Latent Variables in Science: Three Vignettes

Karen Bandeen-Roche
Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Professor of Medicine and Nursing, The Johns Hopkins University

Search for the Hurley-Dorrier Chair of Biostatistics
Johns Hopkins Bloomberg School of Public Health

September 17, 2008

Purpose

- **Vision:** Communicate what I contribute to scientific inquiry

- **Mission**

  — Report on work in a particular area of focus

  — **Brief** overview of other work

  — Metaphor for philosophy on statistical science

Philosophy on Statistical Science

- **A spectrum**
  — **Discipline**

  — **Collaboration** with other science fields

- **Impact potential:** span across the spectrum

- **Span = an especial strength of Johns Hopkins**
  — Department, School, University

Outline

- One slide: Research scope
- Latent variable modeling
  - — What, why, how
  - — Mode for doing science: Do data bear out theoretic predictions?
  - — Vignette 1: Theory operationalization
  - — Vignette 2: If data don't bear out theoretic predictions: How not?
  - — Vignette 3: Translation from latent to observed
- Areas needing discovery

Research Scope

- **Aging, visual health, brain health**
  - — Cohort studies
  - — Programs: Older Americans Independence Center
    Alzheimer's Disease Research Center
    Epi/Biostat of Aging Training Program
  - — Statistical work: longitudinal / multivariate data analysis
- **Multivariate failure time analysis**
  - — Association modeling
  - — Competing risks
- **Latent variable modeling**

Latent Variables: What?

- *Underlying: not directly measured. Existing in hidden form but capable of being measured indirectly by observables*

  — Ex/ Pollution source contributions to an airshed

  — Ex/ Syndromal type

  — Ex/ Integrity of physiological regulation of systemic inflammation

- <u>Some favorite books</u>: Bartholomew (1988), Bollen (1989), McCutcheon (1987), Skrondal & Rabe-Hesketh (2004)

- <u>Model</u>: A framework linking latent variables to observables

Latent Variables: What?
Integrands in a hierarchical model

- Observed variables (i=1,...,n): Y_i=M-variate; x_i=P-variate
- Focus: response (Y) distribution = $G_{Y|x}(y|x)$; x-dependence
- Model:

— Y_i generated from <u>latent</u> (underlying) U_i:

$$F_{Y|U,x}(y|U=u,x;\pi) \qquad \textit{(Measurement)}$$

— Focus on distribution, regression re U_i:

$$F_{U|x}(u|x;\beta) \qquad \textit{(Structural)}$$

> Overall, **hierarchical model**:

$$F_{Y|x}(y|x) = \int F_{Y|U,x}(y|U=u,x)\,dF_{U|x}(u|x)$$

Application: Post-traumatic Stress Disorder Ascertainment

- PTSD
  - — Follows a qualifying traumatic event
    - > ***This study***: ***personal assault***, ***other personal injury/trauma***, ***trauma to loved one***, ***sudden death of loved one*** = "x", along with sex
  - — Criterion endorsement of symptoms related to event ⇒ diagnosis
    - > Binary report on 17 symptoms = "Y"
- Study (Chilcoat & Breslau, *Arch Gen Psych*, 1998)
  - — Telephone interview in metropolitan Detroit
  - — n=1827 with a qualifying event
  - — Analytic issues
    - > Nosology
    - > Does diagnosis differ by trauma type or gender?
    - > *Are female assault victims particularly at risk?*

Latent Variable Models: What / How
Latent Class Regression (LCR) Model

- **Model**:

$$f_{Y|x}(y|x) = \sum_{j=1}^{J} P_j(x,\beta) \prod_{m=1}^{M} \pi_{mj}^{y_m}(1-\pi_{mj})^{1-y_m}$$

- **Structural model**: $[U_i|x_i] = Pr\{U_i=j|x_i\} = P_j(x_i,\beta)$
  — $RPR_j=Pr\{U_i = j|x_i\}/Pr\{U_i = J|x_i\}$; $j=1,\ldots,J$

- **Measurement assumptions** : $[Y_i|U_i]$
  — conditional independence
  — nondifferential measurement
  > ***reporting heterogeneity unrelated to measured, unmeasured characteristics***

- **Fitting**: ML w EM (Goodman, 1974) or Bayesian

- *Posterior* latent outcome information: $Pr\{U_i=j|Y_i,x_i;\theta=(\pi,\beta)\}$

Latent Variable Models: Philosophy

- **Why?**
  — to operationalize / test **theory**
  — to learn about **measurement problems**
  — they **summarize** multiple measures **parsimoniously**
  — to describe population **heterogeneity**

- **Why not?**
  — their **modeling assumptions** may determine scientific conclusions

  — their **interpretation** may be ambiguous
    > nature of latent variables?
    > what if very different models fit comparably?
    > seeing is believing

- **Import:** They are widely used

Vignette 1

Theory Operationalization and Testing

Latent Variable Modeling
Theory operationalization and testing

- **Meaning**
  - — measurement model definition and testing for fit
  - — construct definition and validation
  - — stating, testing implications of scientific hypotheses for latent-observed relationships

- Necessarily collaborative!

- Some collaborations
  - — dry eye syndrome (*with Munoz, Tielsch, West, Schein, **IOVS**, 1997*)
  - — geriatric frailty (*with Xue, Ferrucci, Walston, Guralnik, Chaves, Zeger, Fried, **J Gerontol**, 2006*)
  - — inflammation (*with Walston, Huang, Semba, Ferrucci, submitted*)

Vignette 2

Do data bear out theoretic predictions?

Latent Variable Modeling
Do data bear out theoretic predictions?

- **Commonly used methods for adjudicating fit**
  - — Global fit statistics (*many references*)
    - > thresholds sensitive to study design; black box
  - — Relative fit statistics (*Akaike, 1974; Schwarz, 1978; Lo et al., 2001*)
    - > they're relative
  - — Comparisons of observed and predicted frequencies, associations
    - > Cross-validation (*Cudeck & Browne 1983; Collins & Wugalter 1992*)
    - > Pearson / correlation residuals (*Hagenaars, 1988; Bollen, 1989*)
    - > Posterior predictive distributions (*Gelman et al, 1996*)
    - > Bayesian graphical displays (*Garrett & Zeger, 2000*)
    - > **whether** fit fails, not **how** fit fails

- Common wisdom: LV model assumptions are hard to check
  - — ... or **are** they?

Do data bear out theoretic predictions?

Part 1: Checking empirical reasonableness of the theory

- Rationale
  — If model correct and latent status known, measurement model "easy" to "explicate"

  — If persons can be partitioned into groups such that measurement model holds, model must correctly describe data distribution

- Research question: Suppose we estimate latent status.
  — Might the same idea work?

  — Seems circular?

  — Scientific intuition: Best shot = to randomize

Do data bear out theoretic predictions?

Part 1: Checking empirical reasonableness of the theory

1. FIT MODEL

2. ESTIMATE **posterior probabilities** Θ_i of membership from fit ("hats")

3. **RANDOMLY** ALLOCATE INDIVIDUALS INTO "PREDICTED," I.E. *"PSEUDO-"* CLASSES C_i ACCORDING TO Θ_{i1}, Θ_{i2}, ..., Θ_{iJ}

4. ASSESS ASSUMPTIONS WITHIN PREDICTED CLASSES
   - > $Y_{i1},\ldots,Y_{im}$ not highly associated
   - > Y_i, x_i not highly associated

Bandeen-Roche, Miglioretti, Zeger & Rathouz, 1997;
Huang & Bandeen-Roche, 2004; Wang, Brown & Bandeen-Roche, 2005

Checking the empirical reasonableness of theory

- Does the scheme work?
  — Hardest part: how to formulate what it means for scheme to work

- Notation
  — R_J: "Reasonable" class of LCR models; $\{\pi, ß\} = \phi \in \Phi$

- Formal statement of diagnostic premise: define

$$Z_{i\phi} = \prod_{m=1}^{M} \pi_{mj}^{y_m}(1-\pi_{mj})^{1-y_m} \text{ with prob. } P(x,\beta),\ j=1,\ldots,J$$

— Then (Theorem)

$$\boxed{Pr\{Y_i = y \mid C_i, x_i\}} \xrightarrow{D} Z_{i\phi} \text{ for some } \phi$$

if and only if $f_{Y_i}(y) = f_Y(y) \in R_J$ for each i

Do data bear out theoretic predictions?

Part 2: If not, what can we say about what the model is estimating?

- Under "regularity" assumptions:

  > The distribution of Y can be written as a hierarchical model, except

  $[Y|U^*,x]$, $[U^*|x]$ arbitrary (& specifiable in terms of π^*,β^*)

  > In the long run: No bias in substituting C_i for U_i^*.

  i.e. ***underlying variable distribution has an estimable interpretation even if assumptions are violated***

  and

  regression of C_i on x_i and model-based counterparts eventually equivalent

Model characterization if theory is mistaken
More formal statement

- Under Huber (1967)-like conditions:

—$(\hat{\beta}, \hat{\pi})$ converge in probability to limits (β^*, π^*).

—Y_i asymptotically equivalent in distribution to Y^*, generated as:

i) Generate U_i^* — distribution determined by (β^*, π^*), $G_{Y|x}(y|x)$;

ii) Generate Y^*—distribution determined by (β^*, π^*), $G_{Y|x}(y|x)$, U_i^*

— $\{\Pr[Y_i \leq y | C_i, x_i], i=1,2,...\}$ converges in distribution to $\{\Pr[Y_i^* \leq y | U_i^*, x_i], i=1,2,...\}$, for each supported y.

— C_i converges in distribution to U_i^* given x_i.

Vignette 3

Translation from latent to observed measures

Translation from latent to observed measures

- Goal: Create "scales" for broad analytic use
- Why?

  — Concreteness

  — Seeing is believing

  — Convenience
- What is lacking with existing methods for scale creation?

  — Most yield analyses that differ considerably from LV counterparts
- Target of the current work: Latent class applications

Regression with Latent Variable Scales [what analysis?]
A Staged Approach

- **<u>Step 1</u>**: Fit latent variable measurement model to Y $\Rightarrow \hat{\pi}$
  - For now: Non-differential measurement
- **<u>Step 2</u>:** Obtain predictions O_i given $\hat{\pi}$, Y_i
- **<u>Step 3</u>**: Obtain $\hat{\beta}$ via regression of O_i on x_i
- **<u>Step 4 (rare)</u>**: Fix inferences to account for uncertainty in $\hat{\pi}$

Latent Variable Scale Creation (obtaining O_i)
What do we know?

- **Predominant work:** Latent factor models; linear regression of U on X
  - — $Y = \pi U + \epsilon$; U, $\epsilon \sim$ Normal; ϵ has mean **0** and variance Σ
  - — **Three scaling methods**
    - > **Ad hoc**
    - > **Posterior mean**: O_i as $E[U_i | O_i, \hat{\pi}]$
    - > **"Bartlett" method**: O_i as WLS model fit for "fixed" U_i in

$$Y_i = \hat{\pi} U_i + \epsilon_i, \quad \epsilon_i \sim N(0, \hat{\Sigma});$$

  - — **In Step 3, Bartlett scores yield consistent** $\hat{\beta}$; others don't

Latent Variable Scale Creation (obtaining O_i)
What do we know?

- **Latent class models**
  - **Two methods**

    > **Posterior class assignment**

    - Modal or as "pseudo-class": single or multiple

    > **Posterior probability estimates**:

$h_i = f_{U|Y}(u|Y;\hat{\pi})$; $O_i = h_i$, or logit(h_i), or weighted indicators

— **In Step 3, all are inconsistent for $\hat{\beta}$**

— A **correction**: Croon, *Lat Var & Lat Struct Mod*, 2002
Bolck et al., *Political Analysis*, 2004

Latent Variable Scale Creation (obtaining O_i)
A new proposal

- **Motivation**: Bartlett method
  - — Latent class: $E[Y|U] = \pi S(U)$, where
    - > π: conditional probabilities ("covariates"; design matrix)
    - > $S(U)$: Jx1 with jth element = $\mathbf{1}\{U=j\}$ ("coefficients")
  - — Proposed **Step 2**: Linear regression of Y_i on $\hat{\pi}$, but with Bernoulli family; $O_i = \hat{S}_i$
  - — **A shortcut**: $O_i = \hat{S}_i$ via **ordinary** least squares; COP score
- Proposed **Step 3**: Generalized logit regression of O on x, Normal family

COP Scoring
Does it work in theory?

- **<u>Punch line</u>:** **In Step 3, COP scores yield consistent $\hat{\beta}$** provided data distribution identifiable LCR with non-differential measurement
- Basic ideas

— If π were known: OLS yields unbiased estimator of $\begin{pmatrix} Pr\{U_i = 1\} \\ \vdots \\ Pr\{U_i = J\} \end{pmatrix}$

$$> \begin{pmatrix} Pr\{U_i = 1\} \\ \vdots \\ Pr\{U_i = J\} \end{pmatrix} = \begin{pmatrix} P_1(x_i,\beta) \\ \vdots \\ P_J(x_i,\beta) \end{pmatrix}, \text{ all i, } \Rightarrow \hat{\beta}_{COP} \overset{p}{\rightarrow} \beta$$

— $\hat{\pi} \overset{p}{\rightarrow} \pi$ (marginalization, ML)