Path Analysis

Generally, path analysis is the combination of assumed causal theory with empirical evidence.
“Path analysis, whatever its ultimate fate, will present historians and philosophers of science with a puzzle. Ever since Sewall Wright's classic (1921) article and the ensuing exchanges with his critic, Niles (1922, 1923; Wright, 1923), there has been sharp disagreement about the methods value (Breckler, 1990; Cliff, 1983; Karlin, Cameron, and Chakraborty, 1983; McKim and Turner, 1997). Some consider it one of the best ways to make causal inferences from correlational data; others view it as not only worthless but misleading and not to be used at all (e.g., Freedman, 1987, 1997; Rogosa, 1987). Today, after three generations of discussion by statisticians, psychologists, sociologists, economists, political scientists, geneticists, and philosophers of science, the disagreement still seems about as great and irresoluble as in the Wright-Niles exchange (for historical reviews of path analysis and structural equation models, see Aigner, Hsiao, Kapteyn, and Wansbeek, 1984; Austin and Wolfe, 1991; Bentler, 1980; Bielby and Hauser, 1977; Epstein, 1987; Shipley, 2000).”
Path Analysis

What path analysis **CANNOT** do for you...

- Take non-experimental data and prove whether one variable **actually** causes another.
- Take non-experimental data and prove the direction of causal order between variables.
- Take non-experimental data and distinguish between models that results in identical correlation patterns.

What path analysis **CAN** do for you...

- Provide a graphical way to represent your **assumed** theory.
- Provide a way to empirically estimate the relationships in your assumed theory, in particular to estimate whether the relationships are positive, negative, and importantly to test whether the relationship is zero and hence not supported by the data.
- Provide a way to estimate the assumed causal effect that one variable has on another through its assumed causal effect on other variables.
- Take experimental data (e.g. interventions) and prove whether the experimentally changed variable **actually** causes an outcome.
History

• The original developer of path analysis was a geneticist Sewall Wright in 1934, who was focused on evolutionary biology. Wright S (1934) The method of path coefficients *Annals of the Mathematical Statistics*, 5, 161-215.

• Wright’s work was basically lost until Duncan (1966). Duncan OD (1966) Path analysis: Sociological examples. *American Journal of Sociology*, 72, 1-16. Some mistakes are in this article regarding the calculation of indirect effects. A correction addendum appear in Blalock HM (Ed) (1971)

Several annotated bibliographies of the development of Path analysis can be found:


• Austin JT and Calderon RF (1996) Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural equation modeling*, 3, 105-175.


• Wolfle LM (2003) The introduction of path analysis to the social sciences, and some emergent themes: An annotated bibliography *Structural equation modeling* 10(1), 1-34.
Examples of general theories

Broad theories exist in psychology, sociology, genetics, economics, business, physiology, public health, etc. for how the world works.

- **Attitude model and behavior intentions model** (Ajzen and Fishbein 1980) According to this theory, a person’s behavioral intention is the best predictor of his or her eventual behavior. If someone intends to do one certain thing, he or she will be more than likely to do it. If he or she has no intention to do it, then they will be more than likely not to do it. There are two major components which influence an individual’s intention. They are attitude component and subjective norms.

- **Social-cognitive or social learning theory** (Bandura 1971), i.e. Principles of reinforcement and punishment from behaviorism + People learn by watching others + Cognitive processes mediate social learning

- **Problem behavior theory** (Jessor and Jessor 1977), i.e. posits that problem behaviors co-occur within individuals to form a problem behavior syndrome. This syndrome contributes to a state of transition proneness that provides a means of achieving developmental objectives, such as experimentation with adult roles and identity exploration. Transition proneness is typified by greater involvement in problem behaviors and less participation in conventional behaviors.

- **Expectancy theory** (Porter and Lawlers, 1968) Job performance causes job satisfaction when rewards are positively contingent on performance

- Many many more
Example of Social Cognitive Theory


![Figure 1. Proposed model: Correlates of unhealthy weight-control behaviors among adolescents.](image-url)
Example of Social Cognitive Theory

Figure 4. Final model testing among adolescent girls: Correlates of unhealthy weight-control behaviors.

BMI = body mass index. * p < .01.

Figure 5. Final model testing among adolescent boys: Correlates of unhealthy weight-control behaviors.

BMI = body mass index. * p < .01.
Example of Problem behavior theory

Parental Permissiveness  
Parental Monitoring  
Marks in School  
Parent Orientedness  
Tolerance of Deviance  
Cigarette Smoking  
Marijuana Use  
Alcohol Misuse  
Drug - Driving  
High-Risk Driving  
Drinking Driving

Figure 1. Conceptual model of sequence of associations among content areas.
Example of Problem behavior theory

Figure 2. Hypothetical model predicting problem driving.
Example of Problem behavior theory

\[ \chi^2 = 61.67, \ df = 23, \ p < .000, \ RMSEA = 0.030, \ GFI = .99, \ CFI = .99 \]

Figure 3. The final alternative model predicting problem driving.
Causal analysis

There is an entire methodological/philosophical area of science devoted to causal theory. This area of study concentrates on the theory of how a researcher can conclude whether X ACTUALLY causes Y.

Concepts such as potential outcome and counterfactuals have been developed to carefully think through what it fundamentally means for one thing to cause another.

SEE: Additional causation.pdf file
SEE: Judea Pearl’s website

http://bayes.cs.ucla.edu/jp_home.html

Simple guidelines: To infer that X is a cause of Y.

- The association is consistent across studies
- The alleged cause precedes the effect in time
- The alleged cause is plausible
- The direction of relation is correctly specified (reciprocal causation?)
- The relationship does not disappear when external variables are held constant
Typical use of regression

Given an outcome of interest $Y$, and a list of several predictor variables $X$

- Perform a multiple regression of this outcome on all “potentially relevant” predictors. That is, put all the variables in the model at once and display a table at the end showing the coefficient estimates and associated p-values for whether those coefficients are zero or not.

- Or, a slightly more advanced thing to do empirically is to recognize that correlation between the predictors can result in some combinations of predictors “washing each other out”, that is, finding none of them to be significant, and thus deciding to use one of the many stepwise regression techniques that basically considers all possible combinations of variables to include in the model and chooses the “best” one according to some criteria usually involving the $R^2$.

- Multiple regression (as a statistical technique on its own) makes no assumptions about how variables are causally or not causally related to one another.
Path Analysis - Extending the use of regression

- Path analysis provides a framework for the researcher to think more carefully about how the $X$ variables are related to the $Y$ as well as how the $X$ variables are related to each other.

- What if your theory tells you that the predictor variables are actually causing one another?

- Why stop with considering just one outcome, $Y$, why not consider some of the $X$’s as outcomes (perhaps intermediate outcomes) too?

- Now you are thinking like path analysis.
Review of regression

Standardized regression:

Given variables $Y, X_1, and X_2$, we can standardize each of these variables and get

$$Y^s = \frac{Y - \text{mean}_Y}{\text{stddev}(Y)}, \quad X_1^s = \frac{X_1 - \text{mean}_X_1}{\text{stddev}(X_1)}, \quad X_2^s = \frac{X_2 - \text{mean}_X_2}{\text{stddev}(X_2)}.$$ 

Then the standardized multiple regression is

$$Y^s = \beta_1 X_1^s + \beta_2 X_2^s + \epsilon$$

The Ordinary Least Squares estimates are

$$\hat{\beta}_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{(1 - r_{x_1x_2}^2)} = r_{yx_1 \cdot x_2} \frac{\sqrt{1 - r_{yx_2}^2}}{\sqrt{1 - r_{x_1x_2}^2}}$$

$$\hat{\beta}_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{(1 - r_{x_1x_2}^2)}$$

Notice that the regression coefficients are simply functions of the bivariate correlations $r_{yx_1}, r_{yx_2}, r_{x_1x_2}$. It is not necessary to have the individual observations in order to estimate regression coefficients.

Notice that the regression coefficients are just a scaled version of the partial correlation, $r_{yx_1 \cdot x_2}$. If the partial correlation is zero, the coefficient will be zero.
Review of regression

• Simple linear regression (i.e. only one predictor),

\[ Y^s = \beta_1 X^s_1 + \epsilon \]

recall that \( \hat{\beta}_1 = r_{yx_1} \), that is, the standardized regression coefficient equals the simple bivariate correlation.

• Partial correlation - Correlation between two variables after adjusting for another variable (or set of variables)

\[ r_{xy \cdot w} = \frac{r_{xy} - r_{xw} r_{yw}}{\sqrt{(1 - r_{xw}^2)(1 - r_{yw}^2)}} \]

• Note that like simple correlations the partial correlation between X and Y is the same as the partial correlation between Y and X when the same variables are being conditioned upon, i.e. \( r_{xy \cdot w} = r_{yx \cdot w} \)
Review of regression - Assumptions

Drop the superscripts s notation, we will assume standardized regression coefficients (most of the time)

\[ Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

Assumptions that must be true for OLS estimates to be unbiased for the true model parameters

1. \( Y, X_1, \) and \( X_2 \) are measured without error (i.e. reliability equal to 1)
2. \( \epsilon \) is independent of \( X_1, \) and \( X_2 \)
3. The relationship specified is correct, that is, e.g. \( Y \) is linearly related to \( X_1, \) and \( X_2 \)
Interpretation of $\beta_1$ in multiple regression

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

The regression model above implies that

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_1 x_1 + \beta_2 x_2,$$

i.e. the expected value (or mean value) of $Y$ given that we know $X_1 = x_1$ and $X_2 = x_2$ is $\beta_1 x_1 + \beta_2 x_2$.

Hence, $E(Y|X_1 = x_1 + 1, X_2 = x_2) - E(Y|X_1 = x_1, X_2 = x_2) = \beta_1$

Without any causal assumptions made, the correct interpretation of $\beta_1$ is: We can expect to see an average difference of $\beta_1$ standard deviations in the $Y$ value for individuals who differ on the $X_1$ value by 1 standard deviation but do not differ on the $X_2$ variable.

This interpretation does not hint at any causal relationship because no causal assumptions were made.
Interpretation of $\beta_1$ in a path analysis model

Multiple regression is used to estimate the paths in the following path diagram

The interpretation for $\beta_1$ is:

If we could intervene in the population and increase the $X_1$ variable for each individual by one standard deviation while making sure that each individual’s $X_2$ variable does not change, then we would expect the average $Y$ value after the intervention to be $\beta_1$ standard deviations higher than it was before the intervention.

Full blown causal interpretation. Interpretation is based on assumptions.

Interpretation of $\beta_2$ is analogous.
Symbols

- $X \rightarrow Y$ “X causes Y” If X is changed (intervened upon) then Y will change
- $X \leftarrow Y$ “Y causes X” If Y is changed (intervened upon) then X will change
- $X \leftrightarrow Y$ “X and Y are simply correlated” but nothing is assumed about direction, sometimes called “spurious relationship”. X and Y might have a common cause(s) not included in the model explicitly.
- $X \iff Y$ “X and Y are reciprocally causing each other” X causes Y and Y causes X and consequently X causes itself (at a later time) and Y itself (at a later time). Sometimes called a “feedback loop”. Assumed stability of paths through infinite looping. Used in econometrics.

Path analysis models are made up of these symbols.

We will focus on the First 3 types of paths.
Basic idea of path analysis

Based on assumed causal relationships, bivariate correlation between any two variables can be broken down into a series of effects: direct causal effects, indirect causal effects, and noncausal or spurious components.

Consider

\[
x_2 = \beta_{23}x_3 + \epsilon_2
\]
\[
x_1 = \beta_{12}x_2 + \beta_{13}x_3 + \epsilon_1
\]

The effects in this model are estimated using the following two regression equations:

Notice that the bivariate correlation between \( x_1 \) and \( x_3 \) can be reproduced from these standardized regression coefficients \( \beta \). That is, \( \rho_{13} = \beta_{13} + \beta_{12}\beta_{23} \)
Effect Decomposition of a Bivariate Relationship

In Path Analysis we distinguish 3 types of causal effects

1. **direct** - the influence of one variable on another that is unmediated by any other variable, i.e. each single headed arrow represents a direct effect

2. **indirect** - effect that is mediated by at least one intervening variable

3. **total causal effect** - sum of the direct and indirect

Total effect = Total causal effect + spurious effect

Note the Total effect is estimated by the simple bivariate regression of Y on X.

Total causal effect = direct effect + indirect effect

Spurious effect is then Total effect - Total causal effect.
Calculation of indirect effects

Path Multiplication Rule - The value of the effect associated with a compound path is the product of its path coefficients (this works for standardized regression coefficient or unstandardized)

\[ \text{education} \rightarrow \text{income} \rightarrow \text{conservatism} \]

Unstandardized regression coefficient of Income on Education is \( \beta_1 = 1000 \) $/year, and the regression of conservatism (a 5 point Likert scale) on income yields a regression slope of \( \beta_2 = 0.0002 \) points/$

**What is the indirect effect of education on conservatism?**

If education goes up 1 year, income goes up $1000

If income goes up $1000 then conservatism goes up \( 0.0002 \times 1000 = 0.2 \) points

So, the indirect effect of a 1 year increase in education through income on conservatism is a .2 increase in the conservatism scale.
Estimation for Path Analysis

• If the path analysis model is recursive, i.e. does NOT have
  – feedback loops, e.g.
  – correlated errors, e.g.

Then all the paths can be estimated using a series of multiple regressions estimated via OLS

• If the path analysis model has feedback loops and/or has correlated errors then a method that estimates the paths simultaneously must be used. For example maximum likelihood for the multivariate vector of all observed variables

• Note that the simultaneous ML method can also be used when the model is recursive. This is simpler than performing several different regressions because it is done all in one step.
Exercise Illness Example from Kline p.124
Unstandardized estimates

fitness

exercise

hardiness

stress

illness

Unstandardized estimates

4410.39

14.40

.11

27.13

-8.84

-7.56

284.04

e1

317898.36

e2

41.81

e3

-.39

-.04

.32

-12.15

.40

.00

41.81
Significance of Unstandardized Estimates

Regression Weights

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>S.E.</th>
<th>C.R.</th>
<th>P</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>fitness &lt;- exercise</td>
<td>0.108589</td>
<td>0.013164</td>
<td>8.249151</td>
<td>0.000000</td>
<td></td>
</tr>
<tr>
<td>fitness &lt;- hardiness</td>
<td>0.395956</td>
<td>0.230363</td>
<td>1.718836</td>
<td>0.085644</td>
<td></td>
</tr>
<tr>
<td>stress &lt;- hardiness</td>
<td>-0.392844</td>
<td>0.088732</td>
<td>-4.427302</td>
<td>0.000010</td>
<td></td>
</tr>
<tr>
<td>stress &lt;- fitness</td>
<td>-0.039637</td>
<td>0.019892</td>
<td>-1.992595</td>
<td>0.046306</td>
<td></td>
</tr>
<tr>
<td>stress &lt;- exercise</td>
<td>-0.001434</td>
<td>0.005493</td>
<td>-0.261035</td>
<td>0.794066</td>
<td></td>
</tr>
<tr>
<td>illness &lt;- stress</td>
<td>27.125089</td>
<td>4.520979</td>
<td>5.999826</td>
<td>0.000000</td>
<td></td>
</tr>
<tr>
<td>illness &lt;- fitness</td>
<td>-8.835471</td>
<td>1.743768</td>
<td>-5.066885</td>
<td>0.000000</td>
<td></td>
</tr>
<tr>
<td>illness &lt;- exercise</td>
<td>0.317618</td>
<td>0.479014</td>
<td>0.663067</td>
<td>0.507288</td>
<td></td>
</tr>
<tr>
<td>illness &lt;- hardiness</td>
<td>-12.145922</td>
<td>7.938450</td>
<td>-1.530012</td>
<td>0.126014</td>
<td></td>
</tr>
</tbody>
</table>
Standardized estimates

- Exercise
- Fitness
- Stress
- Illness
- Hardiness

Correlation Coefficients:
- Exercise to Fitness: 0.39
- Fitness to Illness: -0.26
- Exercise to Illness: -0.22
- Hardiness to Fitness: 0.08
- Hardiness to Stress: -0.03
- Stress to Illness: 0.07
- Hardiness to Exercise: -0.01
- Fitness to Stress: -0.07
- Illness to Exercise: 0.03
- Illness to Fitness: -0.11
- Exercise to Hardiness: 0.16
- Fitness to Hardiness: -0.03
- Stress to Illness: 0.29
- Stress to Exercise: -0.18
- Hardiness to Stress: -0.07
- Health Illness: 0.18

Error Terms:
- e1
- e2
- e3
Total, Direct, and Indirect Causal Effects
Total Causal Effects

Endogenous variables | Exogenous variables | Purely Exogenous variables

<table>
<thead>
<tr>
<th>Standardized</th>
<th>hardness</th>
<th>exercise</th>
<th>fitness</th>
<th>stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>fitness</td>
<td>0.081774</td>
<td>0.392453</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>stress</td>
<td>-0.231709</td>
<td>-0.056951</td>
<td>-0.108853</td>
<td>0.000000</td>
</tr>
<tr>
<td>illness</td>
<td>-0.162546</td>
<td>-0.084876</td>
<td>-0.291862</td>
<td>0.290874</td>
</tr>
</tbody>
</table>
Direct and Indirect Causal Effects

<table>
<thead>
<tr>
<th>Implied Correlations</th>
<th>Standardized Direct Effects - Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Effects</td>
<td></td>
</tr>
<tr>
<td>Standardized Total Effects</td>
<td></td>
</tr>
<tr>
<td>Direct Effects</td>
<td></td>
</tr>
<tr>
<td>Standardized Direct Effects</td>
<td></td>
</tr>
<tr>
<td>Indirect Effects</td>
<td></td>
</tr>
<tr>
<td>Standardized Indirect Effects</td>
<td></td>
</tr>
<tr>
<td>Modification Indices</td>
<td></td>
</tr>
<tr>
<td>Bootstrap Summary</td>
<td></td>
</tr>
<tr>
<td>Bootstrap Distributions</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Standard Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group number 1</td>
<td></td>
</tr>
</tbody>
</table>

Standardized Direct Effects - Estimates

<table>
<thead>
<tr>
<th></th>
<th>hardness</th>
<th>exercise</th>
<th>fitness</th>
<th>stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>fitness</td>
<td>0.081774</td>
<td>0.392453</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>stress</td>
<td>-0.222807</td>
<td>-0.014231</td>
<td>-0.108853</td>
<td>0.000000</td>
</tr>
<tr>
<td>illness</td>
<td>-0.073871</td>
<td>0.033805</td>
<td>-0.260200</td>
<td>0.290874</td>
</tr>
</tbody>
</table>

Standardized Indirect Effects - Estimates

<table>
<thead>
<tr>
<th></th>
<th>hardness</th>
<th>exercise</th>
<th>fitness</th>
<th>stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>fitness</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>stress</td>
<td>-0.008901</td>
<td>-0.042720</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>illness</td>
<td>-0.088675</td>
<td>-0.118682</td>
<td>-0.031663</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Results
Direct and Indirect Causal Effects

- Direct effects - Direct effects are those estimates on the arrows, i.e. .392 is the direct effect of exercise on fitness while holding hardiness constant. What is help constant depends on what else is assumed to have a direct effect on the outcome. In this case, only exercise and hardiness are assumed to have direct causal effects on fitness.

- Indirect effects - Indirect effects are the sum of all the possible indirects causal paths going from one cause to one effect, i.e. -.1186 is the total indirect effect of exercise on illness. That is,

\[ .392 \times (-.260) + .392 \times (-.109) \times .291 + (-.014) \times .291 = -1.1184 \]

(off just a little due to rounding)
To get $R^2$ values turn on “squared multiple correlations” under Analysis properties in AMOS. $R^2$ does not change whether causal variables are simply correlated compared to causing one another. Often users incorrectly think that if earlier causal relationships are included this leads to larger $R^2$. 
Standarized/Unstandarized Coefficients

Unstandardized: ↔ is a covariance

Standardized: ↔ is a correlation

\[
\hat{\text{Corr}}(X, Y) = \frac{\hat{\text{Cov}}(X,Y)}{\sqrt{\hat{\text{Var}}X}\sqrt{\hat{\text{Var}}Y}}
\]

Unstandardized regression coefficient: \( \hat{\beta}_{Y \text{on} X} \)

Standardized regression coefficient: \( \hat{\beta}_{Y \text{on} X} \frac{\sqrt{\hat{\text{Var}}X}}{\sqrt{\hat{\text{Var}}Y}} \)
Testing the overall fit of the path model

Test for the fit of the proposed model covariance structure compared to the saturated model.

\[ H_0 : \Sigma = \Sigma(\theta) \]
\[ H_A : \Sigma = S \]

Create the likelihood ratio test statistics and compare to a chi-square distribution with degrees of freedom \( \frac{p(p+1)}{2} - \# \) of parameters estimated.

If every variable has an arrow (single or double headed) connecting it to every other variable, then the model is saturated. This means the estimated covariances will be equal to the observed covariances.

For the Illness/Exercise example d.f. = \( 5 \times 6 / 2 - 15 = 0 \).

It is not useful to consider the chi-square test for a saturated model, instead what we are often interested in is determining whether a model with some relationships fixed to zero fits as well as the saturated model. This amounts to testing regression coefficients equal to zero.
Counting the number of parameters to be estimated

1. All direct effects (single head arrows)
2. All Covariances (double head arrows)
3. Variance of all purely exogenous variables (this includes all error terms since they are also exogenous)

From the Exercise Illness example in Kline

→ $Illness = \beta_1 \text{Fitness} + \beta_2 \text{Stress} + \beta_3 \text{Exercise} + \beta_4 \text{Hardy} + \epsilon_1$

  Implied Variance of Illness:
  \[
  \text{Var}(Illness) = \beta_1^2 \text{Var}(Fit) + \beta_2^2 \text{Var}(Stress) + \beta_3^2 \text{Var}(Ex) + \beta_4^2 \text{Var}(Hardy) + \text{Var}(\epsilon_1) +
  2\beta_1\beta_2 \text{Cov}(Fit, Stress) + 2\beta_1\beta_3 \text{Cov}(Fit, Ex) + 2\beta_1\beta_4 \text{Cov}(Fit, Hardy) + 2\beta_2\beta_3 \text{Cov}(Stress, Ex) +
  2\beta_2\beta_4 \text{Cov}(stress, Hardy) + 2\beta_3\beta_4 \text{Cov}(Ex, Hardy)
  \]

→ $Fitness = \beta_5 \text{Exercise} + \beta_6 \text{Hardy} + \epsilon_2$

  Implied Variance of Fitness:
  \[
  \text{Var}(Fit) = \beta_5^2 \text{Var}(Ex) + \beta_6^2 \text{Var}(Hardy) + 2\beta_5\beta_6 \text{Cov}(Ex, Hardy) + \text{Var}(\epsilon_2)
  \]

→ $Stress = \beta_7 \text{Fitness} + \beta_8 \text{Exercise} + \beta_9 \text{Hardy} + \epsilon_3$

  Implied Variance of Stress:
  \[
  \text{Var}(Stress) = \beta_7^2 \text{Var}(Fit) + \beta_8^2 \text{Var}(Ex) + \beta_9^2 \text{Var}(Hardy) + \text{Var}(\epsilon_3) +
  2\beta_7\beta_8 \text{Cov}(Fit, Ex) + 2\beta_7\beta_9 \text{Cov}(Fit, Hardy) + 2\beta_8\beta_9 \text{Cov}(Ex, Hardy)
  \]

15 parameters = 9 direct effects + 1 covariance + 5 variances (i.e. 2 exogenous observed variables + 3 error variances)
Exercise/Illness example

Model Fit Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>NPAR</th>
<th>CMIN</th>
<th>DF</th>
<th>P</th>
<th>CMIN/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default model</td>
<td>11</td>
<td>5.92143</td>
<td>4</td>
<td>.20509</td>
<td>1.48036</td>
</tr>
<tr>
<td>Saturated model</td>
<td>15</td>
<td>.00000</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independence model</td>
<td>5</td>
<td>165.49898</td>
<td>10</td>
<td>.00000</td>
<td>16.54990</td>
</tr>
</tbody>
</table>
Common cause path model - testing partial correlation zero Example (from Kline p. 29)

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>X shoe-size</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y vocabulary breadth</td>
<td>.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>W age</td>
<td>.8</td>
<td>.6</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ r_{xy \cdot w} = \frac{.5 - .8 \times .6}{\sqrt{(1 - .8^2)(1 - .6^2)}} = \frac{.02}{.48} = .04 \]

The association between shoe-size and vocabulary breadth is explained by the common cause of age. PUT PATH DIAGRAM

Similar to role of Negative Affectivity found in Vassend O, Skrondal A (1999) "The role of negative affectivity in self-assessment of health" Journal of Health Psychology, 4(4), 465-482. Where the authors interpreted results to indicate that the observed correlations among the variables in question are, to a large extent, the spurious product of reporting differences associated with NA.
Mediation - Total and Partial

Generally when we talk about mediation we are asking whether the causal relationship between two variables can be broken down into a series of intermediate causal paths.
Mediation - Total and Partial

Logical Steps to testing for a mediator:

1. Test if there is a significant relationship between X and Z (i.e. is there a significant total causal effect of X on Z), record this effect $\tau$

2. Test if there is significant relationship between X and Y (i.e. is there a significant total causal effect of X on Y), record this effect $\alpha$.

3. Estimate the effect that X and Y have simultaneously on Z, record the Y on Z effect as $\beta$, record the X on Z effect as $\tau'$. 

If $\tau'$ is not significantly different than zero, then the causal effect of X on Z is fully mediated by Y. If $|\tau'| < |\tau|$ but $\tau'$ is still statistically significant, then we say that the relationship between X and Z is partially mediated by Y.

Sometimes see $\frac{|\tau| - |\tau'|}{|\tau|}$ presented as the percent of the total causal effect of X on Z explained by the mediator Y.
### Mediation - Total and Partial

<table>
<thead>
<tr>
<th></th>
<th>Z</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z unhealthy wt control</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X teased</td>
<td>.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Y depression</td>
<td>.8</td>
<td>.6</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\tau = .5, \alpha = .6, \beta = \frac{.8-.5*6}{1-.6^2} = .78, \tau' = \frac{.5-.8*6}{1-.6^2} = .03
\]
Action theory and Conceptual theory
Mediation in interventions

Given that an intervention or clinical trial (experiment) has been conducted where a program or drug or treatment has been given to some people and not others, it is possible to make statements about actual causal effects rather than assumed causal effects. The ideas and estimation are identical to what we’ve done so far with path analysis, the only difference is how the data were generated in the first place (i.e. by an experiment).

Entire area of study called “Evaluation Theory” revolves around how to decide when and how a policy or intervention or action has made a positive change.

Two critical aspects in the design and implementation of intervention

1. Conceptual theory - what are the causes of the outcome

2. Action Theory - what kind of intervention is needed to positively change those causes
Mediation designs for tobacco prevention research

David P. MacKinnon*, Marcia P. Taborga, Antonio A. Morgan-Lopez

Department of Psychology, Arizona State University, Tempe, AZ 85287-1104, USA

Received 9 April 2002; accepted 24 June 2002

Abstract

This paper describes research designs and statistical analyses to investigate how tobacco prevention programs achieve their effects on tobacco use. A theoretical approach to program development and evaluation useful for any prevention program guides the analysis. The theoretical approach focuses on action theory for how the program affects mediating variables and on conceptual theory for how mediating variables are related to tobacco use. Information on the mediating mechanisms by which tobacco prevention programs achieve effects is useful for the development of efficient programs and provides a test of the theoretical basis of prevention efforts. Examples of these potential mediating mechanisms are described including mediated effects through attitudes, social norms, beliefs about positive consequences, and accessibility to tobacco. Prior research provides evidence that changes in social norms are a critical mediating mechanism for successful tobacco prevention. Analysis of mediating variables in single group designs with multiple mediators are described as well as multiple group randomized designs which are the most likely to accurately uncover important mediating mechanisms. More complicated dismantling and constructive designs are described and illustrated based on current findings from tobacco research. Mediation analysis for categorical outcomes and more complicated statistical methods are outlined. © 2002 Published by Elsevier Science Ireland Ltd.

Keywords: Tobacco; Adolescents; Mediation; Indirect effects; Prevention; Methodology

Fig. 1. One mediator model illustrating the incorporation of action theory and conceptual theory for tobacco use prevention. The mediator variable for this illustration is social norms.

Fig. 2. One independent variable, six-mediator model illustrating the incorporation of action theory and conceptual theory for tobacco use prevention.
Suppression Effect

The saying that correlation does not prove causation should be complimented by saying that a lack of correlation does not disprove causation (Bollen, 1989, p.52)

<table>
<thead>
<tr>
<th># of errors</th>
<th>boredom</th>
<th>intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td># of errors</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>boredom</td>
<td>.35</td>
<td>1</td>
</tr>
<tr>
<td>intelligence</td>
<td>0</td>
<td>.7</td>
</tr>
</tbody>
</table>

Consider the following path diagram:

\[
\tau' = \frac{0 - .35 \times .7}{1 - .7^2} = -0.48
\]

When \(|\tau'| > |\tau|\), this is an indication that Y is a suppressor variable, that is, it is masking the relationship between X and Z. You don’t see a bivariate relationship unless you adjust for the suppressor variable.

In the example above boredom is suppressing the relationship between intelligence and the number of errors made.
Equivalent Models

Two models are equivalent if they are covariance equivalent, i.e. if every covariance matrix generated by one model (through some choice of parameters) can also be generated by the others.

For a detailed study of equivalent models check out:


Equivalent Models

Empirically we can choose whether models in Column 1 are better than Column 2 (or vica versa). But empirically we cannot choose which model within each column is better because they are equivalent.

| Column 1 | Column 2 |
A team of occupational health researchers is interested in decreasing the amount of dust inhaled by wood shop employees in Minnesota. They have found that there is large variability across employees in the amount of dust they inhale. Because the researchers are ultimately interested in some sort of intervention or change in policy, they want to build a causal model to explain what is causing individuals to have different levels of dust inhalation. In a study of a total of 100 employees at 10 different shops (10 employees from each shop) the following four variables were collected:

- **percent of time dust control used during one day** \( X_1 \)
- **personal beliefs about dangers of dust** \( X_2 \)
- **amount of dust inhaled during one day** \( Y \)
- **age** \( X_3 \)

where \( X_1 \) was observed for each employee by the occupational health researchers on the set study day and \( Y \) was measured using a dust monitoring device worn around the neck of each employee on the set study day. \( X_2 \) and \( X_3 \) were self-report measures by each employee with \( X_2 \) being the sum of 4 Likert items with higher values indicating a strong belief that dust is dangerous. All simple correlations between these four variables based on the report of the 100 employees are statistically significant.
1. Construct a causal path model which could reasonably explain how the three variables $X_1$, $X_2$, and $X_3$ cause $Y$. For every variable that you make an outcome in your model make sure to attach an error. Also, give labels to all your direct paths, e.g. $\beta_1$, $\beta_2$, etc. (Note: Answers may vary)

2. Give a justification for your model (based on your understanding of the subject matter). That is, give an explanation for each direct path you drew and each you did not draw (i.e. if you do not draw a direct relationship between two variables you should explain why not).

3. Based on your model
   (a) Does $X_1$ have any indirect effect on $Y$, if so what is its value in terms of $\beta$’s
   (b) Does $X_2$ have any indirect effect on $Y$, if so what is its value in terms of $\beta$’s
   (c) Does $X_3$ have any indirect effect on $Y$, if so what is its value in terms of $\beta$’s

4. Pick any one of the direct paths you drew and give an interpretation of its associated $\beta$.

5. Assume you knew that the quality of the air ventilation system at the different shops varied, and that this variable (which wasn’t measured) has a direct causal influence on how much dust is inhaled by employees in the different shops. Does not including this variable in the model bias the effect estimates in your model? Why or why not.
four variables were collected:
percent of time
dust control used
during one day
X1
personal beliefs
about dangers of
dust
X2
amount of dust
inhaled during
one day
Y
age
X3
To include $W$ or not to include $W$

Our interest is in estimating the total causal effect of $X$ on $Y$, i.e. $X \rightarrow Y$

In certain situations, the causal effect is estimated by $\hat{\beta}$ obtained from fitting the simple regression equation

$$Y = \beta X + \epsilon$$

and in some other situations, the causal effect is estimated by $\hat{\beta}_1$ obtained from fitting the regression equation including variables $W$

$$Y = \beta_1 X + \beta_2 W + \epsilon$$

The question is, when does $W$ need to be included or not.....
To include $W$ or not to include $W$

Our interest is in estimating the total causal effect of $X$ on $Y$, i.e. $X \rightarrow Y$

If there is a variable $W$ such that one of the following models is true:

It is not necessary to include $W$ in the model in order to unbiasedly estimate the total causal effect of $X$ on $Y$.

Note that if $W$ is included, the direct causal effect of $X$ on $Y$ will be different than if $W$ wasn’t included. But the same total causal effect will arise with or without $W$. That is, when $W$ is included the total causal effect equals the direct plus indirect causal effects.
To include W or not to include W

Our interest is in estimating the total causal effect of X on Y, i.e. $X \rightarrow Y$

The variable W is a **confounder** in each of the following models

Then W **must** be included in the model in order to unbiasedly estimate the causal effect that X has on Y.
To include W or not to include W

Our interest is in estimating the total causal effect of X on Y, i.e. $X \rightarrow Y$

If there is a W such that one of the following models is true

In each case the partial correlation between Y and W is zero once X is taken into account, thus there will be no difference whether W is included or not
To include $W$ or not to include $W$

Our interest is in estimating the total causal effect of $X$ on $Y$, i.e. $X \rightarrow Y$

If there is a $W$ such that the following models is true

\[
\begin{align*}
W & \leftarrow X \\
Y & \rightarrow X
\end{align*}
\]

In this case the simple correlation between $X$ and $W$ is zero thus the estimate of the causal effect of $X$ on $Y$ is the same whether $W$ is included or not. Note, the overall $R^2$ would be higher if $W$ was included but again it would not effect the causal effect of $X$ on $Y$. 
To include $W$ or not to include $W$

Our interest is in estimating the total causal effect of $X$ on $Y$, i.e. $X \rightarrow Y$

If there is a $W$ such that one of the following models is true

In these cases where $W$ is caused by $Y$, $W$ should not be included in the regression of $Y$ on $X$ in order to estimate the total causal effect of $X$ on $Y$. Including $W$ would result in biased estimates.
Moderator effects

• Same idea as an interaction effect.

• Does the relationship between variables differ across groups, if so then the group moderates the relationship.

• Can be examined using multiple group analysis. 6.10 of Kline.

• Perform chi-square test with direct paths constrained to be the same across the different groups, the perform chi-square allowing paths to be different. Then perform the chi-square difference test.