Latent variable models and surveys: reflections on the Montreal workshop

Mary Thompson

Department of Statistics & Actuarial Science

University of Waterloo

Outline

- the complex surveys working subgroup on weighting and estimation
- Montreal workshop May 4-6, 2005, Latent Variable Models and Complex Surveys in Social Sciences and Health Sciences Research
- current paradigm
- multilevel models
- future directions

Weighting and estimation group

Scientific leader: Chris Skinner, Southampton University

Members included: Biemer, Thompson, Bellhouse, Carle, Chantala, Christ, Kolenikov, Kovacevic, Munk, Roberts, Thomas, Wu, Wang

Focus on complex surveys and multilevel models

treal workshop on LVMSS and complex su

Sponsorship:

- Centre de Recherches Mathématiques (CRM)
- National Program on Complex Data Structures (NPCDS)
- SAMSI
- Statistics Canada

Program

- A. Skrondal and S. Rabe-Hesketh: tutorial on Generalized Latent Variable Modeling; pseudo maximum likelihood estimation of multilevel generalized linear models; multilevel latent growth modeling with latent- trait-dependent dropout: A cluster randomized sex education intervention
- B. Muthén and T. Asparouhov (special session): Survey Data Modeling with Mplus
- K. Bollen: Overview of LVMSS program and outcomes
- C. Skinner: The use of survey weights in multilevel modelling: an overview

Program, continued

- M. Kovacevic and R. Huang: Fitting linear mixed models from survey data
- K. Chantala and C. Suchindran: Adjusting for unequal selection probability in multilevel models: a comparison of software packages
- S. Christ: Inference in structural equation modeling using samples with unequal probabilities of selection and misspecified models
- G. Roberts and M. Kovacevic: Structural equation modeling with Statistics Canada's survey data
- J. N. K. Rao: MCMC methods for correlated binary data
- A. Cyr: IRT modeling for latent variables in complex surveys

Program, continued

- A. Sacker, P. McDonough, R.D. Wiggins and M. Bartley: Employment and self-rated health trajectories in the UK and the US
- H. Ariizumi: Effects of unemployment on health status in the NPHS
- R. D. Wiggins: The development and assessment of a quality of life measure in the context of research on aging
- A. C. Carle: Data quality and nonsampling error in the American Community Survey ... a latent variable model assessment
- J. Olsen: Statistical models for surveys and other studies of small groups

Analysis of complex survey data

- population surveys with stratified multistage designs
- longitudinal surveys
- non-response and attrition
- measurement error
- models for dependence
  - multilevel models
  - causal interpretations
  - etc.

Current paradigm

- Choose system of estimating equations or equivalent at the population level
- For point estimation, replace population sums by weighted sample sums, and solve resulting estimating equations

(Binder, 1983; Skinner, 1989)

Example

Model with group random effect

The contribution of group j to the likelihood:

$$l_j(\theta, \psi) = \int \phi(\eta_j \mid x_{ij}; \psi) \prod_i f(y_{ij} \mid x_{ij}, \eta_j; \theta) d\eta_j$$

The pseudo log likelihood (Skinner, 2005):

$$\log L(\theta) = \sum_{j=1}^{m} w_j \log \int \left(\exp\{\sum_{i=1}^{n_j} w_{i|j} \log f(y_{ij} \mid x_{ij}, \eta_j; \theta)\} \right) \phi(\eta_j \mid x_{ij}; \psi) d\eta_j$$

Basic paradigm continued

For mean squared error estimation, use sandwich estimator
Alternative: use resampling (e.g. bootstrap of Rao and Wu, 1988)
Capture the results of resampling in a set of resampling weight variables
These yield:

- estimated distribution of $\hat{\theta}$
- estimated distribution of the estimating function for sandwich MSE estimator for $\hat{\theta}$ (Binder et al, 2003; Rao)

Advantage of resampling

Suppose the weights are *calibrated* to the totals of x.

The appropriate estimator of variance of a mean of y is not $v(y)$ (provided by software unaware of the calibration) but $v(e)$, e = residual from regression of y on x.

If the resampling weights are also calibrated to the totals of x, the resampling variance will approximate $v(e)$.

Multilevel models

(a unifying framework)
Pseudo log likelihood as before

Pfeffermann et al (1998) PWIGLS (linear multilevel models)

Asparouhov (2002-2005) MPML

Rescalings of $w_{i|j}$ to reduce biases

Top level sandwich estimator OK for large cluster sample sizes

Comparisons

Chantala, Blanchette and Suchindran

- National Longitudinal Survey of Adolescent Health
- BMI vs HR_WATCH (TV or computers)
- SAS, MLWIN, LISREL, Mplus, Stata GLLAMM
- Stata and SAS programs for constructing MLM weights available for use with MLWIN, LISREL, Stata GLLAMM, Mplus

Comparisons

Kovacevic and Huang

- Workplace and Employee Survey
- small sample sizes within clusters
- e.g. log(wage) vs hiedu, nonprft, etc.
- simulation comparison of several non-iterative methods
- point estimates of regression coefficients, variance components
- Huang and Hidoroglou (2003):use of weights in variance component estimation and simplest rescaling of $w_{i|j}$

Future directions

- MITACS and NPCDS investigators
- the Statistics Canada research program
- weighting and estimation issues

MITACS/NPCDS investigators

- event history analysis in longitudinal data (with censoring, truncation, and heterogeneity connected with design structure)
- modelling with cross-nested random effects
- algorithms for the creation of replication variance estimators
- imputation for complex non-response patterns
- plausible value methodology (multiple imputation) in item response theory
- etc.

Statistics Canada

- applications, algorithms and advice
- e.g. NLSCY users
- accounting for attrition
- etc.

Weighting and estimation

- examining the paradigm
- refining the paradigm
- moving away from descriptive inference

Estimation methods

A Scott, November 2005: “Although most statistical packages now have special survey modules that can be used to carry out weighted analyses of the data from such studies, weighting tends to be inefficient in situations . . . where large differences among the weights are typical. Fully efficient likelihood methods exist for some special designs but, apart from the case of simple stratified sampling, these methods require special software and are difficult to implement. Moreover, there are questions about their robustness to model breakdown . . .”