievement.

variables between Ability and Achievement. The second step in the sequential regression, in which Achievement is regressed on Family Background and Ability, operationalizes the model in Figure 11.15, and because there are no intervening variables between Ability and Achievement, the regression coefficient for Ability estimates the total effect of Ability on

Achievement. Finally, the model shown in Figure 11.16, the third step in the sequential regression, provides the estimate of the total effect of Motivation on Achievement.

Interpretation

Let's take a few minutes to interpret these findings and, at the same time, further understand the relation between simultaneous and sequential regression. Focus on Motivation in Figure 11.12. The path model and Table 11.1 suggest that Motivation's effects on Achievement are primarily indirect, not direct. Motivation influences Achievement by influencing the courses students take in high school. Highly motivated students take more academically oriented courses, and these courses, in turn, improve their Achievement. In contrast, Ability's effects on Achievement are primarily direct. A portion of the effect of Ability is indirect, through Motivation and Coursework—more able students are more highly motivated and take more academic coursework, on average, than less able students—but the majority of the effect is direct: more able students also have higher academic Achievement. Again, the simultaneous regressions focused on direct effects and the sequential regressions focused on total effects.

I hope this discussion has illustrated some of the heuristic beauty of path models. They allow us to focus on both direct and indirect effects. Indirect effects, also known as mediating effects, are often vital for understanding how an influence comes about. *How* does Motivation affect Achievement? One important way is by influencing the courses students choose to take in high school. More motivated students take more academic coursework, and this coursework raises achievement. We generally miss understanding these indirect effects when we analyze our data with ordinary MR without path models. When you conduct path analysis, make sure to calculate and interpret all three types of effects. When you find a direct effect and wonder how it comes about, try incorporating several plausible mediating variables in a path model to see if you can understand how these effects happen. Suppose you find, for example, that physical abuse affects children's later social status. You may wonder whether these children's social behaviors (e.g., aggression) mediate, and thus partially explain, this effect. That is, are abused children more likely to be aggressive, with the aggression leading to a reduction in their subsequent social status (Salzinger, Feldman, Ng-Mak, Mojica, & Stockhammer, 2001)?

Path analysis has other advantages over multiple regression. A figure often makes it more obvious than does a table of regression coefficients exactly what are the presumed causes and the presumed effects. I think that the obviousness of the figural, causal assumptions in path analysis makes it more likely that the researchers will consider causal assumptions, as well as the basis for making these assumptions (theory and previous research). If nothing else, the drawing of the path model is at least an informal theory of cause and effect. As already discussed, path analysis makes use of the different stories told by simultaneous and sequential regression. For these reasons, I believe that path analysis (and SEM) is often the best method of analysis for nonexperimental research.

SUMMARY

We have covered a lot of material in this chapter, and I hope the chapter has both covered new ground and made clear some loose ends from our adventures in MR. This chapter formally introduced path analysis, which is the simplest form of structural equation modeling, or SEM.

We introduced the chapter with a simple model involving Ability, Motivation, and Achievement. Our initial, agnostic model simply showed the correlations among these three

variables, a less than satisfying solution since it did not inform our research question of interest, which was understanding the influence of Motivation on Achievement. Thinking through our research interest and using a combination of theory, logic, and previous research, we were able to make some general causal statements: (1) if Motivation and Achievement are causally related, Motivation affects Achievement, rather than the reverse, and (2) Ability may affect both Motivation and Achievement. These statements, a weak causal ordering, were translated into a path model in which Ability was assumed to affect both Motivation and Achievement and Motivation was assumed to affect Achievement. The correlations, notably, were not used to draw the paths. We now had three unknowns (three paths) and three pieces of data (the correlations), and through the use of algebra we were able to generate equations for and solve for the paths.

Although we can solve for the paths using algebra, for simple recursive models the paths are equal to the standardized or unstandardized coefficients from a series of simultaneous regressions. For the three-variable model, we regressed Achievement on Ability and Family Motivation, with the β's providing estimates of the standardized paths from Ability and Motivation (or the *b*'s estimating the unstandardized paths). A second regression of Motivation on Ability provided the estimate of the path from Ability to Motivation. The influences of the disturbances (or residuals) were estimated by $\sqrt{1-R^2}$ from each regression equation. Disturbances represent all other influences on a variable besides the variables in the model, and were symbolized by variables enclosed in circles or ovals.

What evidence was used to make the inferences of causality? It was not the correlations. Instead, we focused on formal and informal theory, time precedence, an understanding of the phenomenon being studied, and logic. At a more formal level, three conditions are required to make a valid inference of cause and effects: there must be a functional relation between the variables, the cause must precede the effect in time (either actually or logically), and the relation must be nonspurious.

We dealt with some jargon you are likely to encounter in path analysis. Measured variables, those measured in your research, are symbolized by rectangles. Unmeasured, or latent variables, are symbolized by circles or ovals. Disturbances represent unmeasured variables not considered in the model; disturbances may also be referred to as residuals or errors. Recursive models have arrows flowing in only one direction, whereas nonrecursive models have feedback loops, or arrows pointing in two directions. Just-identified models are those for which we have just enough information to solve for the paths, and overidentified models are those for which we have more information than we need and can thus estimate some of the paths in more than one way. Underidentified models are those for which we have more paths than we have information to estimate the paths; they are therefore not solvable without the addition of extra constraints. Exogenous variables are presumed causes, variables with no paths pointing towards them. Endogenous variables are presumed effects; they have paths pointing to them in the model. Most of this jargon is summarized in Figure 11.17.

We conducted a path analysis using the data from Chapter 9, where the data were used to highlight the differences in findings from simultaneous and sequential regression. We developed a model of the effects of Family Background, Ability, Motivation, and Coursework on Achievement based on theory, time precedence, previous research, and logic. Paths and disturbances were estimated via a series of simultaneous multiple regressions. Given the accuracy of the model, the results suggested that Ability and Coursework had strong effects on Achievement, Family Background had a small effect, and Motivation had no appreciable effect. Further inspection of the model showed that Motivation had a strong effect on the Coursework students take in high school, so Motivation should have an indirect effect on Achievement through Coursework. We were able to calculate these indirect effects by multiplying together the two paths. We added this indirect effect to the direct effect to estimate the

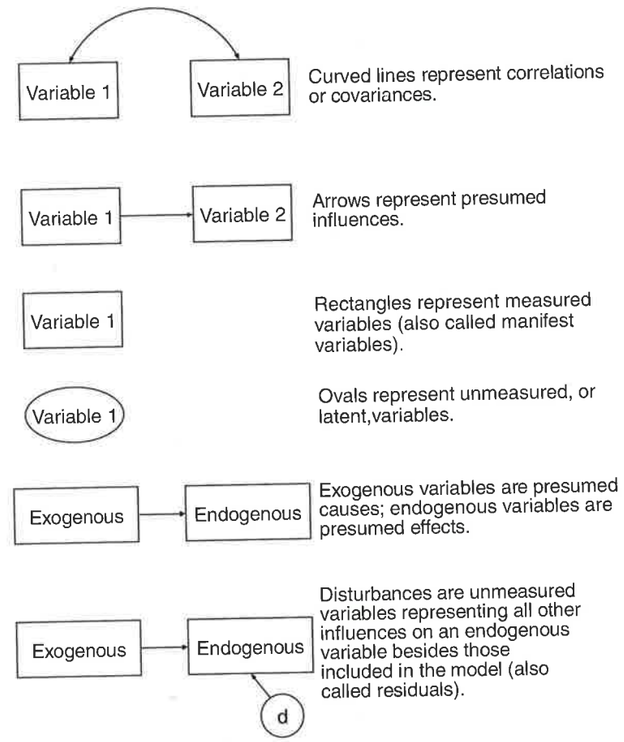

Figure 11.17 Quick summary of some of the jargon of path analysis.

total effect of Motivation on Achievement. When we focused on the total effect, Motivation did indeed have an influence on Achievement and one that makes sense: more motivated students, it appears, take more advanced coursework, and this coursework, in turn, improves their achievement.

An easier way to estimate total effects is through sequential regression. To do so, we regressed Achievement on Family Background and then added Ability, then Motivation, and then Coursework. The β associated with each variable, when entered, represents its total standardized effect. Thus, when Motivation was added to the model, its β was .096, its total effect. This procedure works because the total effects are the same whether or not there are intervening variables between the variable of interest and the outcome. If we remove the intervening variables, the total effects are equal to the direct effects. We then estimated the indirect effects by subtracting the direct from the total effects.

In addition to illustrating the basics of path analysis, this chapter tied together a major loose end of Part 1, the apparent inconsistency between simultaneous and sequential regression results. I argued that path analysis is particularly useful because it allows us to focus on both direct and indirect effects and that indirect effects are useful in explaining how an effect works. Intervening or mediating variables can thus be added to models to help understand how an effect comes about. Path models are also useful because they make explicit what is

too often left vague: the researcher's theory of how variables are causally related. In my opinion, path analysis is the best use of MR for explanatory, nonexperimental research.

EXERCISES

1. Table 11.2 shows the means, standard deviations, and correlations among the variables used in this chapter's example. Reanalyze the five-variable path model. (For users of SPSS, the file "motivate 5 var path.sps" on the Web site (www.tzkeith.com) shows how to analyze such a matrix using this program.) Calculate all paths and disturbances to create a table of direct, indirect, and total effects. Make sure your results match mine.
2. Construct a path model using the variables Family Background, 8th-grade GPA, 10th-grade Self-Esteem, 10th-grade Locus of Control, and 10th-grade Social Studies achievement test scores. How did you make the decisions on which variable affected which? Which of these decisions were the most difficult? What sources could you use to better inform your decisions?
3. What is the identification status of your model from Exercise 2: just-identified, overidentified, or underidentified? If your model is underidentified, see if you can make it into a just-identified model so that you can estimate it.
4. Select the variables BYSES, BYGrads, F1Cncpt2, F1Locus2, and F1TxHStd from the NELS data. Check the variables (e.g., descriptive statistics) to make sure you understand the scales of the variables. Also make sure that any values that should be coded as missing values are so coded.
5. Estimate your model using the variables from NELS (Exercise 4). Calculate the direct effects and disturbances, and put them into your model. Calculate total effects and create a table of direct, indirect, and total effects. Interpret the model; focus on direct, indirect, and total effects.
6. Compare your model and interpretation with others in your class. How many classmates drew the model in the same way you did? How many drew it differently? What difference did these different models make in results and interpretation?
7. Curtis Hansen tested a path model of the influences on accidents among chemical industry workers (1989). A simulated version of a portion of the data are on the website (www.tzkeith.com) under chapter 11 (e.g., "Hansen accident data.sav"; the file is also available in other formats). A guiding question for our analysis might be: what are the

Table 11.2 Means, Standard Deviations, and Correlations among the School Learning Variables

	Family Background	*Ability*	*Motivation*	*Coursework*	*Achievement*
N	1000	1000	1000	1000	1000
Mean	0	100	50	4	50
SD	1	15	10	2	10
Family Background	1				
Ability	.417	1			
Motivation	.190	.205	1		
Coursework	.372	.498	.375	1	
Achievement	.417	.737	.255	.615	1

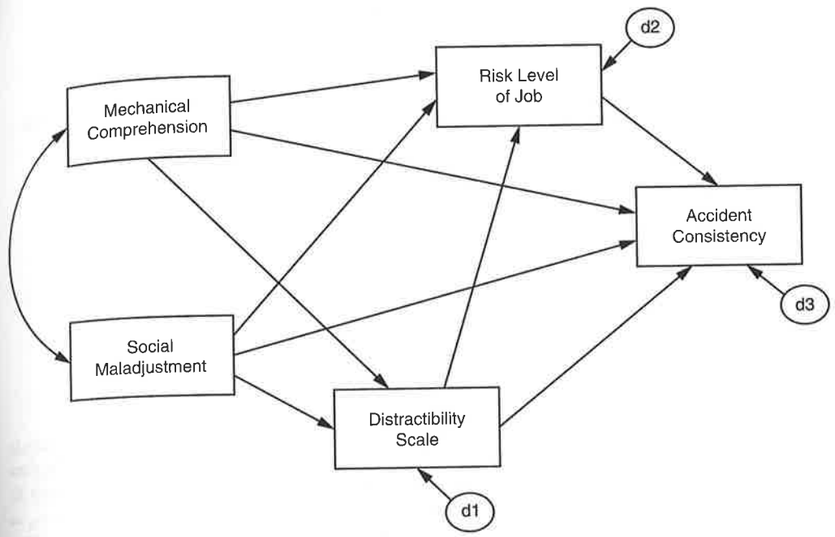

Figure 11.18 Path model of the presumed effects of abilities, personality characteristics, and job characteristics on number and consistency of accidents in an industrial setting.

relative effects of abilities, personality characteristics, and job characteristics on workers' accident rates? Figure 11.18 shows a model designed to answer this question.

Mechanical Comprehension (Mechanic in the data file) was a measure of workers' understanding of mechanical reasoning. Social Maladjustment (Maladjust) was a 50-item scale derived from the MMPI and designed (by Hansen) to assess general social maladjustment. These two variables are exogenous variables in the model. The Distractibility Scale (Distractibility), also derived from the MMPI, was designed to assess distractibility, and especially "neurotic-anxious" (Hansen, 1989, p. 83) characteristics that should lead to distractibility. The Risk Level of the Job (Risk) was a rating of the "responsibility and accident potential" (Hansen, p. 84) of each possible job on a 1 to 35 scale. The final endogenous outcome variable was Accident Consistency (Accident), a measure of the number of accidents for a worker plus the number of years in which each worker had an accident.

Estimate the model shown in the figure using multiple regression analysis. What is the identification status of the model? Calculate the direct effects and disturbances and put them in your model. Calculate total effects on Accident Consistency and create a table of direct, indirect, and total effects. Interpret the results. What were the important effects on accident consistency? Were there meaningful indirect effects? If so, interpret them. Which variable(s) had the strongest total effect on accident consistency?

Notes

1 Here's more detail in solving the paths using algebra. The three equations were

$r_{13} = b + ac$,
$r_{23} = c + ab$, and
$r_{12} = a$. We can rearrange these equations to solve for paths a, b, and c:

$b = r_{13} - ac$,
$c = r_{23} - ab$, and
$a = r_{12}$.

We will solve the equation for b by substituting the third and second equations (for a and c, respectively) into the first equation:

$$\begin{aligned} b &= r_{13} - r_{12}(r_{23} - r_{12}b) \\ &= r_{13} - (r_{12}r_{23} - r_{12}^2 b) \\ &= r_{13} - r_{12}r_{23} + r_{12}^2 b \\ b - r_{12}^2 b &= r_{13} - r_{13}r_{23} \\ b(1 - r_{12}^2) &= r_{13} - r_{12}r_{23} \\ b &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} \end{aligned}$$

See if you can use the same approach to solve for c.

2 The other method of developing equations to solve for paths is called the first law of path analysis (Kenny, 1979, p. 28). The correlation between Y (a presumed effect) and X (r_{xy}) is equal to the sum of the product of each path (p) from all causes of Y times the correlation of those variables with X: $r_{yx} = \sum p_{yz} r_{xz}$. Using the first law, the correlation between Motivation and Achievement is $r_{32} = br_{12} + cr_{22}$, which reduces to $r_{32} = br_{12} + c$ (description and equation adapted from Kenny, 1979, p. 28). The advantage of the first law is that it can be used to generate equations for any type of model, whereas the tracing rule works only with simple recursive models.

3 These rules are that standardized coefficients above .05 could be considered small; those above .10, moderate; and those above .25, large. These rules apply primarily to manipulable influences on school learning.

4 Kline (2011) adds a fourth condition, that the direction of the presumed causation is correctly specified (p. 98). This is a little more specific than we want to get right now, and we will deal with this problem in the next chapter.

5 Total effects are sometimes referred to as total causal effects. It is also possible to subtract the total causal effects from the original correlation to determine the noncausal (or spurious) portion of the correlation.

12
Path Analysis
Dangers and Assumptions

Path analysis is not magic; it does not prove causality. It does not make a silk purse out of a sow's ear; it cannot turn poor data into valid causal conclusions. Like multiple regression, there are assumptions underlying path analysis and the use of multiple regression to estimate paths. Like multiple regression, path analysis is open to abuse. This chapter will discuss these assumptions and the dangers of path analysis; it will also discuss how to avoid the dangers of the method.

ASSUMPTIONS

Because we have so far been using multiple regression to estimate path models, it should not be surprising that the basic assumptions of multiple regression also apply to path analysis. As discussed in Chapter 9, these include the following:

1. The dependent variable is a linear function of the independent variables. In addition, the causal direction in the model must be correct.
2. Each person (or other observation) should be drawn independently from the population.
3. The errors are normally distributed and relatively constant for all values of the independent variable.