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
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 \((i=1,...,n): \ Y_i=\text{M-variate}; \ x_i=\text{P-variate}\)

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

- 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 \text{latent (underlying)} \(U_i\):
      \[
      F_{Y|U,x}(y|U=u,x;\pi) \quad \text{(Measurement)}
      \]
    - Focus on distribution, regression re \(U_i\):
      \[
      F_{U|x}(u|x;\beta) \quad \text{(Structural)}
      \]
    - 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:**
  \[ f_{y|x}(y|x) = \sum_{j=1}^{J} P_j(x, \beta) \prod_{m=1}^{M} \pi_{jm}^{ym}(1-\pi_{jm})^{1-ym} \]

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

- **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?

**Predominant work:** 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 $E[U_i|O_i, \hat{\beta}]$

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

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

- **In Step 3, Bartlett scores yield consistent } \hat{\beta} \text{; 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:
      $$h_i = f_{U|Y}(u|Y; \theta); O_i = \logit(h_i)$$ or weighted
    - In Step 3, all are biased for $\hat{\beta}$

  Bolck et al., *Political Analysis*, 2004
Latent Variable Scaling (obtaining $O_i$)
A new proposal

- **Motivation**: Bartlett method
  - $[Y|U] \sim \text{product Bernoulli, } p = \pi S(U)$

  $> Y, p: M \times 1$ vectors (outcomes)

  $> \pi: M \times J$ matrix of conditional probabilities (design matrix)

  $> S(U): J \times 1$ vector with jth element $= 1\{U=j\}$ ("coeffs")

- Proposed **Step 2**: GLM of $Y_i$ on $\hat{p}$ 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

- **Punch line**: In Step 3, COP scores yield consistent \( \hat{\beta} \).

- **Basic ideas of proof**
  - **If \( \pi \) were known**: OLS yields unbiased estimator of
    \[
    \begin{pmatrix}
    Pr\{U_i = 1\} \\
    \vdots \\
    Pr\{U_i = J\}
    \end{pmatrix}
    = \begin{pmatrix}
    P_1(x_p \beta) \\
    \vdots \\
    P_J(x_p \beta)
    \end{pmatrix}, \text{ all } i, \Rightarrow \hat{\beta}_{COP} \xrightarrow{p} \beta
    \]
  - \( \hat{\pi} \xrightarrow{p} \pi \) (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” – $\tau = 0.10$; “Imprecise” – $\tau = 0.30$

- $M = 4, 8$

- $n = 500, 1000$

- 1 covariate; $\beta_0 = 0$; $\beta_1 = 0.5$

- Lots of secondary simulations to compare COP scores, full LV
## Simulation Study Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Precise, m=4, n=500</th>
<th>Imprecise, m=4, n=1000</th>
<th>Imprecise, m=8, n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E $\hat{\beta}_1$</td>
<td>SE$_{rat}$</td>
<td>Cov</td>
</tr>
<tr>
<td>Modal class</td>
<td>0.48</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>Pseudo-class</td>
<td>0.47</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>Posterior-GLM</td>
<td>1.66</td>
<td>0.98</td>
<td>0.59</td>
</tr>
<tr>
<td>Croon corrected</td>
<td>0.51</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>COP score</td>
<td>0.51</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>LCR</td>
<td>0.51</td>
<td>0.99</td>
<td>0.95</td>
</tr>
</tbody>
</table>

- $n=500$ vs $1000$, $m=8$: negligible difference

- Power = slightly highest for LCR; others $\sim$ 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 $\beta_1$
  — 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!)

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LCR</td>
<td>COP</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.17 (-3.61,-2.73)</td>
<td>-3.12 (-3.51,-2.73)</td>
</tr>
<tr>
<td>Vision</td>
<td>2.05 (1.33, 2.76)</td>
<td>2.15 (1.72, 2.59)</td>
</tr>
<tr>
<td>Age (yr)</td>
<td>0.75 (0.21, 1.29)</td>
<td>0.72 (0.28, 1.17)</td>
</tr>
<tr>
<td>Sex</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

— 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 than Croon correction
  — More accurate/precise and clearly interpretable inferences
    than commonly practiced alternatives