Analysis of Multiple Discrete Surrogates

- methods review, especially latent class regression

Jan. 27, 2006

SLAM Working Group

Yi Huang

Department of Biostatistics,

Johns Hopkins Bloomberg School of Public Health

Scientific Questions

- Data arising as multiple correlated discrete variables are common in bio-medical applications. e.g.
  - Multiple questions are asked to uncover latent depression
  - Multiple indicators for evaluating underlying functioning
- How to assess the latent variable through those multiple correlated indicators?
- How to quantify the association between the latent variable and risk factors?
- How to quantify the direct effect from covariates to outcome after adjusting for this underlying latent variable?

Example: multiple discrete Outcome

Observed Pattern frequency

Y pattern	frequency	percent	Y pattern	frequency	percent
000000	239	5.26	**101011**	**662**	**14.57**
000001	300	6.60	110001	112	2.46
001011	109	2.40	110011	152	3.34
100001	321	7.06	111001	148	3.26
100011	259	5.70	**111011**	446	9.81
101001	335	7.37	**111111**	**78**	**1.72**

Outline

- Review typical methods to deal with multiple correlated discrete outcomes.
  - Summarize then Analyze (STA)
  - Analyze then Summarize (ATS)
  - Summarize and Analyze (SAA)

- Introduce Huang and Bandeen-Roche's latent class regression model (2000).
  - Review model contruction: LCA → LCR-1 → LCR-2
  - Identifiability, estimation, diagnosis
  - Selection: the number of latent classes

Notation

- Multiple correlated discrete outcomes:
  - $Y_1, Y_2, \ldots Y_K$ (k=1, … K)
  - To ease the presentation, let Y to be binary.
  - Ordinal case is presented in Huang's thesis.
- Covariates:
  - $X_1, X_2, \ldots X_p$ (p=1, … P) – primary confounders and risk factors, related to latent variable
  - $Z_1, Z_2, \ldots Z_K$ (k=1, … K) – covariates matrix related to measured indicators $Y_1 \ldots Y_K$.

$$\begin{bmatrix} 1, \quad 1, \ldots\ldots \ 1 \\ Z_{11}, Z_{12}, \ldots Z_{1M} \\ \ldots\ldots\ldots\ \ldots \\ Z_{L1}, Z_{L2}, \ldots Z_{LM} \end{bmatrix}$$

  - Two sets of X and Z: mutually exclusive, or overlap.

Summarize then Analyze (STA)

- Scoring Analysis
  - Easiest and most commonly used in questionnaire responses
- Summary score = $\sum_{k=1}^{K} \omega_k Y_k$
  - Equal weights: Summation or average item ratings
  - Unequal weights: principle component analysis, and others
- Analysis:
  - Regression: summary score ~ covariates.
- Cons:
  - **Differential associations** between X and Ys are **masked**.
  - Ignore potential direct confounding effect from item-specific Z on item-specific outcome.

Analyze then Summarize (ATS)

- Estimation function methods
  - Godambe (1960), Durbin (1960), Wedderburn(1974)
  - Liang and Zeger (1986) extended quasi-score function from univariate responses to multivariate correlated responses.
- GEE1 - β : parameter of interest, α : nuisance.
  - Separate estimating function for **mean parameter β** and **association parameter α** in covariance matrix.
  - $\hat{\beta}$ is consistent, even specify covariance matrix wrongly.
- GEE2 – β and α : parameter of interest.
  - Joint model and estimate β and α
  - GEE2 is more efficient than GEE1.
  - $\hat{\beta}$ is **NOT** consistent if covariance matrix is wrongly specified.
- More appropriate for nicely defined correlated data.

Continue: ATS

- Random Effect Model (Laird and Ware, 1982, 1984)
  - Aim for **individual based** model inference and result interpretation.
  - Assumption: correlation among multiple responses arise from natural heterogeneity across people
  - Heterogeneity – subject-specific regression coefficients, e.g. α_i, and $\boldsymbol{\alpha_i \sim N(a, \sigma^2)}$
- Marginal Model (Heagerty and Zeger, 1996)
  - Aim for **population average based** model inference and result interpretation.
  - Outcome logistic regression (for each item) ~ $X\beta$,
  - **Joint** model: log[odds ratio matrix] ~ $Z\alpha$
  - For ordinal items: Proportional odds model and log[Global odds ratio matrix]

SAA - Latent Variable

- an unobservable variable.
- Realist - existing variable, measure indirectly from manifest variables $\mathbf{Y}$.
- Instrumentalist - a summary construct (η) which can explain all the association among item responses.

$$Pr(\mathbf{Y}_i = \mathbf{y}_i | \eta_i = j) = \Pi_{k=1}^{K} Pr(Y_{ik} = y_{ik} | \eta_i = j) \tag{1}$$

- η_i: latent class-membership
- Index: i – individual;

  j – class membership;

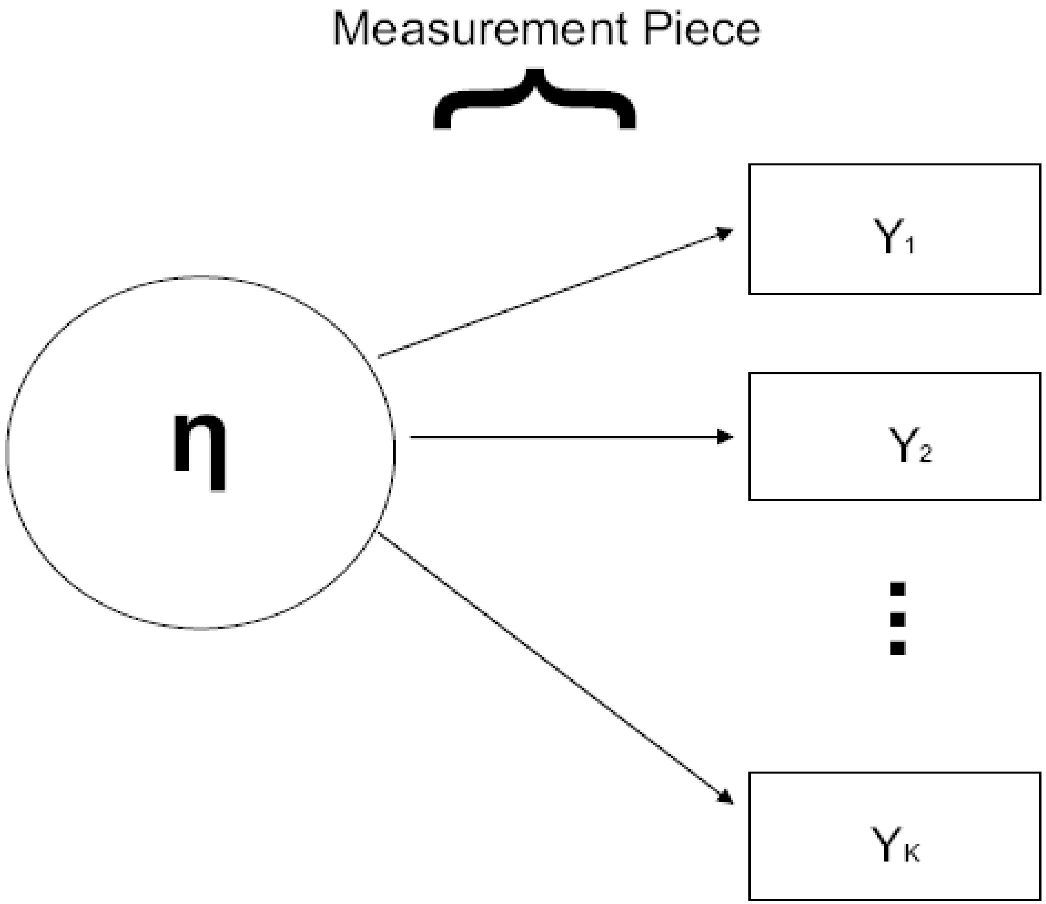
  k – item reponses.

Latent Class Model (LCA)

Pros of LCA vs.

Latent Trait Model (LTM)

1. Y are all binary, latent classes are interpretable as summary of outcome patterns
2. No requirement of the known distribution of the latent variable.



Applications

LTM: determine univariate scales of ability

LCA: cluster analysis tool to find homogeneous groups

Continue - LCA

- Key idea - realist
  - target population = comprised of finite sub-populations
  - responses = imperfect indicator of the subpopulation to which a subject belongs.

LCA could be considered as a tool to help cluster people into depression groups according to their item response patterns.

- Key idea - Instrumentalist
  - Describing associations among the 6 binary responses.
  - Describing the patterns in which multiple positive responses co-occur.

LCA: Assumptions

1. Internal homogenous

$$Pr(Y_{ik} = y|\eta_i = j) = P_{kj}(y)$$

2. Item responses' conditional independency | pressure class:

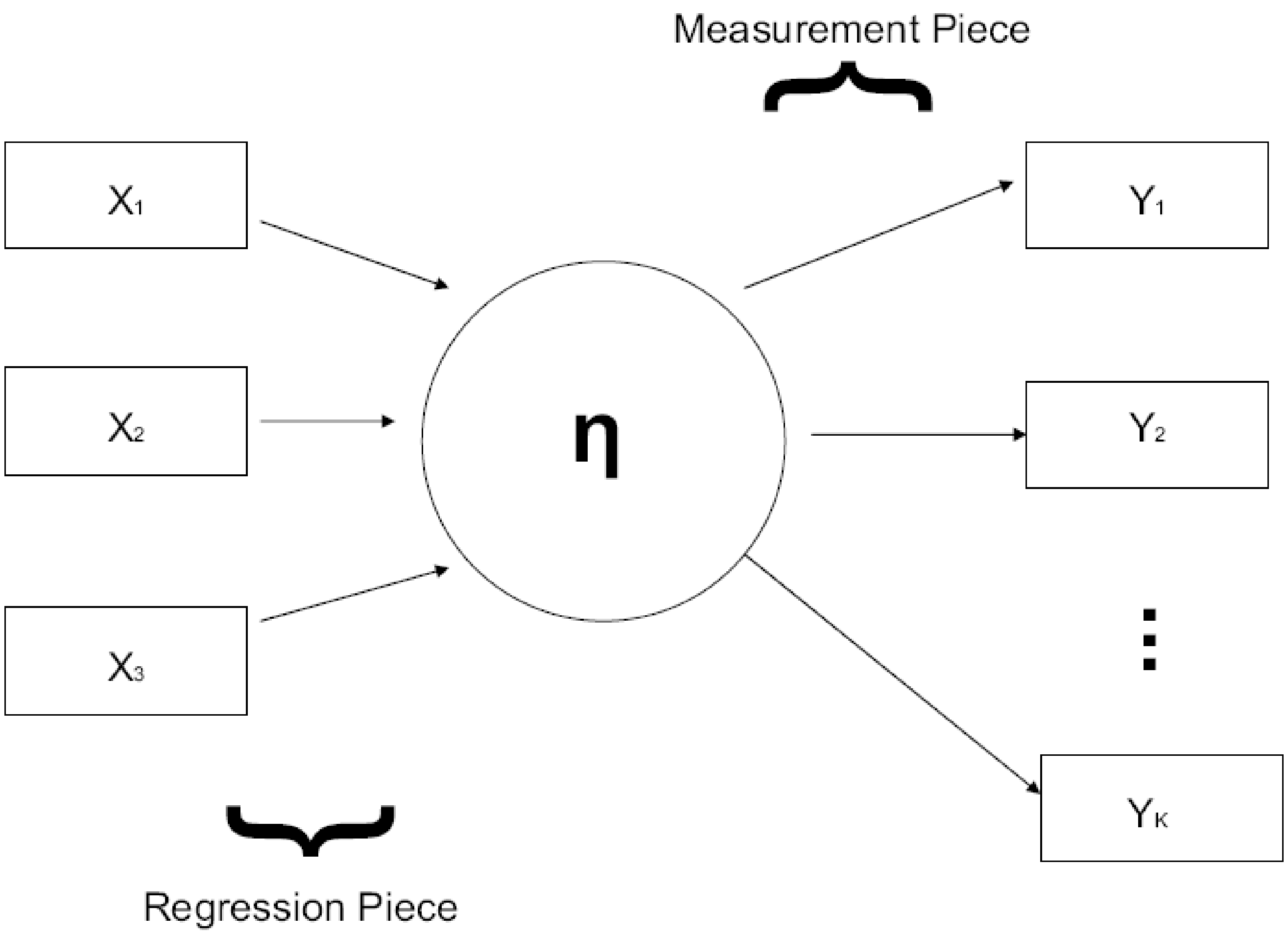
$$Pr(\mathbf{Y}_i = \mathbf{y}_i|\eta_i = j) = \Pi_{k=1}^{K} Pr(Y_{ik} = y_{ik}|\eta_i = j)$$

LCA: Model & Parameters

- **K-latent-class LCA** with binary responses:

$$
\begin{aligned}
Pr(\mathbf{Y}_i = \mathbf{y}_i) \quad &= \quad \textstyle\sum_{j=1}^{J} \pi_j \prod_{k=1}^{K} P_{kj}^{y_{ik}} (1 - P_{kj})^{1-y_{ik}}, \\
\text{with,} \quad & \quad \textstyle\sum_{j=1}^{J} \pi_j = 1 \\
\text{Parameters:} \quad P_{kj}, \quad & k = 1, \ldots, K; j = 1, \ldots, J. \\
\pi_j \quad & = Pr(\eta_i = j), j = 1, \ldots, J-1.
\end{aligned}
$$

Latent Class Regression (LCR -1)



LCR-1: Assumptions

1. Item responses are **conditionally independent** given class membership.

2. **Internal homogeneity**

3. **Non-differential measurement condition** (often, for LCR)
   - the effect of covariates on responses is totally mediated by latent class membership.

$$Pr(Y_{ik} = y|\eta_i = j, \mathbf{X}_i) = Pr(Y_{ik}|\eta_i = j)$$

$$i = 1, \ldots, n; k = 1, \ldots, K; j = 1, \ldots, J$$

LCR-1: Model & Parameters

- **J-latent-class regression model** (LCR) with binary responses:

$$
\begin{aligned}
Pr(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i) &= \textstyle\sum_{j=1}^{J} \pi_j(\beta_j^T x_i) \prod_{k=1}^{K} P_{kj}^{y_{ik}} (1 - P_{kj})^{1-y_{ik}}, \\
\pi_j(\beta_j^T x_i) &= Pr[\eta_i = j|\mathbf{X}], \text{ and } \textstyle\sum_{j=1}^{J} \pi_j(\beta_j^T x_i) = 1
\end{aligned}
$$

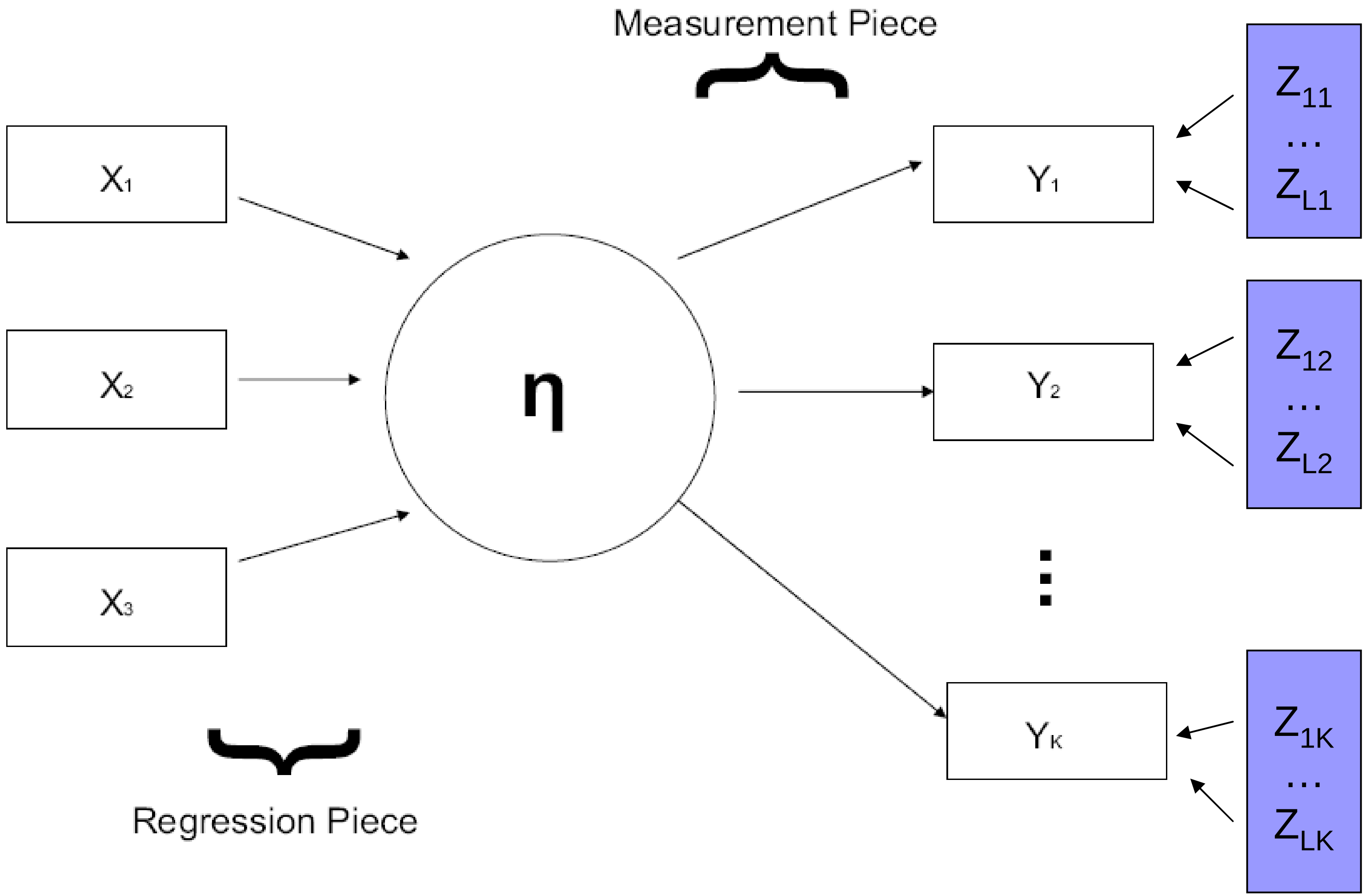
Parameters: $P_{kj} = Pr(Y_{ik} = 1|\eta_i = j), k = 1, \ldots, K.$

$\beta_{jp}, \quad j = 1, \ldots, J-1; p = 0, \ldots, P$

- Ignoring the effect of $\mathbf{X}$s - J-latent-class analysis. (π_j, P_{kj})

(Dayton, Macready 1988; Bandeen-Roche, 1999)

Huang & Bandeen Roche: LCR -2



LCR-2: Model Assumptions

- Relax: internal homogeneity & nondifferential measurement, to

  Conditioning on class membership, responses are only associated with $\mathbf{z}_i$

$$
\begin{aligned}
Pr(Y_{i1} = y_{i1}, \ldots, Y_{iK} = y_{iK} | \eta_i = j, X_i, Z_i) &= \\
Pr(Y_{i1} = y_{i1}, \ldots, Y_{iK} = y_{iK} | \eta_i = j, Z_i)
\end{aligned}
$$

- Class membership probabilities are associated with $\mathbf{x}_i$ only:

$$
Pr(\eta_i = j | X_i, Z_i) = Pr(\eta_i = j | X_i)
$$

- Conditional independency

$$
\begin{aligned}
Pr(Y_{i1} = y_{i1}, \ldots, Y_{iK} = y_{iK} | \eta_i = j, Z_i) &= \\
\Pi_{k=1}^{K} Pr(Y_{ik} = y_{ik} | \eta_i = j, Z_{ik})
\end{aligned}
$$

LCR-2: Model & Parameters

- J - latent classes:

$$
\begin{aligned}
Pr(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i, Z_i) &= \sum_{j=1}^{J} \pi_j(X_i \beta_j) \prod_{k=1}^{K} P_{ikj}^{y_{ik}} (1 - P_{ikj})^{1-y_{ik}}, \\
\pi_j(X_i \beta_j) &= Pr[\eta_i = j | \mathbf{X}], \text{ and } \sum_{j=1}^{J} \pi_j(X_i \beta_j) = 1 \\
P_{ikj} &= Pr(Y_{ik} = 1 | \eta_i = j, Z_{ik}) = logit^{-1}(\gamma_{kj} + Z_{ik} \alpha_k)
\end{aligned}
$$

- Parameters:

$$
\begin{aligned}
&\gamma_{kj}, \quad j = 1, \ldots, J-1; k = 1, \ldots, K \\
&\beta_{jp}, \quad j = 1, \ldots, J-1; p = 0, \ldots, P \\
&\alpha_{lk}, \quad l = 1, \ldots, L; k = 1, \ldots, K.
\end{aligned}
$$

LCR-2: Identifiability

- **Globally identifiable** $\hookrightarrow$

  $\forall \theta, \theta' \in \Theta : f(y|\theta) = f(y|\theta'), \forall y \in Y \iff \theta = \theta'$

  Y: the distribution support.

- **Locally identifiable** at θ_0 $\hookrightarrow$

  $\exists \aleph$, such that $\forall \theta' \in \aleph : f(y|\theta_0) = f(y|\theta'), \forall y \in Y \iff \theta_0 = \theta'$.

  $\aleph$ is a open neighborhood of θ_0.

- Methods to demonstrate local identifiability - developed by Goodman, and Bandeen-Roche et al.

(Goodman, 1974; Bollen, 1989; Bandeen-Roche, 1997; Casella,2002;)

LCR-2: Estimation

- **Maximum Likelihood Approach** - Often use EM algorithm
  - Standard error estimation:(Matrix of observed Fisher Information)$^{-1}$.
  - No assumption on prior.
- **Bayesian Approach** - use MCMC algorithm.
  - Display the posterior distribution for parameters.
  - CI: posterior interval.
- **Cons for both**: computationally intensive; local maximum problem; depend on the fully specified likelihood function.

LCR -2 : Diagnosis

- Failure of Person- χ^2 test & likelihood ratio goodness-of-fit test.
  - P_{ikj}= $Pr(Y_{ik} = 1 | \eta_i = j, Z_{ik})$ **Not** homogenuous in j^{th} class, so could not use Poisson approximation.
  - Parameter space(a smaller model) - as a special case where a subset of parameters of a larger model is set to the boundary of their parameter space.
- Deviance residual is developed to evaluate fit.
- Similar to Bandeen-Roche (1997), propose a pseudo-class membership procedure to check model assumptions specifically.

Selection: number of classes

- **AIC**: popular, but favor bigger models, (theoretically) not consistent, need refit models.
- **BIC**: popular, sometime favor smaller models, (theoretically) consistent, need refit models.
- Connection – factor analysis & LCA
  - Number of factors and number of latent classes – the number of dimensions needed to characterize the systematic part of the response distribution.
- **Thm 5.1** – construct sample **pseudo-residual correlation matrix**, then follow principle component analysis to choose the number of latent classes to began with.

Thanks !

Your comments and questions

are welcome after Brian’s follow-up discussion.