

A World View of Models

- Today's topics:
 - Distinguishing Models you Know: A New World View
 - What kind of predictors?
 - What kind of outcomes?
 - How many piles of variance?
 - The Big Picture
 - Wrapping Up...

Organizing the Models You Know

- There are several major dimensions by which statistical models you know can be organized:
 - What kind of predictors?
 - Continuous or categorical predictors?
 - Observed (square) or latent (circle) predictors?
 - What kind of outcome variable?
 - Normal/continuous? Non-normal/categorical? Another kind?
 - Just one outcome or many? Intermediate outcomes?
 - Observed (square) or latent (circle) outcome variable?
 - How many 'piles of variance' per outcome variable?
 - Just one residual error term or multiple error terms?
 - How many dimensions of sampling need to be controlled for?

The General Linear Model (GLM)

	Categorical X's	Continuous X's	Both Types of X's
Univariate Y	ANOVA	Regression	ANCOVA/ Regression
Multivariate Y's	Profile Analysis or MANOVA	Canonical Correlation	MANCOVA/ Canonical

- The GLM family of models are a special case for when the outcomes are normally distributed, with no intermediate outcomes, there is only one pile of variance (error term) per outcome, and all predictors and outcomes are observed (not latent).
- They get called different names, depending on what kind of X's. Yet distinct model names for different types of predictors (continuous, categorical) are not used in more complex models.

What kind of outcome? *Generalized vs. General*

- *Generalized* linear models are models for *non-normally* distributed outcomes
 - Useful for outcomes that will never be normally distributed
 - Use link functions to transform outcomes into something more continuous and non-bounded
 - Binary data → Logit (log odds) link or probit (standard normal) link
 - Count data → Poisson link
 - Censored data → Tobit link
 - Many other possible link functions
 - The transformed outcome can then be predicted from a typical linear model of predictors (new $y = B_0 + B_1X_1 + B_2X_2$)
 - Still put in whatever kind of predictors (continuous or categorical)

What kind of outcome? *Generalized vs. General*

- General linear models are a special case of generalized linear models when the 'link function' = 'identity'
 - This means multiply outcome by 1, since it's already continuous and unbounded, and assume normally distributed errors
- Which should you choose?
 - "Could be normally distributed, but just isn't" – general model with a transformation of the DV (e.g., SQRT, LN) or robust estimator
 - "Not ever going to be normally distributed" – *generalized* model
- A note about distributional assumptions:
 - Distributional assumptions apply to the distribution of the model residuals – thus, they concern the OUTCOMES, not the predictors
 - Link functions in generalized models change the assumed distribution of the residuals into something besides continuous normal (which is likely to be more reasonable given the distribution of the outcome)

Variants of the Generalized Linear Model for Observed Variables

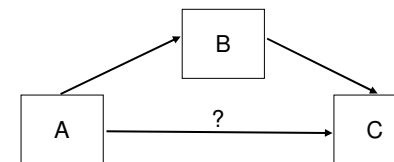
	Categorical X	Continuous X	Categorical & Continuous X
Single Categorical Y	Chi-Square, Loglinear Models	Discriminant Analysis; Logistic or Multinomial Regression	Logistic or Multinomial Regression
Single Continuous Y	Linear Regression, ANOVA	Linear Regression	Linear Regression, ANCOVA
Multiple Continuous Y's	Profile Analysis, MANOVA	Canonical Correlation	Canonical Correlation, MANCOVA

'Linear' vs. 'Non-linear'

- If people ever say to you "you should try a non-linear model", ask them to clarify "what kind of non-linear?"
- Do they mean...
 - A generalized model instead of a general model for an outcome that is not normally distributed (and not ever going to be)?
 - A model with a 'non-linear' effect of a predictor?
 - e.g., $y = B_0 + B_1 \text{age} + B_2 \text{age}^2$ → still linear in the parameters (combine additively), although not linear in the variables
 - A model that is truly non-linear in the parameters?
 - e.g., a negative exponential growth function for decline with age:
 $y = B_0 + B_1 \exp(-B_2 \text{age})$
where B_0 = asymptote, B_1 = change, B_2 = exponential rate of decline
 - Pry still others out there, too...

Intermediate Outcomes?

- The multivariate extensions of the GLM can handle multiple outcomes at once (manova, canonical, etc)
- What about the case of 'intermediate outcomes'?
 - These are often called 'mediators'



Here, variable B is both an outcome (of variable A) and a predictor (of variable C).

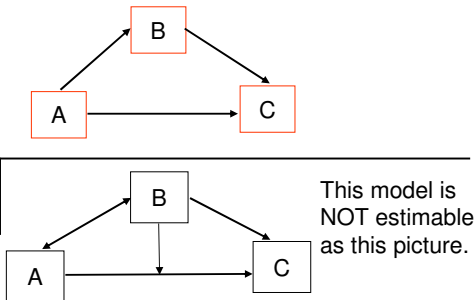
The logic implied is that A causes B, which then causes C.

Typically we assess to what extent the A→C relationship is reduced after accounting for B→C.

'Mediator' vs. 'Moderator'

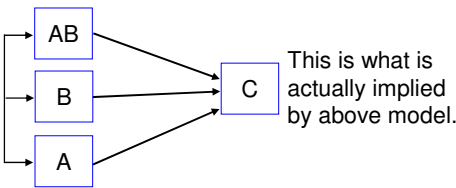
Mediational model:

- A causes B, which **then** causes C
- B is an outcome of A and a predictor of C
- Direct/indirect effects



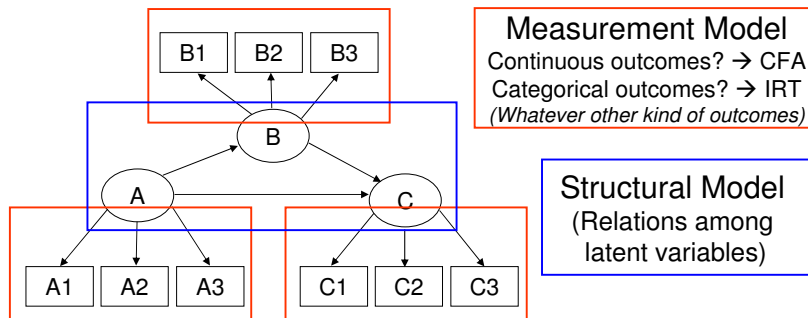
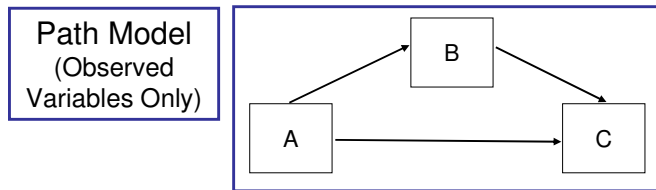
Moderator model:

- B impacts the A→C relationship
- B is a predictor of C, and is **correlated** with A
- Moderation is represented by an interaction effect



Observed vs. Latent Outcomes

- No intermediate outcomes, observed outcomes only?
→ *Regression/Canonical (GLM)*
- With intermediate outcomes, observed outcomes only?
→ *Path Analysis (multiple simultaneous regressions)*
- Latent outcomes (intermediate/not) or latent predictors?
→ *Structural Equation Model (SEM)*
- SEM = correlation/regression among latent variables
 - Each latent variable is defined by a *measurement* model, or a latent factor predicting observed outcomes
 - That latent factor then becomes an IV or DV in the *structural* model of the relations among the latent factors



Organizing the Models You Know

- There are several major dimensions by which statistical models you know can be organized:
 - What kind of predictors?
 - Continuous or categorical predictors?
 - Observed (square) or latent (circle) predictors?
 - What kind of outcome variable?
 - Normal/continuous? Non-normal/categorical? Another kind?
 - Just one outcome or many? Intermediate outcomes?
 - Observed (square) or latent (circle) outcome variable?
 - How many 'piles of variance' per outcome variable?
 - Just one residual error term or multiple error terms?
 - How many dimensions of sampling need to be controlled for?

Why multiple ‘piles of variance’ (or “variance components”)?

- Variance in your outcome is specified as a function of the dimensions of sampling:
 - Sample 100 subjects? Only need an error term (pile of variance in Y) that allows for differences across subjects
 - Sample 100 subjects over 5 time points? Now need at least two error terms that allow for subject differences and time differences
 - Sample 100 kids in 10 schools? Now need at least two error terms that allow for kid differences and school differences
 - Sample 100 kids in 10 schools over 5 time points? At least three error terms...

Why multiple ‘piles of variance’ (or “variance components”)?

- Piles of variance pertaining to higher-order sampling dimensions are known as ‘random effects’
 - Because the intent is to generalize across a population of that sampling dimension...
 - Instead of estimating specific differences (fixed effects) among the units, we estimate the variance across units instead
- Models that contain both fixed and random effects (fixed effects of predictors, >1 pile of variance) are called...
 - Mixed models (“general linear mixed model”)
 - Hierarchical linear models
 - Multilevel models (MLM)
 - Random coefficients models

Variants of the Generalized Linear (Mixed) Model for Observed Variables

	GLM: One error term	MLM: Multiple error terms
Single Categorical Y	Logistic or Multinomial Regression	Logistic or Multinomial Mixed Model
Single Continuous Y	Linear Regression, ANCOVA	Linear Mixed Model
Multiple Continuous Y's	Canonical Correlation, MANCOVA	Multivariate Linear Mixed Model

Ties Between Models

- Random effects = latent variables
 - Differences between sampling units of unknown origin, to be predicted by between-unit differences on predictor variables
- Therefore, many models involving ‘random effects’ can be re-cast as structural equation models
 - Growth models of individual differences in level and change
 - Multilevel → Random Effects, SEM → Latent Factors
 - Measurement models for continuous indicators
 - Latent trait in CFA = random intercept in general linear mixed model
 - Measurement models for categorical indicators
 - Theta in IRT = random intercept in generalized linear mixed model
 - Categorical latent variables?
 - Mixture models, latent class analysis, diagnostic classification models

The Big Picture

- General Linear Models are a special case of MLM or SEM with only fixed effects (only one pile of variance per outcome), no intermediate outcomes, and all outcomes and predictors are observed
- Once you add intermediate outcomes, you move into path analysis
- In adding latent variables, you move into SEM (or maybe into MLM, if you specify your latent variables as random effects instead)
- Once your outcomes are not normal, continuous variables, you move into generalized models (for observed or latent variables)
- The biggest umbrella framework I know of:
 - Skrondal & Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. Chapman and Hall: Boca Raton, FL.
 - Pretty much everything is special case of this framework somehow.
- Mplus is about the only program capable of estimating nearly everything within this giant umbrella framework.

What this means for SEM...

- Structural Equation Modeling will be focused on:
 - Path analysis: Relations among observed variables
 - Measurement models to build latent variables (thetas):
 - Confirmatory factor models for continuous indicators
 - Item Response (generalized) models for categorical indicators
 - Structural models: Relations among latent variables
 - Relationships attenuated for measurement error
 - Complex models involving intermediate outcomes
 - Tests of global fit and of all specific parameters of interest
 - What else? How about Multilevel Structural models:
Relations among latent variables at multiple levels of analysis
 - Structural models among each set of variance piles per sampling dimension (over time, over people, over groups, etc)

Wrapping Up...

- Statistical models can be broadly organized along a few dimensions:
 - General (normal) vs. Generalized (not normal)
 - Observed (measured) vs. Latent (unmeasured) variables
 - One dimension of sampling (one error term per outcome) vs. multiple dimensions of sampling (multiple error terms)
- Seemingly disparate families of statistical models share more commonalities than they appear:
 - Because random effects can be conceptualized as latent variables, there are points of intersection among multilevel (mixed) models and structural equation models
 - Let me and Jim help you find your way in both worlds ☺