Improved Latent Variable-based Outcomes for Subsequent Regression Analysis

---

Karen Bandeen-Roche and Janne Petersen
Professor of Biostatistics and Medicine and Ph.D. Candidate
Johns Hopkins Bloomberg Sch. Pub. Health, Univ. of Copenhagen

Methods and Applications in Modern Statistics
Workshop to Celebrate David Ruppert
Keystone Resort, Colorado
June 1, 2008

ABSTRACT

Latent variable models have long been utilized by behavioral scientists to summarize constructs that are represented by multiple variables or are difficult to measure, such as health practices and psychiatric syndromes. They are regarded as particularly useful when measurable variables are highly imperfect surrogates for the construct of inferential interest; among numerous criticisms, they are criticized as being overly abstract and computationally intensive. We propose a new strategy for developing latent measurement model-based "indices" for subsequent use in regression modeling. Unlike most existing strategies, it yields approximately unbiased estimators for regression parameters vis a vis full latent variable regression. Small sample performance properties are evaluated. The methods are illustrated using data on vision and adverse functioning in older adults. It is hoped that, by counter-balancing strengths and weaknesses of latent variable modeling, the findings will improve the utility of latent variable-based approaches for scientific investigations.

Introduction: Statistical Problem

- Observed variables (i=1,...,n): Y_i=M-variate; x_i=P-variate

- Focus: response (Y) distribution = $G_{Y|x}(y|x)$; x-dependence

- Modeling issue: flexible or theory-based?
  - — Option 1 - Flexible: $g_m(E[Y_{im}|x_i]) = f_m(x_i)$, m=1,...,M

  - — Option 2 - Theory-based:
    - > Y_i generated from latent (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:

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

The particular latent variable model at issue for this work: Latent Class Regression (LCR) Model

- **Model**:

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

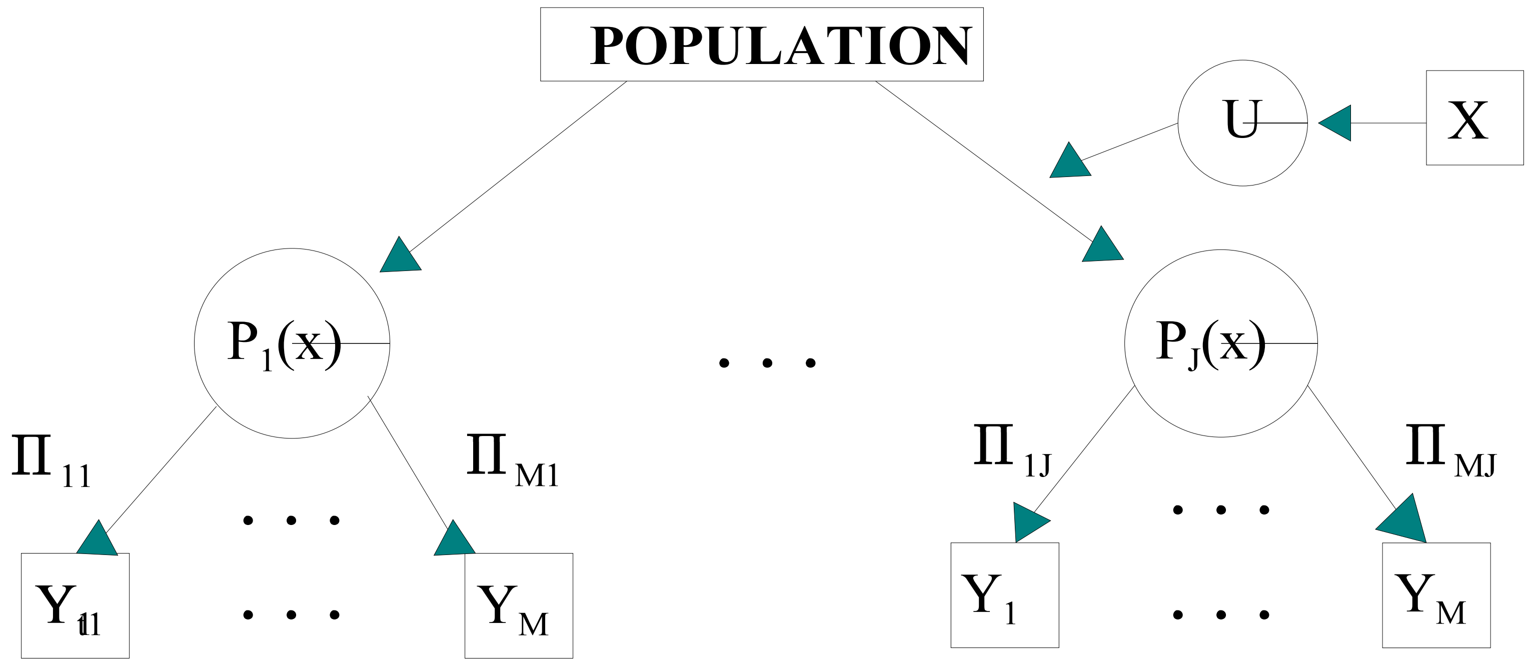
- **Structural model assumption** : $[U_i|x_i] = Pr\{U_i=j|x_i\} = P_j(x_i, \beta)$
  — Generalized logit link: $RPR_j=Pr\{U_i = j|x_i\}/Pr\{U_i = J|x_i\}$; $j=1,\ldots,J$

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

- **Fitting**: ML w EM; robust variance

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

Model 1
Latent Class Regression



> $P_j(x) = \Pr\{U = j|x\}$
> $\pi_{mj} = \Pr\{Y_m=1|U = j\}$

References: Dayton & Macready 1988, van der Heidjen et al., 1996; Bandeen-Roche et al., 1997

Latent Variable Scaling
A Three-Stage Approach

While behavioral health researchers favor latent variable modeling, they frequently aim to develop an index and then use it as an observed variable in subsequent regressions rather than fit a "full-blown" latent regression in a single step. That is:

- **Step 1**: Fit full latent variable measurement model $\Rightarrow \hat{\pi}$
  — For now: Non-differential measurement

- **Step 2:** Obtain predictions O_i given $\hat{\pi}$, Y_i

- **Step 3**: Obtain $\hat{\beta}$ via regression of O_i on x_i

- **Step 4 (rare)**: Fix inferences to account for uncertainty in $\hat{\pi}$

Latent Variable Scaling (obtaining O_i)
What do we know?

- **<u>Predominant work</u>:** Latent Factor models

  — $U \sim \text{Normal}$; $[Y|U] \sim \pi U + \epsilon$, $\epsilon \sim N(0,\Sigma)$

  — **Three scaling methods**

  > **Ad hoc**

  > **Posterior mean**: O_i as $E[U_i|O_i, \hat{\pi}]$

  > **"Bartlett" method**: Weighted least squares, U_i "fixed"

  $Y_i = \hat{\pi} U_i + \epsilon_i$, $\epsilon_i \sim N(0, \hat{\Sigma})$; O_i as WLS model fit for U_i

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

Latent Variable Scaling (obtaining O_i)
What do we know?

- **Latent Class models**
  — **Two scaling methods**

  > **Posterior class assignment**
  - Modal or as "pseudo-class": single or multiple

  > **Posterior probability estimates**:

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

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

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

Latent Variable Scaling (obtaining O_i)
A new proposal

- **Motivation**: Bartlett method
  - — [Y|U] ~ product Bernoulli, p = πS(U)
    - > Y, p: Mx1 vectors (**outcomes**)
    - > π: MxJ matrix of conditional probabilities (design matrix)
    - > S(U): Jx1 vector with jth element = **1**{U=j} ("coeffs")
  - — Proposed **Step 2**: GLM of Y_i on $\hat{\pi}$ with **linear** link, Bernoulli family; $O_i = \hat{S}_i$
  - — ML for GLM can be written as IRWLS
  - — **A shortcut**: $O_i = \hat{S}_i$ via **ordinary** least squares; COP score

COP Scoring
Theory

- Proposed **Step 3**: GLM of O on x with **gen. logit** link, Normal family

- **<u>Punch line</u>:** **In Step 3, COP scores yield consistent $\hat{\beta}$.**

- **Basic ideas of proof**

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

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

  — $\hat{\pi} \xrightarrow{p} \pi$ (marginalization, ML); then, uniform integrability

Simulation Study

- Basic template: 2 classes; $\pi = \begin{pmatrix} \tau & 1-\tau \\ & \vdots & \\ \tau & 1-\tau \end{pmatrix}$

  —**2 measurement scenarios**: "**Precise**"–τ=0.10; "**Imprecise**"–τ=0.30

- M=4, 8
- n=500, 1000
- 1 covariate; $\beta_0 = 0$; $\beta_1 = 0.5$
- Lots of secondary simulations to compare COP scores, full LV

Simulation Study
Results

Method	Precise, m=4, n=500			Imprecise, m=4, n=1000			Imprecise, m=8, n=1000		
	$E\hat{\beta}_1$	SE_{rat}	Cov	$E\hat{\beta}_1$	SE_{rat}	Cov	$E\hat{\beta}_1$	SE_{rat}	Cov
Modal class	0.48	1.00	0.95	0.30	0.96	0.68	0.37	1.03	0.83
Pseudo-class	0.47	0.98	0.95	0.24	0.97	0.50	0.33	1.03	0.76
Posterior-GLM	1.66	0.98	0.59	0.33	0.96	0.71	0.62	0.98	0.92
Croon corrected	0.51	NA	NA	0.49	NA	NA	0.47	NA	NA
COP score	0.51	0.97	0.95	0.51	0.98	0.96	0.49	1.00	0.94
LCR	0.51	0.99	0.95	0.52	0.98	0.96	0.49	1.02	0.95

- n=500 vs 1000, m=8: negligible difference
- Power = slightly highest for LCR; others = ~ comparable except pseudo
  — Relative efficiency re LCR: $\geq$ 0.89

Simulation Study
COP Score Performance in Secondary Runs

- Findings similar in many cases:
  - — 3 classes
  - — $\beta_0 \neq 0$, different β_1
  - — different measurement models
  - — continuous versus binary x

- Multiple (4) covariates

  - — Accuracy of mean model estimation maintained

  - — Accuracy of standard errors compromised

    > For moderate $|\beta_1|$: coverages ~ within 0.02 of 0.95

    > With large $|\beta_1|$: coverages as low as 0.83

Application

IADL Functioning in the Salisbury Eye Evaluation (SEE) Study

- **Study**: Salisbury Eye Evaluation (SEE; West et al. 1997)
  — Representative of community-dwelling elders
  — n=2520; 1/4 African American
  — This talk: A convenience sample of n=1329

- **Question of interest**: Is worse vision associated with worse IADL functioning independently of age (and sex)?

  — IADL (Y): Indicators of difficulty **shopping**, **preparing meals**, **doing light housework**, and **using the phone**

  — Vision (primary X): Visual acuity (logMAR)

Application Findings

- Two class model (questionable fit–apparent differential measurement by sex!)

Coefficient	Model 1		Model 2	
	LCR	COP	LCR	COP
Intercept	-3.17 (-3.61,-2.73)	-3.12 (-3.51,-2.73)	-2.91 (-3.44,-2.34)	-3.02 (-3.47,-2.57)
Vision	2.05 (1.33, 2.76)	2.15 (1.72, 2.59)	2.00 (1.21, 2.78)	2.11 (1.68, 2.55)
Age (yr)	0.75 (0.21, 1.29)	0.72 (0.28, 1.17)	0.72 (0.17, 1.26)	0.71 (0.27, 0.15)
Sex	NA	NA	-0.68 (-1.34,-0.03)	-0.17 (-0.63, 0.28)

— Re green estimates: many other methods closer to LCR

Discussion

- **Finding:** Proposal of a novel latent class "index"
  - — applicable in multi-stage analysis (index 1st then regression)
  - — yields consistent regression coefficient estimators (theory)
  - — achieves accurate small sample performance (simulation)

- **Gaps**
  - — Inference to account for uncertainty due to first stage
  - — Estimation target, correction if differential measurement

- **Contribution**
  - — First such index for latent class analysis
  - — More easily implemented that Croon correction
  - — More accurate/precise and clearly interpretable inferences than commonly practiced alternatives