Finding Our Way: An Introduction to Path Analysis

David L Streiner, PhD[1]

Path analysis is an extension of multiple regression. It goes beyond regression in that it allows for the analysis of more complicated models. In particular, it can examine situations in which there are several final dependent variables and those in which there are "chains" of influence, in that variable A influences variable B, which in turn affects variable C. Despite its previous name of "causal modelling," path analysis cannot be used to establish causality or even to determine whether a specific model is correct; it can only determine whether the data are consistent with the model. However, it is extremely powerful for examining complex models and for comparing different models to determine which one best fits the data. As with many techniques, path analysis has its own unique nomenclature, assumptions, and conventions, which are discussed in this paper.

(Can J Psychiatry 2005;50:115–122)

Information on author affiliations appears at the end of the article.

Highlights

- Path analysis can be used to analyze models that are more complex (and realistic) than multiple regression.
- It can compare different models to determine which one best fits the data.
- Path analysis can disprove a model that postulates causal relations among variables, but it cannot prove causality.

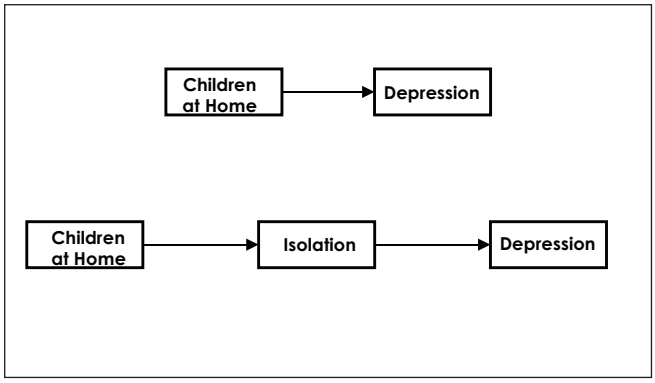
Key Words: *path analysis, structural equation modelling, multiple regression*

One of the first things we learn in introductory statistics is that there are 2 types of variables: independent variables (IVs) and dependent variables (DVs). The distinction between them is in most cases relatively clear and straightforward: we want to see what effect the IVs (sometimes called predictor variables in multiple regression) have on the DV. For example, if we compare antidepressant medication with cognitive-behavioural therapy to see which leads to greater remission in depressive symptoms, the IV is the type of treatment, and the DV is, perhaps, the score on a depression scale. In selecting candidates for medical school, the DV would be successful graduation, or not, and the IVs could be previous grade point average, a numerical grade assigned to letters of reference, qualifying examination scores, and so on.

Unfortunately, the real world refuses to fall into just these 2 categories of variables. Life insists on being more complicated, and there are situations in which, if we are asked whether a certain variable is a DV or an IV, we would have to say, "Yes." Let's assume, for example, that we found that women with young children at home suffer more from depression than women matched for age and marital status who do not have kids at home. One hypothesis is that children at home cause women to suffer from depression, based on the well-known fact that kids can drive us crazy. This is the model shown on the top of Figure 1—children at home have a direct effect on depression. However, another hypothesis is that staying at home leads to social isolation and that the isolation leads to dysphoria. This is the model on the bottom of Figure 1.

In both models, it is obvious that children at home is the IV and depression is the DV. But what do we call isolation? It is a DV with respect to children, but it is an IV as regards depression. In fact, the problem goes beyond mere terminology; the deeper problem is how we should analyze the lower model. The issue becomes more complicated as our models become more complex. For example, we can posit that other factors may influence depression directly, for example, hormonal changes or past depression and family history of depression; that children at home cause other stresses that in turn affect the woman's mood; and that there can be outcomes in addition to dysphoria. If we want to look at all these variables simultaneously, we need both new terminology and a new analytic strategy.

Figure 1 Two possible models of the effects of children at home on depression



The name for the strategy comes from the pictorial representation of the models themselves. In the bottom of Figure 1, we are describing a path from children to isolation to depression. More complicated models may have more paths, or paths that lead through more variables. Not surprisingly, then, this technique is called path analysis. Path analysis is an extension of multiple regression that allows us to examine more complicated relations among the variables than having several IVs predict one DV and to compare different models against one another to see which one best fits the data. Before we begin, though, a disclaimer about both the statistics and the terminology. In the past (and occasionally now, among the benighted), this technique was referred to as causal modelling, because it was believed that it could be used to uncover causal pathways among variables—which factors were responsible for what outcomes. As we'll see later, this is a laudable goal but one that cannot be solved by statistical methods, no matter how powerful. From a statistical vantage point, models that map out totally bizarre routes that can never exist in reality may look as good—or even better—than models that conform more closely to reality. Causality can be proven only through the correct research design (for example, longitudinal studies or experiments), and no amount of statistical legerdemain can pull cause and effect out of a cross-sectional or cohort study.

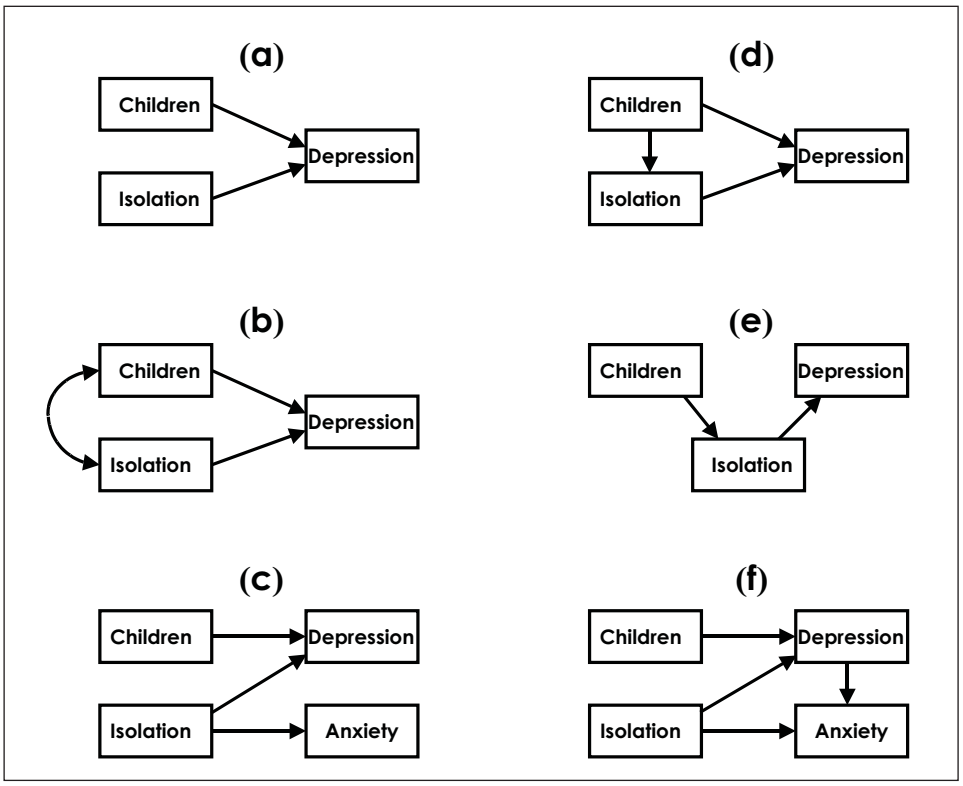
Some Terminology and Drawing Conventions

To return to the naming of the variables. In path analysis (and in its more sophisticated counterpart, structural equation modelling, which will be discussed in a later paper), we completely avoid the confusion about IVs and DVs by the simple expedient of not using those labels. Rather, we use the terms exogenous and endogenous variables:

- Exogenous variables have straight arrows emerging from them and none pointing to them (except from error terms).
- Endogenous variables have at least one straight arrow pointing to them.

The rationale for these terms is that the causes of (or factors that influence) exogenous variables are determined outside the model that we're examining; whereas the factors affecting endogenous variables exist within the model itself. In a

Figure 2 Some possible path models



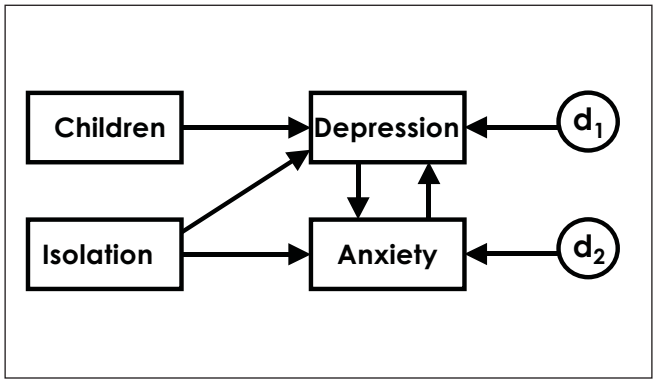
moment, we'll see why there is that restriction on the shape (not the moral fibre) of the arrow. Thus, in the bottom of Figure 1, having children is an exogenous variable, and both social isolation and depression are endogenous ones. Let's take a look at some other possible models.

In Figure 2a, we are hypothesizing that exogenous variables of having children and being isolated independently influence depression and that children and isolation are not correlated with each other. This would be called an independent model, reflecting the lack of correlation between the 2 exogenous variables. If we believe that the variables are correlated, we draw a curved, 2-headed arrow between them, as in Figure 2b, where a curved, double-headed arrow reflects a correlation (or covariance, a term that will be defined shortly) between variables. This is the run-of-the-mill multiple regression model, although we often have more than just 2 exogenous or predictor variables. Also, because as Meehl said, "everything correlates to some extent with everything else" (1, p 204), we customarily assume that the correlations are present—so I won't bother to draw the curved arrows (except when necessary). In Figure 2c, we are extending the model by looking at 2 endogenous variables, depression and anxiety. We are saying that having children affects only depression but not anxiety, while isolation influences both depression and anxiety.

The models on the right side of Figure 2 are called mediated or indirect ones, because the exogenous variables act on an endogenous variable, at least in part, through their influence on an intermediary (endogenous) variable. Thus, in Figure 2d, having children influences depression directly, as does isolation, but having children also affects isolation, which then acts on depression; in other words, being isolated would cause women to suffer from depression even in the absence of children, but the presence of children exacerbates this effect. This is somewhat different from Figure 2e, which is the same as the bottom part of Figure 1 in that isolation is due solely to the rug rats and would not exist in women if not for them. Finally, in Figure 2f, we see that one final endogenous variable (depression) can also affect another one (anxiety). This by no means exhausts the possibilities. Paths can be much longer and involve more intermediate steps, and there could be (and often are) more variables brought into the picture—both literally and figuratively.

Two more terms need to be introduced. In Figure 3, note that pointing toward the 2 endogenous variables are small circles with d's in them. The d's are short for disturbance, which in path analysis is analogous to the error term tacked at the end of regression equations. As with the error term, the disturbance term captures 2 things: 1) imprecision in the measurement of the endogenous variable, because all our measurement tools are subject to some degree of error; and 2) all the other factors affecting the endogenous variable that we didn't measure

Figure 3 Adding disturbance terms and making the model nonrecursive



because of oversight, lack of time, ignorance of their importance, laziness, or whatever. Because every endogenous variable must have a disturbance term associated with it, we often don't bother to draw it, to keep the drawing simpler, but if it's not explicitly drawn, it's implicitly present.

Finally (yes!), Figure 3 differs from Figure 2f in one subtle but important way. In Figure 2f, one endogenous variable (depression) affects a second (anxiety), which is logical, but the second does not affect the first. It likely makes more clinical sense for there to be a path in both directions; that is, depression affects anxiety, and anxiety in turn is depressogenic. This latter model, with a feedback loop, is called nonrecursive, whereas path models that lead inexorably in one direction are referred to as recursive. (No, I didn't inadvertently reverse the terms. They are, for some unfathomable reason, completely and utterly counterintuitive. You'll just have to live with that.) There is one cardinal rule for attempting to analyze nonrecursive models: Don't! Although they are often a more accurate reflection of the way things work in the real world, the analytic problems they produce are horrendous. Even more troubling, these problems are not immediately apparent from most computer printouts, especially to neophytes, so they are not aware that they are getting into a quagmire.

As we go through some examples, 2 other terms will keep appearing—variances and covariances. (Because they were probably first mentioned in introductory statistics, I'm considering them to be old terms, so I haven't violated my statement about the number of new ones I'm introducing.) Variance is simply the amount of variation in a variable from one person to another and, formally, is the square of the standard deviation (SD). A covariance is the first cousin of a correlation. A correlation tells us whether, if variable A goes up, variable B also goes up (a positive correlation), goes down (a negative correlation), or changes in a way unrelated to variable A (a zero correlation). When we calculate a correlation, both variables are transformed into standard scores that have a mean of 0 and an SD of 1. Covariance is exactly the same

thing, except that the variables aren't transformed but remain in their original units of measurement. For various arcane reasons, it is often better to calculate multivariable statistics by using the covariances among variables rather than their correlations. The reason for our obsession with them is that they are all we have to work with in determining how variables influence each other. When we've done a study measuring several variables on each person, what we end up with is a matrix consisting of the variances of each variable along the main diagonal (the rows running from the upper left to the lower right corners) and the covariances in all the cells off the main diagonal. The job of path analysis is to determine whether there are any meaningful patterns in these data.

A Multiple Regression Example

Let's begin to see what path analysis can do by using a different example and framing it first in multiple regression terms. In previous articles in this series, you have been introduced to a disorder not yet found in the DSM—photonumerophobia, which is the fear that our fear of numbers will come to light. We may want to see whether a person's score on a scale of photonumerophobia (the PNP scale, which ironically is scored using numbers) can be predicted from 3 factors: overall level of anxiety (ANX), high school math grade (HSM), and the discrepancy between what the person estimated last year's taxes to be and what the tax people said it actually was (TAX). That is, we are hypothesizing that photonumerophobia is related to a person's overall anxiety level (we would expect a positive correlation) as well as to doing poorly in math class (a negative correlation) and to screwing up one's income tax returns (again a positive correlation: the greater the mistake, the more severe the phobia). If we wrote this out as a regression, it would look like

$$PNP = b_0 + b_1 ANX + b_2 HSM + b_3 TAX + Error$$

where b_0 is the intercept, and the other b's are the slopes for each IV. The means and SDs of the variables and the correlations among them are shown in Table 1, as are the unstandardized regression weights (b) and the standardized ones (β). (Very briefly, the b weights are used with the data in their original units of measurement; the β weights after each variable have been standardized to have a mean of 0 and an SD of 1. For a more complete description of the difference, see 2,3). The multiple correlation, R, is 0.638, and therefore R^2 (the square of R, reflecting the proportion of variance in the DV accounted for by the equation) is 0.407.

In the conventions of path analysis, each IV or predictor variable (now called an exogenous variable) would be shown as a rectangle, with another rectangle for the DV, or endogenous variable, as in the left side of Figure 4. Because computer programs suffer from severe organic brain damage, it's necessary to draw the curved arrows reflecting correlations among the variables and the disturbance term. When we run the program, we'll get reams of paper, but most of the results can be summarized in a diagram, shown in the right side of Figure 4. Very reassuringly, the correlations among the predictors (next to the curved arrows) agree with the results from the regression analysis; as do the β weights, which are printed near the paths (the straight arrows). Finally, the number over the right side of PNP is the value of R^2, which again is identical to the results from the multiple regression. Also reassuring from a theoretical standpoint is that the signs of the paths correspond to what we hypothesized.

Let's review what we've accomplished so far. We've taken a problem that is easily handled with inexpensive, well-known, and widely available computer programs (for multiple regression) and shown how we can get identical results by using an expensive and arcane one (for path analysis and structural equation modelling). If this is all there is, it reflects progress that we can easily live without. Fortunately, though (especially for the purveyors of the software), path analysis can do much more. For example, we can postulate other hypotheses about the relations among the variables and see whether they are better or worse in accounting for the variance in the PNP scale. Two examples are shown in Figure 5. In Figure 5a, it's hypothesized that, rather than the variables acting separately on PNP, there is truly a path, with anxiety

Figure 4 Hypothesized model (a) and results (b) of the regression equation

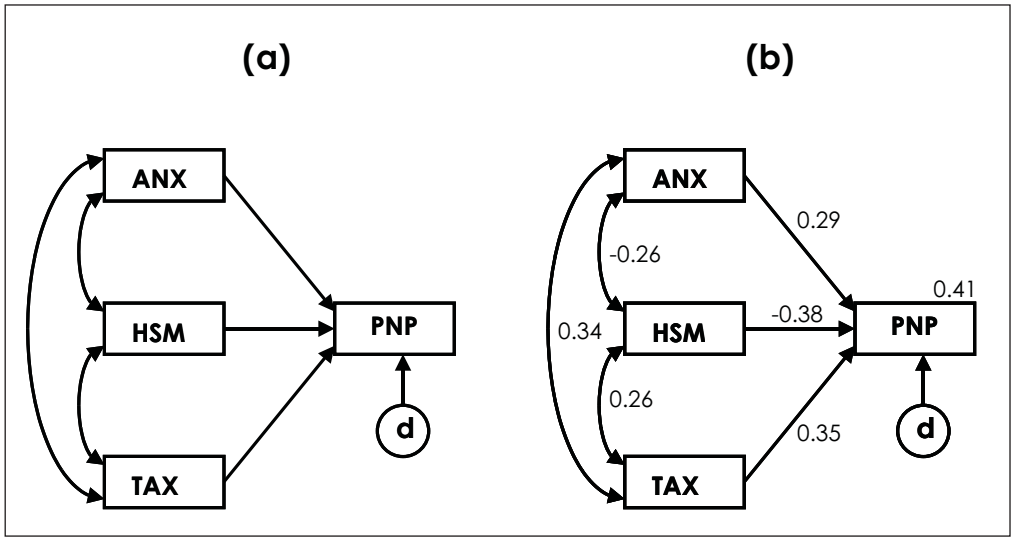
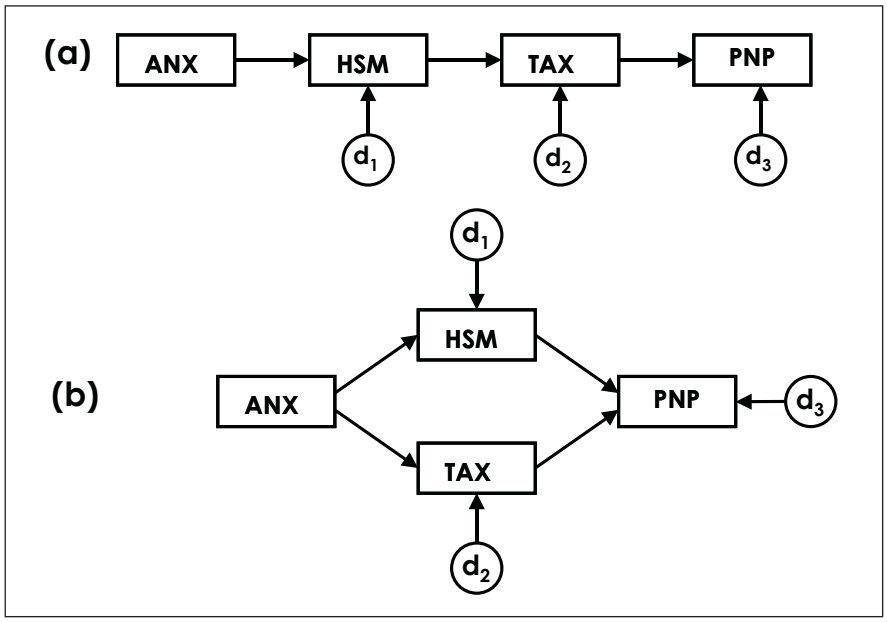


Table 1 Means, SDs, correlations among the variables, and unstandardized and standardized regression weights for the relation between photonumerophobia (PNP), anxiety (ANX), high school math grade (HSM), and income tax discrepancy (TAX).

	PNP	ANX	HSM	TAX	Mean	SD	b	β
PNP	1.000	0.509	–0.366	0.346	26.79	7.33		
ANX		1.000	–0.264	0.338	20.33	5.17	0.414	0.292
HSM			1.000	0.260	74.69	5.37	–0.517	–0.379
TAX				1.000	1983.23	525.49	0.005	0.346

Figure 5 Other possible models



leading to poor math grades in high school, which results in mistakes in completing tax forms and culminates in photonumerophobia. In Figure 5b, the hypothesis is that anxiety results in both poor grades and computational errors in filling out taxes and that both of these together lead to photonumerophobia.

Fitting the Model

The issue now is how to choose among the models. The first place to look is at the path coefficients themselves. As mentioned above, they are standardized regression weights, identical to the β weights of multiple regression. Their sign should correspond to what the model predicts. If we expect anxiety to be positively correlated with PNP, then a path coefficient with a negative sign would lead to an increase in our own anxiety level, as well as to rejection of the model. Moreover, the β weight should be statistically significant. We aren't able to determine this from the diagram, but the accompanying printed output will tell us whether it's significant or not.

There are several (actually, a myriad) fit indices that tell you how well the model fits the data. Because these are also used in structural equation modelling, and because of space limitations in this paper, they will be discussed in a subsequent paper.

Model Specification

What can account for a poorly fitting model? The most likely cause is model misspecification. This can occur in several ways: the inclusion of variables that are not related to any of the endogenous variables, the omission of crucial variables, and the use of paths that connect variables not in fact related to each other. Ideally, if an extraneous variable has been included, its path coefficient will be low and nonsignificant, signalling that the model should be rerun without that

variable. The same situation should apply if a path is drawn between variables that aren't related to each other. Unfortunately, there is nothing that can tell us when we've omitted some crucial variable(s); further, if the omitted variables are correlated with some in the model, the fit indices can still be quite high (4). Protecting yourself against this requires knowledge of the literature and a reasonable hypothesis regarding what can affect the outcomes.

If we attempt to apply these fit indices to the model in Figure 4, we'll see somewhat unusual results. All the indices that should be above 0.90 are in fact 1.00—a perfect fit! Moreover, the χ^2_{GoF}, which should be low, is as low as it can get: 0.00 (albeit with 0 degrees of freedom [df]). We know the model makes sense, but no prediction model is this good. Common sense would tell us that each of these variables is measured with some degree of error, so that even if the theoretical model were a perfect reflection of reality (in which case the next Nobel prize would be mine), the match between the data and the model wouldn't be perfect. Moreover, the R^2 is only 0.41, so our "perfect" model isn't accounting for most of the variance in PNP. What's going on?

The problem has to do with the identification of the model. Identification refers to how many things we have to estimate (such as the path coefficients and correlations) in relation to how much information we can derive from the data information in terms of the observed variances of the variables and the covariances among them. In this regression model, the amount of information is exactly equal to the number of paths we have to estimate; it is "just identified." We can see this from the fact that, as we said in the previous paragraph, the df = 0. So let's see what's meant by identifying the model. We'll start off with a simple example: What are the values of the unknown terms in the following equation?

$A + B = 10$

The answer is indeterminate, in that there are an infinite number of possibilities. A can be 10 and B can be 0, or A can be 9 and B can be 1, or A can be 938 and B can be –928, and so on, ad infinitum. We do not have enough information to solve the equation, so we say that it is underidentified. In other words, we cannot derive unique values for the 2 unknowns (and, by analogy, the various parameters in the model). If we know that A = 7, then there is only one possible solution, and the model is said to be just identified, which is good. Even better is an overidentified model: we have more information than we need (analogous, for those who remember high school math, to having more simultaneous equations than unknown variables), so we can test different models against one another. Equally important, we have seen that we cannot get any fit indices from models that are just identified but we can with overidentified models.

How many parameters can we estimate in a model? If there are k variables, then we have k variances and $([k^2 - k] / 2)$ covariances, for a total of $([k^2 + k] / 2)$ pieces of information. In Figure 4, there are 4 variables, so we can estimate a maximum of $([4^2 + 4] / 2) = 10$ parameters. The next question, then, is how many parameters do we want to estimate in this model? To begin with, there are the 3 paths from the exogenous variables (ANX, HSM, and TAX) to the endogenous one (PNP). Next, there are 3 covariances (or correlations) among the exogenous variables. Then, there is the variance of the disturbance term (d). Finally, there are the variances of the exogenous variables themselves, for a total of 10. The df is the difference between the number of parameters we can estimate and the number of parameters we want to estimate. Note that df is not related to the number of subjects in the study; simply increasing the sample size will not solve the problem of a model that is under- or just identified (although it does have other benefits, which I'll discuss later).

Two questions arise from this. First, why don't we want to estimate the variance of PNP (or any other endogenous variable, for that matter)? The answer is that the purpose of path analysis is to find out what affects the endogenous variables, that is, how the exogenous variables work together (their covariances) and which paths are important (determined in part by the variances of the exogenous variables). Because the model postulates that the endogenous variables are determined or influenced solely by the exogenous ones, they are not free to vary on their own but only in response to the exogenous variables. Consequently, we are not concerned with estimating PNP's variance. In general, then, we want to estimate 1) the paths, 2) the covariances among the exogenous variables, and 3) the variances of the exogenous variables but not the variances of the endogenous ones.

Types of Paths

The second question is "How can we increase the df to turn an underidentified model into an overidentified one?" The answer arises from the fact that paths come in 2 flavours: free and fixed. What we've been looking at so far are free paths, that is, paths free to take on any value that best fits the data; in most cases, this is what we are most interested in. We can increase the df by setting various paths to some predefined value—in other words, by fixing them so that the program won't have to estimate them from the data. The most efficient way is to set some paths equal to zero, that is, to simplify the model by dropping some of the hypothesized links. Depending on the variable that's dropped from the model, this can sometimes increase df by 2 or more, since it is not necessary to estimate either the path or the variance of the term and there may be fewer covariances to estimate. Needless to say, this can be a draconian method if all the variables and paths are

deemed essential to the model, but it may at times be necessary. For example, in the discussion about Figure 4, I said that we should not have anxiety influencing depression and depression affecting anxiety, logical as it may seem, because that would result in a nonrecursive model. Deciding which arrow to eliminate may be based on our theories of anxiety and depression or, perhaps, on research that may indicate which condition is generally present first or, in the absence of these, on just a guess—not an ideal situation by any means, but one imposed by the technique.

This need for parsimony should also serve as a warning, which I'll return to later: path analysis is a model-testing approach, not a model-building one. You should not throw in any variable that's available and draw every conceivable path, just to see what comes out. There should be a sound rationale for the model, based on theory and research, or even on a strong hunch. Don't be scared off by the term "theory," though. The rationale does not have to rival Einstein's theory of general relativity in its complexity or degree of development. A postulate as simple as "isolation leads to dysphoria" is a theory, albeit not one that will immortalize its developer.

Another technique to reduce the df is to fix the value of a path to be equal to 1. For example, the paths from the disturbance terms to the endogenous variables are automatically fixed by the program to be 1. (This is because we do not have enough information to estimate both the variance and the path coefficient of the d's, for the same reason, explained above, that we had difficulty when trying to estimate the values of A and B that add up to 10. Because we are usually more interested in the magnitude of the error, as reflected by its variance, than we are in the path coefficient, the latter is fixed, although we can overrule the program if we wish.)

Sample Size

As we just mentioned, the df depend solely on the number of variables and parameters in the model, not on the number of subjects. An underidentified model will remain underidentified even if we use 10 times the number of people. The sample size becomes important in accurately estimating the values of the paths, variances, and covariances. Because all these are parameters, there is a standard error (SE) associated with each one, as well as a *z* test, which is the ratio of the parameter to the SE. If the sample size is too small, the estimates of the parameters are unstable, reflected in large SEs and nonsignificant *z* tests for their significance. Klein (5) recommends a minimum of 10 cases for every parameter that's estimated and 20 if you can find them; 5 are too few. Don't forget that there are often 2 or 3 parameters for every variable, so path analysis is much hungrier for subjects than are other multivariable techniques such as multiple regression or factor analysis, which usually require "only" 10 subjects per variable.

Assumptions

Because path analysis is an extension of multiple linear regression, many of the same assumptions hold for the 2 techniques. First, as the name implies, the relations among the variables must be linear. Second, there should be no interactions among the variables (although we can add a new term that reflects the interaction of 2 variables). Third, the endogenous variables must be continuous (although you can get away with a minimum of 5 categories if you have ordinal data) and relatively normally distributed, with skewness and kurtosis coefficients below 1. Fourth, it is assumed that the covariances among the disturbance terms are zero (equivalent to the assumption of uncorrelated errors among the predictor variables in regression), although more advanced variants of path analysis can deal with violations of this assumption. Finally, as mentioned previously, path analysis is quite sensitive to the specification of the model; including irrelevant variables, or more seriously, omitting relevant ones, can drastically affect the results.

Interpretation and Model Building

As I mentioned above (but cannot mention too often, although you may dispute this), path analysis is a technique for testing models, not for building them. It does not make sense to draw all possible paths, close your eyes, and press the compute button. You may get results (assuming your model isn't underidentified), but it would be fatuous to believe them, or even to hope that they will be replicated. Similarly, most computer programs for path analysis can print out modification indices that tell you how the model can be improved, for example, by including covariances between variables or error terms or by including paths between variables that weren't specified in the model. The major criterion for accepting or rejecting the suggestions is your theory. If the changes make theoretical sense, try them out; if there is no theoretical justification for them, ignore them. Changing your model simply to improve the fit may result in a model that is neither sensible nor reproducible.

Finally, having a model that fits the data doesn't prove that the model is correct. There may be better models that you haven't tested. More important, often changing the direction of an arrow, or even a series of arrows, may result in models that are statistically equivalent. For example, if we ran the model shown in Figure 5b, we would find that R^2 was 0.434 (admittedly, with not-too-impressive fit statistics). Changing the direction of the arrow between ANX and TAX (and moving the disturbance term as required) results in an identical value for R^2 and all the fit statistics. Obviously, both can't be

correct; they simply fit the data equally well. This fact also illustrates the point that path analysis cannot be used to establish causality: that is done through the design of the study, not its analysis.

Following in the Paths of Others

Assuming that the readers of this article will more likely be consumers than doers of path analysis, the following are points to bear in mind, akin to the Convoluted Reasoning and Anti-Intellectual Pomposity (CRAP) detectors in *PDQ Statistics* (2).

1. Are the signs of the paths correct, and are they statistically significant? Don't be fooled by pretty diagrams and the bottom line; each element of the model has to make sense.

2. Does the final model presented in the paper make sense? That is, does it appear as if the model were derived from some coherent theoretical and (or) empirical base, or could the paths have been drawn simply to improve the fit of the model?

3. Was the sample size sufficient? Add up the number of paths, the number of curved arrows, the number of exogenous variables, and the number of disturbance terms, and multiply by 10. If the sample size is less than this, interpret the results with a considerable degree of scepticism.

Summary

Path analysis is a powerful statistical technique that allows for more complicated and realistic models than multiple regression with its single dependent variable. However, the increased sophistication places additional demands on users. Mastering a new computer program and new terminology is perhaps the easiest part; the more demanding requirement is that far greater attention must be paid to the underlying model, in terms of including as many relevant variables as possible, weeding out irrelevant ones, and specifying relations among the variables. This requires solid knowledge of the literature and a model that makes clinical and theoretical sense. However, the rewards are being able to test these models and to compare different models against one another. Path analysis also provides a stepping stone to an even more sophisticated and useful technique—structural equation modelling—which will be discussed in a subsequent article.

References

1. Meehl P. Why summaries of research on psychological theories are often uninterpretable. Psychol Rep 1990;66:195–244.
2. Norman GR, Streiner DL. PDQ statistics. 3rd ed. Toronto (ON): BC Decker; 2003.
3. Norman GR, Streiner DL. Biostatistics: the bare essentials. 2nd ed. Toronto (ON): BC Decker; 2000.
4. Tomarken AJ, Waller NG. Potential problems with "well fitting" models. J Abnorm Psychol 2003;112:578–98.
5. Klein RB. Principles and practice of structural equation modeling. New York: Guilford; 1998.

Manuscript received April 2004 and accepted May 2004.
This is the 24th article in the series on Research Methods in Psychiatry. For previous articles please see Can J Psychiatry 1990;35:616–20, 1991;36:357–62, 1993;38:9–13, 1993;38:140–8, 1994;39:135–40, 1994;39:191–6, 1995;40:60–6, 1995;40:439–44, 1996;41:137–43, 1996;41:491–7, 1996;41:498–502, 1997;42:388–94, 1998;43:173–9, 1998;43:411–5, 1998;43:737–41, 1998;43:837–42, 1999;44:175–9, 2000;45:833–6, 2001;46:72–6, 2002;47:68–75, 2002;47:262–6, 2002;47:552–6, 2003;48:756–61.
[1]Assistant Vice President, Research and Director, Baycrest Centre for Geriatric Care; Professor, Department of Psychiatry, University of Toronto, Toronto, Ontario.
Address for correspondence: Dr DL Streiner, Director, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, ON M6A 2E1
e-mail: dstreiner @ klaru-baycrest.on.ca

Résumé : Trouver notre voie : une introduction à l'analyse de dépendance

L'analyse de dépendance est une extension de la régression multiple. Elle va plus loin que la régression en ce qu'elle permet l'analyse de modèles plus compliqués. En particulier, elle peut examiner des situations où se trouvent plusieurs variables dépendantes finales et celles où il y a des « chaînes » d'influence, quand la variable A influence la variable B, qui à son tour influe sur la variable C. Malgré son ancien nom de « modélisation causale », l'analyse de dépendance ne peut pas servir à établir la causalité, ou même à déterminer si un modèle spécifique est exact; elle peut seulement déterminer si les données sont conformes au modèle. Toutefois, elle est très efficace pour examiner les modèles complexes et comparer différents modèles, afin de déterminer lequel correspond le mieux aux données. Comme dans bien des techniques, l'analyse de dépendance a ses propres nomenclature, hypothèses et conventions, que cet article décrit.