

## Path Analysis

Generally, path analysis is the combination of **assumed** causal theory with empirical evidence.

Slide 1

## 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.

Slide 2 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.

Slide 3

## 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 Wolfe LM (1991) Annotated bibliography of structural equation modeling: Technical work. *British Journal of Mathematical and Statistical Psychology*, 44, 93-152.
- Austin JT and Calderon RF (1996) Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural equation modeling*, 3, 105-175.
- Wolfe LM (1999) Sewall Wright on the method of path coefficients: An annotated bibliography. *Structural equation modeling*, 6, 280-291.
- Wolfe 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.

Slide 4

## 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 persons 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 individuals 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

From Neumark-Sztainer D, Wall MM, Story M, Perry C (2003) "Correlates of unhealthy weight-control behaviors among adolescents: Implications for prevention programs", *Health Psychology*, 22(1), 88-98.

Slide 5

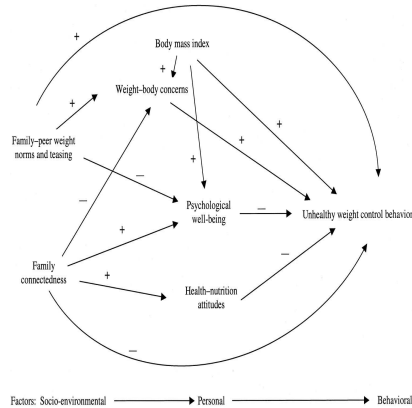


Figure 1. Proposed model: Correlates of unhealthy weight-control behaviors among adolescents.

## Example of Social Cognitive Theory

From Neumark-Sztainer D, Wall MM, Story M, Perry C (2003) "Correlates of unhealthy weight-control behaviors among adolescents: Implications for prevention programs", *Health Psychology*, 22(1), 88-98.

Slide 6

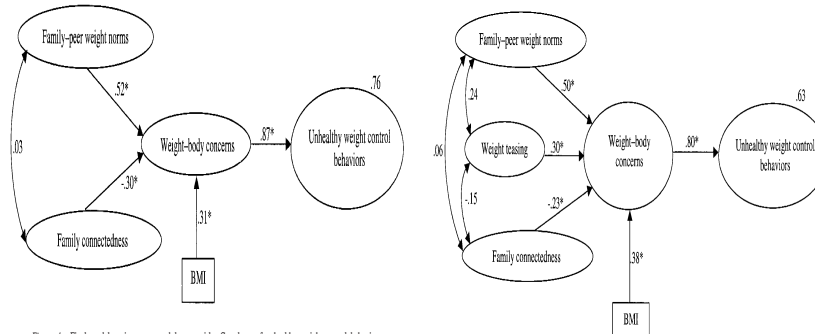


Figure 4. Final model testing among adolescent girls: Correlates of unhealthy weight-control behaviors. BMI = body mass index. \*  $p < .01$ .

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

From Raymond C, Shope, JT (2004) "Adolescent Problem Behavior and Problem Driving in Young Adulthood" *Journal of Adolescent Research*, Vol. 19 No. 2, 205-223.

208 JOURNAL OF ADOLESCENT RESEARCH / March 2004

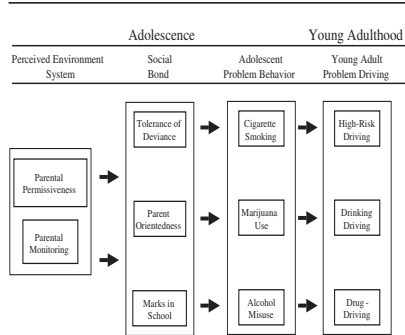


Figure 1. Conceptual model of sequence of associations among content areas.

Slide 7

## Example of Problem behavior theory

From Raymond C, Shope, JT (2004) "Adolescent Problem Behavior and Problem Driving in Young Adulthood" *Journal of Adolescent Research*, Vol. 19 No. 2, 205-223.

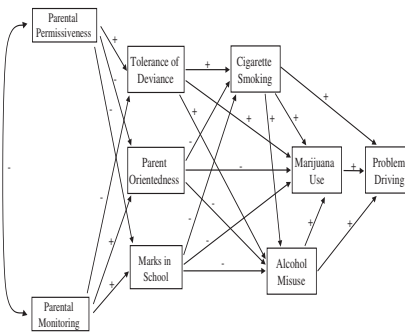


Figure 2. Hypothetical model predicting problem driving.

Slide 8

## Example of Problem behavior theory

From Raymond C, Shope, JT (2004) "Adolescent Problem Behavior and Problem Driving in Young Adulthood" *Journal of Adolescent Research*, Vol. 19 No. 2, 205-223.

Slide 9

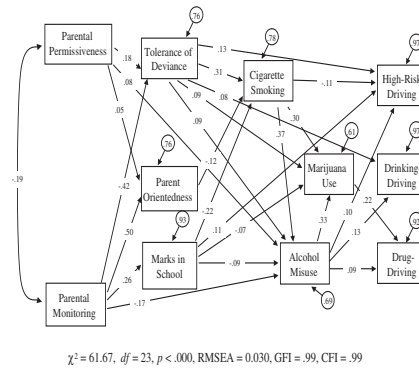


Figure 3. The final alternative model predicting problem driving.

## 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.

Slide 10

SEE: Additional causation.pdf file

SEE: Judea Pearl's website

[http://bayes.cs.ucla.edu/jp\\_home.html](http://bayes.cs.ucla.edu/jp_home.html)

Simple guidelines: To infer that X is a cause of Y (from an introductory statistical textbook by Moore and McCabe)

- 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  $\mathbf{X}$

- 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.

Slide 11

## Path Analysis - Extending the use of regression

- Path analysis provides a framework for the researcher to think more carefully about how the  $\mathbf{X}$  variables are related to the  $Y$  as well as how the  $\mathbf{X}$  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.**

Slide 12

## Review of regression

Standardized regression:

Given variables  $Y, X_1, \text{ and } X_2$ , we can standardize each of these variables and get

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

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

Slide 13

## Review of regression

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

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

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.  $r_{xy \cdot w} = r_{yx \cdot w}$

Slide 14

## Review of regression - Assumptions

Drop the superscripts 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  $X_2$

Slide 15

## 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  $\beta_1 x_1 + \beta_2 x_2$ .

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

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.

Slide 16

## Interpretation of $\beta_1$ in a path analysis model

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

Slide 17

The interpretation for  $\beta_1$  is:

If we could intervene in the population and increase the X1 variable for each individual by one standard deviation while making sure that each individual's X2 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.
- $X \rightleftarrows Y$  "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.

Slide 18

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

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

$$\begin{aligned}x_2 &= \beta_{23}x_3 + \epsilon_2 \\x_1 &= \beta_{12}x_2 + \beta_{13}x_3 + \epsilon_1\end{aligned}$$

Notice that the bivariate correlation between  $x_1$  and  $x_3$  can be reproduced from these standardized regression coefficients  $\beta$ . That is,  $\rho_{13} = \beta_{13} + \beta_{12}\beta_{23}$

Slide 19

## 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.**

Slide 20

## 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)

education ----> income -----> 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  $0.0002 \times 1000 = .2$  points

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

Slide 21

## Estimation for Path Analysis

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

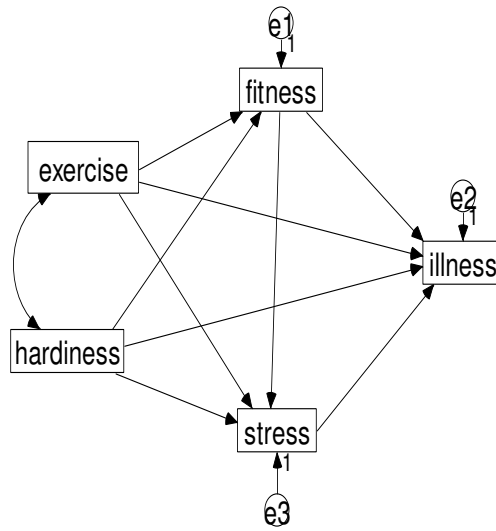
Then all the paths can be estimated using a series of multiple regressions estimated via OLS

- 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
- Not 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.

Slide 22

Exercise Illness Example from Kline p.117

Slide 23



Unstandardized estimates

Slide 24



## Significance of Unstandardized Estimates

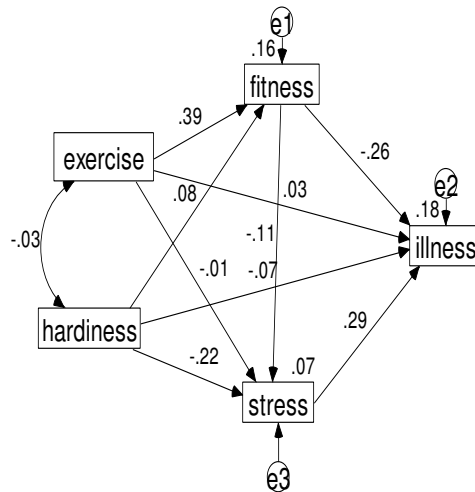
Slide 25

The screenshot shows a software window titled "C:\Documents and Settings\melanie\Desktop\old laptop\illnessexercise". The "Regression Weights" table is displayed, showing the following data:

	Estimate	S.E.	C.R.	P	Label
fitness <- exercise	0.108589	0.013164	8.249151	0.000000	
fitness <- hardiness	0.395956	0.230363	1.718836	0.085644	
stress <- hardiness	-0.392844	0.088732	-4.427302	0.000010	
stress <- fitness	-0.039637	0.019892	-1.992595	0.046306	
stress <- exercise	-0.001434	0.005493	-0.261035	0.794066	
illness <- stress	27.125089	4.520979	5.999826	0.000000	
illness <- fitness	-8.835471	1.743768	-5.066885	0.000000	
illness <- exercise	0.317618	0.479014	0.663067	0.507288	
illness <- hardiness	-12.145922	7.938450	-1.530012	0.126014	

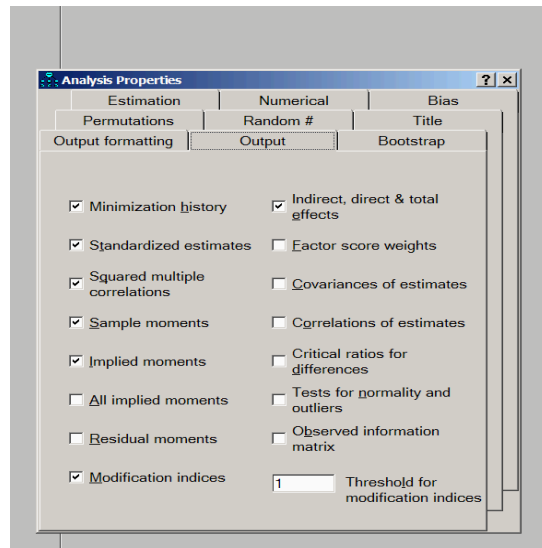
## Standardized estimates

Slide 26



## Total, Direct, and Indirect Causal Effects

Slide 27



## Total Causal Effects

Slide 28

The 'Standardized Total Effects - Estimates' table is displayed as follows:

	hardiness	exercise	fitness	stress
fitness	0.081774	0.392453	0.000000	0.000000
stress	-0.231709	-0.056951	-0.108853	0.000000
illness	-0.162546	-0.084876	-0.291862	0.290874

Endogenous variables

Exogenous variables

Purely Exogenous variables

## Direct and Indirect Causal Effects

Slide 29

The screenshot shows a software window titled 'C:\Documents and Settings\melanie\Desktop\old laptop\illnessexercise'. The window displays a table of standardized direct effects. The table has four columns: 'hardiness', 'exercise', 'fitness', and 'stress'. The rows represent the effects on 'fitness', 'stress', and 'illness'.

	hardiness	exercise	fitness	stress
fitness	0.081774	0.392453	0.000000	0.000000
stress	-0.222807	-0.014231	-0.108853	0.000000
illness	-0.073871	0.033805	-0.260200	0.290874

The screenshot shows a software window titled 'C:\Documents and Settings\melanie\Desktop\old laptop\illnessexercise'. The window displays a table of standardized indirect effects. The table has four columns: 'hardiness', 'exercise', 'fitness', and 'stress'. The rows represent the indirect effects on 'fitness', 'stress', and 'illness'.

	hardiness	exercise	fitness	stress
fitness	0.000000	0.000000	0.000000	0.000000
stress	-0.008901	-0.042720	0.000000	0.000000
illness	-0.088675	-0.118682	-0.031663	0.000000

## 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 held 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,
 
$$.392 * (-.260) + .392 * (-.109) * .291 + (-.014) * .291 = -.1184$$
 (off just a little due to rounding)

Slide 30

Slide 31

- Logical steps in testing for mediation
- Action theory and Conceptual theory  
Mediation in interventions

MacKinnon DP, Taborga MP, Morgan-Lopez AA (2002) "Mediation designs for tobacco prevention research" *Drug and Alcohol Dependence*, 68, S69-S83.

- Estimating proportion of Total causal effect due to indirect vs. direct effects

Example in a clinical trial

Tollefson GD, Sanger TM (1997) "Negative Symptoms: A path analytic approach to a double-blind, placebo- and haloperidol-controlled clinical trial with olanzapine" *American Journal of Psychiatry*, 154(4), 466-474.

- Total causal effect versus direct effect
- Confounding
- Equivalent models

Slide 32

### Papers warning about the problems with path analysis

Both of these papers are strongly negative against the practice of path analysis. Their main theme is: For non-experimental data need to be very aware that **actual** causal relationships are not proven by path analysis. Both of these papers are very good to read.

- Stone-Romero EF and Rosopa PJ (2004) "Inference problems with hierarchical multiple regression-based tests of mediating effects", *Research in Personnel and Human Resources Management*, 23, 249-290.
- Freedman DA (1987) "As others see us: A case study in path analysis", *Journal of Educational Statistics*, 12(2) 101-128.