CHAPTER 8

PATH ANALYSIS

In the previous chapter, we discussed in detail one of the main purposes of multiple regression—that being *prediction*. In this chapter, we present a discussion of another use of multiple regression—regression as a technique for providing *explanations* of possible causal relationships among a set of variables. Path analysis is actually one of two techniques that are classified under the broad heading of causal modeling. Following a brief introduction to causal modeling, and the distinctions between the two major types of causal modeling, we present a detailed discussion of path analysis, focusing on appropriate uses and proper interpretations of the technique.

SECTION 8.1 PRACTICAL VIEW

Purpose

Regression can be used to establish the possibility of cause-and-effect relationships among a set of variables (Sprinthall, 2000). Using regression analysis in this manner constitutes a specific set of statistical analysis techniques known as causal modeling. ***Causal modeling*** techniques examine whether a pattern of intercorrelations among variables "fits" the researcher's underlying theory of which variables are causing other variables (Aron & Aron, 1997). It is important to remember, however, that in causal modeling we are attempting to draw causal inferences from correlational data—the degree of confidence in the validity of causal inference from correlational data is typically much weaker than inference drawn from data resulting from a well-designed experimental study where the important concept of random assignment to treatments has been incorporated (Tate, 1992). Conclusions drawn from causal modeling with correlational data must be confined to the following limitation: The results of causal modeling are valid and unbiased *only if* the assumed model adequately represents the *real* causal processes (Tate, 1992).

In causal modeling, the causal interrelationships are examined among a set of variables that have been logically ordered on the basis of time (Sprinthall, 2000). Logically, a causal variable must precede any variable that it supposedly affects—this establishes the causal ordering of the variables (Sprinthall, 2000). There are two types of causal modeling techniques: path analysis and structural equation modeling (the latter will be described at the end of this section). ***Path analysis*** begins with the researcher developing a diagram with arrows connecting variables and depicting the *causal flow*, or the direction of cause-and-effect. The precursor to path analysis is a simpler version of causal modeling in which the only effects represented are direct causal effects. Path analysis has a substantial advantage over the simpler model in that both *direct* and *indirect* causal effects can be estimated. We will first examine the simplest form of causal modeling, followed by a presentation of the more involved form—path analysis.

As we have mentioned, the simplest version of the causal modeling technique is one in which only direct causal effects are represented (Tate, 1992). This version is quite similar to multiple regression, as discussed in the previous chapter. The direct causal effect of an IV (X) on a DV (Y) is defined as the amount of change in Y resulting from a unit change in X, holding constant all other causal determinants of Y. The causal model is represented by a single regression equation in which the IVs are the causal determinants of the DV. For example, we might want to determine the direct causal paths of three IVs on a single DV. Assume we wanted to investigate the effects of a country's location in the world (*region*), its status as a developing nation (*develop*), and the number of doctors per 10,000 people (*docs*) on male life expectancy (*lifexpm*). Since we are assuming only direct causal paths, our single equation would simply attempt to explain the direct causes of each of the three IVs (*region*, *develop*, *docs*) on the DV (*lifexpm*).

The development of a causal model is probably the most difficult aspect of conducting any causal modeling study (Tate, 1992). The specification of the model is a formal declaration of the researcher's beliefs regarding the causal links among the variables. What was the basis of our decision to order the four variables as described above? These beliefs are typically influenced by several sources of information, including the research literature, formal and informal theories, personal observations and experiences with the phenomenon of interest, expert opinions, and last but certainly not least, common sense and logic (Tate, 1992). Specification of a hypothesized model is often complicated by several sources of difficulty:

- the vagueness of many theories in social science research;
- the potentially infinite number of possible causal determinants which are often posited in the related research literature; and
- the complexity of nearly all phenomena of interest in social science research (Tate, 1992)—which has, of course, been discussed on several occasions in this text.

The specified causal model can be represented in two ways: as an equation or in diagrammatic form. The assumed causal model, when stated as an equation, is often referred to as a ***structural equation***, and is typically stated in its standardized form. If we were to "define" our variables using z-score coefficients—i.e., the standardized form where z_1 = *region*, z_2 = *develop*, z_3 = *docs*, and z_4 = *lifexpm*—the structural equation for our working example would be:

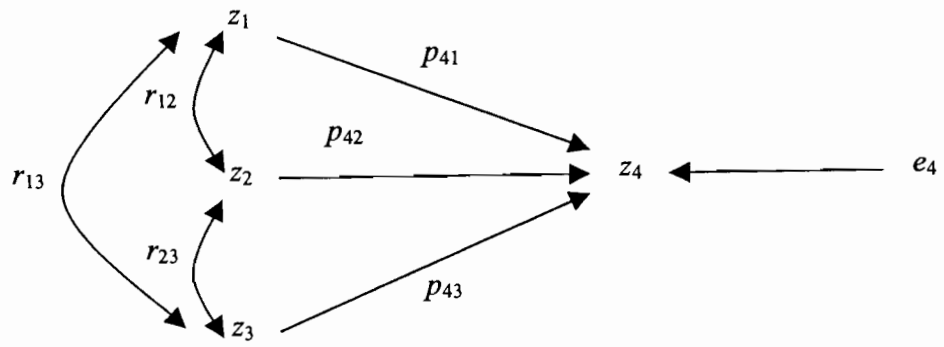
$$z_4 = p_{41}z_1 + p_{42}z_2 + p_{43}z_3 + e_4 \qquad \text{(Equation 8.1)}$$

In this structural equation, the direct causal effects are represented by the p coefficients, often called ***path coefficients*** or ***structural coefficients***. These coefficients are analogous to standardized regression coefficients, β, resulting from a multiple regression analysis (Agresti & Finlay, 1997) and their interpretation is similar (Aron & Aron, 1997 & 1999; Tate, 1992; Asher, 1983). In other words, they are interpreted as the estimated change in the DV, expressed in standard deviation units, associated with a one standard deviation change in each IV, holding the other IVs constant. The subscripts that accompany the path coefficients indicate the direction of causation with the first subscript indicating the variable being determined and the second indicating the direct cause (Tate, 1992). The zs indicate the standardized raw score value on each variable. The final component in the structural equation is the residual or e_i. This residual term, called the ***disturbance term*** in causal modeling parlance, represents the composite effect of any other direct determinants of z_4, which have not been included in the causal model, plus any measurement error in z_4 (Tate, 1992; Tatsuoka, 1988).

Although we are really dealing with three IVs and one DV in our working example, it is not accurate to refer to them as such when conducting a causal modeling study. In the specific language of causal modeling, the variable that is being explained by the model (i.e., the DV, the effect, or, in our example, z_4) is referred to as the ***endogenous variable***, while all variables not explained by the model (i.e., the IVs, the causes, or z_1, z_2, and z_3) are referred to as ***exogenous variables*** (Tate, 1992; Tatsuoka, 1988). Endogenous variables are assumed to have their variance explained by the exogenous variables included in the model; whereas, the variability of exogenous variables is assumed to be explained by other variables outside the causal model under consideration (Pedhazur, 1982).

The second way that a specified causal model can be portrayed is with a path diagram. A ***path diagram*** is a pictorial representation of the theoretical explanations of cause-and-effect relationships among a set of variables (Agresti & Finlay, 1997). The path diagram for our simple working example is shown in Figure 8.1. It is important to note that a path diagram is not necessary for causal modeling analysis, but is helpful is presenting the results of the analysis (Pedhazur, 1982).

Figure 8.1 Sample Path Diagram for a Single Equation Causal Model.

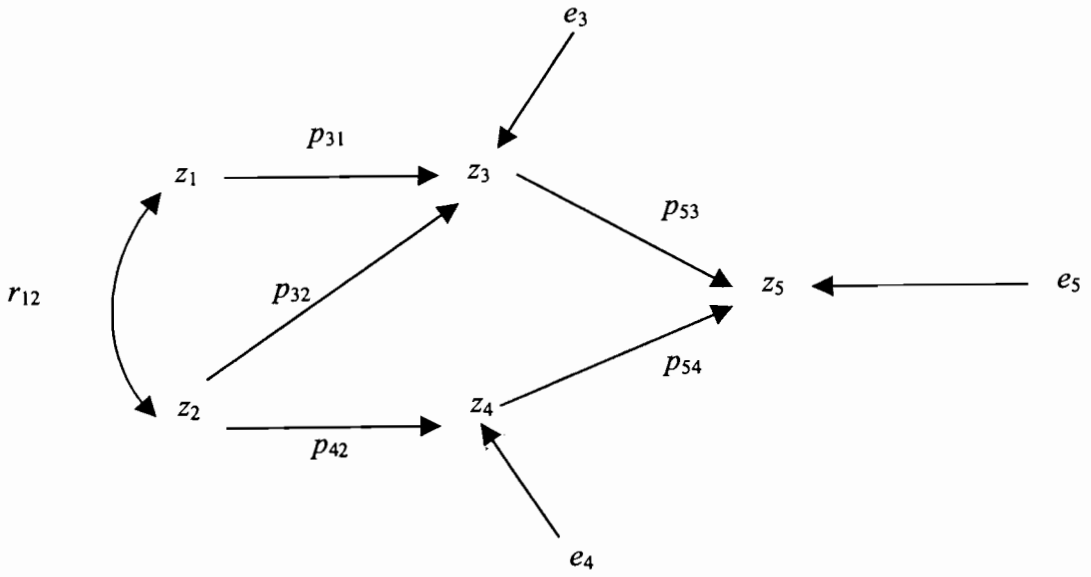


The direct causal effects of the exogenous variables z_1, z_2, and z_3 on the endogenous variable z_4 are shown with straight arrows, with the arrowheads indicating the assumed direction of causation (Tate, 1992). These arrows are often referred to as *causal paths* and are labeled with the associated path coefficients. Notice that the effect of the disturbance term is also included. Finally, the curved, double-headed arrows simply represent the bivariate correlations between exogenous variables in the model.

As we've previously mentioned, path analysis builds upon this simpler version by modeling both direct *and* indirect causal effects among the variables. An *indirect effect* occurs when a variable affects an endogenous variable through its effect on some other variable, known as an *intervening variable* (Agresti & Finlay, 1997). As in any causal modeling analysis, the first step is to specify the model of direct causal links among variables. This model will then imply indirect and total causal effects—this is a critical element that is missing in the simpler, single equation model previously discussed (Tate, 1992). Another distinct advantage of path analysis over the single equation model is the fact that it is now possible to test the overall fit of the model to the data in order to ascertain if the model (theory) is consistent with the observed correlations (actual data). The method by which we assess model fit will be described momentarily and elaborated upon in Section 8.3. If serious inconsistencies between the model and the data exist, it is recommended that the model be revised prior to describing any of the causal effects (Tate, 1992). It is important to note that consistency between the model and the observed correlations does not prove the validity of the model, but does represent support of the model (Tate, 1992).

We can now expand our single equation model into a path model. To do so, we decide to add another variable to our model—the number of deaths per 1,000 people (*deathrat*). Our initial path model is presented in Figure 8.2. The arrows indicate that *deathrat* and *docs* (although, we will actually use the natural log of *docs*, or *lndocs*) are the only important direct causal determinants of *lifexpm*. Furthermore, we hypothesize that *region* and *develop* have a direct causal effect on *deathrat*, and that *develop* has a direct causal effect on *lndocs*. Notice that in our path model, *lndocs* has changed from an exogenous (unexplained) variable to an endogenous (explained) variable.

Figure 8.2 Path Diagram for the Initial Model (Male Life Expectancy).



z_1 = *region*
z_2 = *develop*
z_3 = *deathrat*
z_4 = *lndocs*
z_5 = *lifexpm*

Model specification in path analysis becomes a much more convoluted process than in the single equation model (Tate, 1992). Correct specification in the single equation model necessitates that we have identified all causal effects of the lone endogenous variable. In our path analysis model, we have likely investigated the possible causes of our ultimate variable of interest—male life expectancy. After all, it is this variable in which we are most interested, in terms of explaining or describing its causal determinants. However, correct specification of the overall model in path analysis requires that *each* endogenous variable in the model is correctly specified. In other words, we have assumed that the model for *lifexpm* is correctly specified, but what about the models for *deathrat* and *lndocs*? We must ensure that the models for these additional endogenous variables are also correctly specified in order for our overall model to be valid and accurate (Tate, 1992).

Notice that we have also made informed decisions about excluding certain paths from the model. Specifically, our model has four missing paths—those from z_1 to z_4, z_1 to z_5, z_2 to z_5, and z_3 to z_4. It is important to note that these missing paths should also be consistent with theory and the associated litera-

ture (Tate, 1992). For example, excluding the path from *region* (z_1) to *lifexpm* (z_5) indicates the belief that *region* only affects *lifexpm* through its effect on *deathrat* (i.e., an indirect effect). Tate (1992) notes that a final theoretical path model should be "represented as much by the excluded paths as by the included paths."

Finally, when we are reasonably comfortable with our theoretical model, we can formally represent that model with a system of structural equations. This system of structural equations must include one for *each* endogenous variable in the model. For our current example, as depicted in Figure 8.2, these equations would be:

$$z_3 = p_{31}z_1 + p_{32}z_2 + e_3 \qquad \text{(Equation 8.2)}$$

$$z_4 = p_{42}z_2 + e_4 \qquad \text{(Equation 8.3)}$$

$$z_5 = p_{53}z_3 + p_{54}z_4 + e_5 \qquad \text{(Equation 8.4)}$$

One should notice that in a path diagram, indirect effects are identified by a chain of two or more straight arrows all going in the same direction (Tate, 1992). The value of an indirect path coefficient is determined by finding the product of all path coefficients in the chain. In Figure 8.2 for example, the paths from *region* to *deathrat* (p_{31}) and from *deathrat* to *lifexpm* (p_{53}) combine together to produce an indirect effect of *region* on *lifexpm* (equal to $p_{31}p_{53}$).

Multiple regression analysis provides the values for the unbiased estimates of the path coefficients. In order to obtain the coefficients, a separate regression run must be completed for each structural equation, each including only the direct causal effects for its associated endogenous variable. Using Equation 8.2 as an example, in order to obtain p_{31} and p_{32}, one must regress z_3 (*deathrat*) on z_1 (*region*) and z_2 (*develop*). In other words, a multiple regression analysis is conducted with *deathrat* as the DV and *region* and *develop* as the IVs. Similar procedures are then conducted for the two remaining structural equations.

Probably the most crucial part of the analysis in a causal modeling study is the assessment of model fit. Before the obtained estimates of path coefficients can be used to describe the causal effects among the variables, one should determine whether or not the model is consistent with the observed, empirical correlations among the variables. This is typically accomplished by obtaining the ***reproduced correlations***—those logically implied by the hypothetical or theoretical model—and comparing them to the empirical correlations (Agresti & Finlay, 1997; Tate, 1992). The reproduced correlations, therefore, are the bivariate correlations that *would* be produced *if* the causal model were correctly specified. If the observed and the reproduced correlations are reasonably close (say, within roughly .05 of each other), it can be assumed that the model is consistent with the empirical data (Tate, 1992). Larger discrepancies indicate that the model is not consistent with the data and model revisions should be considered. Unfortunately, the reproduced correlations, and subsequent comparisons to observed correlations, cannot be obtained via computer analysis and must be computed by hand. The procedures for doing so are described in detail in Section 8.3.

Earlier in this chapter, we alluded to a second type of causal modeling strategy, and we briefly introduce it here. This second type of causal modeling offers several advantages over path analysis. ***Structural equation modeling***, sometimes referred to as *latent variable modeling*, also involves diagrams with arrows showing causal flows among variables. However, one major advantage is that the computer analysis procedure provides an overall indication of the fit between the model and the theory. We have briefly mentioned, and will see later in some detail, how this assessment of model fit must be done by hand in path analysis. A second major advantage of structural equation modeling over path analysis is that it can incorporate latent variables. A *latent variable* is a variable that cannot actually be

measured but can only be *approximated* with actual measures (Aron & Aron, 1997). For example, "intelligence" is a latent variable. We would be hard pressed to find a single measure for intelligence, but we can approximate measures for intelligence by obtaining values on several observable variables such as IQ, performance on academic achievement tests, etc. In structural equation modeling, a diagram is set up such that latent variables are combinations of observable, measurable variables. Path diagrams in structural equation modeling are much more involved, incorporating several additional components over and above those included in a path analysis. A disadvantage of structural equation modeling is that standard statistical analysis software packages (such as SPSS) are not able to conduct the required procedures. Special statistics programs are required in order to conduct this type of analysis. One such program, LISREL, takes its name from the purpose of the technique—that is, to uncover linear structural relations. Discussions of structural equation modeling and the LISREL program are beyond the scope and purpose of this text. If interested in reading further about structural equation modeling, brief descriptions and examples are included in Aron and Aron (1997 & 1999) and Johnson and Wichern (1998). If detailed information is required, the reader is directed to Tabachnick and Fidell (1996), Long (1983), and Pedhazur (1982).

Sample Research Questions

Returning to our working example for path analysis, we can specify our research questions for the study as follows:

(1) Is our model—which describes the causal effects among the variables "region of the world," "status as a developing nation," "number of deaths," "number of doctors," and "male life expectancy"—consistent with our observed correlations among these variables?
(2) If our model is consistent, what are the estimated direct, indirect, and total causal effects among the variables?

SECTION 8.2 ASSUMPTIONS AND LIMITATIONS

Since path analysis is essentially an extension and specific application of multiple regression, the assumptions that were discussed in the previous chapter are also appropriate here. As a reminder of those assumptions, we simply list them here:

(1) The independent variables are fixed (i.e., the same values of the IVs would have to be used if the study were to be replicated).
(2) The independent variables are measured without error.
(3) The relationship between the independent variables and the dependent variable is linear (in other words, the regression of the DV on the combination of IVs is linear).
(4) The mean of the residuals for each observation on the dependent variable over many replications is zero.
(5) Errors associated with any single observation on the dependent variable are independent of (i.e., not correlated with) errors associated with any other observation on the dependent variable.
(6) The errors are not correlated with the independent variables.
(7) The variance of the residuals across all values of the independent variables is constant (i.e., homoscedasticity of the variance of the residuals).
(8) The errors are normally distributed.

If additional specific information regarding the assumptions associated with multiple regression is required, the reader is advised to revisit Chapter 7.

As we have previously mentioned, valid causal inference requires the correct specification of the structural equation(s) in a path analysis. If, *and only if*, the model is correctly specified, the estimates of the various causal effects will be accurate and unbiased (Tate, 1992). In contrast, any specification errors that exist will result in the estimates of causal effects to be biased to some unknown degree. In order to use multiple regression in a manner to estimate the path coefficients, the following assumptions regarding correct model specification must be met:

(1) The model must accurately reflect the actual causal sequence.
(2) The structural equation for each endogenous variable includes all variables that are direct causes of that particular endogenous variable (i.e., variables that are not included in the model, and whose effects are therefore assumed to be "captured" by the residuals, are also assumed not to be correlated with any of the determinant variables).
(3) There is a one-way causal flow in the model (i.e., there can be no reciprocal causation between variables).
(4) The relationships among variables are assumed to be linear, additive, and causal in nature; any curvilinear relations, etc., are to be excluded.
(5) All exogenous variables are measured without error (Tate, 1992; Pedhazur, 1982).

The reader should note that assumptions #1 through #4 for path analysis deal directly with the specification of the model which, as we have previously mentioned, can be based on a combination of factors (theory, experience, research literature, opinion, etc.). As we have seen with previous techniques, assumption #5 is largely an issue of research design and data collection.

We would be remiss if we did not discuss several limitations of path analysis. Earlier in the chapter, we referred to the fact that with path analysis we are attempting to estimate and describe causal relationships through the use of correlational data. Because of this fact, the degree of confidence we can have in the causal inferences drawn from the results of the analysis is bound to be much less than the confidence in inferences drawn from an experimental study.

Furthermore, if it is concluded that a model is not consistent with the empirical data, the model has been *misspecified*, which is a matter of degree. This degree of misspecification is subjective, to say the least, and must be evaluated by the researcher. Tate (1992) describes this limitation in the following manner:

> "A model which omits several relatively unimportant causes, ignores a real but weak causal feedback, and is based on measures with some modest measurement error may still produce estimates which are technically biased but still reasonable (and valuable) approximations to the true causal effects. On the other hand, completely misleading conclusions may result from a model which is perfect in every way except for the omission of a single important variable" (p. 319).

There is no statistical test that will definitively indicate whether or not the misspecification is within reasonable limits—those decisions are left to the researcher.

Due to the above limitations, it has been suggested (Tate, 1992) that the use of "conditional" statements in reporting the results of a path analysis study is warranted. For example, one might state obtained results in the following manner: "If this model accurately reflects reality, the estimated causal effects are ..." This serves as an appropriate reminder to ourselves—and to the readers of our research reports—of the limitations associated with drawing causal inferences from correlational data.

Methods of Testing Assumptions

With respect to the initial eight assumptions associated with the use of multiple regression analysis, a thorough discussion of the methods of assessing the tenability of those assumptions was presented in Chapter 7. As a reminder to the reader, these assumptions may be assessed through the use of routine data screening procedures (see Chapter 3), but are most appropriately tested through inspection of bivariate scatterplots and more accurately through inspection of the residuals plots (see Chapter 7). Recall that residuals plots may be used to assess assumptions of linearity, normality, and constant variance (homoscedasticity).

The method of assessing the validity of the assumptions specific to path analysis differs greatly from the assessment of assumptions for statistical inference (Tate, 1992). No statistical procedures exist for evaluating these assumptions since they deal specifically with the degree to which the causal model has been correctly specified. There is no empirical test that can tell us the extent to which we have selected and described the correct model. In order to evaluate these five assumptions, Tate (1992) suggests that we focus our attention on the credibility, reasonableness, and utility of a proposed model. In other words,

- a model should be plausible to those who are expert in the particular field of inquiry;
- the results should be reasonable within the context of the current research literature; and
- a model should be useful in predicting future events.

The responsibility for assessing the assumptions in this manner ultimately rests with the researcher and his/her subjective judgments.

SECTION 8.3 PROCESS AND LOGIC

The Logic Behind Path Analysis

You will recall that in Chapter 7 we provided a brief overview of the calculations involved in conducting a multiple regression analysis. The same logic and associated calculations hold true here, since we are again applying a regression analysis, albeit within the analysis of a causal model as opposed to a straightforward multiple regression. Therefore, we will again be calculating the β coefficients (i.e., the standardized versions in order to represent the path coefficients) for each causal determinant, the squared multiple correlation (R^2) for each structural equation, and the associated significance tests. This will be done in the same manner as described in the previous chapter.

Typically, we reserve this section of each chapter to explain the logic behind the calculations of each technique, without overwhelming the reader with mathematical equations and hand calculations, since the calculations are obtained via computer analysis. However, you will recall that we mentioned earlier that the assessment of model fit in a path analysis can only be accomplished through the use of hand calculations. The assessment of model fit is conducted by obtaining the reproduced correlations and comparing them to the empirical correlations, then evaluating them against the difference criterion of .05. Again, if all reproduced and observed correlations are relatively close to each other, the model is consistent with the empirical data; in other words, the model "fits" the data.

One commonly used approach to determining the reproduced correlations between two variables (and, therefore, among all variables in the set) involves the identification of all legitimate paths between the variables in the model in a process referred to as *path tracing* (Tate, 1992) or *path decomposition* (Pedhazur, 1982). ***Path tracing*** is a process that results in a correlation coefficient for each path, which

is equal to the product of all coefficients in the path. A key is that one may only use legitimate paths, which are those paths that do not violate any of the following three rules:

(1) no path may pass through the same variable more than once,
(2) no path may go backward on an arrow after going forward on another arrow (although it is acceptable to go forward on an arrow after *first* going backward), and
(3) no path may include more than one double-headed curved arrow (Tate, 1992).

To illustrate this process, refer to Figure 8.3, which represents the same model as in Figure 8.2 but which now includes the path coefficients resulting from our regression analysis. If we wanted to obtain the reproduced correlation between z_1 and z_3, the legitimate paths are:

Path	Component
z_1 to z_3	p_{31}
z_1 to z_2 to z_3	$r_{12}p_{32}$

Therefore, the resulting equation for the reproduced correlation (symbolized by $\hat{r}$) between z_1 and z_3 is represented by the following equation:

$$\hat{r}_{13} = p_{31} + r_{12}p_{32}$$

Making the appropriate substitutions of path coefficients, we now have:

$$\hat{r}_{13} = (-.395) + (-.621)(-.123) = -.319$$

As another example, let us consider the reproduced correlation between z_1 and z_5. The legitimate paths are:

Path	Component
z_1 to z_3 to z_5	$p_{31}p_{53}$
z_1 to z_2 to z_3 to z_5	$r_{12}p_{32}p_{53}$
z_1 to z_2 to z_4 to z_5	$r_{12}p_{42}p_{54}$

The resulting equation is obtained for $\hat{r}$ between z_1 and z_5:

$$\hat{r}_{15} = p_{31}p_{53} + r_{12}p_{32}p_{53} + r_{12}p_{42}p_{54}$$

Again, making the appropriate substitutions gives us:

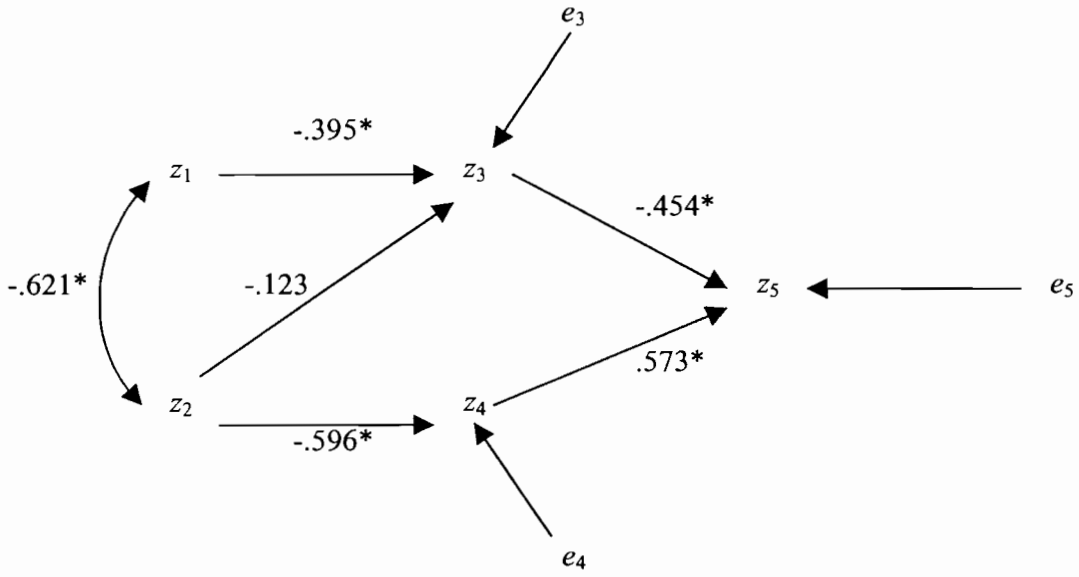
$$\hat{r}_{15} = (-.395)(-.454) + (-.621)(-.123)(-.454) + (-.621)(-.596)(.573) = .356$$

Correlation decompositions such as these would be determined for all possible bivariate correlations in the model, with the exception of those between exogenous variables. The complete set of path decompositions and reproduced correlations for the model shown in Figure 8.3 is presented in Table 8.1. It is recommended that the reader reproduce these results in order to practice the identification of legitimate paths in a path model, while adhering to the path tracing rules.

The reader will notice that each path component in Table 8.1 includes an abbreviated "label" (i.e., D, I, S, or U). It is important to note the conceptual differences among the various types of path components—this is ultimately important when attempting to describe the direct, indirect, and total causal effects in a model (Tate, 1992). Causal effects are represented by paths consisting only of direct

causal links—in other words, only straight arrows—that flow in only one direction. These causal effects may be *direct* (a causal path consisting of only one link; denoted "D" in Table 8.1) or *indirect* (consisting of two or more links; denoted "I" in Table 8.1). For example, the $\hat{r}_{15}$ decomposition shown above includes the indirect effect of z_1 on z_5, mediated through z_3 ($p_{31}p_{53}$).

Figure 8.3 Path Diagram for the Initial Model (Male Life Expectancy), Including Path Coefficients.



* Significant at the .05 level.

z_1 = *region*
z_2 = *develop*
z_3 = *deathrat*
z_4 = *lndocs*
z_5 = *lifexpm*

Any path components resulting from paths that have reversed causal direction at some point are called ***spurious effects*** (denoted "S" in Table 8.1), indicating that the relationship is caused by a common third factor (Tate, 1992). The paths may or may not include a double-headed curved arrow. For example, in the decomposition of $\hat{r}_{35}$, the component $p_{31}r_{12}p_{42}p_{54}$ represents a spurious effect—in other words, portions of r_{35} are <u>not</u> due to *either* direct or indirect causal effects of z_3 on z_5. Note that any path between two endogenous variables, which includes a curved arrow, will always represent a spurious effect (Tate, 1992).

Finally, in any model that contains more than one exogenous variable, as does Figure 8.3, the associated unexplained correlations among them will result in a degree of "undeterminability" with respect to the resolution of the direct and indirect effects of exogenous variables on endogenous variables (Tate, 1992). Since a model such as this does not explain the relationship among exogenous variables, we must recognize that this unanalyzed portion (denoted "U" in Table 8.1) may represent some degree of causal effect that has not been included in the model. In this situation, the total causal effect on an endogenous variable must be accompanied by a note specifying that there exists some uncertainty due to the unanalyzed component.

Table 8.1 Path Decompositions for the Initial Model (Male Life Expectancy) Shown in Figure 8.3.

Reproduced Correlation	Path Decomposition
$\hat{r}_{13}$	$p_{31} + r_{12}p_{32}$ (D) (U)
$\hat{r}_{14}$	$r_{12}p_{42}$ (U)
$\hat{r}_{15}$	$p_{31}p_{53} + r_{12}p_{32}p_{53} + r_{12}p_{42}p_{54}$ (I) (U) (U)
$\hat{r}_{23}$	$p_{32} + r_{12}p_{31}$ (D) (U)
$\hat{r}_{24}$	p_{42} (D)
$\hat{r}_{25}$	$p_{32}p_{53} + p_{42}p_{54} + r_{12}p_{31}p_{53}$ (I) (I) (U)
$\hat{r}_{34}$	$p_{32}p_{42} + p_{31}r_{12}p_{42}$ (S) (S)
$\hat{r}_{35}$	$p_{53} + p_{32}p_{42}p_{54} + p_{31}r_{12}p_{42}p_{54}$ (D) (S) (S)
$\hat{r}_{45}$	$p_{54} + p_{42}p_{32}p_{53} + p_{42}r_{12}p_{31}p_{53}$ (D) (S) (S)

Once all of the reproduced correlations have been obtained for a path model (see Table 8.2), they are displayed adjacent to the observed correlations. Those reproduced correlations that have a difference greater than .05 from the empirical correlations are indicated with an asterisk (see Table 8.3). Any differences that are substantially larger than the .05 criterion indicate that the model is not consistent with the empirical data and revisions to the model are warranted prior to describing any of the causal effects. This method of testing for model fit is only possible when there are one or more missing paths in the model—if all possible paths are included, the reproduced correlations will *always* be exactly equivalent to the observed correlations (Tate, 1992)—by definition, the fit of the model will be perfect. *Recall that if one goal of any analysis is a parsimonious solution, we should always have some missing paths in a model.*

If it is determined that a model does not "fit" the data, consideration should be given to retaining included paths and incorporating excluded paths. This is accomplished by first testing all missing paths for each endogenous variable in the model (Tate, 1992). In our working example, we originally regressed z_4 on z_2 but chose to exclude the regression of z_4 on z_1 and z_3. In order to test the missing paths for z_4, we must regress z_4 on *all* of its direct causal determinants (z_1, z_2, and z_3). Similarly, we would then regress z_5 on z_1, z_2, z_3, and z_4. Support for adding any originally excluded paths is indicated by a significant path coefficient (β) in the computer output. Support for the original model is indicated by any nonsignificant path coefficients. Second, we would want to examine empirical support for all paths that we initially chose to include. This is also accomplished by examining the significance of each path coefficient—significance denotes that the model (at least, that particular coefficient) is supported by the

data. If a path coefficient is not statistically significant, one should consider dropping it from the model *unless there is strong theoretical support for its inclusion* (Tate, 1992).

Table 8.2 Calculations of Reproduced Correlations for the Initial Model (Male Life Expectancy) Shown in Figure 8.3.

$\hat{r}_{13}$	$= p_{31} + r_{12}p_{32}$ $= (-.395) + (-.621)(-.123) = \mathbf{-.319}$ (D) (U)
$\hat{r}_{14}$	$= r_{12}p_{42}$ $= (-.621)(-.596) = \mathbf{.370}$ (U)
$\hat{r}_{15}$	$= p_{31}p_{53} + r_{12}p_{32}p_{53} + r_{12}p_{42}p_{54}$ $= (-.395)(-.454) + (-.621)(-.123)(-.454) + (-.621)(-.596)(.573) = \mathbf{.356}$ (I) (U) (U)
$\hat{r}_{23}$	$= p_{32} + r_{12}p_{31}$ $= (-.123) + (-.621)(-.395) = \mathbf{.122}$ (D) (U)
$\hat{r}_{24}$	$= p_{42}$ $= (-.596) = \mathbf{-.596}$ (D)
$\hat{r}_{25}$	$= p_{32}p_{53} + p_{42}p_{54} + r_{12}p_{31}p_{53}$ $= (-.123)(-.454) + (-.596)(.573) + (-.621)(-.395)(-.454) = \mathbf{-.397}$ (I) (I) (U)
$\hat{r}_{34}$	$= p_{32}p_{42} + p_{31}r_{12}p_{42}$ $= (-.123)(-.596) + (-.395)(-.621)(-.596) = \mathbf{-.073}$ (S) (S)
$\hat{r}_{35}$	$= p_{53} + p_{32}p_{42}p_{54} + p_{31}r_{12}p_{42}p_{54}$ $= (-.454) + (-.123)(-.596)(.573) + (-.395)(-.621)(-.596)(.573) = \mathbf{-.495}$ (D) (S) (S)
$\hat{r}_{45}$	$= p_{54} + p_{42}p_{32}p_{53} + p_{42}r_{12}p_{31}p_{53}$ $= (.573) + (-.596)(-.123)(-.454) + (-.596)(-.621)(-.395)(-.454) = \mathbf{.606}$ (D) (S) (S)

In assessing the fit of our model in Figure 8.3, it can be seen from Table 8.3 that six of the ten reproduced correlations have differences greater (and substantially so!) than .05. Upon examination of the significance tests for missing paths resulting from the supplemental regression runs as described in the previous paragraph, it was determined that several paths should be added—specifically, the paths from z_1 to z_4 (p_{41}), from z_2 to z_5 (p_{52}), and from z_3 to z_4 (p_{43}). Additionally, because its beta coefficient was not significant, it was decided that the path from z_2 to z_3 (p_{32}) be removed from the model. The resulting revised path diagram, including path coefficients, is presented in Figure 8.4.

Once a model has been revised, the fit should again be reassessed. The path decompositions for our revised model are shown in Table 8.4. Calculation of the subsequent reproduced correlations is presented in Table 8.5. Reproduced correlations for the revised model are once again compared to the empirical correlations (see Table 8.3). This model obviously results in a much better fit than the initial model—only one of the reproduced correlations exceeds the .05 criterion. Had the fit *not* substantially improved, this process would continue until an adequate fit of the model to the empirical data has been achieved.

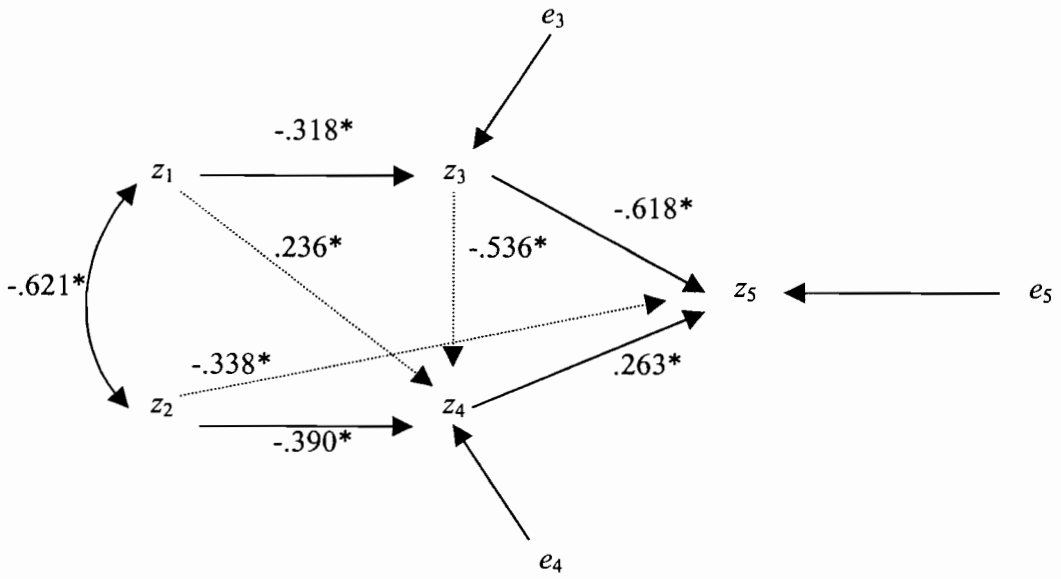
Table 8.3 Observed and Reproduced Correlations for the Initial (Figure 8.3) and the Revised (Figure 8.4) Models (Male Life Expectancy).

	z_1	z_2	z_3	z_4	z_5
	Observed Correlations				
z_1	1.000				
z_2	-.621	1.000			
z_3	-.318	.125	1.000		
z_4	.648	-.596	-.655	1.000	
z_5	.592	-.564	-.835	.870	1.000
	Reproduced Correlations (Initial Model)				
z_1	1.000				
z_2	-.621	1.000			
z_3	-.319	.122	1.000		
z_4	.370*	-.596	-.073*	1.000	
z_5	.356*	-.397*	-.495*	.606*	1.000
	Reproduced Correlations (Revised Model)				
z_1	1.000				
z_2	-.621	1.000			
z_3	-.318	.197*	1.000		
z_4	.648	-.643	-.688	1.000	
z_5	.578	-.630	-.866	.906	1.000

* Difference between reproduced and observed correlation is greater than 0.05.

At this point, we are satisfied with the fit of our model to the associated empirical data and can describe the causal effects of the variables and their correlations. A table that summarizes the causal effects of a model is typically presented in published research. Such a summary for our revised model is presented in Table 8.6. The reader should note that the indirect effects listed in the table are simply the sum of all indirect effects as identified in the path decompositions. The total effects consist of the sum of the direct and indirect effects.

Figure 8.4 Path Diagram for the Revised Model (Male Life Expectancy), Including Path Coefficients.



* Significant at the .05 level.

Note. Revised paths are shown with dashed arrows.

z_1 = *region*
z_2 = *develop*
z_3 = *deathrat*
z_4 = *lndocs*
z_5 = *lifexpm*

Interpretation of Results

As somewhat indicated, interpretation of the SPSS output for a path analysis is quite extensive since it requires several hand calculations. Once you have conducted all the regression analyses for the initial path model, the path (β) coefficients with the respective level of significance should be noted within the path model. These coefficients indicate the estimated change in the respective endogenous variables and are used to calculate the reproduced correlations through path decomposition. Reproduced correlations are then compared to the empirical correlations to test the model fit. If any reproduced correlations exhibit more than a .05 difference from the empirical correlations, the model is not consistent with the empirical data and should be revised. Once a consistent model has been generated, the specific causal effects for each endogenous variable are determined with respect to direct, indirect and total effects. Utilizing the path decompositions is imperative in this process, since both the direct and indirect effects are identified for each path. The reader should note that a path may have several indirect effects; consequently the sum of these indirect values represents the overall indirect effect for the path. The total effect is also calculated by adding the direct and indirect effects for each path. Finally, R^2 is interpreted to indicate the amount of variance in each endogenous variable that is explained by its structural model.

Table 8.4 Path Decompositions for the Revised Model (Male Life Expectancy) Shown in Figure 8.4.

Reproduced Correlation	Path Decomposition
$\hat{r}_{13}$	p_{31} (D)
$\hat{r}_{14}$	$p_{41} + r_{12}p_{42} + p_{31}p_{43}$ (D) (U) (I)
$\hat{r}_{15}$	$p_{31}p_{53} + r_{12}p_{42}p_{54} + p_{41}p_{54} + p_{31}p_{43}p_{54} + r_{12}p_{52}$ (I) (U) (I) (S) (U)
$\hat{r}_{23}$	$r_{12}p_{31}$ (U)
$\hat{r}_{24}$	$p_{42} + r_{12}p_{41} + r_{12}p_{31}p_{43}$ (D) (U) (U)
$\hat{r}_{25}$	$p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{31}p_{43}p_{54} + r_{12}p_{41}p_{54}$ (D) (I) (U) (U) (U)
$\hat{r}_{34}$	$p_{43} + p_{31}p_{41} + p_{31}r_{12}p_{42}$ (D) (I) (S)
$\hat{r}_{35}$	$p_{53} + p_{43}p_{54} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$ (D) (S) (S) (S) (S)
$\hat{r}_{45}$	$p_{54} + p_{43}p_{53} + p_{41}p_{31}p_{53} + p_{41}r_{12}p_{52} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{43}p_{31}r_{12}p_{52}$ (D) (S) (S) (S) (S) (S) (S)

Continuing with our example (see Figure 8.1) that seeks to investigate the causal effects among the variables "region of the world," "status as a developing nation," "number of deaths," "number of doctors," and "male life expectancy," we screened data for missing cases and outliers. Outliers were identified by calculating Mahalanobis distance and conducting **`Explore`**. Figure 8.5 presents these results and indicates that cases #29 and #56 should be eliminated from further analysis since they exceed the chi square criterion of 20.516 (df=5). Variables were then evaluated for normality by creating a scatterplot matrix (see Figure 8.6). Since the variable of doctors (*docs*) is curvilinearly related to male life expectancy, it was transformed to *lndocs* by taking its natural log. Next, multivariate normality and homoscedascticity were assessed by creating a residuals plot (see Figure 8.7). The fairly consistent spread of residuals indicates that the test assumptions are fulfilled. Prior to conducting the regression analyses, a correlation matrix (see Figure 8.8) was created since these empirical correlations will be needed later to test model fit. Applying the initial model, the first series of regression analyses was conducted. Since this model has three endogenous variables, the following analyses were run: z_3 on z_1 and z_2 (see Figure 8.9); z_4 on z_2 (see Figure 8.10); and z_5 on z_3 and z_4 (see Figure 8.11). Prior to interpreting the path coefficients, one should review the tolerance statistic for each exogenous variable included in each regression analysis in order to determine if multicollinearity can be assumed. If tolerance is greater than .1, one may proceed with interpreting the path coefficients. For our example, tolerance statistics were all adequate. Path (beta) coefficients from the output were transferred to the path diagram of the

Table 8.5 Calculations of Reproduced Correlations for the Revised Model (Male Life Expectancy) Shown in Figure 8.4.

$\hat{r}_{13} = p_{31}$
$= (-.318) = \mathbf{-.318}$
(D)

$\hat{r}_{14} = p_{41} + r_{12}p_{42} + p_{31}p_{43}$
$= (.236) + (-.621)(-.390) + (-.318)(-.536) = \mathbf{.648}$
(D) (U) (I)

$\hat{r}_{15} = p_{31}p_{53} + r_{12}p_{42}p_{54} + p_{41}p_{54} + p_{31}p_{43}p_{54} + r_{12}p_{52}$
$= (-.318)(-.618) + (-.621)(-.390)(.263) + (.236)(.263) + (-.318)(-.536)(.263) + (-.621)(-.338) = \mathbf{.578}$
(I) (U) (I) (I) (U)

$\hat{r}_{23} = r_{12}p_{31}$
$= (-.621)(-.318) = \mathbf{.197}$
(U)

$\hat{r}_{24} = p_{42} + r_{12}p_{41} + r_{12}p_{31}p_{43}$
$= (-.390) + (-.621)(.236) + (-.621)(-.318)(-.536) = \mathbf{-.643}$
(D) (U) (U)

$\hat{r}_{25} = p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{31}p_{43}p_{54} + r_{12}p_{41}p_{54}$
$= (-.338) + (-.390)(.263) + (-.621)(-.318)(-.618) + (-.621)(-.318)(-.536)(.263) + (-.621)(.236)(.263) = \mathbf{-.630}$
(D) (I) (U) (U) (U)

$\hat{r}_{34} = p_{43} + p_{31}p_{41} + p_{31}r_{12}p_{42}$
$= (-.536) + (-.318)(.236) + (-.318)(-.621)(-.390) = \mathbf{-.688}$
(D) (I) (S)

$\hat{r}_{35} = p_{53} + p_{43}p_{54} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$
$= (-.618) + (-.536)(.263) + (-.318)(.236)(.263) + (-.318)(-.621)(-.338) + (-.318)(-.621)(-.390)(.263) = \mathbf{-.866}$
(D) (I) (S) (S) (S)

$\hat{r}_{45} = p_{54} + p_{43}p_{53} + p_{41}p_{31}p_{53} + p_{41}r_{12}p_{52} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{43}p_{31}r_{12}p_{52}$
$= (.263) + (-.536)(-.618) + (.236)(-.318)(-.618) + (.236)(-.621)(-.338) + (-.390)(-.338) + (-.390)(-.621)(-.318)(-.618) + (-.536)(-.318)(-.621)(-.338) = \mathbf{.906}$
(D) (I) (S) (S) (I) (S) (S)

initial model as seen in Figure 8.3. All coefficients were significant with the exception of p_{32}. Reproduced correlations were then calculated through the path decompositions (see Table 8.2) and resulted in seven of the reproduced correlations differing from the empirical correlations by more than .05 (see Table 8.3). Since this initial model did not fit the empirical data, regression analyses were conducted on the following missing paths: z_4 on z_1, z_2, and z_3 (see Figure 8.12), and z_5 on z_1, z_2, z_3, and z_4 (see Figure 8.13). Evaluation of the path coefficients and respective levels of significance indicate that only the following paths were significant: z_3 on z_1; z_4 on z_1, z_2, and z_3; and z_5 on z_2, z_3, and z_4. Consequently, regression analyses were conducted again to include only those significant paths for each endogenous

variable in order to obtain "final" path coefficients (non-significant paths are excluded, therefore changing the values of the significant paths). These analyses are presented in Figures 8.14 and 8.15. The revised model with respective path coefficients is displayed in Figure 8.4. Calculation of reproduced correlations through path decompositions (see Table 8.5) and subsequent comparison to the empirical correlations (see Table 8.3) indicate the revised model fits the empirical data. Utilizing calculations for the direct and indirect effects from Table 8.5, we summarize the causal effects of the revised model in Table 8.6. In addition, R^2 is noted for each endogenous variable within this summary table. R^2 can be found in the final (accepted) regression analyses for each endogenous variable (see Figures 8.12, 8.14, and 8.15). For example, causal effects of *region, develop, deathrat*, and *lndocs*, explain 93.2% (R^2=.932) of variance in *lifeexpm*.

Table 8.6 Summary of Causal Effects for Revised Model (Male Life Expectancy).

		Causal Effects		
Outcome	Determinant	Direct	Indirect	Total
Death Rate	Region	-.318*	–	-.318
(R^2 = .101)	Developing Status	–	–	–[+]
Doctors	Region	.236*	.170	.406[+]
(R^2 = .738)	Developing Status	-.390*	–	-.390[+]
	Death Rate	-.536*	-.075	-.611
Male Life Expectancy	Region	–	.304	.304[+]
(R^2 = .932)	Developing Status	-.338*	-.103	-.441[+]
	Death Rate	-.618*	-.141	-.759[+]
	Doctors	.263*	.463	.726

* Direct effect is significant at the .05 level.

[+] Total effect may be incomplete due to unanalyzed components.

Figure 8.5 Outliers Determined by Mahalanobis Distance.

Extreme Values

			Case Number	Value
MAH_1	Highest	1	56	24.06622
		2	29	20.82071
		3	72	20.00161
		4	30	19.16860
		5	43	16.08589
	Lowest	1	35	.82904
		2	63	.86684
		3	40	.90204
		4	33	1.04412
		5	42	1.12388

Cases #56 and #29 exceed the $\chi^2(5)$=20.516 criteria and should be eliminated from further analysis.

Figure 8.6 Scatterplot for Model (Male Life Expectancy) Variables.

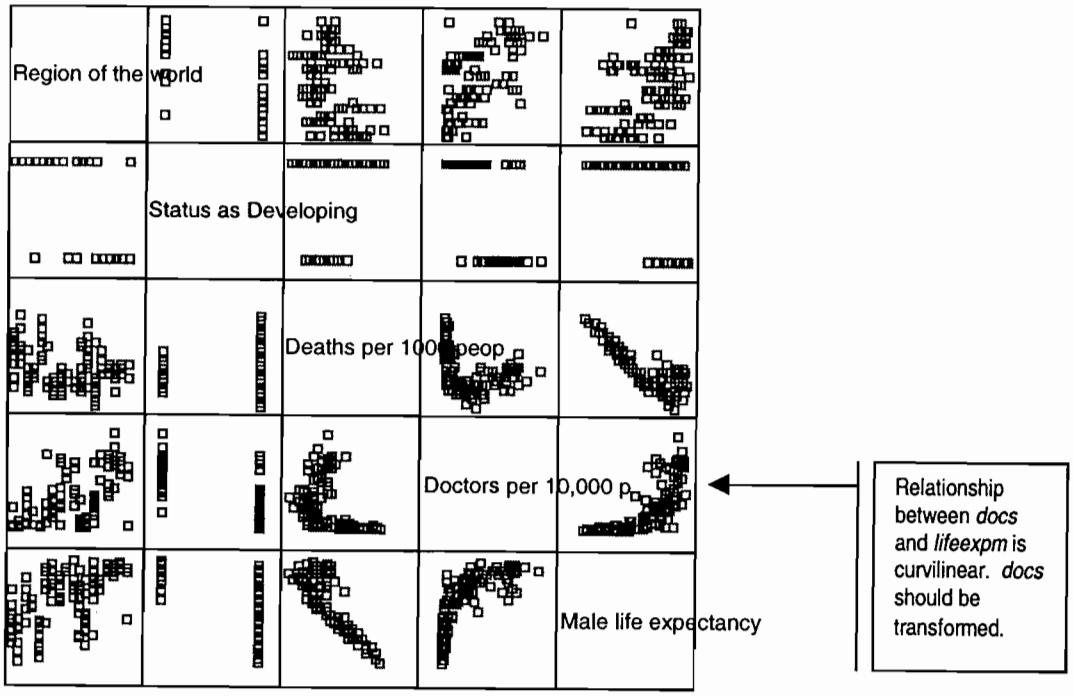


Figure 8.7 Residuals Plot for Model (Male Life Expectancy) Variables.

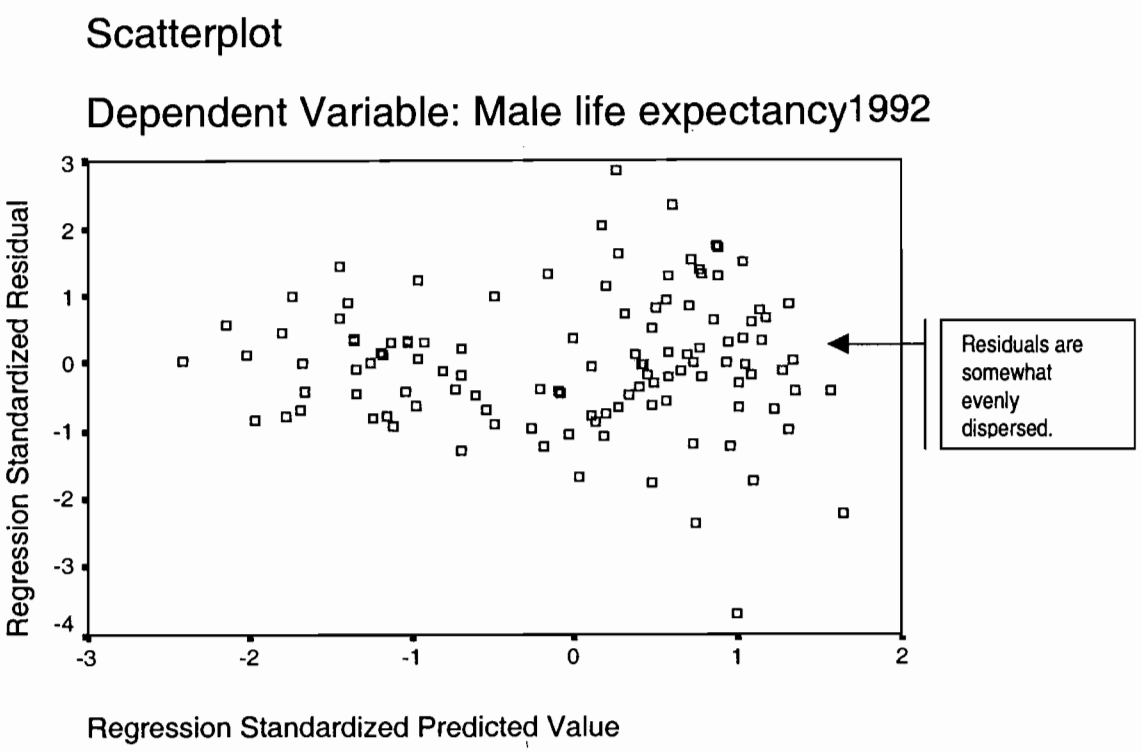


Figure 8.8 Correlation Matrix for Model (Male Life Expectancy) Variables.

Correlations

		Region of the world	Status as Developing Country	Deaths per 1000 people, 1992	Natural log of doctors per 10000	Male life expectancy 1992
Region of the world	Pearson Correlation	1.000	-.621**	-.318**	.648**	.592**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	120	120	119	119	120
Status as Developing Country	Pearson Correlation	-.621**	1.000	.125	-.596**	-.564**
	Sig. (2-tailed)	.000	.	.176	.000	.000
	N	120	120	119	119	120
Deaths per 1000 people, 1992	Pearson Correlation	-.318**	.125	1.000	-.655**	-.835**
	Sig. (2-tailed)	.000	.176	.	.000	.000
	N	119	119	119	118	119
Natural log of doctors per 10000	Pearson Correlation	.648**	-.596**	-.655**	1.000	.870**
	Sig. (2-tailed)	.000	.000	.000	.	.000
	N	119	119	118	119	119
Male life expectancy 1992	Pearson Correlation	.592**	-.564**	-.835**	.870**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.
	N	120	120	119	119	120

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 8.9 Regression Output for *deathrat* (z_3) on *region* (z_1) and *develop* (z_2).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.332[a]	.110	.095	4.40

a. Predictors: (Constant), DEVELOP, REGION

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	14.499	1.697		8.542	.000		
	REGION	-.356	.101	-.395	-3.513	.001	.607	1.648
	DEVELOP	-1.352	1.237	-.123	-1.093	.277	.607	1.648

a. Dependent Variable: DEATHRAT

Path coefficient of *deathrat* on *develop* is NOT significant.

Figure 8.10 Regression Output for *lndocs* (z_4) on *develop* (z_2).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.596[a]	.356	.350	1.2596

a. Predictors: (Constant), DEVELOP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.202	.242		13.207	.000		
	DEVELOP	-2.216	.276	-.596	-8.036	.000	1.000	1.000

a. Dependent Variable: LNDOCS

Path coefficient is significant.

Figure 8.11 Regression Output for *lifeexpm* (z_5) on *deathrat* (z_3) and *lndocs* (z_4).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935[a]	.874	.872	3.55

a. Predictors: (Constant), DEATHRAT, LNDOCS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	67.029	1.342		49.959	.000		
	LNDOCS	3.631	.277	.573	13.093	.000	.571	1.752
	DEATHRAT	-1.002	.097	-.454	-10.370	.000	.571	1.752

a. Dependent Variable: LIFEEXPM

Path coefficients are significant.

Figure 8.12 Regression Output of Missing Paths: *lndocs* (z_4) on *region* (z_1), *develop* (z_2), *deathrat* (z_3).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.859[a]	.738	.731	.8131

a. Predictors: (Constant), REGION, DEATHRAT, DEVELOP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.904	.403		9.681	.000		
	REGION	7.172E-02	.020	.236	3.643	.000	.550	1.818
	DEVELOP	-1.450	.230	-.390	-6.310	.000	.601	1.663
	DEATHRAT	-.187	.018	-.536	-10.563	.000	.892	1.121

a. Dependent Variable: LNDOCS

All path coefficients are significant.

Figure 8.13 Regression Output of Missing Paths: *lifeexpm* (z_5) on *region* (z_1), *develop* (z_2), *deathrat* (z_3) and *lndocs* (z_4).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.966[a]	.933	.930	2.62

a. Predictors: (Constant), LNDOCS, DEVELOP, REGION, DEATHRAT

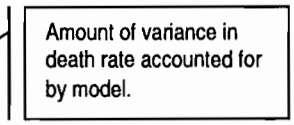
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	79.189	1.757		45.083	.000		
	REGION	6.446E-02	.067	.033	.960	.339	.493	2.030
	DEVELOP	-7.677	.861	-.326	-8.913	.000	.446	2.244
	DEATHRAT	-1.366	.080	-.618	-17.011	.000	.451	2.218
	LNDOCS	1.571	.302	.248	5.199	.000	.262	3.815

a. Dependent Variable: LIFEEXPM

Path coefficient of *lifeexpm* on *region* is not significant and should not be included.

Figure 8.14 Regression Output of Significant Paths: *deathrat* (z_3) on *region* (z_1).



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.318[a]	.101	.093	4.41

a. Predictors: (Constant), REGION

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	12.857	.790		16.273	.000		
	REGION	-.286	.079	-.318	-3.626	.000	1.000	1.000

a. Dependent Variable: DEATHRAT

Figure 8.15 Regression Output of Significant Paths: *lifeexpm* (z_5) on *develop* (z_2), *deathrat* (z_3) and *lndocs* (z_4).

Amount of variance in male life expectancy accounted for by model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.965[a]	.932	.930	2.62

a. Predictors: (Constant), LNDOCS, DEVELOP, DEATHRAT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	79.811	1.632		48.899	.000		
	DEVELOP	-7.963	.808	-.338	-9.852	.000	.506	1.977
	DEATHRAT	-1.364	.080	-.618	-17.002	.000	.451	2.217
	LNDOCS	1.665	.286	.263	5.823	.000	.293	3.417

a. Dependent Variable: LIFEEXPM

Final path coefficients.

Writing Up Results

Since the process of path analysis typically revises an initial model that was generated, the summary of results should first discuss the initial model. The path diagram with the path coefficients should be presented. Significant coefficients should be indicated with an asterisk. Within the results summary, you should state that reproduced correlations were calculated to check the model fit as well as indicate how many of the reproduced correlations were not consistent with the empirical correlations. This narration should also include a description of how the revised model was derived (i.e., theory, logic, analysis of missing paths). You should then present the revised model in narrative and pictorial form (path diagram). Your summary should also discuss the significance of the revised model path coefficients. A table that compares the empirical and reproduced correlations for the initial and revised models should be presented. Since the revised model should be a good fit, indicate that reproduced correlations are consistent with empirical correlations in the narrative. A final component of a path analysis results section is a summary table of the causal effects for the revised model. This table should include the direct, indirect, and total effects for each endogenous variable. A flag is used to indicate total effects that may be incomplete due to unanalyzed components. This table should also present R^2 for each endogenous variable. The amount of total effect (direct and indirect) for each endogenous variable should be discussed in the results summary, beginning with the endogenous variable of most interest.

Figure 8.16 summarizes the components of the results narrative for path analysis as well as how it is supported by numerous tables and figures.

Figure 8.16 Steps for Presenting Path Analysis Results.

Results Narrative		Tables/Figures
1. Present initial model: variables and flow.	→	Summarize initial model in path diagram.
2. Describe any data elimination and/or transformation.		
3. Discuss significance of path coefficients.	→	Present path coefficients in path diagram.
4. Describe how reproduced correlations were not consistent with empirical correlations.	→	Create table that compares empirical correlations to reproduce correlations for the initial model.
5. Describe process of revising model.		
6. Present revised model: variables, flow, and significant path coefficients.	→	Summarize revised model in path diagram (including path coefficients).
7. Describe how reproduced correlations were consistent with empirical correlations.	→	Create table that compares empirical correlations to reproduce correlations for the revised model.
8. Discuss causal effects for each endogenous variable: total causal effects and R^2.	→	Create table of causal effects (direct, indirect, and total) for each endogenous variable.

The following results narrative applies the output from Figures 8.5 – 8.15. Due to space constraints, we will utilize applicable figures and tables that were previously presented in the text.

> A path analysis was conducted to determine the causal effects among the variables of region of the world *(region, z_1)*, status as a developing nation (*develop, z_2*), number of deaths per 1,000 people (*deathrat, z_3*), number of doctors per 10,000 people (*docs, z_4*), and male life expectancy (*lifexpm, z_5*). Prior to the analysis, two outliers were removed. In addition, the variable of *docs* was transformed by taking its natural log. The initial model, presented in Figure 8.3, was not consistent with the empirical data. More specifically, six of the reproduced correlations exceeded a difference of .05. Tests of the missing paths in the initial model indicated that three

additional paths would significantly contribute to the model: *region* on *docs, develop* on *lifeexpm,* and *deathrat* on *docs.* In addition, the non-significant path of *develop* on *deathrat* was removed from the model. Thus, a revised model was generated and is presented in Figure 8.4. Recomputation of reproduced correlations for the revised model indicated consistency with the empirical correlations as only one reproduced correlation exceeding a difference of .05 (see Table 8.3). All path coefficients were significant at the .05 level. The direct, indirect, and total causal effects of the revised model are presented in Table 8.6. The outcome of primary interest was male life expectancy; the determinant with the largest total causal effect was death rate (-.759). The remaining determinants of male life expectancy as indicated by total causal effect were number of doctors (.726), status as a developing nation (-.441), and region (.304). This model explained approximately 93% of variance in male life expectancy. The primary determinant of the number of doctors was death rate (-.611) with region (.406) and status as a developing nation (-.390) following. Approximately 74% of variance in the number of doctors was explained by the model. The primary determinant of death rate was region (-.318), which explained approximately 10% of variance in death rate.

SECTION 8.4 SAMPLE STUDY AND ANALYSIS

This section provides a complete example that applies the entire process of conducting path analysis: development of model and research questions, data screening methods, test methods, interpretation of output, and presentation of results. The SPSS data set of *country.sav* is utilized.

Problem

In this example, we are interested in developing a causal model for explaining infant mortality. More specifically, we will investigate the causal effects among the following variables: number of doctors per 10,000 people (*docs,* z_1), gross domestic product (*gdp,* z_2), death rate per 1,000 people (*deathrat,* z_3), birth rate per 1,000 (*birthrat,* z_4), and infant mortality per 1,000 live births (*infmr,* z_4). Utilizing logic and theory, we develop the path model shown in Figure 8.17. Specific research questions generated are:

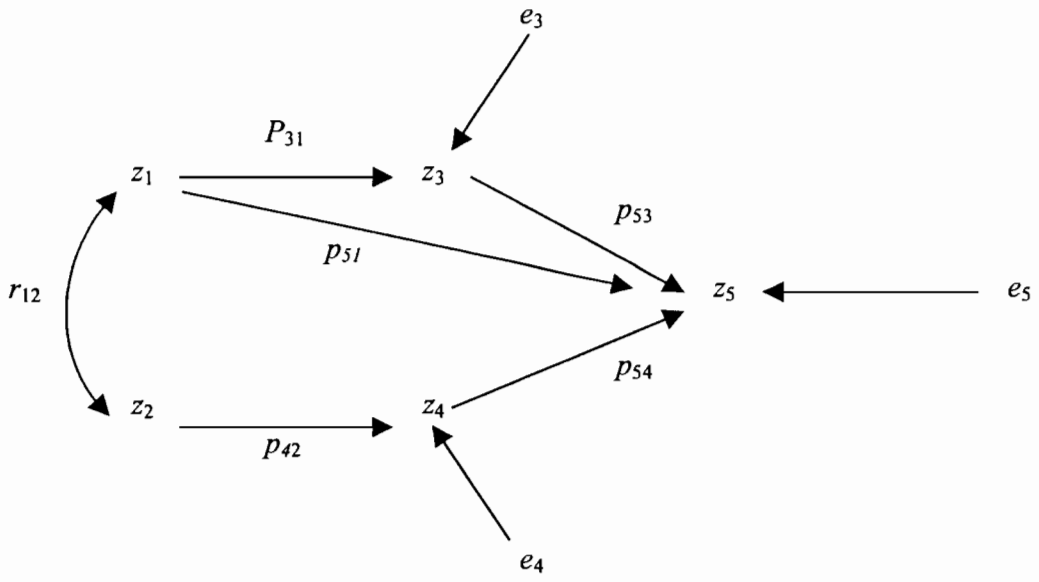
(1) Is our model—which describes the causal effects among the variables "number of doctors," "gross domestic product," "death rate per 1,000," "birth rate per 1,000," and "infant mortality per 1,000 live births"—consistent with our observed correlations among these variables?

(2) If our model is consistent, what are the estimated direct, indirect, and total causal effects among the variables?

Method, Output, and Interpretation

Since path analysis requires a great deal of interpretation throughout the process of conducting the analysis, we have combined discussion of methods, output, and interpretation in this section.

Figure 8.17 Path Diagram for the Initial Model (Infant Mortality).



z_1 = *docs (lndocs)*
z_2 = *gdp (lngdp)*
z_3 = *deathrat*
z_4 = *birthrat*
z_5 = *infmr*

Figure 8.18 Outliers Determined by Mahalanobis Distance.

Extreme Values

			Case Number	Value
MAH_1	Highest	1	72	14.74252
		2	101	14.71959
		3	37	12.30083
		4	91	12.15138
		5	81	12.02676
	Lowest	1	30	.89494
		2	50	.98185
		3	53	1.18165
		4	45	1.26494
		5	93	1.30022

No cases exceed $\chi^2(5)=20.516$.

Figure 8.19 Scatterplot for Model (Infant Mortality) Variables.

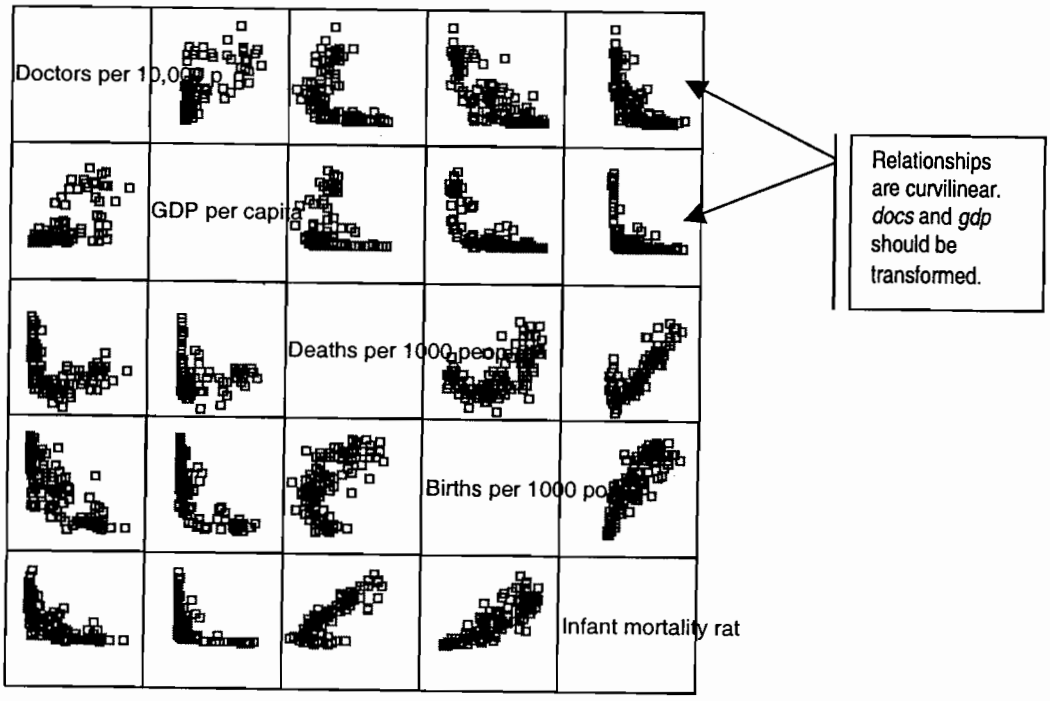


Figure 8.20 Residuals Plot for Model (Infant Mortality)Variables.

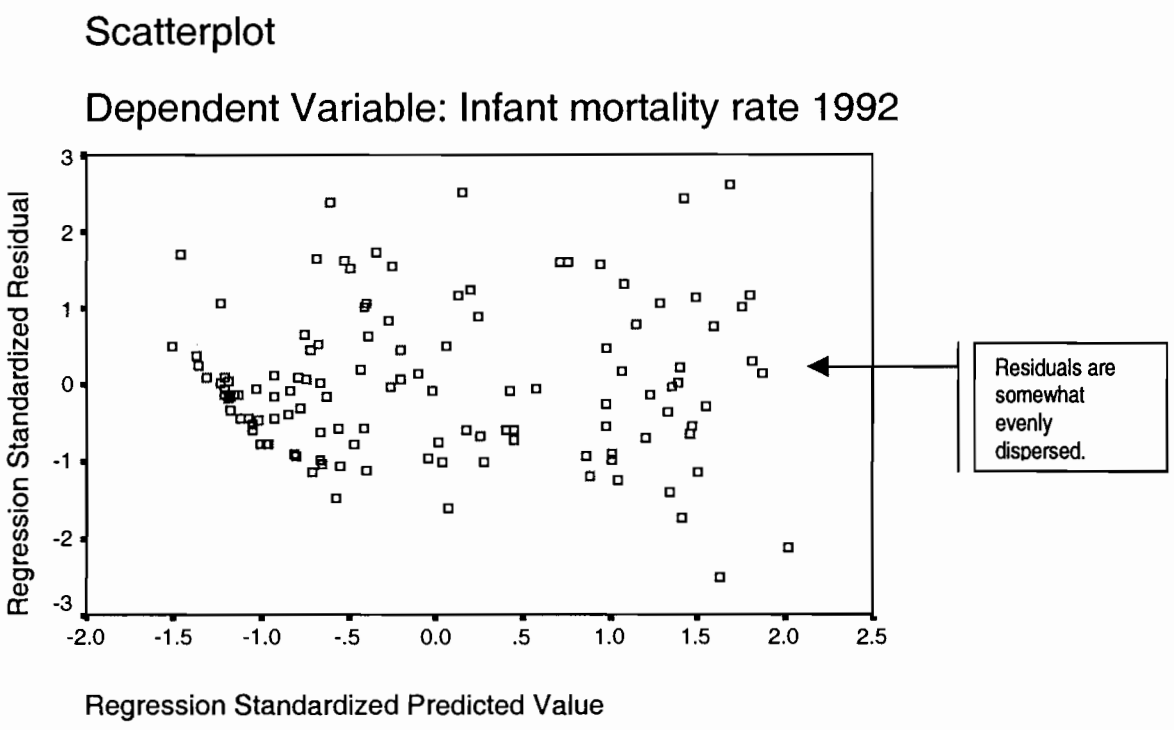


Figure 8.21 Correlation Matrix for Model (Infant Mortality Variables)

Correlations

		LNDOCS	LNGDP	DEATHRAT	BIRTHRAT	INFMR
LNDOCS	Pearson Correlation	1.000	.824**	-.643**	-.821**	-.831**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	121	121	120	120	121
LNGDP	Pearson Correlation	.824**	1.000	-.512**	-.803**	-.809**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	121	122	121	121	122
DEATHRAT	Pearson Correlation	-.643**	-.512**	1.000	.568**	.780**
	Sig. (2-tailed)	.000	.000	.	.000	.000
	N	120	121	121	121	121
BIRTHRAT	Pearson Correlation	-.821**	-.803**	.568**	1.000	.862**
	Sig. (2-tailed)	.000	.000	.000	.	.000
	N	120	121	121	121	121
INFMR	Pearson Correlation	-.831**	-.809**	.780**	.862**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.
	N	121	122	121	121	122

**. Correlation is significant at the 0.01 level (2-tailed).

Once the path model was generated, all model variables were screened for missing data outliers and tested for assumptions. Identification of outliers was done by conducting a preliminary **Regression** to calculate Mahalanobis distance. The **Explore** procedure was completed to determine if any cases exceeded the chi square criterion of 20.516 (*df*=5). No outliers were found (see Figure 8.18). Test assumptions were assessed by creating a scatterplot matrix and a residuals plot. The scatterplot matrix indicated that the variables *docs* and *gdp* were curvilinear (see Figure 8.19). These variables were transformed by taking the natural log of each. The residuals plot was then created with the transformed variables and demonstrated fair dispersion (see Figure 8.20). With normality and homoscedasticity assumed, a correlation matrix was then created for all the model variables (see Figure 8.21). Finally, the following series of **Regression** analyses were conducted for the three endogenous variables: z_3 on z_1 (see Figure 8.22); z_4 on z_2 (see Figure 8.23), and z_5 on z_1, z_3, and z_4 (see Figure 8.24). All tolerance statistics were greater than .1. Path coefficients can be seen in the path diagram (see Figure 8.25); coefficients were then used to calculate the reproduced correlations through the path decompositions, which are displayed respectively in Tables 8.7 and 8.8. A comparison of the reproduced correlations to the empirical correlations shows that six of the reproduced correlations differ by more than .05 from the empirical correlations (see Table 8.9). Consequently, we concluded that our initial model was not consistent with the empirical data. Analysis of missing paths, which essentially include all possible paths for each endogenous variable, were conducted for the following: z_3 on z_1, z_2, z_4 (see Figure 8.26); z_4 on z_1, z_2, z_3 (see Figure 8.27), and z_5 on z_1, z_2, z_3, and z_4 (see Figure 8.28). Analysis of missing paths for *deathrat* (z_3) reveals no additional paths that would contribute to the model. Evaluation of missing paths for *birthrat* (z_4) indicates that the path from *lndocs* (z_1) would significantly contribute to the model. Finally, analysis of missing paths for *infmr* (z_5) indicates two revisions: removal of the path from *lndocs* and the addition of the path from *lngdp*. In order to obtain the accurate path coefficients for our revised model, regression analysis must be conducted again utilizing only the appropriate paths: z_3 on z_1; z_4 on z_1 and z_2 (see Figure 8.29), and z_5 on z_2, z_3, and z_4 (see Figure 8.30). Since paths did not change for z_3,

the results from the original analysis may be used (see Figure 8.22). The revised model is presented in Figure 8.31. Reproduced correlations were calculated, as defined by the path decompositions (Tables 8.10 & 8.11), and were compared to the empirical correlations (see Table 8.9). Only one reproduced correlation exceeded the criterion of a .05 difference. Thus, we concluded that the revised model is consistent with empirical data. The final step was to calculate the direct, indirect, and total effects for each endogenous variable. These causal effects are presented in Table 8.12.

Figure 8.22 Regression Output for *deathrat* (z_3) on *lndocs* (z_1).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.643[a]	.414	.409	3.48

a. Predictors: (Constant), LNDOCS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	13.127	.439		29.875	.000		
	LNDOCS	-1.874	.205	-.643	-9.131	.000	1.000	1.000

a. Dependent Variable: DEATHRAT

Path coefficient is significant.

Figure 8.23 Regression Output for *birthrat* (z_4) on *lngdp* (z_2).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.803[a]	.645	.642	7.92

a. Predictors: (Constant), Natural log of GDP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	81.791	3.511		23.296	.000		
	Natural log of GDP	-6.977	.475	-.803	-14.697	.000	1.000	1.000

a. Dependent Variable: Births per 1000 population, 1992

Path coefficient is significant.

Figure 8.24 Regression Output for *infmr* (z_5) on *lndocs* (z_1), *deathrat* (z_3), and *birthrat* (z_4).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.937[a]	.879	.876	15.196

a. Predictors: (Constant), BIRTHRAT, DEATHRAT, LNDOCS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-28.326	9.337		-3.034	.003		
	LNDOCS	-4.104	1.713	-.148	-2.395	.018	.273	3.662
	DEATHRAT	3.770	.402	.396	9.379	.000	.585	1.710
	BIRTHRAT	1.720	.187	.521	9.202	.000	.326	3.067

a. Dependent Variable: INFMR

All path coefficients are significant.

Table 8.7 Path Decompositions for the Initial Model (Infant Mortality) Shown in Figure 8.17.

Reproduced Correlation	Path Decomposition
$\hat{r}_{13}$	p_{31} (D)
$\hat{r}_{14}$	$r_{12}p_{42}$ (U)
$\hat{r}_{15}$	$p_{51} + p_{31}p_{53} + r_{12}p_{42}p_{54}$ (D) (I) (U)
$\hat{r}_{23}$	$r_{12}\,p_{31}$ (U)
$\hat{r}_{24}$	p_{42} (D)
$\hat{r}_{25}$	$p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{51}$ (I) (U) (U)
$\hat{r}_{34}$	$p_{31}r_{12}p_{42}$ (S)
$\hat{r}_{35}$	$p_{53} + p_{31}p_{51} + p_{31}r_{12}p_{42}p_{54}$ (D) (I) (S)
$\hat{r}_{45}$	$p_{54} + p_{42}r_{12}p_{51} + p_{42}r_{12}p_{31}p_{53}$ (D) (S) (S)

Table 8.8 Path Decompositions and Calculation of Reproduced Correlations for the Initial Model (Infant Mortality) Shown in Figure 8.17.

$\hat{r}_{13} = p_{31}$
$= (-.643) = \mathbf{-.643}$
(U)

$\hat{r}_{14} = r_{12}p_{42}$
$= (.824)(-.803) = \mathbf{-.662}$
(U)

$\hat{r}_{15} = p_{51} + p_{31}p_{53} + r_{12}p_{42}p_{54}$
$= (-.148) + (-.643)(.396) + (.824)(-.803)(.521) = \mathbf{-.748}$
(D) (I) (U)

$\hat{r}_{23} = r_{12}p_{31}$
$= (.824)(-.643) = \mathbf{-.530}$
(U)

$\hat{r}_{24} = p_{42}$
$= (-.803) = \mathbf{-.803}$
(D)

$\hat{r}_{25} = p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{51}$
$= (-.803)(.521) + (.824)(-.643)(.396) + (.824)(-.148) = \mathbf{-.750}$
(I) (U) (U)

$\hat{r}_{34} = p_{31}r_{12}p_{42}$
$= (-.643)(.824)(-.803) = \mathbf{.425}$
(S)

$\hat{r}_{35} = p_{53} + p_{31}p_{51} + p_{31}r_{12}p_{42}p_{54}$
$= (.396) + (-.643)(-.148) + (-.643)(.824)(-.803)(.521) = \mathbf{.713}$
(D) (I) (S)

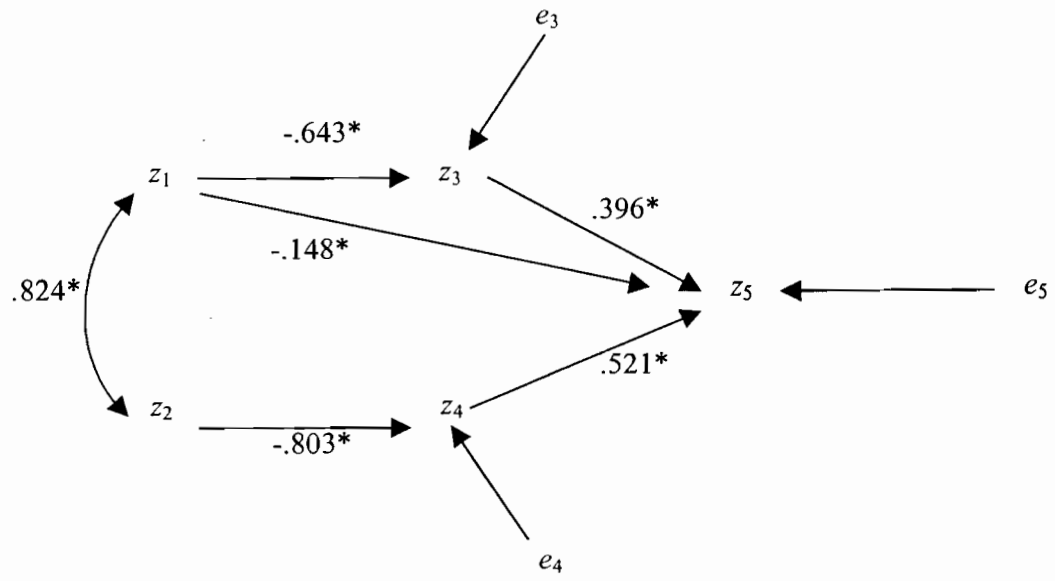
$\hat{r}_{45} = p_{54} + p_{42}r_{12}p_{51} + p_{42}r_{12}p_{31}p_{53}$
$= (.521) + (-.803)(.824)(-.148) + (-.803)(.824)(-.643)(.396) = \mathbf{.787}$
(D) (S) (S)

Table 8.9 Empirical and Reproduced Correlations for the Initial Model (Figure 8.17) and the Revised Model (Figure 8.31)

	z_1	z_2	z_3	z_4	z_5
			Observed Correlations		
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.512	1.000		
z_4	-.821	-.803	.568	1.000	
z_5	-.831	-.809	.780	.862	1.000
			Reproduced Correlations (Initial Model)		
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.530	1.000		
z_4	-.662*	-.803	.425*	1.000	
z_5	-.748*	-.750*	.713*	.787*	1.000
			Reproduced Correlations (Revised Model)		
z_1	1.000				
z_2	.824	1.000			
z_3	-.643	-.530	1.000		
z_4	-.820	-.803	.528	1.000	
z_5	-.824	-.817	.768	.738*	1.000

* Difference between reproduced and observed correlation is greater than 0.05.

Figure 8.25 Path Diagram for the Initial Model (Infant Mortality), Including Path Coefficients



z_1 = *docs (lndocs)*
z_2 = *gdp (lngdp)*
z_3 = *deathrat*
z_4 = *birthrat*
z_5 = *infmr*

Figure 8.26 Regression Output of Missing Paths: *deathrat* (z_5) on *lndocs* (z_1), *lngdp* (z_2), and *birthrat* (z_4).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.647[a]	.419	.404	3.50

a. Predictors: (Constant), Births per 1000 population, 1992, Natural log of GDP, Natural log of doctors per 10000

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	9.503	3.689		2.576	.011		
	LNDOCS	-1.899	.412	-.652	-4.605	.000	.250	4.001
	LNGDP	.346	.404	.116	.856	.394	.272	3.682
	BIRTHRAT	3.700E-02	.047	.107	.792	.430	.277	3.616

a. Dependent Variable: DEATHRAT

Path coefficients for *lngdp* on *deathrat* and *birthrat* on *deathrat* are NOT signficant and should not be included.

Figure 8.27 Regression Output of Missing Paths: *birthrat* (z_4) on *lndocs* (z_1), *lngdp* (z_2), and *deathrat* (z_3).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.851[a]	.725	.718	6.93

a. Predictors: (Constant), Deaths per 1000 people, 1992, Natural log of GDP, Natural log of doctors per 10000

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	60.082	5.040		11.921	.000		
	LNDOCS	-3.851	.814	-.459	-4.732	.000	.252	3.966
	LNGDP	-3.425	.738	-.399	-4.639	.000	.320	3.126
	DEATHRAT	.145	.184	.050	.792	.430	.584	1.712

a. Dependent Variable: BIRTHRAT

Path coefficient for *deathrat* on *birthrat* is NOT significant and should not be included.

Figure 8.28 Regression Output of Missing Paths: *infmr* (z_5) on *lndocs* (z_1), *lngdp* (z_2), *deathrat* (z_5), and *birthrat* (z_4).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.946[a]	.895	.891	14.202

a. Predictors: (Constant), Births per 1000 population, 1992, Deaths per 1000 people, 1992, Natural log of GDP, Natural log of doctors per 10000

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	25.204	15.399		1.637	.104		
	LNDOCS	-.452	1.820	-.016	-.248	.805	.211	4.732
	LNGDP	-6.946	1.646	-.245	-4.219	.000	.270	3.706
	DEATHRAT	3.896	.377	.410	10.338	.000	.581	1.721
	BIRTHRAT	1.402	.190	.425	7.374	.000	.275	3.636

a. Dependent Variable: INFMR

Path coefficient for *infmr* on *lndocs* is NOT significant and should not be included.

Figure 8.29 Regression Output for Significant Paths: *birthrat* (z_4) on *lndocs* (z_1) and *lngdp* (z_2).

Amount of variance in birth rate accounted for by model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.851[a]	.723	.719	6.92

a. Predictors: (Constant), Natural log of doctors per 10000, Natural log of GDP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	61.796	4.545		13.597	.000		
	LNDOCS	-4.149	.720	-.494	-5.761	.000	.321	3.116
	LNGDP	-3.393	.736	-.396	-4.610	.000	.321	3.116

a. Dependent Variable: BIRTHRAT

Final path coefficients for revised model.

Figure 8.30 Regression Output for Significant Paths: *infmr* (z_5) on *lngdp* (z_2), *deathrat* (z_3), and *birthrat* (z_4).

Amount of variance in infant mortality accounted for by model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.945[a]	.893	.890	14.368

a. Predictors: (Constant), Births per 1000 population, 1992, Deaths per 1000 people, 1992, Natural log of GDP

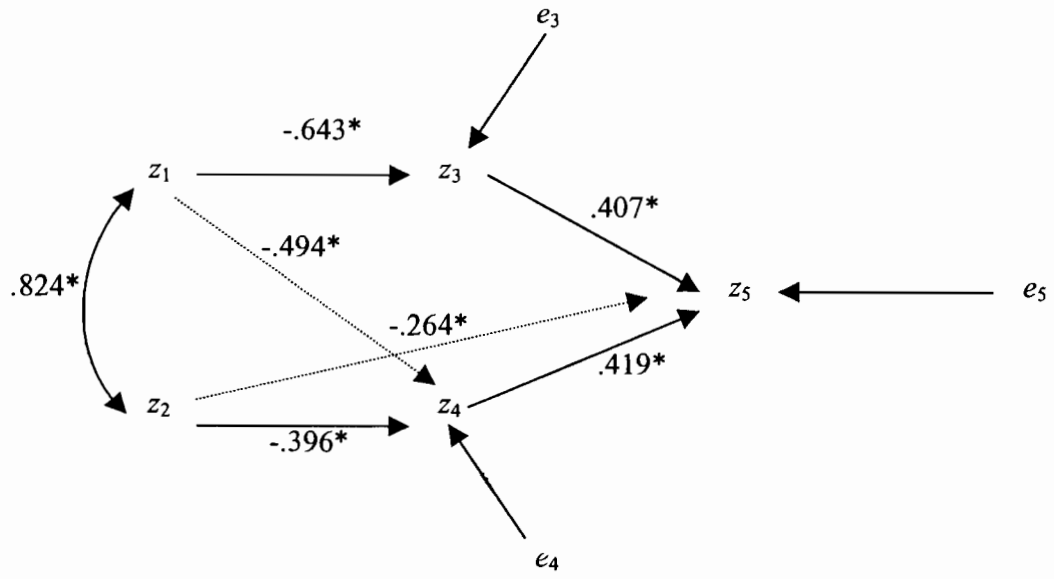
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	30.347	15.347		1.977	.050		
	LNGDP	-7.512	1.455	-.264	-5.163	.000	.351	2.852
	DEATHRAT	3.792	.344	.407	11.012	.000	.669	1.495
	BIRTHRAT	1.374	.175	.419	7.864	.000	.322	3.106

a. Dependent Variable: INFMR

Final path coefficients for revised model.

Figure 8.31 Path Diagram for the Revised Model (Infant Mortality), Including Path Coefficients.



* Significant at the .05 level.

Note. Revised paths are shown with dashed arrows.

z_1 = *docs (lndocs)*
z_2 = *gdp (lngdp)*
z_3 = *deathrat*
z_4 = *birthrat*
z_5 = *infmr*

Presentation of Results

The following summary of results applies the output from Figures 8.18 – 8.31. The reader should note that due to space constraints, we have referenced appropriate figures and tables that were previously presented in the text.

A path analysis was conducted to determine the causal effects among the variables of number of doctors per 10,000 people (*docs,* z_1), gross domestic product (*gdp,* z_2), number of deaths per 1,000 people (*deathrat,* z_3), birth rate per 1,000 people (*birthrat,* z_4) and infant mortality per 1,000 live births (*infmr,* z_5). Prior to the analysis, the variables of *docs* and *gdp* were transformed by taking the natural log. The initial model, presented in Figure 8.17, was not consistent with the empirical data. More specifically, six of the reproduced correlations exceeded a difference of .05. Tests of the missing paths in the initial model indicated that two additional paths would significantly contribute to the model: *birthrat* on *docs* and *infmr* on *gdp*. In addition, the nonsignificant path of *infmr* on *docs* was removed from the model. Thus, a revised model was generated and is presented in Figure 8.31. Computation of reproduced correlations for the revised model indicated consistency with the empirical correlations as only one reproduced correlation exceeded a difference of .05 (see Table 8.9). All path coefficients were significant at the

.05 level. The direct, indirect, and total causal effects of the revised model are presented in Table 8.12. The outcome of primary interest was infant mortality; the determinant with the largest total causal effect was number of doctors (–.469). The remaining determinants of infant mortality, as indicated by total causal effect, were gross domestic product (.430), birth rate (.419), and death rate (.407). This model explained approximately 89.3% of variance in infant mortality. The primary determinant of the birth rate was number of doctors (.494) followed by gross domestic product (.396). Approximately 72.3% of variance in the birth rate was explained by the model. The primary determinant of death rate was the number of doctors (–.643), which explained approximately 41.4% of variance in death rate.

Table 8.10 Path Decompositions for the Revised Model (Infant Mortality) Shown in Figure 8.31.

Reproduced Correlation	Path Decomposition
$\hat{r}_{13}$	p_{31} (D)
$\hat{r}_{14}$	$p_{41} + r_{12}p_{42}$ (D) (U)
$\hat{r}_{15}$	$p_{31}p_{53} + p_{41}p_{54} + r_{12}p_{52} + r_{12}p_{42}p_{54}$ (I) (I) (U) (U)
$\hat{r}_{23}$	$r_{12}\,p_{31}$ (U)
$\hat{r}_{24}$	$p_{42} + r_{12}p_{41}$ (D) (U)
$\hat{r}_{25}$	$p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{41}p_{54}$ (D) (I) (U) (U)
$\hat{r}_{34}$	$p_{31}p_{41} + p_{31}r_{12}p_{42}$ (S) (S)
$\hat{r}_{35}$	$p_{53} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$ (D) (I) (S) (S)
$\hat{r}_{45}$	$p_{54} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{41}p_{31}p_{53}$ (D) (S) (S) (S)

Table 8.11 Path Decompositions and Calculation of Reproduced Correlations for the Revised Model Shown in Figure 8.31.

$\hat{r}_{13} = p_{31}$
$= (-.643) = \mathbf{-.643}$
(D)

$\hat{r}_{14} = p_{41} + r_{12}p_{42}$
$= (-.494) + (.824)(-.396) = \mathbf{-.820}$
(D) (U)

$\hat{r}_{15} = p_{31}p_{53} + p_{41}p_{54} + r_{12}p_{52} + r_{12}p_{42}p_{54}$
$= (-.643)(.407) + (-.494)(.419) + (.824)(-.264) + (.824)(-.396)(.419) = \mathbf{-.824}$
(I) (I) (U) (U)

$\hat{r}_{23} = r_{12}\, p_{31}$
$= (.824)(-.643) = \mathbf{-.530}$
(U)

$\hat{r}_{24} = p_{42} + r_{12}p_{41}$
$= (-.396) + (.824)(-.494) = \mathbf{-.803}$
(D) (U)

$\hat{r}_{25} = p_{52} + p_{42}p_{54} + r_{12}p_{31}p_{53} + r_{12}p_{41}p_{54}$
$= (-.264) + (-.396)(.419) + (.824)(-.643)(.407) + (.824)(-.494)(.419) = \mathbf{-.817}$
(D) (I) (U) (U)

$\hat{r}_{34} = p_{31}p_{41} + p_{31}r_{12}p_{42}$
$= (-.643)(-.494) + (-.643)(.824)(-.396) = \mathbf{.528}$
(S) (S)

$\hat{r}_{35} = p_{53} + p_{31}p_{41}p_{54} + p_{31}r_{12}p_{52} + p_{31}r_{12}p_{42}p_{54}$
$= (.407) + (-.643)(-.494)(.419) + (-.643)(.824)(-.264) + (-.643)(.824)(-.396)(.419) = \mathbf{.768}$
(D) (S) (S) (S)

$\hat{r}_{45} = p_{54} + p_{42}p_{52} + p_{42}r_{12}p_{31}p_{53} + p_{41}p_{31}p_{53}$
$= (.419) + (-.396)(-.264) + (-.396)(.824)(-.643)(.407) + (-.494)(-.643)(.407) = \mathbf{.738}$
(D) (S) (S) (S)

Table 8.12 Summary of Causal Effects for Revised Model (Infant Mortality) Shown in Figure 8.31.

Outcome	Determinant	Causal Effects		
		Direct	Indirect	Total
Death Rate	Doctors	-.643*	–	-.643
(R^2 = .414)	GDP	–	–	–[+]
Birth Rate	Doctors	-.494*	–	-.494[+]
(R^2 = .723)	GDP	-.396*	–	-.396[+]
	Death Rate	–	–	–[+]
Infant Mortality	Doctors	–	-.469	-.469[+]
(R^2 = .893)	GDP	-.264*	-.166	-.430[+]
	Death Rate	.407*	–	.407[+]
	Birth Rate	.419*	–	.419[+]

* Direct effect is significant at the .05 level.
[+] Total effect may be incomplete due to unanalyzed components.

SECTION 8.5 SPSS "HOW TO"

This section presents the steps for examining and conducting path analysis using linear **Regression**. For an extensive discussion on multiple regression, the reader is referred to Chapter 7. To open the Regression Dialogue Box, select the following:

Analyze
Regression
Linear

Linear Regression Dialogue Box (see Figure 8.32)

Once in this box, click the first endogenous variable to be analyzed and move it to the Dependent Box. For our example, the first endogenous variable is *deathrat*. Click each exogenous variable that has been identified as having a causal path to the specific endogenous variable and move to the Independent(s) Box. For our initial model, we only predicted that *lndocs* would have causal effect on *deathrat*. For method, select Enter; this is the default. Next, click **Statistics**.

Linear Regression Statistics Dialogue Box (see Figure 8.33)

Within this box, select **Model Fit** and **Collinearity Diagnostics**. Then click **Continue**. Click **OK**.

Figure 8.32 Linear Regression Dialogue Box.

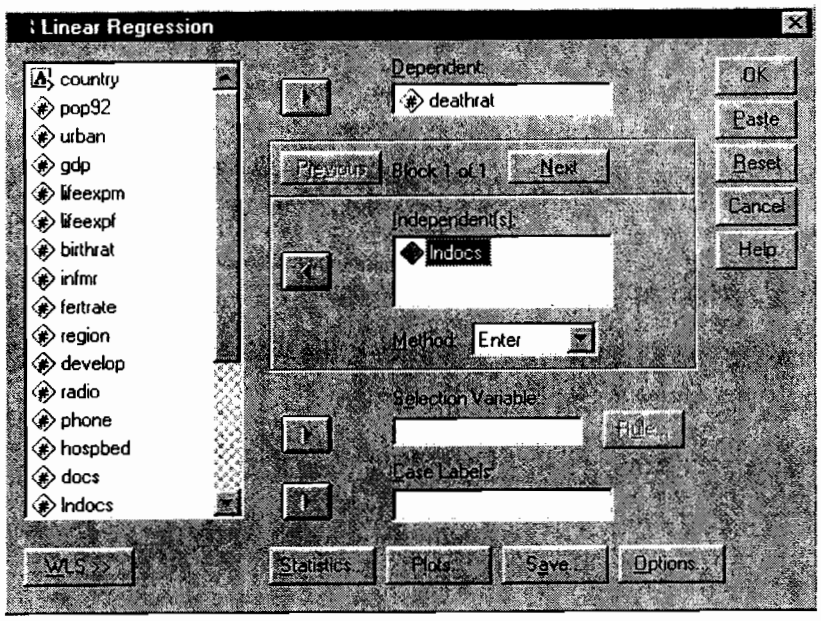
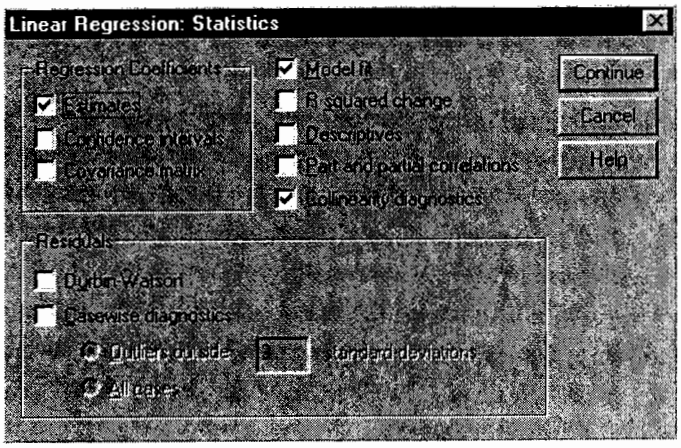


Figure 8.33 Linear Regression Statistics Dialogue Box.



This process of regression analysis would be conducted for each endogenous variable within the initial model. For our initial model for infant mortality, the following three analyses were conducted:

Analysis	Endogenous Variable	Exogenous Variables
1	*deathrat*	*lndocs**
2	*birthrat*	*lngdp**
3	*infmr*	*lndocs*, deathrat*, birthrat**

*Indicates significant path coefficients at the .05 level.

Since the initial model was not consistent with the empirical data, the following analyses were conducted to explore the significance of paths missing from the initial model:

Analysis	Endogenous Variable	Exogenous Variables
4	*deathrat*	*lndocs*, lngdp, birthrat*
5	*birthrat*	*lndocs*, lngdp*, deathrat*
6	*infmr*	*lndocs, lngdp*, deathrat*, birthrat**

*indicates significant path coefficients at the .05 level

Since these analyses revealed that only some of the paths were significant, the following analyses were conducted to determine path coefficients for those paths that were significant:

Analysis	Endogenous Variable	Exogenous Variables
7	*birthrat*	*lndocs*, lngdp**
8	*infmr*	*lngdp*, deathrat*, birthrat**

*indicates significant path coefficients at the .05 level

Since the very first analysis produced the path coefficient of *deathrat* on *lndocs*, this analysis did not need to be repeated. The reader should note that this example required eight regression analyses to create an appropriate path model—this is quite common in path analysis.

Summary

Path analysis allows the researcher to determine causal effects among numerous variables. This technique is not exploratory in nature; rather, the researcher is testing the legitimacy of a causal model that has been based upon logic, theory, and/or experience. This causal model is depicted in a path diagram, in which effects between variables are represented by arrows. A straight line with a single arrowhead represents a direct effect, while a curved line with two arrowheads represents the bivariate correlation between two variables. An indirect effect occurs when a variable intervenes between the effect of two variables. Although a path model seeks to explain the causal determinants (i.e., IVs or, in path analysis, referred to as exogenous variables) of one variable (i.e., the DV or the endogenous variable), a model may examine several endogenous variables due to indirect effects.

Once a causal model has been developed, numerous regression analyses are conducted to determine path (beta) coefficients. To test the model fit, one must calculate the reproduced correlations for each path. Reproduced correlations are calculated through the development and application of path decompositions. If several reproduced correlations differ from empirical correlations by more than .05, then the model is not consistent with the empirical data. To revise the model, one examines missing paths within the model. Utilizing only paths that are significant, the model is revised and once again tested by comparing the empirical and reproduced correlations. Once a model has very few reproduced correlations that significantly differ from the empirical, the model is said to be consistent with empirical data. Figure 8.34 provides a checklist for conducting path analysis.

Figure 8.34 Checklist for Conducting Path Analysis.

I. Develop Path Model
- a. Create path diagram.
- b. Develop path decompositions.

II. Screen Data
- a. Missing Data?
- b. Multivariate Outliers?
  - ❑ Run preliminary Regression to calculate Mahalanobis' Distance.
    1. **Analyze...Regression...Linear.**
    2. Identify a variable that serves as a case number and move to Dependent Variable box.
    3. Identify all appropriate quantitative variables and move to Independent(s) box.
    4. **Save.**
    5. Check **Mahalanobis'**.
    6. **Continue.**
    7. **OK.**
    8. Determine chi square χ^2 critical value at $p<.001$.
  - ❑ Conduct **Explore** to test outliers for Mahalanobis chi square χ^2.
    1. **Analyze...Descriptive Statistics...Explore**
    2. Move *mah_1* to Dependent Variable box.
    3. Leave Factor box empty.
    4. **Statistics.**
    5. Check **Outliers.**
    6. **Continue.**
    7. **OK.**
  - ❑ Delete outliers for subjects when χ^2 exceeds critical χ^2 at $p<.001$.
- c. Linearity, Normality, Homoscedasticity?
  - ❑ Create Scatterplot Matrix of all model variables.
    - ❑ Scatterplot shapes are not close to elliptical shapes→reevaluate univariate normality and consider transformations.
  - ❑ Run Normality Plots with Tests within **Explore.**
  - ❑ Run preliminary Regression to create residual plot.
    1. **Analyze...Regression...Linear**
    2. Move primary endogenous variable (DV) to Dependent Variable box.
    3. Move exogenous variables (IVs) to Independent(s) Variable box.
    4. **Plot.**
    5. Select ZRESID for y-axis.
    6. Select ZPRED for x-axis.
    7. **Continue.**
    8. **OK.**
  - ❑ If residuals are clustered at the top, bottom, left, or right area in plot→ reevaluate univariate normality and consider transformations.

(Figure 8.34 is continued on the next page.)

Figure 8.34 Checklist for Conducting Path Analysis. (*Continued*)

III. Conduct Multiple Regression Analyses for Path Analysis

a. Run Regression using **Linear Regression** for each endogenous variable.
   1. **Analyze...** **Regression...** **Linear**
   2. Move endogenous variable (DV) to Dependent Variable box
   3. Move exogenous variables (IVs) to Independent(s) box
   4. Select **Enter**.
   5. **Statistics.**
   6. Check **Model Fit** and **Collinearity Diagnostics.**
   7. **Continue**
   8. **OK.**

b. Interpret tolerance.
c. If tolerance for each exogenous variable is greater than .1, interpret path (beta) coefficient for each path.
d. Transfer path coefficients to path diagram.
e. Calculate reproduced correlation coefficients through path decompositions.
f. Compare reproduced correlations to empirical correlations.
g. If only a few reproduced correlations differ from the empirical correlations by more than .05, your model is fairly consistent with the empirical data. Proceed with step IV.
h. If several reproduced correlations differ from the empirical correlations by more than .05, evaluate missing paths to determine if other paths may significantly contribute to the model. Analyze missing paths by following the steps beginning with III.a.
i. Once significant paths have been determined, conduct regression analysis using only the significant paths. Analyze significant paths by following the steps beginning with III.a.

IV. Summarize Results

a. Describe path model.
b. Present path diagram.
c. Describe any data elimination or transformation.
d. Present path coefficients in path diagram.
e. Describe how reproduced correlations were not consistent with empirical correlations.
f. Describe process for revising model.
g. Present revised path diagram with path coefficients.
h. Describe how reproduced correlations were consistent with empirical correlations.
i. Present table that compares empirical correlations to reproduced correlations for both the initial and revised models.
j. Discuss causal effects for each endogenous variable: total causal effects and R^2.
k. Present table of causal effects (direct, indirect, and total) for each endogenous variable.