PSY792F SEM

Week 2 – Path Analysis

Mark A. Prince, PhD, MS

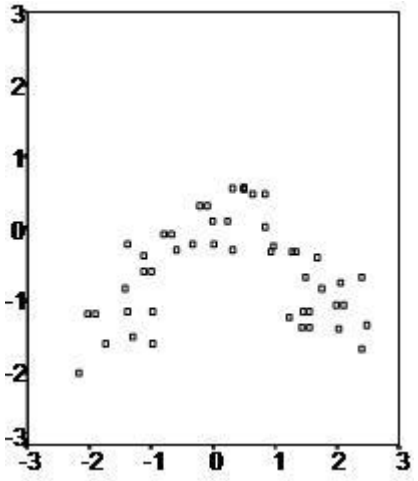
Baseline Exam Results

- Research area and types of data/analyses
  - A lot of variability
    - Some mentions of nested data, latent variables, longitudinal data, mediation/moderation, mixture models
    - Overall, I think we have the right topics on the syllabus
- Regression assumptions
  - A few people knew this, but we have to review
- Non-normal data
  - Very few people knew anything about this.
  - I am not going to cover it properly because there were only a few people that reported needing it
  - I'm willing help those interested learn how to model count or censored data.
- Factor Analysis
  - Surprisingly you all knew this better than regression.
  - We will review key gaps in knowledge in the second unit.

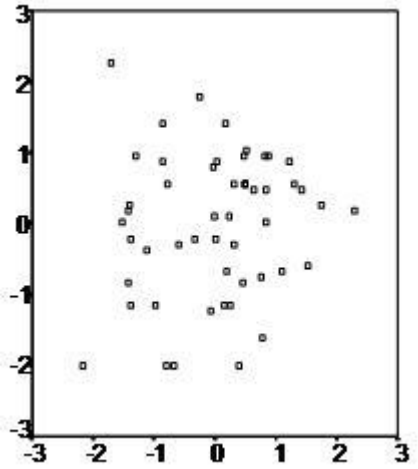
Linear Regression Assumptions

- At least two variables of ratio or interval scale
- Linear relationship (expected value of residuals at every point is 0)
  - Relationship between the independent and dependent variables expected to be linear
    - Check with scatter plots
    - Model comparisons with quadratic and other non-linear models
  - No Outliers – this will influence linearity
  - If violated – you will be more likely to make a type II error

Non-linear

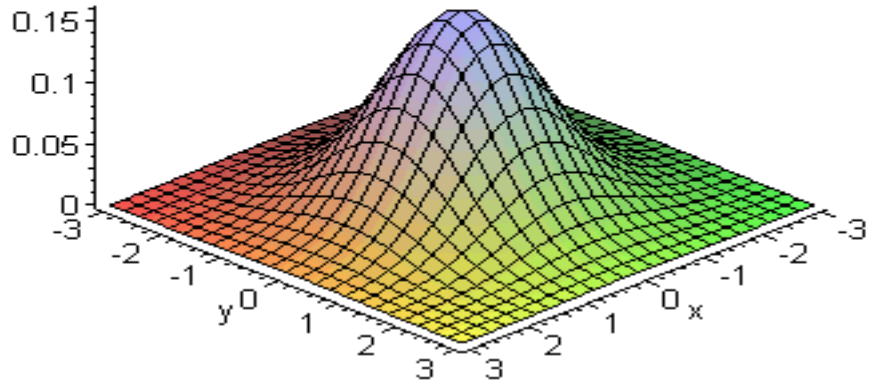


Linear



Linear Regression Assumptions

- Multivariate normality
  - The dependent variable needs to be normally distributed
  - Check with a histogram and a fitted normal curve, a Q-Q-Plot, or the Kolmogorov-Smirnof test.
  - When the data is not normally distributed a non-linear transformation, e.g., log-transformation, **might** fix this issue
    - Consider other distributions (e.g., negative binomial, Poisson, gamma)
  - If violated – you will be more likely to make a type II error

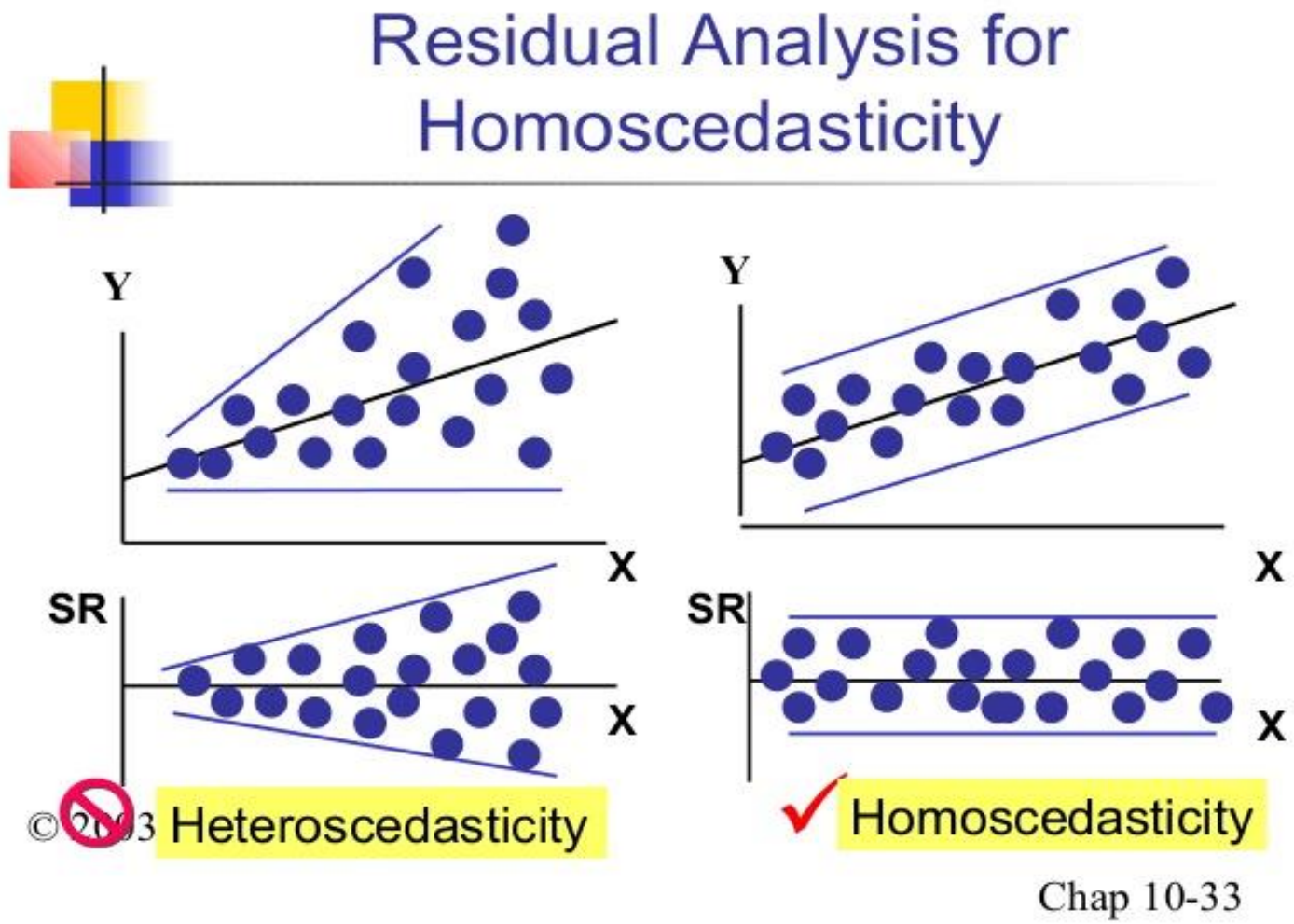


Linear Regression Assumptions

- No or little multicollinearity
  - Multicollinearity occurs when the independent variables are not independent from each other.
    - Check with correlation matrix: r > .79 is considered multicollinear
    - Consider factor analysis or using total score vs. subscale scores or a latent variable
    - If violated - you will be more likely to make a type II error
- Independence of observations (expected correlation between residuals of any two observations is 0)
  - No auto-correlation
    - Each participant should be independent
    - No nested data
    - Need multi-level modeling to address this
    - If violated can cause numerous problems
      - The underling math assumes that events are “disjoint”
        - P(A and B) = P(A)P(B)
      - If independence is violated you need to use “conditional probability”
        - P(B|A) = P(A and B)/P(A)

Linear Regression Assumptions

- Homoscedasticity
- The scatter plot is good way to check for homoscedasticity
  - That is the error terms along the regression are equal
  - Violation can result in Type I or Type II error – essentially violation gives too much weight to cases with extreme values
    - Really it means your model is mis-specified



Testing Assumptions

- See SPSS file with simulated data, syntax, and output
- There are also pdfs of the output and syntax

When to use Path Analysis

- Extension of multiple regression
  - Multiple Dependent Variables
- Use Path Analysis when:
  - All variables are observed
  - You have more than one dependent variable
    - Without Path Analysis you would run many regression analyses separately
    - With Path Analysis you can run them all at the same time
  - You can specify the relationships among dependent variables
    - Are dependent variables expected to be related to each other?
    - Are independent variables expected to be related to each other?
    - Any mediators or moderators of the relationship?
    - Covariates?
      - (what does this mean in a path model?)

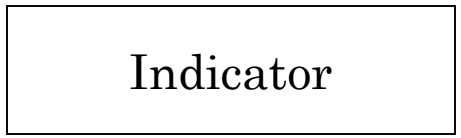
Why use Path Analysis

- Reduce Type I error inflation
  - Simultaneous estimation
- Test more complex models
  - Flexibility
  - Direct and indirect effects
  - Mediation/moderation
  - Multi-group analysis

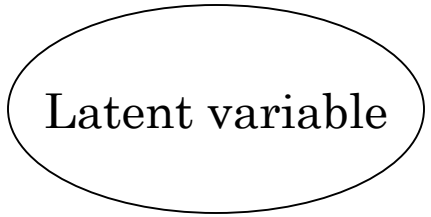
Theoretical Overview

- For Path Analysis all the same assumptions of multiple regression hold
- Still uses OLS regression and Maximum Likelihood estimation
  - *Does everyone know what these are?*
- Describe model with a Path Diagram (see slides to follow)
- **Exogenous and endogenous variables:**
  - Exogenous (think of them as predictors or IVs)
    - Variables where no arrows pointing towards them, except the measurement error term.
    - If exogenous variables are correlated to each other, then a double headed arrow will connect those variables.
  - En**d**ogenous (think of them as outcomes or DVs)
    - Variables may have both the incoming and outgoing arrows.
- Terms:
  - Path coefficient (just a regression coefficient)
  - Disturbance (just a residual)
  - Direct effect – single arrow from X to Y
  - Indirect effect – X predicts Y via M (X-M-Y)
  - Total effect = direct + indirect effects

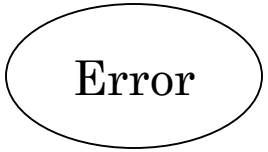
Introduction to Terms/Notation



Observed variable, items in a measure

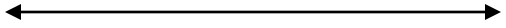


Unobserved variables or factors



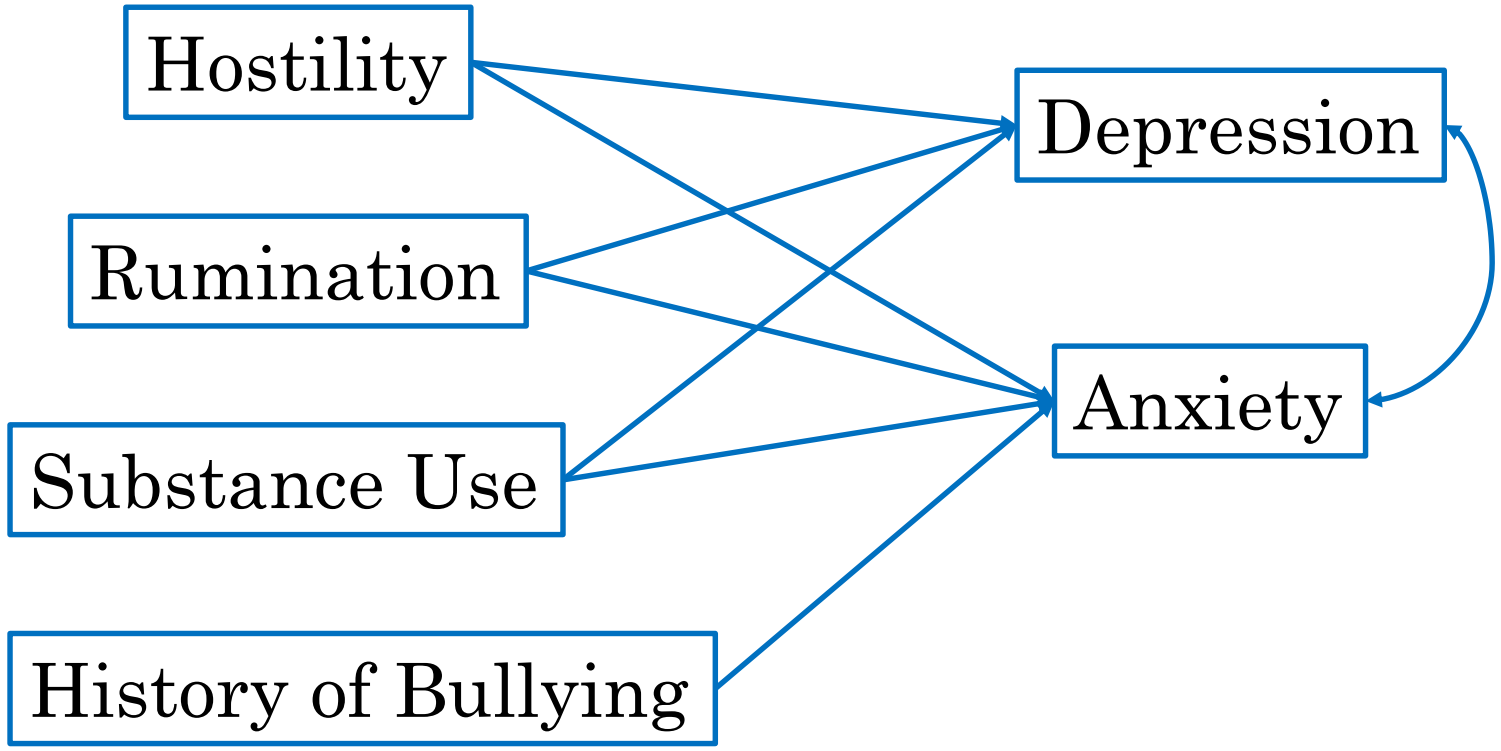
Measurement error

Direct effect, regression coefficient

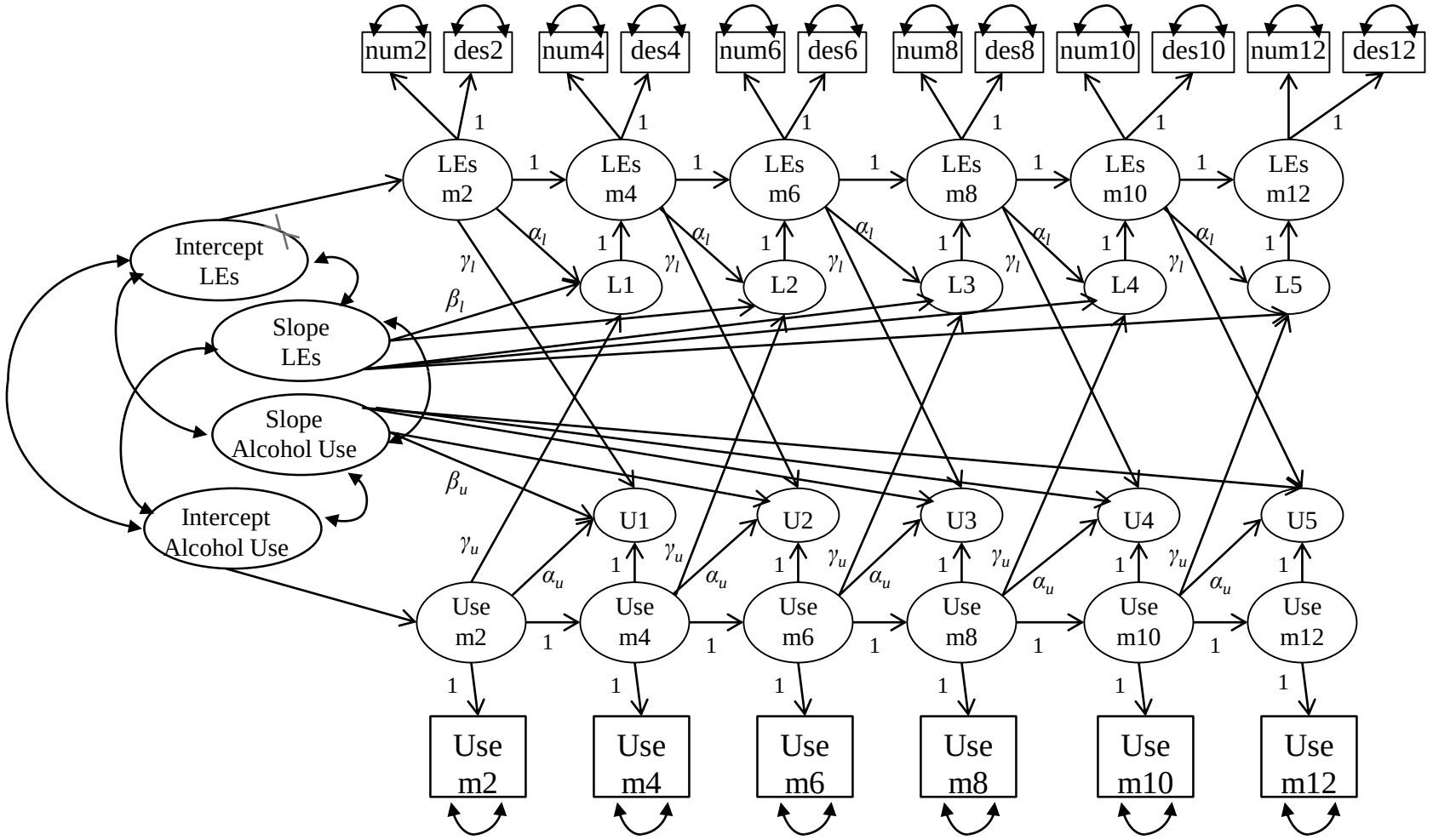


Covariance

Simple Path Diagram



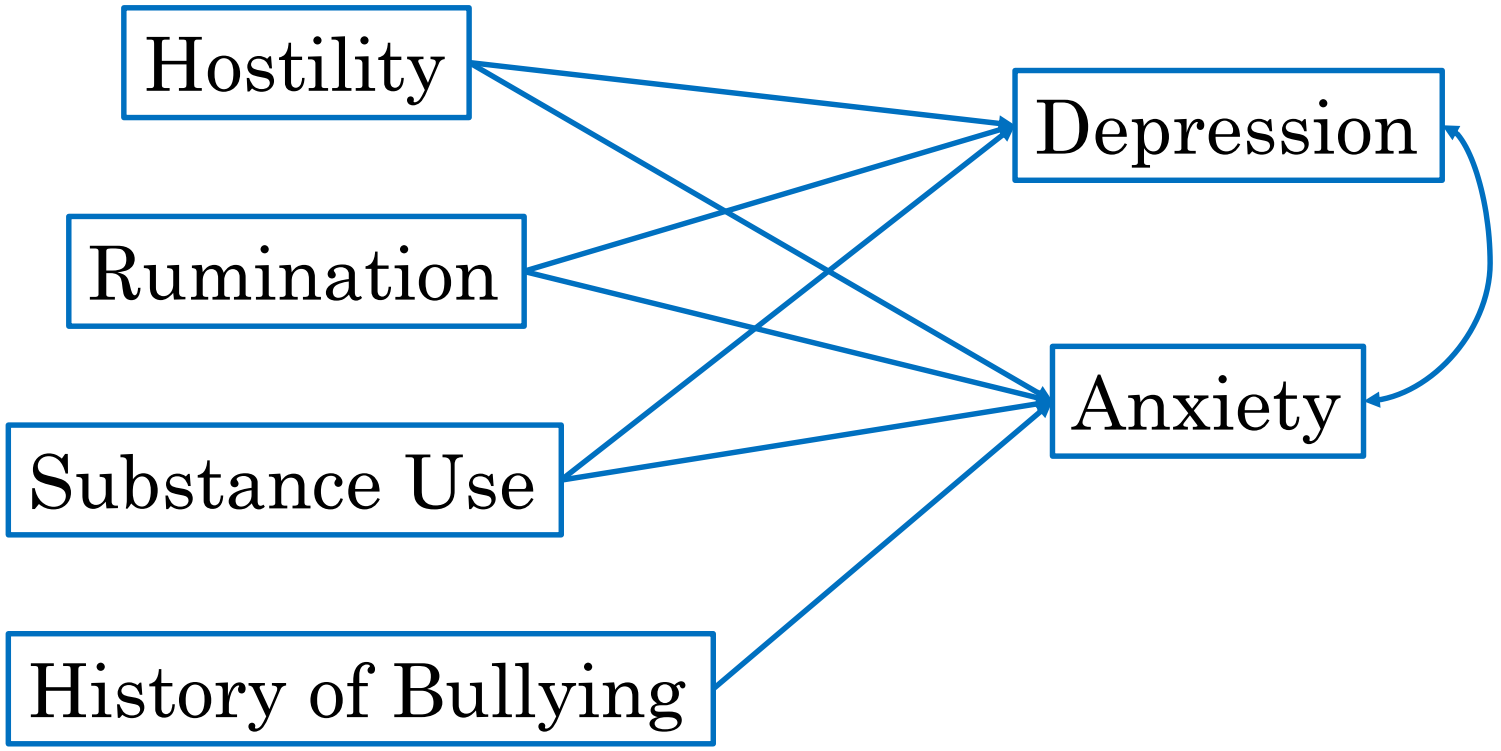
More complex path diagram



Data considerations for Path Analysis

- Wide format
  - Each participant gets a single row
  - Contrast with tall or long format where each participant gets many rows
    - Needed for MSEM
- Data considerations
  - Regression assumptions apply here
  - DVs – Continuous normal? Count? Censored? Binary?
  - Missing data?
  - Multi-group?
  - Bootstrapping?
- Start with a picture

Simple Path Model



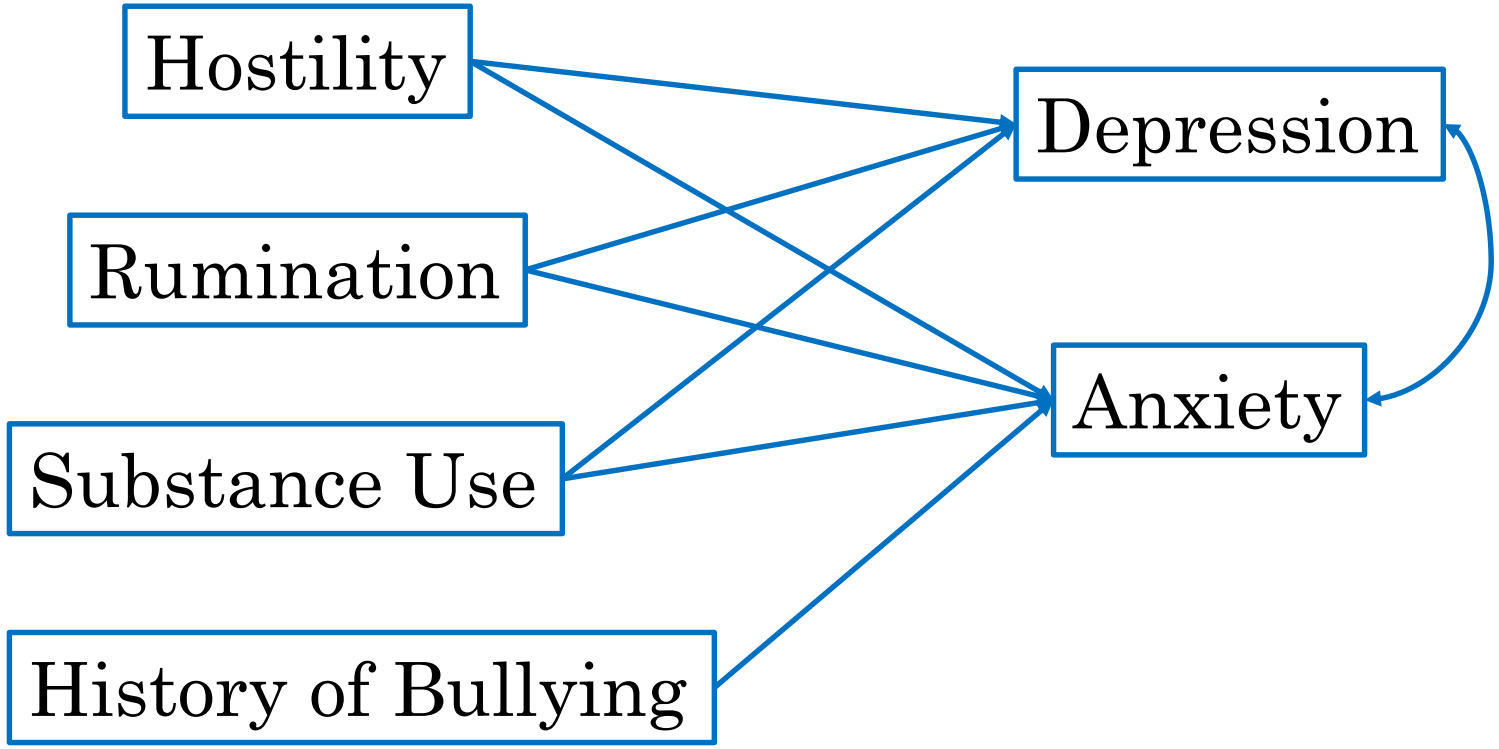
Hostility: continuous
Rumination: continuous
Substance Use: continuous
History of Bullying: continuous
Depression: continuous
Anxiety: continuous

How to write the code

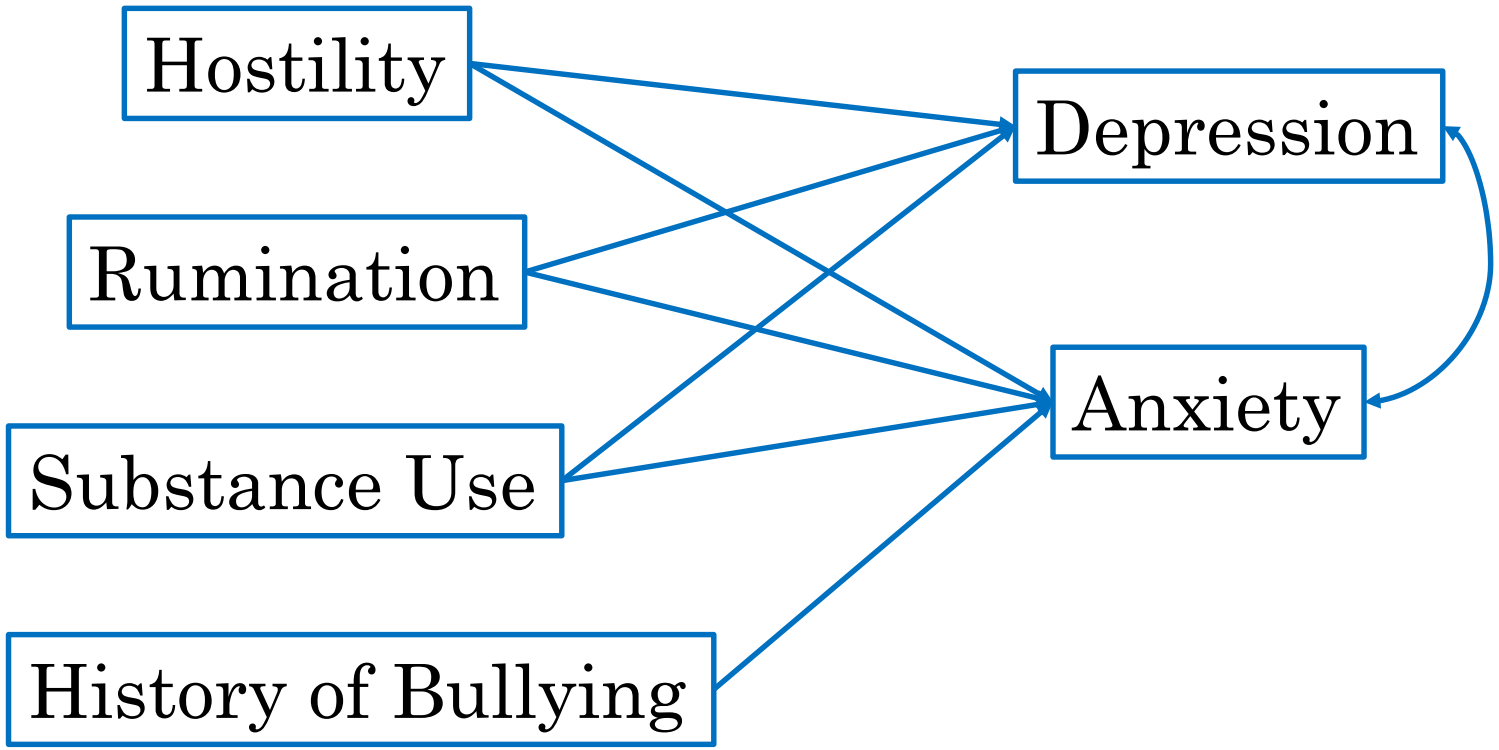
- Mplus – users guide
- Basic commands:
  - In mplus, there are three basic commands
    - **On**: this is a regression command. Y on x; means to regress y onto x.
    - **With**: this is a correlation. Y with x; means correlate the two.
    - **By**: this is to create a factor. Y by x; means that y is a factor and x is the first indicator.
  - The second important point is instituting constraints
    - Y with x (1); the (1) means to impose a constraint. Anything that has the (1) following it will all be constrained to equality. You can only have one (#) on a line.
  - The last important point is variances / means
    - Just listing the variable name = variance or residual variance
    - Listing variable name in [ ] = mean or intercept.
    - Dep1@0 says to fix the variance of dep1 to 0
    - [dep1@0] says to fix the mean of dep1 to 0

How to write the code

- Model statement
  - Use picture as a guide
    - Depression on hostility;
    - Depression on rumination;
    - Depression on substance use;
    - Anxiety on hostility;
    - Anxiety on rumination;
    - Anxiety on substance use;
    - Anxiety on history of bullying;
    - Depression with anxiety;



```
TITLE:    Path Analysis All Continuous Variables
DATA:
FILE IS ex3.11.dat;

VARIABLE:
NAMES ARE
depression anxiety HoB Host rumination Sub;

USEVARIABLES ARE
depression anxiety HoB Host rumination sub;

MODEL:
depression on host;
depression on rumination;
depression on sub;

anxiety on host;
anxiety on rumination;
anxiety on sub;
Anxiety on HOB;

depression with anxiety;

Output: sampstat stdyx cinterval;
```



How to read the results

- See output file
- Sample statistics
- Model fit
  - Chi-square
  - RMSEA
  - CFI/TLI
  - SRMR
- Model results
  - Parameter estimates
    - Unstandardized regression coefficients/correlations
      - Means, intercepts, variances, residual variances
    - Standardized regression coefficients/correlations
    - R-square
    - Confidence Intervals (un)standardized

How to write up the results

- Analysis plan
  - Data decisions
    - e.g., transformed variables, how variables were coded or re-coded, used summary scores vs. subscale scores, how the data were collected (e.g., number of time points)
  - Model choice
    - Based on hypotheses
  - Model fit
    - Overall
      - Criteria used, citations
    - Comparative
      - Criteria used, citations
      - Sequence of models testing, citations or rationale
      - Model building considerations, citations or rationale

Comparing multiple models

- If you don't specify all the paths in your model a priori (e.g., some exploratory hypotheses) you can compare multiple models
- Modindices
  - An output statement that will suggest paths to add
  - May not make sense, so add them thoughtfully
- Run a series of models and compare model fit
  - Select the most parsimonious model with the best fit that makes theoretical sense

How to write up the results part 2

- Results Section
  - Parameter estimates that relate to your hypotheses
    - Factor loadings
    - Regression coefficients
    - Correlations
  - No interpretation of results, but explain effect
    - High values of X associated with high values of Y
    - Use “such that”

How to interpret the results

- Discussion
  - Summary of results
  - Longitudinal vs. cross-sectional?
  - What does it mean?
  - Strengths?
  - Limitations?

Now you try…

- Pick a dataset
  - Test the assumptions
  - Pull the data into mplus
  - Run a couple path models