Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

Lecture 1 (January 4, 2007): Introduction of latent modelling

XH Andrew Zhou

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Applications of latent variable models

Latent variable modeling can be used to represent the following phenomena:

- "True" variables measures with errors
- Hypothetical constructs
- Unobserved heterogeneity
- Missing data
- Potential outcomes
- Latent responses underlying categorical variables
- generate flexible multivariate distributions
- Combine information about individual units from different sources

”True” continuous variable measured with error

- Latent variables can represent the "true scores" of a continuous variable measured with error.
- Let y_j be the measured score of unit j with error. The classical measurement error model:

$$y_i = \eta_i + \epsilon_j,$$

where η_j is the true continuous score of unit j, and ϵ_j is the measurement error. The measurement errors have zero means and are uncorrelated with each other and the true scores. The true score is defined as the expected value of the measured variable for a subject,

$$\eta_j = E(y_j)$$

over imagined replications. This is the expected value definition of latent variables.

A Path Diagram

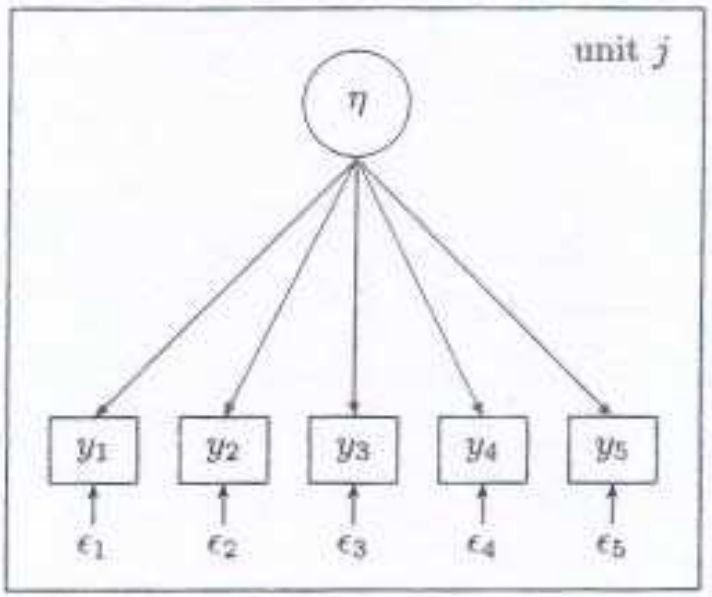

Figure 1.1 *Path diagram of the classical measurement model*

- The circle represents the latent variable; the rectangles represent the observed measurements; and the arrows represent linear relations.

”True” continuous variable measured with error, cont

- Typically, multiple measurements are taken for each unit and the measurement model becomes:

$$y_{ij} = \eta_j + \epsilon_{ij}.$$

”True” categorical variable with error

- When the true variable is constructed as categorical, such as medical diagnosis of disease, the observed measurements are also categorical, measurement error then take the form of misclassification.
- Measurement models with categorical latent and measured variables are known as latent class models.

Diagnosis of myocardial infarction, a latent class model

Rindskopf and Rindskopf (1996) analyzed data from a coronary care unit in New York city to rule out myocardial infarction or "heart attach". Each Each patient was assessed on 4 diagnostic criteria:

Q-wave : a dummy variable for presence of absence of a Q-wave on the ECG

History : a dummy variable for presence of a previous heart attach

LDH - lactate dehydrogenase: a dummy variable for having a elevated level of LDH

CPK-MB : a dummy variable for presence or absence of a CPK-MB isoenzyme

Since a patient has either MI or not, it is reasonable to specify 2 latent classes.

A basic assumption of measurement models

- For both continuous and categorical variables, a basic assumption is the conditional independence. That is, measurements are conditionally independent given the latent variable. i.e. the dependence among measurements is solely due to their common association with the latent variable.

Hypothetical constructs

- In contrast to true variables measured with error that are assumed to exist, hypothetical constructs have an exclusively epistemological status.
- A construct is an intellectual device by means of which one construes events.
- A construct is a simply a concept.
- Relationships between constructs provide inductive summaries of observed relationships as a basis for elaborating networks of theoretical laws.
- The words that scientists use to denote constructs, for example, "anxiety" and "intelligence", have no real counterparts in the world of observables; they are only heuristic devices for exploring observables.
- Since hypothetical constructs do not correspond to real phenomena, it follows that they cannot be measured directly even in principle.
- A construct is operationally defined in terms of a number of items or indirect "indicators" such as answers in an intelligence test.

Hypothetical constructs, cont

- The relationship between the latent construct and the observed indicators can be modeled using a common factor model:

$$y_{ij} = \lambda_i \eta_j + \epsilon_{ij},$$

where η_j is the latent trait variable or "common factor", representing the hypothetical construct, for subject j, λ_i is a factor loading for item i ($i = 1, \ldots, I$) and ϵ_{ij} are unique factors. Here η_j is independent of ϵ_{ij}.

Hypothetical constructs in psychological research

- Most of research in psychology is concerned with hypothetical constructs such as "self-esteem", "personality", and "life satisfaction".
- In medicine, examples of hypothetical constructs include "depression" and "quality of life".

Categorical hypothetical constructs

- Sometimes it is more natural to consider categorical constructs or typologies.
- In psychology, "stages of change" (pre-comtemplation, contemplation, preparation, action, maintenance, and relapse) are though to be useful for assessing whether patients are in their "journal" to change health behaviors such trying to quit to smoking
- In medicine, functional syndromes such as irritable bowel syndrome, which are characterized by a set of symptoms (whose cause is unknown), can be viewed as categorical hypothetical constructs. Here the fact that certain symptoms have high probabilities of occurring together is taken as an indication that they may be caused by the same disorder.

Developments of hypothetical constructs

- Instead of defining hypothetical constructs on the theoretical grounds, they are sometimes derived from an exploratory factor analysis for continuous hypothetical constructs and exploratory latent class analysis for categorical hypothetical constructs.
- Factor analysis may provide a means of evaluating theory or of suggesting revision in theory. It requires, however, that the theory be explicitly specified prior to the analysis of data.

Developments of hypothetical constructs, cont

- Although continuous hypothetical constructs are usually modeled by common factors, several multivariate statistical methods have also been used to explore the "dimensions" underlying data, such as principle component analysis, partial least squares, and canonical correlations, discriminant analysis, and multidimensional scaling.
- Categorical constructs can also be derived using cluster analysis, finite mixture modeling or multidimensional scaling.

Validity measures

- Although hypothetical constructs are useful in many disciplines, methods for investigating construct validity, the measurement and interrelationships among constructs are essential.
- Some argue that construct validity can be investigated by means of structural equation models with latent variables.

Convergent validity

- One advantage of latent variable modeling is that we can investigate the tenability of hypothesized structure, either by assessing model fit or by elaborating the model.
- For example, convergent validity can be assessed by specifying models where indicators designed to reflect a given construct only reflect that construct and not others.
- As an illustration, the upper panel of Figure 1.2 hypothesized a model where the first factor is measured by items 1 to 3, and the second factor is measured by the items 4 to 6.
- If this model is rejected in favor of the model in the lower panel, where item 5 loads on both factors, then convergent validity does not hold.

Convergent validity, cont

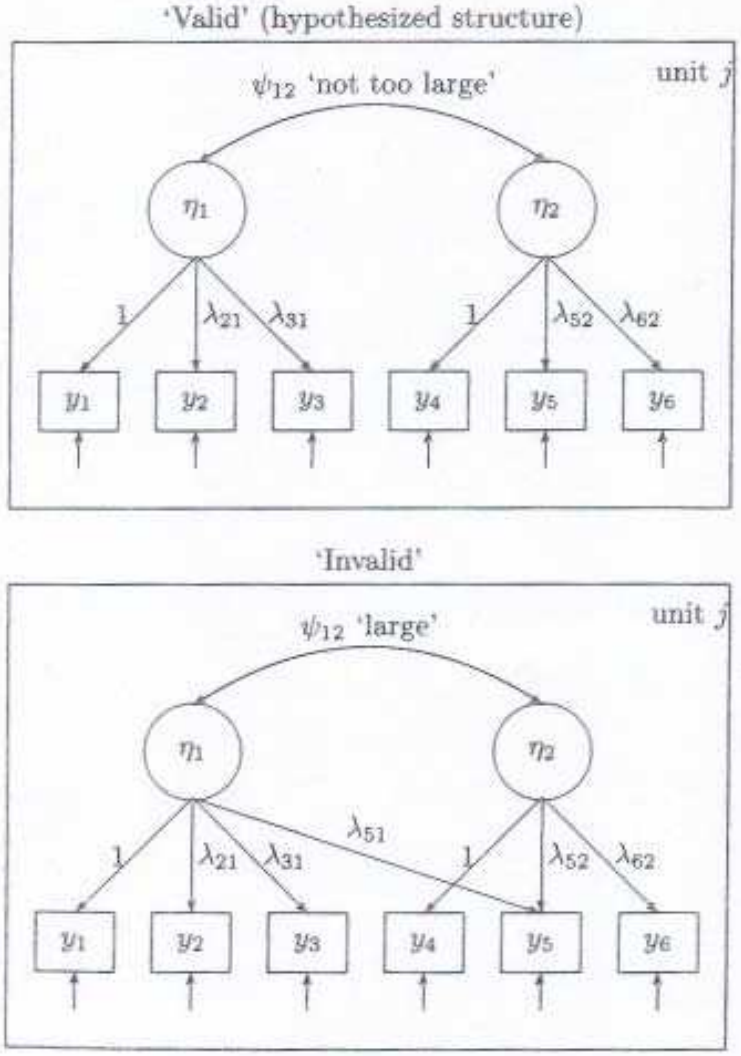

Figure 1.2 *Convergent and discriminant validity*

Construct validity

- The statements connecting constructs with each other, and observable indicators with constructs, constitute a nonmological network (law-like).
- The construct validity can be investigated by means of structural equation models with latent variables, which explicitly specify hypothesized structures.

Construct validity

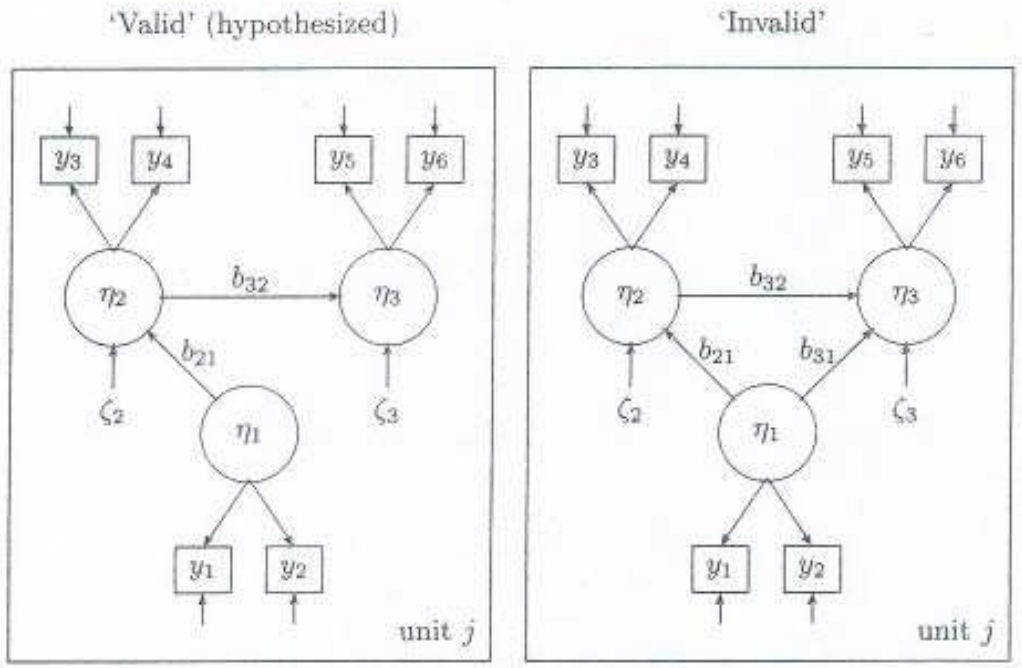

Figure 1.3 *Nomological validity*

- *Nomological validity* is typically assessed by investigating the tenability of

Unobserved heterogeneity

- A major aim of statistical modeling is to "explain" the variability in the response variable in terms of variability in observed covariates.
- However, in practice, not all relevant covariates are observed, leading to unobserved heterogeneity. Including latent variables, refereed to as random effects, is a common way to take unobserved heterogeneity into account.
- Unobserved heterogeneity is not a hypothetical construct since it merely represents the combined effect of all unobserved, and it not given any meaning beyond this.
- However, the random effects in genetic studies occupy an intermediate position as they can be interpreted as shared and unobserved genetic and environmental influences.

Unobserved heterogeneity, cont

- When the units are clustered, shared unobserved heterogeneity may influence "intra-cluster" dependence among he responses, even after conditioning on observed covariates.
- One way of accounting for unobserved heterogeneity is to include a random intercept in a regression model.
- In multilevel or hierarchical data there are often several levels of clustering, such as patients clustered within doctors in hospitals. We can then use latent variables at each of the higher levels to represent unobserved heterogeneity at that level.

An example: a sample three-level model

- A simple 3-level random intercept regression model for patient i cared by doctor j in hospital k can be written as

$$y_{ijk} = \eta_{0ijk} + \beta_1 x_{ijk} + \epsilon_{ijk},$$

$$\eta_{0ijk} = \gamma_{00} + \xi_{jk}^{(2)} + \xi_{k}^{(3)}.$$

- In the level-1 model for y_{ijk}, x_{ijk} is a covariate with "fixed" regression coefficient β_1, and η_{0ijk} is a random intercept with mean γ_{00} and residuals $\xi_{jk}^{(2)}$ and $\xi_{k}^{(3)}$ at levels 2 and 3, respectively.

Latent responses

- Latent variables can represent continuous variable underlying observed "coarsened" responses such as dichotomous or ordinal response.
- For example, in the dichotomous case, the observed response y_i is modeled as resulting from a regression model for an underlying continuous response y_i^*:

$$y_i^* = x_i'\beta + \epsilon_i,$$

with a threshold model

$$y_i = 1 \text{ if } y_i^* > 0, y_i = 0 \text{ if } y_i^* \leq 0.$$

- This model corresponds to a probit model if ϵ_i is a standard normal, a logit model if ϵ_i has a logistic distribution, and a complementary log-log model if ϵ_i has a Gumbel distribution.

Generating flexible distributions

- Latent variables are useful for generating distributions with the desired variable function and shape, or multivariate distributions with a particular dependence structure.