Basics of SEM

- What is SEM?
- SEM vs. other approaches
- Definitions
- Implied and observed correlations
- Identification
- Latent vs. observed variables
- Exogenous vs. endogenous variables
- Multiple regression as a SEM model
- Steps in SEM analysis
- Interpreting output

What is SEM?

- Many names
  - » structural equation modeling
  - » covariance structure analysis (or covariance structure modeling or analysis of covariance structure)
  - » causal modeling
  - » path analysis (with latent variables)
- Several computer programs
  - » LISREL [**LI**near **S**tructural **REL**ationships]
    - the original
  - » EQS [**EQ**uation**S**]
  - » AMOS [**A**nalysis of **Mo**ment **S**tructures]
    - can be integrated with SPSS
  - » CALIS, LISCOMP, RAMONA, SEPATH, and others

SEM: What It Is and What It Isn't

WHAT IT IS:

- Tests hypotheses about relationships between variables
- Very flexible
- Comprehensive: Subsumes many other techniques
    - » multiple regression
    - » confirmatory factor analysis
    - » path analysis
    - » ANOVA

WHAT IT ISN'T:

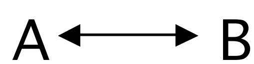
- Only for correlational studies
- A way to test causal hypotheses from correlational data

SEM vs. Other Approaches

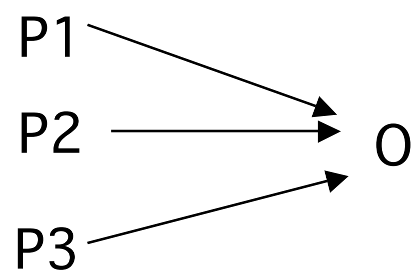
- Similar to standard approaches
  - » based on linear model
  - » based on statistical theory; conclusions valid only if assumptions are met
  - » not a magic test of causality
  - » statistical inference compromised if post hoc tests performed
- Different from standard approaches
  - » Requires formal specification of model
  - » Allows latent variables
  - » Statistical tests and assessment of fit more ambiguous
    - can seem like less of a science; more of an art

Some Definitions

- Model
  - » statement about relationships between variables
- Specification
  - » act of formally stating a model
- Examples
  - » zero-order correlation: 2 variables are related (but no direction specified)



  - » multiple regression: predictors have directional relationship with outcome variable



- Little explicit specification in standard techniques

How SEM Works

- You supply two main things
  - » Formal specification of model
  - » Observed relationship between variables
    - (i.e., a covariance or correlation matrix)
  - » (You also need to supply the number of participants or cases)
- Model implies a set of covariances
- Software tries to reproduce observed covariance matrix
- It does this by estimating parameters in the model
- Software produces two main things:
  - » parameter estimates
  - » information about how well it did in reproducing the covariance matrix

More Definitions

- Parameters
  - » parameters are constants
  - » indicate the nature and size of the relationship between two variables in the population
  - » we can never know the true value of a parameter, but statistics help us make our best guess
- Parameters in SEM
  - » can be specified as “fixed” (to be set equal to some constant like zero)
  - » or “free” (to be estimated from the data)
- Parameters in other techniques
  - » Pearson correlation: one parameter is estimated (r)
  - » Regression: regression coefficients are estimated

An Example

- Model

$$A \xrightarrow{p1} B \xrightarrow{p2} C$$

- Implied correlations
  - » $r_{A,B} = p1$; $r_{B,C} = p2$
  - » $r_{A,C} = p1*p2$
- Observed correlations
  - » $r_{AB} = .4$; $r_{B,C} = .4$; $r_{A,C} = .16$
    - perfect fit
  - » $r_{A,B} = .4$; $r_{B,C} = .4$; $r_{A,C} = .70$
    - unacceptable fit
  - » $r_{A,B} = .4$; $r_{B,C} = .4$; $r_{A,C} = .20$
    - ok fit
- Difference between “ok” and “unacceptable” is a judgment call
  - » no “p < .05” rule for the overall fit

Identification

- Refers to the relationship between what will be estimated (the parameters) and the information used to derive these estimates
- If a model is *identified* it is possible to calculate (estimate) a unique value for every parameter
- If not, the model is *unidentified* or *underidentified*
- Model will be unidentified if

  #Parameters > #Observations

- Can also be *empirically underidentified* depending on data
  - » e.g., with high multicollinearity it's as if you have fewer observed variables

Analogy: Solving Simultaneous Equations

1) $$x + y = 6$$

- » no unique solution (x=5,y=1 or x=4,y=2)
- » not identified

2) $$x + y = 6$$

$$2x + y = 10$$

- » unique solution: x=4, y=2
- » solution perfectly reproduces data (perfect fit to data)
- » "just identified"

3) $$x + y = 6$$

$$3x + 3y = 18$$

- » no unique solution
- » 2nd equation adds no constraints
- » empirically underidentified
- » like multicollinearity

More simultaneous equations

4) $x + y = 6$

$2x + y = 10$

$3x + y = 12$

- No solution perfectly reproduces data
  - » x=4, y=2 works for first two, but gives wrong answer for third equation
- But, can minimize differences between data and predicted outcomes
  - » usually, try to minimize sum of squares differences
  - » e.g., x=4, y=2 gives SS of 4
- Best solution is x=3.0, y=3.3
  - » sum of squared diffs = 0.67
- Unique best solution exists, but will not fit observed data perfectly
- Can measure how well it fits

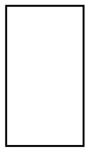
Fit: How Good is the Unique Solution?

- Note that more constraints (more equations) means that it's less likely that our fit will be good
  - » keep in mind when evaluating models
  - » excellent fit less impressive if not very many df
- Fit refers to how much the predicted covariances (or correlations) differ from the observed covariances
  - » small squared differences (residuals) indicate an acceptable fit
  - » i.e., the model is plausible (can't be rejected)
- Two main ways to measure: χ^2 and fit indices
  - » we'll come back to this in a few weeks

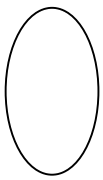
Latent and Observed Variables

- One big advantage of SEM: allows for the use of latent variables
  - » aka factors, constructs
  - » unmeasured (and unmeasurable) “pure” variables
  - » free of measurement error and “unique” factors
  - » represented by circles or ellipses
- In contrast to observed variables
  - » aka manifest or measured variables; indicators
  - » something directly measured (e.g., by a questionnaire)
  - » include measurement error and other variance not related to the “pure” construct of interest
  - » represented by squares or rectangles

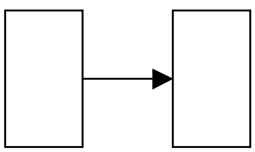
SEM Notation



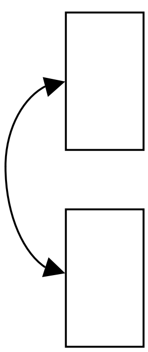
Boxes are used to describe observed variables



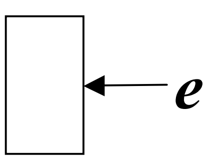
Circles are used to describe latent variables



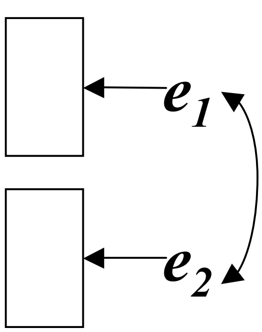
A single-headed arrow between two boxes
represents a causal relation



A double-headed arrow between two boxes
represents a noncausal (unexplained) relation

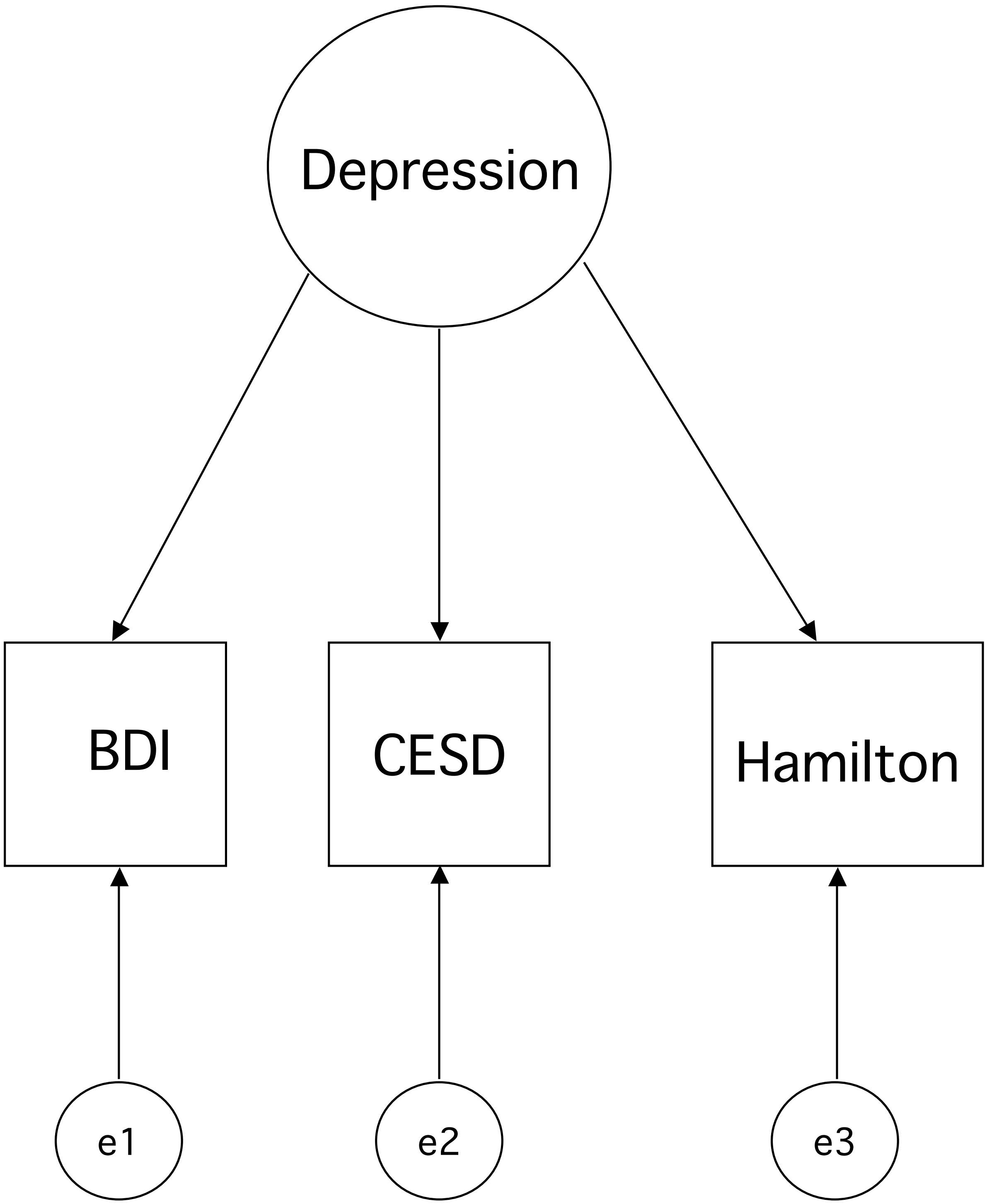


Arrows which do not originate from a box represent residuals



Double-headed arrows between two residuals represent the covariance of those residuals

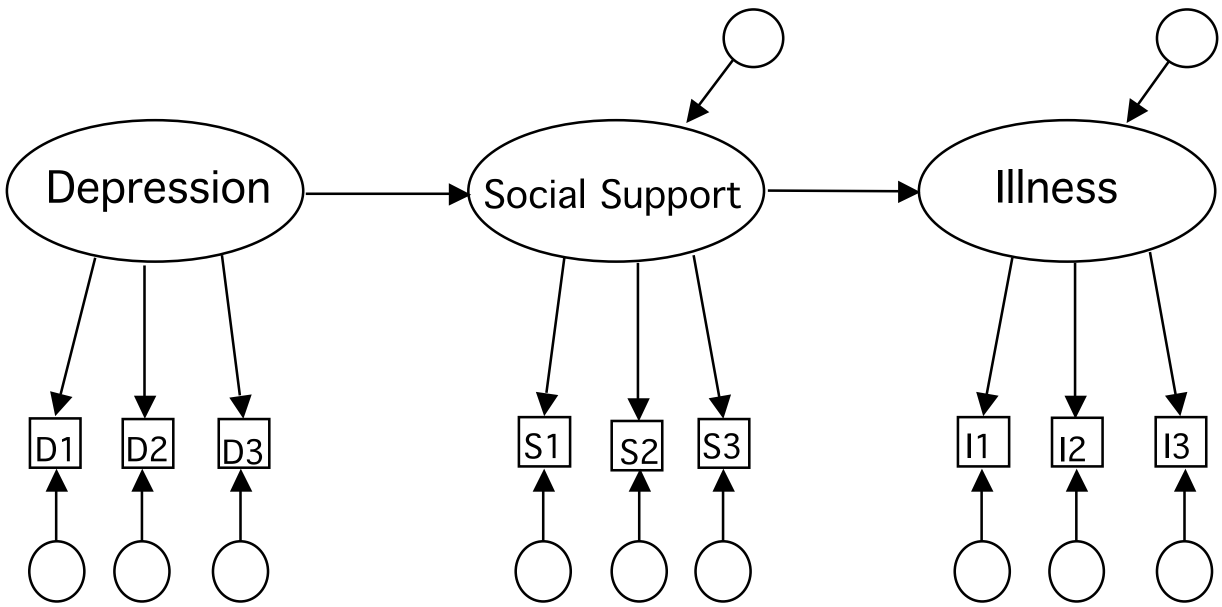
Latent/Observed (cont.)



Exogenous vs. Endogenous

- Exogenous variables
  - » "of external origin"
  - » causes are not included in the model (i.e., no arrows pointing to the variable; only arrows pointing out)
  - » like an IV (ANOVA) or a predictor (regression)
- Endogenous variables
  - » "of internal origin"
  - » represented as the effects of other variables (i.e., at least one arrow pointing to it)
  - » like a DV (ANOVA) or an outcome or criterion variable (regression)
- Endogenous variables can also predict other variables in the model
  - » different than ANOVA and regression
  - » endogenous vars can have arrows pointing in and pointing out

Exogenous vs. Endogenous (cont.)



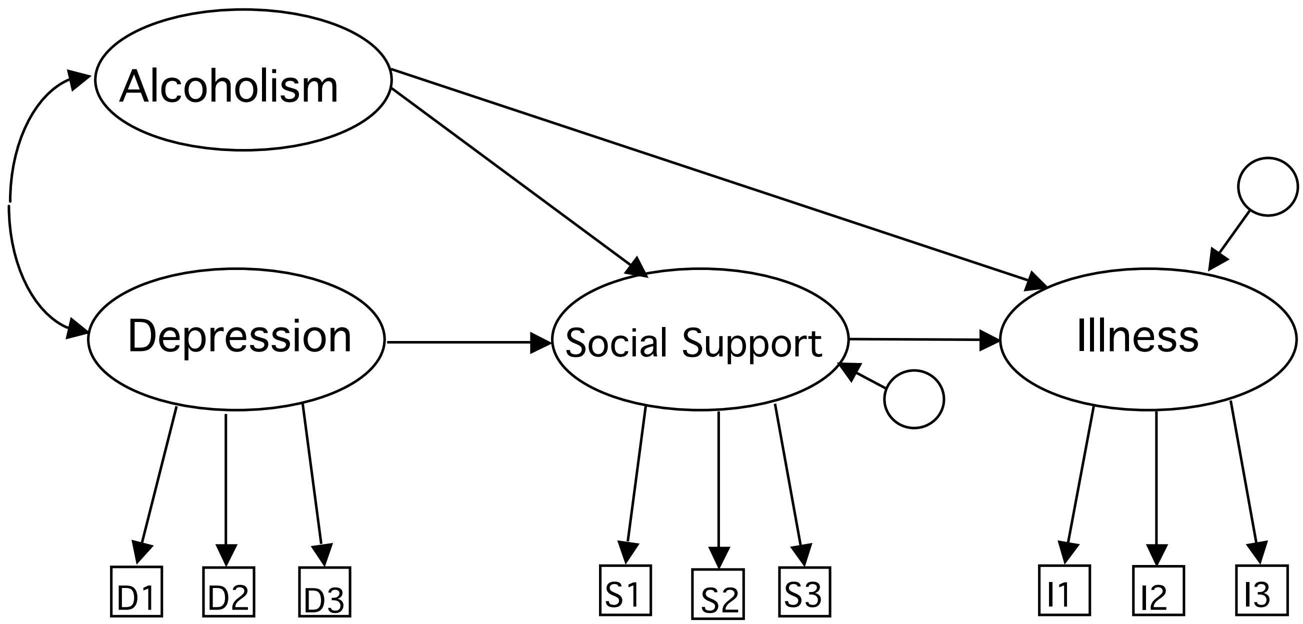
Disturbances

- Every endogenous variable has a disturbance
- These represent all omitted causes, plus any random or measurement error
  - » i.e., all variance that predictors didn't predict
- Also called residuals or error terms
  - » "error term" implies that there are no omitted causes (only error variance)
- Disturbances can be conceptualized as unmeasured (latent) exogenous variables
- They allow us to compute a percent variance explained for each endogenous variable

Types of Associations

- Association
  - » non-directional relationship
  - » the type evaluated by Pearson correlation
- Direct
  - » a directional relationship between variables
  - » the type of association evaluated in multiple regression or ANOVA
  - » the building block of SEM models
- Indirect
  - » Two (or more) directional relationships
  - » V1 affects V2 which in turns affects V3
  - » relationship between V1 and V3 is mediated by V2
- Total
  - » sum of all direct and indirect effects

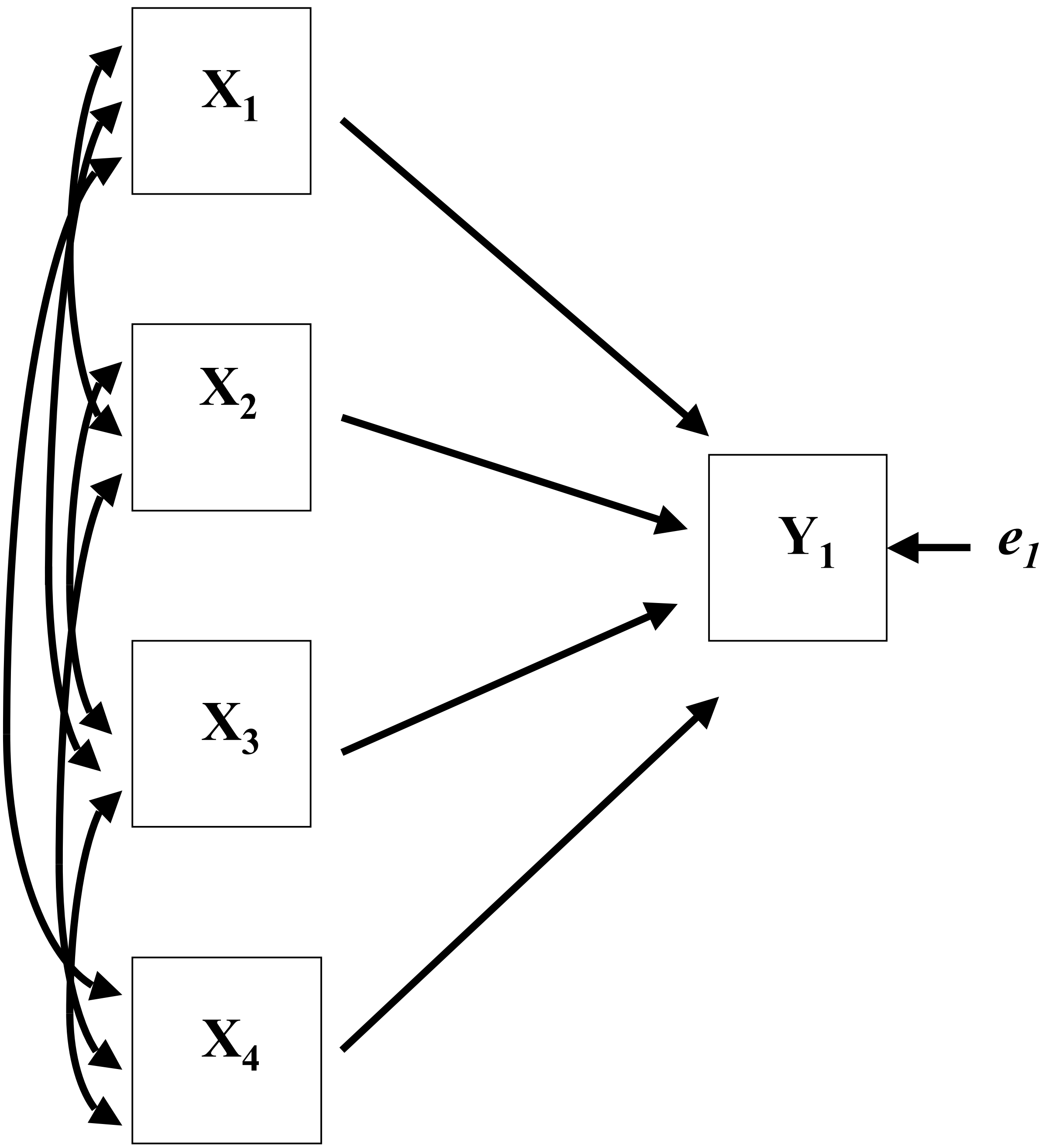
Associations



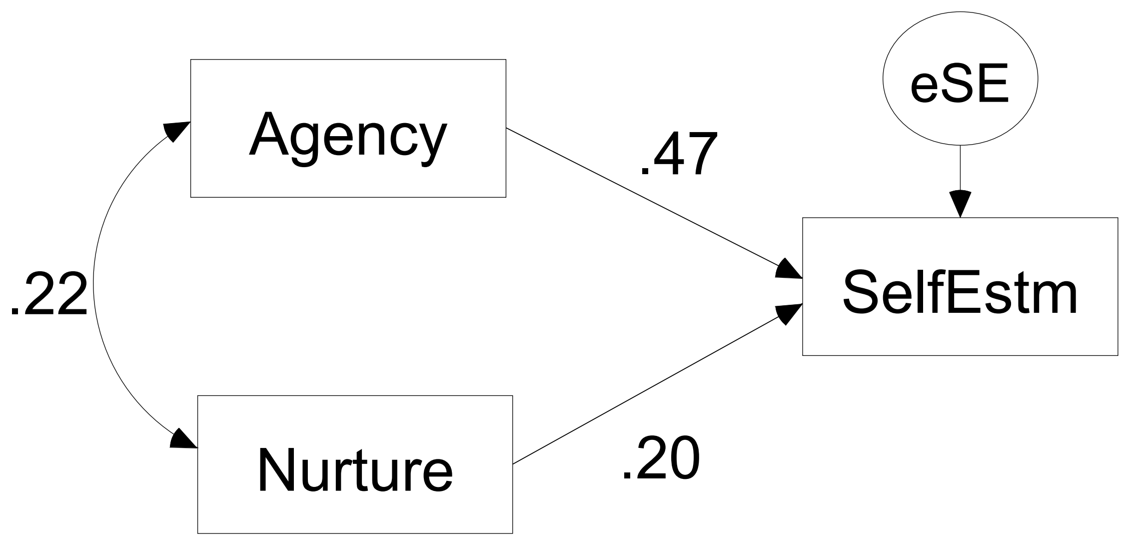
Multiple Regression

- Can run regression analyses using SEM software
- Mathematics/computer algorithm used by SEM is different, but
- Parameter estimates will be identical or very close
- Note that fit will be perfect (number of observations and number of parameters are equal)
- Running in SEM buys you nothing
  - » but, nice analysis to start with (you can check against SPSS or SAS run)
  - » SEM allows multiple DVs
  - » SEM allows two-group (or multi-group) comparisons

Multiple Regression Diagram



Multiple Regression Diagram



Steps in a SEM Analysis

- Step 1: Model specification
  - » usually done by drawing pictures using SEM software
- Step 2: Parameter estimation
  - » SEM software performs this step
  - » Iterative process
  - » Final result is a set of parameters that produce best fit to data possible
- Step 3: Assessment of fit
  - » Software did the best it could, but how good is that?
  - » Variety of ways to assess fit

Computer Software: Preparation

- The three steps in a SEM analysis are easy to remember:
  - » the software ensures that we have a properly specified model before parameters are estimated;
  - » parameter estimates are computed, and provided both on the diagram and in text output; and
  - » fit statistics appear after the parameter estimates. (defaults vary, but software allows the user to change the defaults).

Interpreting Output I

- Listing of model specification
  - » always good to check this
  - » familiarity with syntax more imp than in SPSS
- Listing of observed covariance matrix
- Scan for error messages
  - » e.g., that model did not converge
- Parameter values
  - » unstandardized and standardized
  - » like B's and β's in regression
- Listing of predicted covariance matrix
- Matrix of residuals
- Additional information on fit

Computer Software: EQS

- EQS 6.1 available in 4th floor computer lab
- EQS 6.1 also available on machines in computer lab in SS1
- Academic license: $595
- Go to Multivariate Software home page
  - » http://www.mvsoft.com
- Lisrel has a free, downloadable student versions (limited in terms of # of cases and/or # of variables)
  - » go to www.ssicentral.com
- Amos comes as part of SPSS GradPack (Windows version only)