Some Remarks on Latent Variable Models in Categorical Data Analysis

Alan Agresti

Professor Emeritus

Department of Statistics

University of Florida

USA

Outline

- Selective *historical overview* of some important and/or interesting contributions to the literature about latent variable analysis of multivariate categorical responses

- Summary of main ideas, suppressing details

- Caveat: Enormous literature, so survey is highly selective. Examples selected shaped by my revision work on *Categorical Data Analysis*, 3rd edition (2013).

- Observed variables categorical, but latent variable can be discrete (i.e., latent "classes") or continuous.

Paul Lazarsfeld (1901-1976): Latent structure analysis

Lazarsfeld (1950) introduces *latent class model*, treating a contingency table as a finite mixture of unobserved tables generated under a conditional independence structure.

For categorical $(Y_1, Y_2, \ldots, Y_T)$, model assumes latent categorical Z such that for each possible sequence $(y_1, \ldots, y_T)$ and each category z of Z,

$$\begin{aligned} &P(Y_1 = y_1, \ldots, Y_T = y_T \mid Z = z) \\ &= P(Y_1 = y_1 \mid Z = z) \cdots P(Y_T = y_T \mid Z = z). \end{aligned}$$

Model receives more attention after *Latent Structure Analysis* text written by Lazarsfeld and Henry (1968).

Latent class models: “Local independence”

L & H, p. 22: “The defining characteristic of the latent structure models is the *axiom of local independence*.”

> Within a latent class, responses to different items are independent.

Model fitting? “Accounting equations” (apparently suggested by Mosteller) equate relative frequencies to corresponding marginal probabilities of various orders. Iterative solution using “determinantal method” approximates minimum chi-squared (BAN) estimates.

Anderson (1954) shows asymptotic properties, also in L & H.

p. 13: When $\{Y_t\}$ have $>$ 2 categories, the model has so many restrictions that its practical application seems doubtful.

Neil Henry: e-mail May 15, 2012

> “While Lazarsfeld was many things, he was not a statistician. The people he had working on LSA with him were sociology students with mathematical abilities, but no interest in inferential statistics. ... The papers, mostly unpublished, that I inherited in 1960 were full of these accounting equation solution techniques. Eventually I learned enough history to realize that he had adopted Karl Pearson’s ‘method of moments’ technique of estimation.
>
> MLE was impossible (as a practical estimation technique) in the 40s and 50s, of course.”

Leo Goodman: EM algorithm for ML fitting

Goodman (1974) shows how to fit basic latent class model (discrete latent variable) using maximum likelihood (ML)

- Uses EM algorithm, treating data on Z as missing.
- The E (expectation) step in each iteration calculates pseudo-counts for the unobserved table using working conditional distribution for $(Z \mid Y_1, \ldots, Y_T)$.
- The M (maximization) step treats pseudo counts as data and maximizes the pseudo-likelihood, by fitting the loglinear model of conditional independence (given the latent classes) symbolized by marginal sufficient statistics $(Y_1 Z,\ Y_2 Z, \ldots,\ Y_T Z)$.
- This is an early application of EM, three years before the classic *JRSS-B* paper by Dempster, Laird, and Rubin.

The EM algorithm for ML fitting (continued)

- EM algorithm is computationally simple and stable, and each iteration increases the likelihood.
- However, convergence can be very slow.
- Later methods also include use of Newton-Raphson algorithm (e.g., Haberman 1988 *Sociol. Methodology*).
- Some software (e.g., Latent GOLD) uses EM at first, then switches to Newton-Raphson to speed convergence.
- Problematic issues: log likelihood can have local maxima (especially as number of latent classes increases), and with complex models, identifiability is an issue.

Extension: Using Goodman association models

- Goodman (1979) proposes models that provide structured associations between ordinal variables.
- e.g., *uniform association model*: For expected frequencies $\{\mu_{ij}\}$ in two-way contingency table,

  $$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$$

  Equally-spaced scores (e.g., $\{u_i = i\}$ and $\{v_j = j\}$) imply common value of *local odds ratios*

  $(\mu_{ij}\mu_{i+1,j+1})/(\mu_{i,j+1}\mu_{i+1,j}) = \exp(\beta)$

  for pairs of adjacent rows and adjacent columns.
- Goodman (1981) and others showed such association models and their multivariate generalizations fit well when there is underlying multivariate normal distribution.

Example: Association models with latent classes

With ordinal data, seems natural to have ordered latent classes also, such as by assuming uniform association between them and each response.

Agresti and Lang (1993) modeled agreement among many raters evaluating carcinoma with exchangeble model having same β between each ordinal variable and the latent variable.

Model parameters describe two components of agreement:

- Strength of association between classifications by pairs of raters (governed by size of β)
- Heterogeneity among observers' rating distributions

Solution with 3 latent classes may reflect
Class 1: rater agreement that carcinoma = yes
Class 2: strong disagreement (some raters yes, some no)
Class 3: rater agreement that carcinoma = no

Rasch model: Continuous latent variables

- Rasch model is an *item response model* for a binary response. For subject i with item t

  $$\text{logit}[P(Y_{it} = 1 \mid u_i)] = u_i + \beta_t,$$

  with u_i a latent "ability" measure for subject i.
- Rasch (1961): $\{u_i\}$ fixed, eliminated using conditional ML.
- Tjur (1982): Averaging over u_i in nonparametric manner, Rasch model for T items implies *quasi-symmetry* (QS) loglinear model for observed 2^T table. $\{\hat{\beta}_t\}$ for QS identical to conditional ML estimates for Rasch model.
- Much literature over the years by Hatzinger, Dittrich and colleagues about IRT – loglinear connections.
- Analogs for ordinal IRT models and corresponding ordinal QS loglinear models (e.g., Agresti 1993).

Rasch mixture model for latent abilities

- In Rasch model logit$[P(Y_{it} = 1 | u_i)] = u_i + \beta_t$, often $\{u_i\}$ is now treated parametrically, e.g. N(0, σ^2).
- Lindsay, Clogg, and Grego (1991) instead treat u_i nonparametrically, with finite number q of values,

  $$P(U = a_k) = \rho_k, \quad k = 1, \ldots, q,$$

  for unknown q, $\{a_k\}$ and $\{\rho_k\}$ (*Rasch mixture model*).
- Likelihood increases in q, but reaches maximum when $q = (T+1)/2$.

Other latent variable extensions

- Bartholomew constructed general latent variable models for categorical responses, in analogy with factor analysis.
  “Factor analysis for categorical data”
  (*J. Roy. Statist. Soc. B*, 1980)
  “Latent variable models for ordered categorical data”
  (*J. Econometrics*, 1983).

- These and Christoffersen (1975), Muthén (1977) induce extensive “latent trait” literature in psychometrics that grows together with item response literature in educational statistics.

- For recent survey, see Bartholomew, Knott, and Moustaki, *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. (2011).

Latent mixture for summarizing goodness of fit

- For a model, with sufficiently large n, traditional goodness-of-fit statistics (e.g., chi-squared) reject a model even if in practical terms the lack of fit is unimportant.
- Rudas, Clogg, and Lindsay (1994): For a model for a contingency table with true probabilities $\boldsymbol{\pi}$, express

  $$\boldsymbol{\pi} = (1-\rho)\boldsymbol{\pi}_1 + \rho\boldsymbol{\pi}_2,$$

  with $\boldsymbol{\pi}_1$ the model-based probabilities and $\boldsymbol{\pi}_2$ unconstrained.
- Index of lack of fit is the smallest such ρ possible for which this holds (i.e., fraction of population that cannot be described by the model).
- Recognizes George Box's quote that "All models are wrong, but some are useful." Useful means ρ close to 0.

Latent mixing of logistic regression

Follman and Lambert (1989) analyzed effect of dosage of poison on probability of death of protozoan of a particular genus

Poison			Poison		
Dose	Exposed	Dead	Dose	Exposed	Dead
4.7	55	0	5.1	53	22
4.8	49	8	5.2	53	37
4.9	60	18	5.3	51	47
5.0	55	18	5.4	50	50

Latent mixing of logistic regression (continued)

- For $\pi_i(x)$ = probability of death at log dose level x for genus type i, $i = 1, 2$, and ρ = probability a protozoan belongs to genus type 1 (unknown),

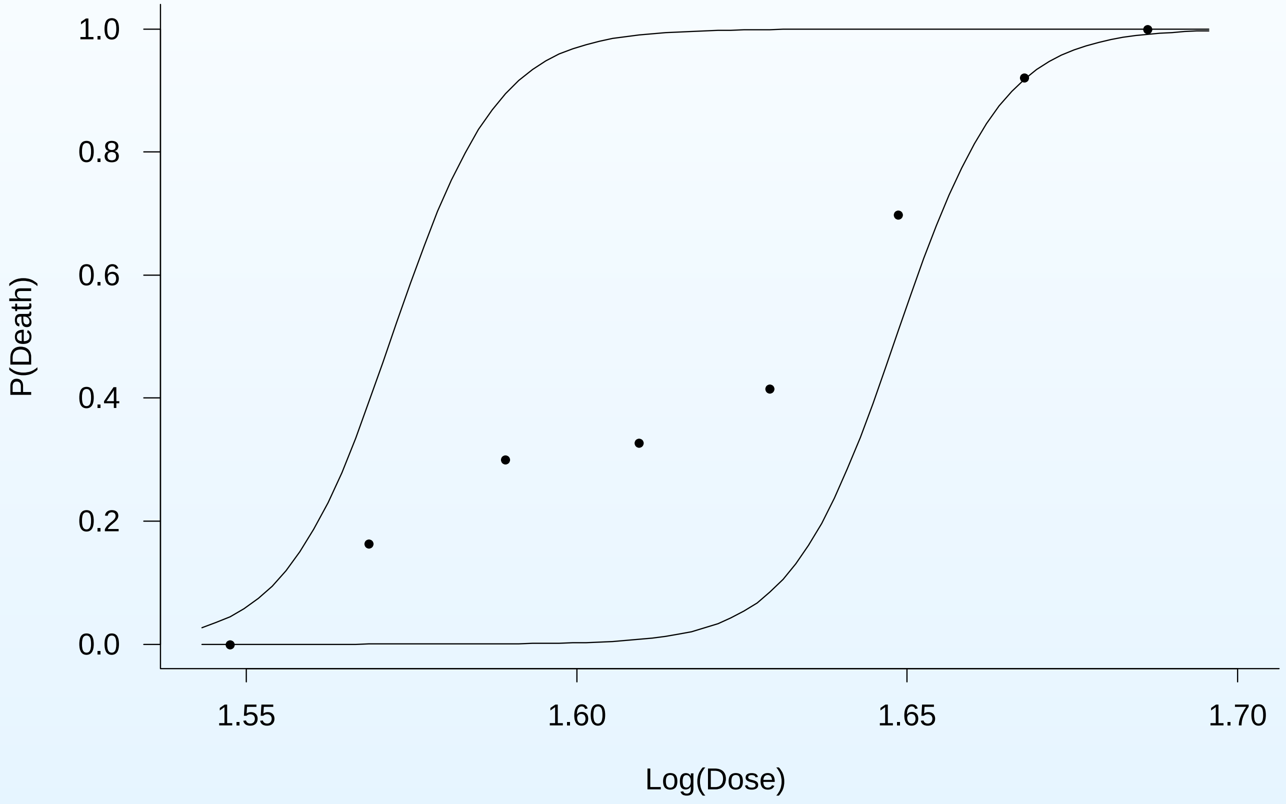
$$\pi(x) = \rho\pi_1(x) + (1-\rho)\pi_2(x), \quad \text{where} \quad \text{logit}[\pi_i(x)] = \alpha_i + \beta x.$$

- Curve for $\pi(x)$ is weighted average of two curves having same logistic shapes but different intercepts.

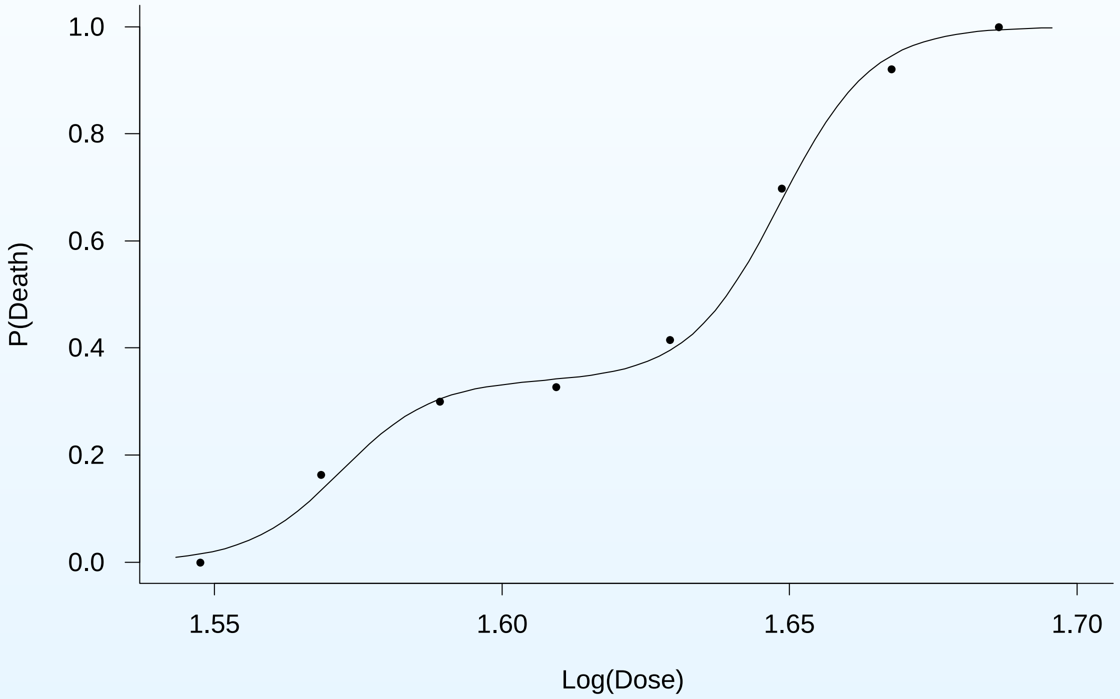
Deviance decreases by 21.3 (df = 2) compared to single logistic regression curve.

(Decreases by only 1.7 when take normal mixture.)

The two separate logistic regression curves



Mixture of two logistic regression curves



Count data latent variable model

- Lambert (1992) also proposed *zero-inflated Poisson* (ZIP) regression model, for applications in which some observations must be zero and others are zero just by chance.

  (e.g., number of times went to gym in last week, frequency of sexual intercourse in past month, number of papers professors publish in a year)

- For overdispersed data, alternative to ZIP model is *zero-inflated negative binomial* (Greene, 1994).

- Alternative *hurdle model* (Mullahy 1986) has logistic model for (zero, positive) outcome and truncated count-data model for positive counts.

Generalized linear mixed models (GLMM)

- Pierce and Sands (1975) proposed logistic regression with a random intercept, assumed to be normally distributed, in never published Oregon State technical report on “Extra-Bernoulli variation in regression of binary data.”
- They used Gauss-Hermite quadrature, still a practical solution for generalized linear mixed models with simple random effects structure.
- Breslow and Clayton (1993) developed *penalized quasi-likelihood* (PQL) as simple alternative to Gauss-Hermite quadrature for more complex random effects structure.
- PQL can be highly biased for categorical response with large variance component (Lin 1997).

Questions in fitting GLMMs

- Still a current research topic, as models are proposed with more complex random effects structure.
  (e.g., multilevel models)

- Zipunnikov and Booth (2013) suggest that higher-order Laplace approximations work better in practice than some methods that, in theory, produce ML
  (such as Monte Carlo EM).

- Bayes approach (e.g., with MCMC) is also used to approximate ML, but how does this work in model having large number of parameters and/or large number of random effects?

  Hot topic: How to choose “objective prior” in high-dimensional problems?

Distribution assumed for random effects important?

- Let y_{it} denote observation t in cluster i, $t = 1, \ldots, T_i$, with random effects $\boldsymbol{u}_i$ for cluster i. For $\mu_{it} = E(Y_{it}|\boldsymbol{u}_i)$, GLMM has form

  $$g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \boldsymbol{z}_{it}^T \boldsymbol{u}_i$$

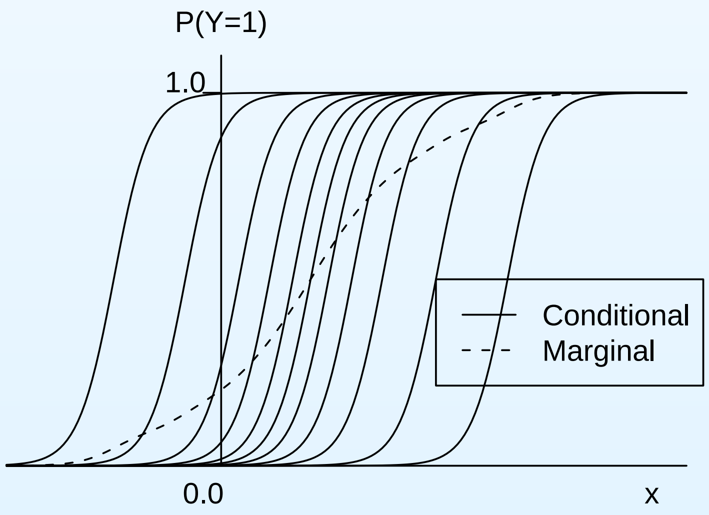
  for link function $g(\cdot)$ and fixed effects $\boldsymbol{\beta}$, and $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.
- Other assumptions for random effects include a nonparametric approach (Aitkin 1999) and a mixture of normal distributions (Molenberghs et al. 2010).
- Ordinary multivariate normal has advantage of natural use in multivariate case and for multilevel models.
- What if we assume normality but actual distribution quite different?

Effect of non-normal random effects

- Most literature shows relatively little effect in bias and efficiency in choosing incorrect random effects distribution (e.g., Neuhaus et al. 1992).
- Accuracy of predicted random effects not much affected by violations. McCulloch and Neuhaus (2011) show different distributions yield different predicted values but have similar MSE performance.
- But, when var($\boldsymbol{u}_i$) depends on covariates, between-cluster effects may be sensitive to misspecification of dist. of $\boldsymbol{u}_i$, because of implied diminution of marginal effect (Heagerty and Zeger 2000).

Logistic random-intercept model

Figure shows the *conditional* subject-specific curves for a random effects model and *marginal* (population-averaged) curve averaging over these.



Effect of non-normal random effects (continued)

- Agresti, Caffo, Ohman-Strickland (2004) found significant efficiency loss for logistic-normal random intercept model when true distribution is two-point mixture, especially when var(u_i) and T are large.

  Example: Follman and Lambert two-point mixture of logistic regression has

  $\hat{\beta} = 124.8$, $SE = 25.2$, $\hat{\beta}/SE = 4.9$,

  whereas normal mixture model has

  $\hat{\beta} = 65.5$, $SE = 19.5$, $\hat{\beta}/SE = 3.4$.

  Relevant sort of example?
  Opinions about legalized abortion in several situations.

Standard models motivated by latent variable models

Example: Probit model, Logistic regression model

- *Tolerance distribution* in dose–response studies:

  A tolerance distribution with cdf F for dosage that induces ‘success’ implies model for ’success probability’ $\pi(x)$

  $$G^{-1}[\pi(x)] = \alpha + \beta x$$

  for standardized cdf G for F.

  G^{-1} is the “link function.”

  $G = \Phi$ (standard normal) gives probit (Bliss, 1935).

  $G =$ standard logistic gives logit (Berkson 1944).

Latent model for probit regression

- *Threshold model*:

  Assumes unobserved continuous response y^* such that we observe $y = 0$ if $y^* \leq \tau$ and $y = 1$ if $y^* > \tau$.

  Suppose $y^* = \alpha + \beta x + \epsilon$, where $\{\epsilon_i\}$ independent from $N(0, \sigma^2)$. Then,

$$P(Y = 1) = P(Y^* > \tau) = P(\alpha + \beta x + \epsilon > \tau)$$

$$= P(-\epsilon < \alpha + \beta x - \tau) = \Phi[(\alpha + \beta x - \tau)/\sigma].$$

  (For identifiability, set $\sigma = 1$ and $\tau = 0$.)

  Thus, probit model results (link function is then Φ^{-1}).

  Logistic regression model (logit link) follows when ϵ has instead a standard logistic distribution.

Latent model for probit and logit (continued)

- *Utility model*:

Let U_0 be *utility* of $y = 0$ and U_1 the utility of $y = 1$. For $y = 0$ and 1, suppose $U_y = \alpha_y + \beta_y x + \epsilon_y$. A particular subject selects $y = 1$ if their $U_1 > U_0$.

If ϵ_0 and ϵ_1 are independent N(0, 1) random variables,

$$P(Y = 1) = P(\alpha_1 + \beta_1 x_1 + \epsilon_1 > \alpha_0 + \beta_0 x_0 + \epsilon_0)$$

$$= P\Big\{(\epsilon_0 - \epsilon_1)/\sqrt{2} < \big[(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)x\big]/\sqrt{2}\Big\} = \Phi(\alpha^* + \beta^* x)$$

where $\alpha^* = (\alpha_1 - \alpha_0)/\sqrt{2}$ and $\beta^* = (\beta_1 - \beta_0)/\sqrt{2}$.

Again, this is probit model.

Get logit model when $\epsilon \sim$ extreme-value distribution.

Ordinal models also result from latent variable models

Regression model for ordered categorical response
(Anderson and Philips 1981, McKelvey and Zavoina 1975)

y = *observed* ordinal response

y^* = *underlying* continuous latent variable,

cdf $G(y^* - \eta)$ with $\eta = \eta(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$

thresholds (cutpoints) $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_c = \infty$ such that

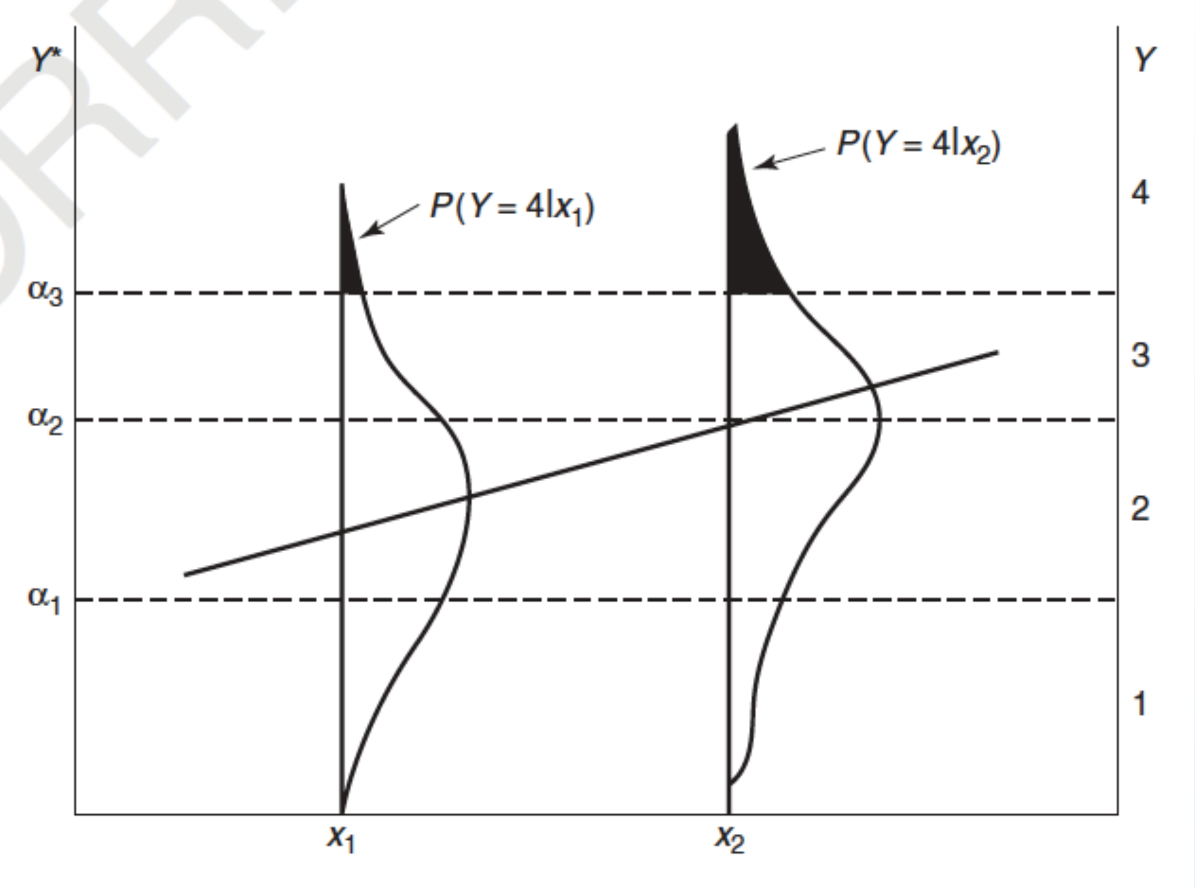
$$y = j \text{ if } \alpha_{j-1} < y^* \le \alpha_j$$

Then

$$P(y \le j \mid \boldsymbol{x}) = P(y^* \le \alpha_j \mid \boldsymbol{x}) = G(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x})$$

$\rightarrow$ model $G^{-1}[P(y \le j \mid \boldsymbol{x})] = \alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}$

Latent variable model for an ordinal response



Latent variable model for ordinal response

Here, G^{-1} is again a link function.

Get cumulative logit model when G = logistic cdf $(G^{-1} =$ logit).

So, cumulative logit (probit) model fits well when regression model holds for underlying logistic (normal) response.

Note: Model is often expressed as

$$\text{logit}[P(y \leq j)] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}.$$

This derivation suggests such models are designed to detect shifts in *location* (center), not *dispersion* (spread).

Model implies conditional distributions of y at different settings of explanatory variables are *stochastically ordered*; i.e., cdf at one setting always above or always below cdf at another setting.

Latent model showing how OLS regression can fail

Suppose

$$y^* = 20.0 + 0.6x - 40z + \epsilon$$

$x \sim$ unif(0, 100), $P(z = 0) = P(z = 1) = 0.50$, $\epsilon \sim N(0, 10^2)$.

For random sample of size $n = 100$, suppose

$y = 1$ if $y^* \leq 20$, $y = 2$ if $20 < y^* \leq 40$, $y = 3$ if $40 < y^* \leq 60$,

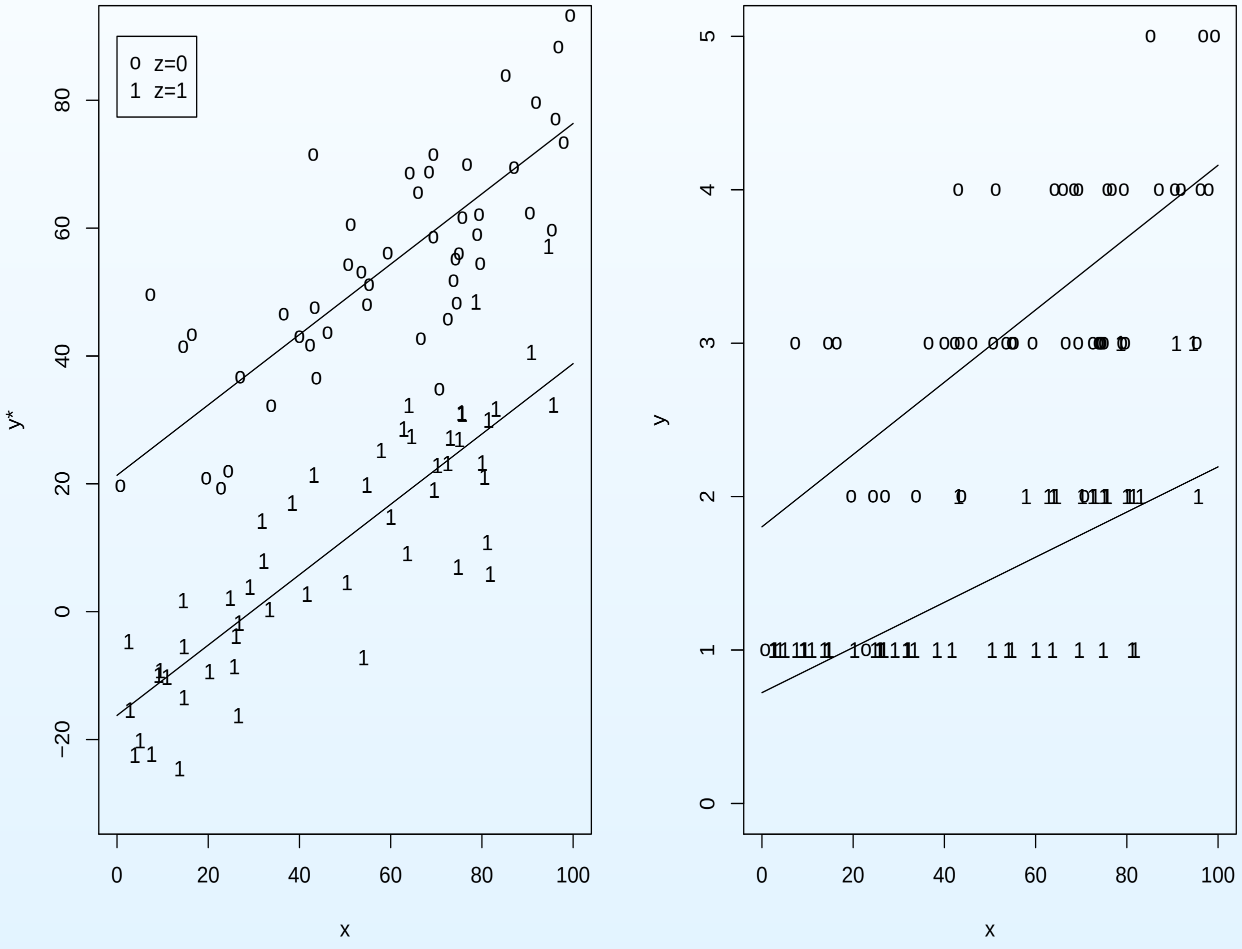
$$y = 4 \text{ if } 60 < y^* \leq 80, \quad y = 5 \text{ if } y^* > 80.$$

When $x < 50$ with $z = 1$, there is high probability that observations fall in lowest category of y. Suppose we fit model

$y = \alpha + \beta_1 x + \beta_2 z + \beta_3 (x \cdot z) + \epsilon$

to investigate effects and possible interactions.

Latent model, and OLS fit to observed data



Floor effect for ordinal data

Because of *floor effect*, least squares line for *observed* data with fixed y scores has slope half as large when $z = 1$ as when $z = 0$.

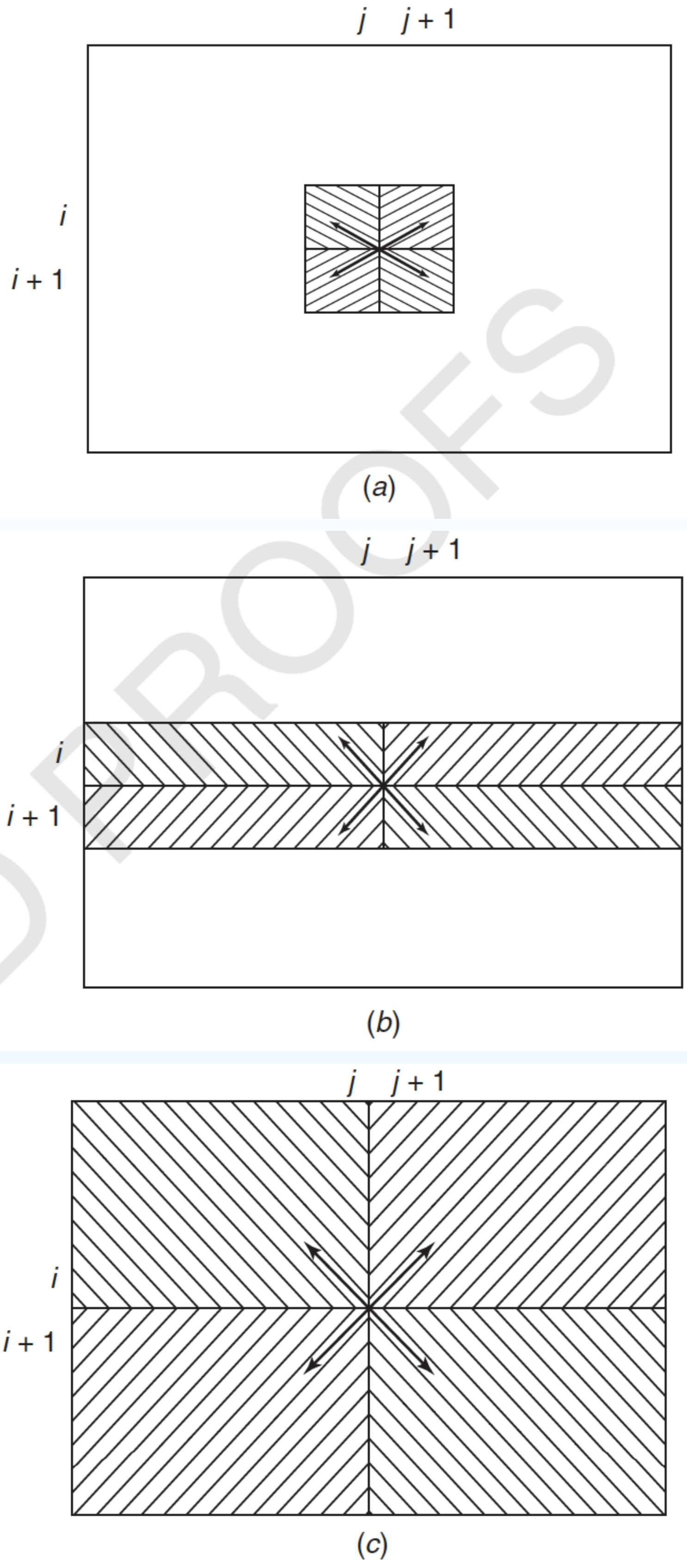
Interaction is statistically and practically significant.

Such spurious effects would not occur in fitting the ordinal model (cumulative logit or cumulative probit).

Note: The ordinal model does not require scores for y values.

Latent variable structure for ordinal contingency tables

- What if assume an underlying continuous variable, but rather than select a particular parametric form (such as logistic or normal), merely assume ordinal structure such as (1) nonnegative *local* log odds ratios, (2) stochastic ordering, (3) nonnegative *global* log odds ratios for binary collapsings of y and x.
- *Order-restricted methods* can fit data subject to such restrictions, test hypotheses such as independence against such alternatives (e.g., Agresti and Coull 2002).
- For significance testing, large-sample distribution theory awkward, relying on *chi-bar-squared* distributions.
- Confidence intervals: Seems to be little research yet.
- How to infer whether particular structure holds, such as uniformly nonnegative local log odds ratios?



(a)

(b)

(c)

Example: Trauma due to subarachnoid hemorrhage

	Outcome				
Treatment group	Death	Vegetative state	Major disabiity	Minor disability	Good recovery
Placebo	59	25	46	48	32
Drug	135	39	147	169	102

- Sample log odds ratios are:
  Local: ($-$0.38, 0.72, 0.10, $-$0.10)
  Global: (0.28, 0.47, 0.32, 0.15)
- For testing H_0: independence (identical distributions) against nonnegative local log odds ratios, P-value = 0.012.
- Strong evidence against H_0, but inappropriate to conclude that the true local log odds ratios are uniformly nonnegative.

A Bayesian solution (with M. Kateri, 2013)

- For a particular prior distribution for multinomial probabilities, construct posterior distribution and evaluate posterior probability over desired set (e.g., nonnegative local log odds ratios).

- For Dirichlet priors, posterior is Dirichlet. Can simulate from it to precisely approximate posterior probabilities.

- Example: For uniform densities over admissible probability values (i.e., Dirichlet with hyperparameters = 1), based on 1,000,000 simulations, posterior probability of nonnegative

  local log odds ratios: 0.014

  global log odds ratios: 0.705
  (closed form in Altham 1969)

Brief summary of some other work

- Anderson and Vermunt (2000, *Sociol. Methodology*): Goodman association model arises when observed $\{Y_t\}$ conditionally independent given latent Z that is conditionally normal (given observed variables).
- Gueorguieva and Agresti (2001, *J. Amer. Statist. Assoc.*): Probit model for joint modeling of clustered binary and continuous responses, based on underlying joint normality.
- Vermunt (2003, *Sociol. Methodology*): Multilevel latent class models, relaxing assumption of local independence.
- Vermunt and Magidson: *Latent GOLD* software fits wide variety of mixture models, including latent class models, nonparametric mixtures of logistic regression, Rasch mixture models, zero-inflated models, multilevel models, continuous latent var's.

Applications generalize scope of models

- Reboussin and Ialongo (2010, *J. Roy. Stat. Soc.*):
  Model drug use among students who suffer from attention deficit hyperactivity disorder (ADHD), using (1) longitudinal latent transition model with latent classes for stages of marijuana use that describes probability of transitioning between stages, (2) cross-sectional latent class model constructs ADHD subtypes and describes influence of subtypes on transition rates.
- Lin et al. (2008, *Biometrics*):
  Model repeated transitions between independence and disability states of daily living using multivariate latent var's. State-specific latent var. represents tendency to remain in state, accounts for correlation among repeated sojourns in same state. Correlation among sojourns across states accounted for by correlation between different latent var's.

Summary and final comments

- Latent variable models have a long and substantial history for categorical data analysis, which we've barely scratched.
- Many methods commonly used by statisticians for categorical data analysis have latent variable justifications.
- Future: Challenges statisticians face with large data sets with huge numbers of variables are especially challenging for latent variable modeling.
- We should not forget the dangers of *reification* – acting as if an assumed latent variable truly measures the characteristic of interest (Gould 1981, *The Mismeasure of Man*).

Some useful references

Bartholomew, D., M. Knott, and I. Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. (see Chapters 4-6)

Goodman, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*

Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent Structure Analysis*.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.*

Websites: www.people.vcu.edu/ nhenry/LSA50.htm (Neil Henry reminiscences)
statisticalinnovations.com/products/aboutlc.html (Latent GOLD)
www.stata.com/meeting/2nasug/lclass.pdf (Stata)
www.stat.rutgers.edu/home/buyske/software.html (R)
support.sas.com/kb/30/623.html (SAS)
www.statmodel.com (Mplus)
spitswww.uvt.nl/ vermunt/ (LEM et al.)
www.john-uebersax.com/stat/ (overview)
faculty.chass.ncsu.edu/garson/PA765/latclass.htm (overview)

Other books: Collins / Lanza (2009), Hagenaars / McCutcheon (2009), Heinen (1996)