Stat 404

Path Analysis

A. A path model:

1. \(Z_1\) and \(Z_2\) are called “EXOGENOUS variables,” which means that their causes lie outside the path model.

2. \(Z_3\) and \(Z_4\) are called “ENDOGENOUS variables,” since they are caused by variables within the model.

3. The two-headed curved arrow represents an unanalyzed correlation. Here \(Z_1\) and \(Z_2\) are correlated but their causal relation is not indicated.

4. PATHS are single-headed arrows. These give the model direction. Here, the theoretically-grounded proposition, “\(Z_3\) is caused by \(Z_1\) and \(Z_2\),” is represented by the arrows labeled \(p_{31}\) and \(p_{32}\) respectively.

   a. Path coefficients are standardized regression coefficients. Thus, for example,

   \[
p_{31} = \hat{\beta}_1 \quad \text{and} \quad p_{32} = \hat{\beta}_2
   \]

   from the regression, \(\hat{Z}_3 = \hat{\beta}_1 Z_1 + \hat{\beta}_2 Z_2\).

   b. The regressions used to estimate the above-depicted path model are as follows:
\[ Z_3 = p_{31}Z_1 + p_{32}Z_2 + p_{3u}u \]
\[ Z_4 = p_{41}Z_1 + p_{42}Z_2 + p_{43}Z_3 + p_{4v}v \]

5. RESIDUAL PATHS (\(p_{3u}\) and \(p_{4v}\)) represent the effects of other causes (not accounted for in the model) on the endogenous variables. These causes are assumed to result from…

a. a large number of outside causes of \(Z_3\) and \(Z_4\), none of which has any great influence on them, or, more likely,…

b. a few major factors, which (in addition to a large number of minor factors) are not correlated with the exogenous variables.

For simplicity, I shall henceforth refer to \(p_{3u}\) as \(u\) and \(p_{4v}\) as \(v\). Note that for the path model depicted above, …

\[
u = \sqrt{1 - R^2_{Z_3,Z_1Z_2}}\
\[
v = \sqrt{1 - R^2_{Z_4,Z_1Z_2Z_3}} .
\]

6. Since Ordinary Least Squares regression is used in estimating the paths, all OLS assumptions that hold for regression models hold for path analysis as well. However, there are a few additional assumptions required when regressions are estimated simultaneously.

B. Assumptions (in addition to those for multiple regression) made when doing path analysis:

1. Variables are linearly and additively related. That is, polynomial and interactive expressions are NOT used in path analyses.

2. Residuals are uncorrelated with variables prior to them in the model. That is, …
\[ E(Z_1u) = E(Z_2u) = 0, \quad \text{and} \quad E(Z_1v) = E(Z_2v) = E(Z_3v) = 0. \]

3. With few exceptions (and only once briefly during this section), causation is assumed to be unidirectional. That is to say, if \( Z_3 \) causes \( Z_4 \), then \( Z_4 \) does not cause \( Z_3 \).

a. Any path model that depicts causation unidirectionally is called a RECURSIVE path model.

b. If all variables in a recursive path model are connected by paths, the model is said to be FULLY RECURSIVE.

c. Note that of all possible causal relations among one’s variables, at least half of these are set to zero in a recursive system. For example, the relations among the four variables in the above path diagram can be organized into a matrix as follows:

<table>
<thead>
<tr>
<th>Cause</th>
<th>( Z_1 )</th>
<th>( Z_2 )</th>
<th>( Z_3 )</th>
<th>( Z_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E )</td>
<td>( Z_1 )</td>
<td>—</td>
<td>( ?^1 )</td>
<td>0</td>
</tr>
<tr>
<td>( f )</td>
<td>( Z_2 )</td>
<td>( ?^1 )</td>
<td>—</td>
<td>0</td>
</tr>
<tr>
<td>( e )</td>
<td>( Z_3 )</td>
<td>( p_{31} )</td>
<td>( p_{32} )</td>
<td>—</td>
</tr>
<tr>
<td>( c )</td>
<td>( Z_4 )</td>
<td>( p_{41} )</td>
<td>( p_{42} )</td>
<td>( p_{43} )</td>
</tr>
</tbody>
</table>

Note that because their relation is left “unanalyzed” in the path diagram, \( Z_1 \) and \( Z_2 \) are not listed as causally related in the table.

\(^1\) Note that because their relation is left “unanalyzed” in the path diagram, \( Z_1 \) and \( Z_2 \) are not listed as causally related in the table.
4. Relaxing these assumptions generally precludes the use of OLS as a means of estimating path coefficients.

C. Decomposing correlation coefficients and generating their corresponding “normal equations” from standardized regression equations.

1. When $X_i$ is standardized (i.e., when $Z_i = \frac{X_i - \bar{X}}{\sigma_X}$ for $i=1,\ldots,n$), then

\[
\hat{\sigma}_Z^2 = \frac{\sum (Z_i - \bar{Z})^2}{n-1} = 1 \quad \text{and} \quad \bar{Z} = \frac{\sum Z_i}{n} = 0 \quad \text{jointly imply that} \quad \sum Z_i^2 = n - 1 \approx n. ^2
\]

2. Now consider the formula that one would use in calculating a correlation coefficient between any pair of these standardized variables:

\[
r_{Z_XZ_Y} = \frac{\sum (Z_X - \bar{Z}_X)(Z_Y - \bar{Z}_Y)}{\sqrt{\sum (Z_X - \bar{Z}_X)^2} \sqrt{\sum (Z_Y - \bar{Z}_Y)^2}} = \frac{\sum Z_XZ_Y}{\sqrt{\sum Z_X^2} \sqrt{\sum Z_Y^2}} = \frac{\sum Z_XZ_Y}{\sqrt{n \times n}} = \frac{\sum Z_XZ_Y}{n}
\]

When given the correlation coefficients associated with a fully recursive path diagram, one can decompose these correlations by (1) substituting into this formula the below expressions for subsequent variables in the path model and then (2) simplifying based on the equivalences, \(\frac{\sum Z_i^2}{n} = 1\) and \(\frac{\sum Z_XZ_Y}{n} = r_{Z_XZ_Y}\).

3. Expressions for the variables in our path model are as follows:

\[
\begin{align*}
Z_1 &= r_{12}Z_2 + e_1 \\
Z_2 &= r_{12}Z_1 + e_2 \\
Z_3 &= p_{31}Z_1 + p_{32}Z_2 + p_{3u}u
\end{align*}
\]

Regarding these first two equations, recall (e.g., from Lab 2) that in bivariate regression, the standardized slope equals the correlation coefficient.

\[\text{Note that the “–1” in “n-1” is ignored in the next pages. It makes no difference in the mathematics that follow.}\]
\[ Z_4 = p_{41}Z_1 + p_{42}Z_2 + p_{43}Z_3 + p_{44}v \]

4. A decomposed correlation matrix:

\[
\begin{array}{cccc}
Z_1 & Z_2 & Z_3 & Z_4 \\
\hline
Z_1 & - & r_{12} & - & - \\
Z_2 & p_{31} + p_{32}r_{12} & p_{32} + p_{31}r_{12} & - & - \\
Z_3 & - & p_{41} + p_{42}r_{12} & p_{42} + p_{41}r_{12} & p_{43} + p_{41}p_{31} + p_{41}p_{32}r_{12} \\
Z_4 & - & - & p_{43}p_{32} + p_{43}p_{31}r_{12} & p_{42}p_{32} + p_{42}p_{31}r_{12} \\
\end{array}
\]

The following subsections show how each cell in this table was obtained.

a. Cell (2,1):

\[
r_{12} = \frac{\sum Z_1Z_2}{n} = \frac{\sum Z_1(r_{12}Z_1 + e_2)}{n} = r_{12}\left(\frac{\sum Z_1^2}{n}\right) + \frac{\sum Z_1e_2}{n} = r_{12}(1) + 0 = r_{12}
\]

A few comments:

i. The equivalence, \( \frac{\sum Z_1e_2}{n} = 0 \), follows from the \( E\{X^T e\} = E\{X^T\}^* E\{e\} \)

assumption and the zero covariance between \( Z_1 \) and \( e_2 \) that it implies.

ii. Neither this “decomposition” nor its counterpart,

\[
r_{12} = \frac{\sum Z_2Z_1}{n} = \frac{\sum Z_2(r_{12}Z_1 + e_1)}{n} = r_{12}\left(\frac{\sum Z_2^2}{n}\right) + \frac{\sum Z_2e_1}{n} = r_{12}(1) + 0 = r_{12},
\]

is particularly enlightening. This is understandable given that the relation

between \( Z_1 \) and \( Z_2 \) is left unanalyzed in the path model.

b. Cell (3,1):
\[ r_{13} = \frac{\sum Z_1 Z_3}{n} = \frac{\sum Z_1 (p_{31} Z_1 + p_{32} Z_2 + u)}{n} \]
\[ = p_{31} \left( \frac{\sum Z_1^2}{n} \right) + p_{32} \left( \frac{\sum Z_1 Z_2}{n} \right) + \frac{\sum Z_1 u}{n} = p_{31} + p_{32} r_{12} \]

c. Cell (3,2):
\[ r_{23} = \frac{\sum Z_2 Z_3}{n} = \frac{\sum Z_2 (p_{31} Z_1 + p_{32} Z_2 + u)}{n} \]
\[ = p_{31} \left( \frac{\sum Z_1 Z_2}{n} \right) + p_{32} \left( \frac{\sum Z_2^2}{n} \right) + \frac{\sum Z_2 u}{n} = p_{31} r_{12} + p_{32} \]

d. Time to pause and get some perspective on what is going on here:

i. We have decomposed \( r_{13} \) and \( r_{23} \) into two parts—a direct effect (indicated with a red arrow) and an unanalyzed effect (indicated with a series of blue arrows, respectively “through” \( Z_2 \) and \( Z_1 \)).

ii. This was performed in two steps:

1) In the former case, \( r_{13} = \frac{\sum Z_1 Z_3}{n} \).

2) The latest (in the time sequence displayed in the path diagram) variable is substituted by the regression equation in which it is the dependent variable.

For example, \( r_{13} = \frac{\sum Z_1 (p_{31} Z_1 + p_{32} Z_2 + u)}{n} \).

3) The equation is then simplified in terms of unknown path coefficients (i.e., p’s) and known correlations (or r’s).
iii. The NORMAL EQUATIONS for an endogenous variable consist of the set of equations in which correlations between a single endogenous variable and all variables causally prior to it have been simplified in this way. For example, the normal equations for the endogenous variable, $Z_3$, are as follows:

$$r_{13} = p_{31} + p_{32}r_{12}$$
$$r_{23} = p_{31}r_{12} + p_{32}$$

Thinking back upon our earlier discussion of finding a unique OLS solution for $\hat{\beta}$, here there likewise exists a unique solution for the unknowns, $p_{31}$ and $p_{32}$, given that these are two equations with two unknowns. (We shall return to this point later within the context of “identification.”)

e. Cell (4,1):

$$r_{14} = \frac{\sum Z_1Z_4}{n} = \frac{\sum Z_1(p_{41}Z_1 + p_{42}Z_2 + p_{43}Z_3 + v)}{n}$$

$$= p_{41}\left(\frac{\sum Z_1^2}{n}\right) + p_{42}\left(\frac{\sum Z_1Z_2}{n}\right) + p_{43}\left(\frac{\sum Z_1Z_3}{n}\right) + \frac{\sum Z_1v}{n}$$

$$= p_{41} + p_{42}r_{12} + p_{43}r_{13} + 0 = p_{41} + p_{42}r_{12} + p_{43}(p_{31} + p_{32}r_{12})$$

$$= p_{41} + p_{42}r_{12} + p_{43}p_{31} + p_{43}p_{32}r_{12}$$
Note that the decomposed correlation’s four parts consist of the direct effect of $Z_1$ on $Z_4$ (i.e., $p_{41}$), the indirect effect of $Z_1$ on $Z_4$ “through” $Z_3$ (i.e., $p_{43}p_{31}$), and two unanalyzed effects ($p_{42}r_{12}$ and $p_{43}p_{32}r_{12}$). These effects are displayed graphically above, with the direct effect in red, the indirect effect in green, and the two unanalyzed effects (conspicuous due to their beginning with the unanalyzed relation between $Z_1$ and $Z_2$) in blue.

f. Cell (4,2):

$$r_{24} = \frac{\sum Z_2 Z_4}{n} = \frac{\sum Z_2 (p_{41}Z_1 + p_{42}Z_2 + p_{43}Z_3 + \nu)}{n}$$

$$= p_{41} \left( \frac{\sum Z_1 Z_2}{n} \right) + p_{42} \left( \frac{\sum Z_2^2}{n} \right) + p_{43} \left( \frac{\sum Z_2 Z_3}{n} \right) + \frac{\sum Z_2 \nu}{n}$$

$$= p_{41}r_{12} + p_{42} + p_{43}r_{23} + 0 = p_{41}r_{12} + p_{42} + p_{43} \left( p_{31}r_{12} + p_{32} \right)$$

$$= p_{41}r_{12} + p_{42} + p_{43}p_{31}r_{12} + p_{43}p_{32}$$

g. Cell (4,3):

$$r_{34} = \frac{\sum Z_3 Z_4}{n} = \frac{\sum Z_3 (p_{41}Z_1 + p_{42}Z_2 + p_{43}Z_3 + \nu)}{n}$$
The four magenta-colored effects are spurious ones between endogenous variables, but that are “explained” by prior variables in the path diagram. Beyond these spurious effects there is only the single direct effect (in red), $p_{43}$.

h. Summarizing: In the above decomposition, correlations among the path model’s variables can be broken down into four types of effects:

i. **Direct** effects (DE): The direct path between two variables.

ii. **Indirect** effects (IE): A path mediated by an **intervening** variable. This type of effect has been previously referred to as “interpretation” or “mediation.”

iii. **Unanalyzed** (or associational) effects (U): A path originating with one exogenous variable that is mediated by another exogenous variable.
iv. **Spurious** effects due to common causes (S): That component of a correlation that is due to each variable’s correlation with a causally prior variable. This type of effect has been previously referred to as “explanation.”

5. Three concluding remarks about fully recursive models:

a. As we have just demonstrated, it is always possible to perfectly reproduce a correlation matrix from the paths in a fully recursive model.

b. Second, the effects depicted in a path model can be classified exclusively as either “real” (i.e., direct or indirect) or “not real” (i.e., unanalyzed or spurious). Note that your path model (i.e., a set of hypotheses grounded in your theory) allows you to clearly differentiate between effects that “are” versus “are not” real. Since social scientists tend to concern themselves with “real effects,” it may be useful to construct tables that list only each path model’s direct, indirect, and total effects. Let’s start with a few definitions:

i. **Total Effect** = Direct Effect + Indirect Effect  
   (That was easy.)

ii. The **direct effect** from $Z_i$ to $Z_j$ consists of the path, $p_{ji}$.  
   (No more; no less.)

iii. The **indirect effect** from $Z_i$ to $Z_j$ is the sum of the products of each set of paths other than $p_{ji}$ that connect $Z_i$ to $Z_j$ unidirectionally but not through an unanalyzed correlation. For example, consider the following path diagram:
iv. The effects table for this diagram is as follows:

<table>
<thead>
<tr>
<th>Variables</th>
<th>Direct</th>
<th>Indirect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent</td>
<td>Independent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_5$</td>
<td>$Z_1$</td>
<td>$p_{51}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$Z_2$</td>
<td>$p_{52}$</td>
<td></td>
</tr>
<tr>
<td>$Z_3$</td>
<td>$Z_1$</td>
<td>$p_{31}$</td>
<td>$p_{31}p_{35}$</td>
</tr>
<tr>
<td></td>
<td>$Z_2$</td>
<td>$p_{32}$</td>
<td>$p_{32}p_{35}$</td>
</tr>
<tr>
<td></td>
<td>$Z_5$</td>
<td>$p_{35}$</td>
<td>0</td>
</tr>
<tr>
<td>$Z_4$</td>
<td>$Z_1$</td>
<td>$p_{41}$</td>
<td>$p_{31}p_{43}$</td>
</tr>
<tr>
<td></td>
<td>$Z_2$</td>
<td>$p_{42}$</td>
<td>$p_{32}p_{43}$</td>
</tr>
<tr>
<td></td>
<td>$Z_5$</td>
<td>$p_{45}$</td>
<td>$p_{35}p_{43}$</td>
</tr>
<tr>
<td>$Z_3$</td>
<td>$p_{43}$</td>
<td>0</td>
<td>$p_{43}$</td>
</tr>
</tbody>
</table>

* Note: Effects in purple are effects added when moving from the 4-variable fully recursive path model diagrammed at the beginning of this section to the above 5-variable fully recursive model.

v. As we shall see shortly, one also can construct such “Effects Tables” for recursive models that are not fully recursive.

c. The third comment on fully recursive models involves the concept of identification.

A fully recursive model is always just identified, which means that there is a single (i.e., unique) solution to the system of equations described by one’s path diagram.

The principle at work here is that a unique solution to a system of equations exists whenever a set of “k” distinct equations has “k” unknowns. To illustrate this principle, let’s work with the following formula:

$$\hat{\beta} = R^{-1}r$$
i. In the three variable case (i.e., when k=2) and prior to inverting R, this expression consists of a system of two equations with two unknowns:

\[
r = \begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix} = R\beta = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}
\]

OR

\[
r_{13} = \hat{\beta}_1 + \hat{\beta}_2 r_{12} = p_{31} + p_{32} r_{12} \\
r_{23} = \hat{\beta}_1 r_{12} + \hat{\beta}_2 = p_{31} r_{12} + p_{32}
\]

Note that these are the normal equations (mentioned previously on page 7 of these Lecture Notes) for the endogenous variable, \( Z_3 \).

ii. To solve these equations one must first obtain the inverse of the R-matrix:

\[
R^{-1} = \begin{bmatrix}
1 & -r_{12} \\
1 - r_{12}^2 & 1 - r_{12}^2 \\
- r_{12} & 1 - r_{12}^2 \\
1 - r_{12}^2 & 1
\end{bmatrix}
\]

iii. Solving for \( \beta \) yields the familiar formulae for standardized slopes in the trivariate case:

\[
\beta = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} p_{31} \\ p_{32} \end{bmatrix} = R^{-1} r = \begin{bmatrix}
1 & -r_{12} \\
1 - r_{12}^2 & 1 - r_{12}^2 \\
- r_{12} & 1 - r_{12}^2 \\
1 - r_{12}^2 & 1
\end{bmatrix} \begin{bmatrix} r_{13} \\ r_{23} \end{bmatrix} = \begin{bmatrix}
r_{13} - r_{23} r_{12} \\
r_{23} - r_{13} r_{12} \\
1 - r_{12}^2 \\
1 - r_{12}^2
\end{bmatrix}
\]

iv. These two formulae are thus the unique solution for \( \beta \) when \( r_{12}, r_{13}, \) and \( r_{23} \) are known.

D. At this point we have only considered fully recursive path models. As just discussed, these models are always just identified (i.e., they afford the same number of equations as
unknowns for each of the model’s endogenous variables). Beyond just identified path models there are also overidentified and underidentified path models.

1. **Overidentification** (the recursive case only)

   a. An overidentified path model is one that yields fewer unknown paths than equations, where these equations are produced by decomposing its known correlations between any of its endogenous variables and each of the variables it is regressed upon. Such decompositions can be done in all recursive path models—the only type of overidentified path models we shall consider. Let us begin with the following path diagram:

   To estimate these paths using OLS, one would find estimates for the standardized slopes (or paths) in the following three regression models:

   \[
   Z_2 = p_{21}Z_1 + u
   \]

   \[
   Z_3 = p_{31}Z_1 + p_{32}Z_2 + v
   \]

   \[
   Z_4 = p_{41}Z_1 + p_{43}Z_3 + w
   \]

   b. Note that \( p_{42}Z_2 \) is missing from the third regression equation. Put differently, one might speak of this recursive model as one in which the path, \( p_{42} \), has been set to
zero. Statisticians refer to each path that has been set to zero as an overidentifying restriction in the path model. For example, the above path diagram depicts a path model with one overidentifying restriction (namely, that $p_{42} = 0$). The more overidentifying restrictions, the fewer unknowns in your path model.

c. Decomposing the correlations between each endogenous variable and each variable prior to it in the path diagram, yields three sets of normal equations:

i. One equation with one unknown:

$$r_{12} = p_{21}$$

ii. Two equations with two unknowns:

$$r_{13} = p_{31} + p_{32}r_{12}$$

$$r_{23} = p_{31}r_{12} + p_{32}$$

iii. Three equations with two unknowns (i.e., something new):

$$r_{14} = p_{41} + p_{43}r_{13}$$

$$r_{24} = p_{41}r_{13} + p_{42}r_{23}$$

$$r_{34} = p_{41}r_{13} + p_{43}$$

iv. The process of decomposition required in obtaining the first two sets of normal equations has already been described on pp. 4-7. The third set is derived as follows:

$$r_{14} = \frac{\sum Z_1 Z_4}{n} = \frac{\sum Z_1(p_{41}Z_1 + p_{43}Z_3 + w)}{n}$$

$$= p_{41} \left( \frac{\sum Z_1^2}{n} \right) + p_{43} \left( \frac{\sum Z_1Z_3}{n} \right) + \frac{\sum Z_1w}{n} = p_{41} + p_{43}r_{13}$$
\[ r_{24} = \frac{\sum Z_2 Z_4}{n} = \frac{\sum Z_2 (p_{41} Z_1 + p_{43} Z_3 + w)}{n} \]

\[ = p_{41} \left( \frac{\sum Z_1 Z_2}{n} \right) + p_{43} \left( \frac{\sum Z_2 Z_3}{n} \right) + \frac{\sum Z_2 w}{n} = p_{41} r_{12} + p_{43} r_{23} \]

\[ r_{34} = \frac{\sum Z_3 Z_4}{n} = \frac{\sum Z_3 (p_{41} Z_1 + p_{43} Z_3 + w)}{n} \]

\[ = p_{41} \left( \frac{\sum Z_1 Z_3}{n} \right) + p_{43} \left( \frac{\sum Z_3^2}{n} \right) + \frac{\sum Z_3 w}{n} = p_{41} r_{13} + p_{43} \]

v. Thus the only unknowns among the three equations are \( p_{41} \) and \( p_{43} \). Unique solutions for these unknowns can be obtained by solving any pair among the equations. However, it is unlikely that the solutions found using each pair of equations will be identical!!! (Note: When using Ordinary Least Squares, the unique solution obtained using \( r_{14} = p_{41} + p_{43} r_{13} \) and \( r_{34} = p_{41} r_{13} + p_{43} \) is used.)

d. Of course, it is possible that the solutions for \( p_{41} \) and \( p_{43} \) are identical among all three pairs of normal equations. However, this will only occur if the assumptions of path analysis are met and if the model’s overidentifying restriction is correctly specified (i.e., if it is true that \( p_{42} = 0 \)). Put differently, \( r_{14} = p_{41} + p_{43} r_{13} \) and \( r_{34} = p_{41} r_{13} + p_{43} \) will only be appropriate in estimating \( p_{41} \) and \( p_{43} \) if specific “components” of \( r_{14} \) and \( r_{34} \) equal zero.
i. Let’s begin by comparing the components of $r_{14}$ between this overidentified model and the just identified model in which $p_{42}$ is also estimated:

Just identified: $r_{14} = p_{41} + p_{42}p_{21} + p_{43}p_{31} + p_{43}p_{32}p_{21}$

Overidentified: $r_{14} = p_{41} + p_{43}p_{31} + p_{43}p_{32}p_{21}$

ii. Now, we can compare the components of $r_{34}$ between the overidentified model and the same just identified model:

Just identified: $r_{34} = p_{41}p_{31} + p_{41}p_{32}p_{21} + p_{42}p_{31}p_{21} + p_{42}p_{32} + p_{43}$

Overidentified: $r_{34} = p_{41}p_{31} + p_{41}p_{32}p_{21} + p_{43}$

iii. Thus, OLS estimation based on $r_{14} = p_{41} + p_{43}r_{13}$ and $r_{34} = p_{41}r_{13} + p_{43}$ will only generate unbiased estimators of $p_{41}$ and $p_{43}$ if each of the following is true:

1) $p_{42}p_{21} = 0$ There is no indirect effect of $Z_1$ on $Z_4$ through $Z_2$.

2) $p_{42}p_{31}p_{21} = 0$ There is no spurious effect of $Z_3$ on $Z_4$ due to $Z_1$’s direct effect on $Z_3$ and its indirect effect on $Z_4$ through $Z_2$.

3) $p_{42}p_{32} = 0$ There is no spurious effect of $Z_3$ on $Z_4$ due to $Z_2$’s direct effects on each.
4) In conclusion, the path model will be misspecified to the extent that these three effects as well as the direct effect of $Z_2$ on $Z_4$ are nonzero, thereby yielding biased estimators of $p_{41}$ and $p_{43}$.

e. Note how this concluding image of path analysis resonates with the dual objectives of scientific research:

i. On the one hand, one seeks the most parsimonious among all possible path models. That is, one seeks “that model” with the most overidentifying restrictions (i.e., with the least parameters being estimated).

ii. On the other hand, one seeks a model that accurately describes the data. This is done in a manner similar to that done with hierarchically-related regression models, only now we compare hierarchically-related path models.

iii. When an overidentifying restriction is introduced in a path model, a testable hypothesis is created. Introducing some new notation for a standardized partial slope within a system of equations, the null hypothesis in this case is as follows:

1) $H_0: \beta_{42,13} = 0$, where $Z_4$ is the associated dependent variable, $Z_2$ is the associated independent variable, and $Z_1$ and $Z_3$ are all other independent variables in the regression of $Z_4$ on $Z_2$.

2) Since $p_{42} = \hat{\beta}_{42,13}$, this hypothesis is easily tested (in a 2-tailed test) by rejecting this null hypothesis if $\left| \frac{\hat{\beta}_{42,13}}{\sigma_{\hat{\beta}_{42,13}}} \right| > z_{a/2}$ (i.e., by dividing the partial slope by its standard error).
3) However, when one’s path model contains more than one identifying restriction, the test is more complex.

f. Testing Overidentified Path Models (a.k.a. finding the most parsimonious, best-fitting path model):

i. Let’s assume that we have randomly sampled n=40 ISU researchers in the social sciences, and that we try to reconstruct their careers using path analysis. We estimate the following path model:

![Diagram of path model]

ii. The path model depicted here has three overidentifying restrictions (namely, 
\[ p_{31} = p_{42} = p_{52} = 0 \].

iii. Imagine that the R-squared values for the fully recursive model are as follows:
\[ R^2_{3,12} = .1539, \ R^2_{4,123} = .1680, \text{ and } R^2_{5,1234} = .5528. \]

iv. Since we know that each residual path, \( e \), is calculated as \( e = \sqrt{1 - R^2} \) for the regression model associated with the endogenous variable toward which its arrow
points, we can calculate the squares of the fully recursive model’s residual paths using the formula, $e^2 = 1 - R^2$:

$$e_3^2 = 1 - R_{3,12}^2 = 1 - .1539 = .8461$$

$$e_4^2 = 1 - R_{4,123}^2 = 1 - .1680 = .8320$$

$$e_5^2 = 1 - R_{5,1234}^2 = 1 - .5528 = .4472$$

v. We can likewise calculate the squares of the restricted model’s residual paths:

$$u^2 = (.9667)^2 = .9345$$

$$v^2 = (.9293)^2 = .8636$$

$$w^2 = (.7084)^2 = .5018$$

vi. Pedhazur (2nd edition, pp. 618-628) describes a goodness of fit statistic (developed by Specht [1975]) for testing whether a restricted model differs significantly from a fully recursive model. Calculation of the statistic for our path model is described below in cookbook fashion:

1) First, calculate the square of each residual path in both the fully recursive and the more restricted models:

<table>
<thead>
<tr>
<th></th>
<th>$Z_3$</th>
<th>$Z_4$</th>
<th>$Z_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Recursive</td>
<td>.8461</td>
<td>.8320</td>
<td>.4472</td>
</tr>
<tr>
<td>Restricted</td>
<td>.9345</td>
<td>.8636</td>
<td>.5018</td>
</tr>
</tbody>
</table>

2) Calculate the goodness of fit statistic, $Q$, by multiplying these squares as follows:

$$Q = \frac{(8461)(8320)(4472)}{(9345)(8636)(5018)} = .7773$$
Note that since the denominator is always larger than the numerator (i.e., the residuals from the restricted model will always be larger than or equal to those from its corresponding fully recursive model), the range of possible values of $Q$ are between zero (all variance in one or more of the endogenous variables is explained within the fully recursive model) and one (the restricted path model explains the same amount of each endogenous variable’s variance as does the fully recursive model).

3) Count the restricted model’s number of overidentifying restrictions, and call this number, $d$. In this case $d = 3$.

4) Calculate $W = -(n - d)\ln(Q)$, which is the goodness-of-fit statistic that we seek. $W$ has a chi-square distribution with $d$ degrees of freedom. In this case, the empirical value of chi-square is as follows:

$$W = -(40 - 3)\ln(0.7773) = 9.32$$

5) Compare $W$ to the critical value of chi-square, $\chi^2_{3.05} = 7.815$, obtained by consulting a chi-square table. Note that the larger the empirical chi-square (i.e., the larger $W$ is), the greater the difference in variance explained between the fully recursive and the reduced models. The conclusion in this case is thus that the former model explains significantly more variance than the latter:

The reduced model has the wrong combination of overidentifying restrictions.

*Another model should be fit.* (A cautionary note: $W$ is likely to be large when one’s sample size is large. For large samples, you may wish to compare the relative sizes of residual terms in the full and restricted models as well as to evaluate how close $Q$ is to 1. In such cases, you may judge the models not
to be different enough to opt for the fully recursive model, despite a “significant” difference between the two.)

6) Note: The alternative hypothesis being tested here is that the generalized proportion of variance explained by the fully recursive model (namely, $R_m^2 = 1 - \left( e_i^2 \times e_i^2 \times e_i^2 \right)$) is significantly larger than the generalized proportion of variance explained by the reduced model (namely, $M = 1 - \left( u^2 \times v^2 \times w^2 \right)$). If there is a significant drop from the former to the latter, too much explained variance was lost when overidentifying restrictions were added (i.e., paths were set to zero) in the reduced model. Null and alternative hypotheses can be written as follows:

$$H_0: R_m^2 = M$$
$$H_A: R_m^2 > M$$

7) Pedhazur (2nd edition, pp. 627-8) also describes a goodness of fit statistic for comparing two restricted path models that are nested one within another (i.e., when all paths in the more restricted path model are estimated in the less restricted path model). The only difference between this case and the one just discussed is that in the formula for $W$, the degrees of freedom subtracted from the sample size equals the difference in the number of overidentifying restrictions in the more restricted model minus the number of such restrictions in the less restricted model. This difference is also the number of degrees of freedom used when finding the critical value for chi-square. The conclusion to be drawn from such a test is whether (when chi-square is large) the model with fewer or (when chi-square is small) with more overidentifying
restrictions is more parsimonious. Of course, subsequent tests might show neither path model to be more parsimonious than a model within which both are nested (e.g., the fully recursive model) or a model that is nested within both. In this case, null and alternative hypotheses are as follows:

\[ H_0: M_1 = M_2 \]
\[ H_A: M_1 > M_2 \]

2. **Underidentification**

a. This discussion will make use of an example modified from Wonnacott and Wonnacott (1970, pp. 172-183).³

b. Consider the following nonrecursive path diagram:

```
Price consumers are willing to pay (Z_p)
P_{QP}
```
```
Quantity available to customers (Z_q)
P_{Qv}
```

After removing variance in Quantity-available due to the Law of Supply, there is still residual variance in Quantity-available that inhibits Price (due to the Law of Demand).

```
Price consumers are willing to pay (Z_p)
P_{Pu} \rightarrow u
```
```
Quantity available to customers (Z_q)
P_{Qv} \rightarrow v
```

After removing variance in Price-willingness due to the Law of Demand, there is still residual variance in Price-willingness that stimulates Quantity (due to the Law of Supply).

³ The modification is the depiction in the above diagram of demand as “the price one is willing to pay” rather than as “the willingness to pay for a given quantity.” Wonnacott and Wonnacott (1970) reverse the direction of \( P_{Qv} \) for this reason. If this were done in the diagram, the depiction would get even messier, requiring effects, \( P_{QP_1} \) and \( P_{QP_2} \).
ii. $P_{PQ}$ should have a negative slope, because (according to the law of demand) the more a commodity is available, the less that careful consumers will be willing to pay for it.

d. This diagram depicts an unidentified (or not identified) path model. Note that its paths would be obtained by estimating the following standardized regressions:

$$Z_Q = p_{QP} Z_P + v$$
$$Z_P = p_{PQ} Z_Q + u$$

Solving the normal equations for the effect of $P$ on $Q$, …

$$r_{PQ} = \frac{\sum Z_P Z_Q}{n} = \frac{\sum Z_P (p_{QP} Z_P + v)}{n} = p_{QP} \frac{\sum Z_P^2}{n} + \frac{\sum Z_P v}{n} = p_{QP} + \theta_{Qu},$$

where $\theta_{Qu} = P_{Qv} \psi_{vu}$

And likewise, …

$$r_{QP} = \frac{\sum Z_Q Z_P}{n} = \frac{\sum Z_Q (p_{PQ} Z_Q + u)}{n} = p_{PQ} \frac{\sum Z_Q^2}{n} + \frac{\sum Z_Q v}{n} = p_{PQ} + \theta_{Pv},$$

where $\theta_{Pv} = P_{Pu} \psi_{uv}$.

Thus for each endogenous variable there is one equation with two (actually three) unknowns, and so the path model is unidentified.

e. Now, imagine that the commodity under discussion is an agricultural product (e.g., corn). Clearly, the quantity of corn on the market is a function of rainfall. And rainfall, in turn, is not directly associated with consumers’ willingness to pay for corn. Accordingly, a modified path diagram would look as follows:
Note that in this diagram we are assuming that rainfall accounts for variations in “quantity of corn” that are not due to changes in price (i.e., the increase in quantity produced by rainfall is not an increase to meet demand).

f. Some comments:

i. According to the diagram, \( p_{pr} = 0 \), since there is no path from rainfall to price.

ii. By omitting the path associated with \( \psi_{uv} \), the researcher must defend an assumption that \( v' \) does not vary with factors (e.g., rainfall) that influence price through quantity. Such a defense might read as follows: “Once rainfall is added into the model, the error term, \( v' \), has been adjusted for the key Quantity-related factor (namely, rainfall) that serves to reduce the Price consumers are willing to pay for corn. Accordingly, I now assume that these errors are unassociated with Price.”

g. The first standardized regression is now expanded as follows:

\[
Z_Q = p'_{QP} Z_P + p_{QR} Z_R + v'
\]

This modification yields normal equations for quantity \((Z_Q)\) that have a unique solution.
Note the assumption here that $\sum Z_p v' = 0$. As explained above, justifying this assumption requires an argument that the rainfall variable accounts for all the variance in the price of corn that is attributable to quantity—a dubious premise given that amounts of sunshine, soil conditions, etc. are also likely to be important factors. Rainfall serves here as an “instrumental variable” (Wonnacott and Wonnacott 1970, p. 191) that, as discussed previously in our discussion on model specification, ensures that the $E(X^T e) = E(X^T) E(e)$ assumption is met. Nevertheless, the idea here is that when quantity is regressed on both rainfall and price, errors from this regression no longer to vary according to rainfall (and thereby, we assume, according to all other factors that might reciprocally influence price through quantity).

The second normal equation is obtained as follows:

$$r_{rQ} = \frac{Z_R Z_Q}{n} = \frac{Z_R (p'_{Qp} Z_p + p_{QR} Z_R + v')}{n}$$

$$= p_{Qp}' \frac{Z_R^2}{n} + p_{QR} \frac{Z_R^2}{n} + \frac{Z_R v'}{n} = p_{Qp}' r_{RP} + p_{QR}$$

Thus we have two equations

$$r_{pQ} = p_{Qp}' + p_{QR} r_{RP}$$

$$r_{rQ} = p_{Qp}' r_{RP} + p_{QR}$$

with two unknowns (namely, $p_{Qp}'$ and $p_{QR}$).
h. Yet at this point the path model is still underidentified. The second endogenous variable (namely, price) must also be regressed on an instrumental variable to make the model just identified.

i. Wonnacott and Wonnacott (1970, pp. 172-181) show that by adding per capita income into the model as an exogenous variable with a single path from it to price, the model becomes just identified. The theoretical rationale for including this variable is that income accounts for variations in the price consumers are willing to pay for corn that are not due to changes in the quantity of corn available.

ii. The path diagram of our now just identified model looks as follows:

With income now in the model, we assume that it accounts for variations in price that are not due to changes in quantity (i.e., the decrease in willingness to pay due to poverty is not a decrease because of oversupply).

iii. Thus the regression equation for estimating the effects on price is now as follows:

$$Z_p = p_{PQ} Z_Q + p_{PI} Z_I + u'$$
The normal equations associated with price are derived identically to how they were derived for quantity.

i. **The order condition for identification**

   i. There are many tests available for judging whether a path model is or is not identified. We shall only discuss the order condition, which must necessarily hold if a regression model is to be identified.

   ii. The condition (Wonnacott and Wonnacott 1970, p. 180): When a set of regression equations is identified, *for each equation the number of exogenous variables excluded from the equation equals at least the number of endogenous independent variables in the equation.*

   iii. Returning to our modified Price-Quantity path model, path coefficients are now estimated using the following two regression models:

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Excluded Exogenous Variable</th>
<th>Included Endogenous Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_Q = p'_p Z_p + p'_Q Z_R + v'$</td>
<td>Income</td>
<td>$\geq$ Price</td>
</tr>
<tr>
<td>$Z_p = p'_p Z_Q + p'_p Z_i + u'$</td>
<td>Rainfall</td>
<td>$\geq$ Quantity</td>
</tr>
</tbody>
</table>

   Thus the order condition for identification is met with these equations. Here each equation has one exogenous variable excluded from it (but included in the other equation) and one endogenous variable included in it as an independent variable.

   iv. If the order condition is not met, one’s system of equations is not identified. However, if it is met, one’s system of equations could still be underidentified. That is, the order condition is a necessary but not a sufficient condition of identification.

j. A few closing comments on the last path model
i. Although this model is identified, it is not recursive, because it does not depict causality unidirectionally. Ordinary Least Squares (OLS) estimation cannot work with such nonrecursive models. Instead, another estimation method like Maximum Likelihood (ML) estimation should be used. Software for testing such models include LISREL and Amos—software packages that afford direct tests of hypothesis regarding correlations among errors.

ii. All recursive path models are identified, however (Johnston 1972, p. 369). This is the case because errors are (assumed) unassociated with prior variables in a recursive system. In a nonrecursive system the notion of “prior” is not always clearly defined. For more detail on this topic see Johnston (1972 or 1984).

k. Underidentified path models are hopeless (Pedhazur 1997, p. 805). In them parameter estimation is impossible, unless they are modified in one of two ways:

i. Exogenous instrumental variables are added into the model, or

ii. paths are removed from the model.

E. Some references:


