MRC Chalk Talk

What You Need to Know about

Multilevel Modeling of Complex Survey Data

Yoonsang Kim, PhD

Institute for Health Research and Policy

University of Illinois at Chicago

Focus of the Chalk Talk

- Complex survey design
- Longitudinal survey or multi-stage survey design rather than repeated cross-sectional surveys
- Regression models

    Multilevel modeling approach

    Covariance structure approach

Why longitudinal survey?

- Follow up samples of people or households over periods of time
- To help identify causality: longitudinal surveys can shed lights on causality than mere association
- To increase accuracy of recall, and avoid recall bias and 'telescoping'

  Telescoping: survey respondents report events as having taken place within a reference period when in fact they took place longer ago.

Examples of longitudinal survey

- **Panel Survey of Income Dynamics**
  Long-term panel survey. To collect behavioral, attitudinal, and circumstantial data on a range of social and economic issues
- **Health and Retirement Study**
  To follow age-eligible individuals (50 years and older) and their spouses/partners as they transition from active working to retirement, measuring aging-related changes, financial status, health, and retirement planning
- There are more...

Weaknesses of longitudinal surveys

- Panel conditioning

  The way respondents report changes because of their experience of the first wave. After the 1st wave, sample members have already experienced the survey and therefore have a very good idea of exactly what it consists of, what kinds of questions will be asked.

- Sample attrition

  The proportion of sample units that respond at every wave may be low. Participation in a longitudinal survey requires considerable commitment from respondents.

- <u>Weighting</u>

  For a t-wave longitudinal survey, there can be $2^t - 1$ possible populations and $2^t - (t+1)$ longitudinal populations, thus that many sets of weights! (that are designed to make the set of persons who responded to x waves representative of such population)

- <u>Tracking and tracing</u>

  Need to retain the ability to contact sample members at each wave, requiring an administrative system and a program of operations to be put in place

(Lynn, 2009)

Complex survey design

Stratification, Clustering, Weighting

- Multistage sampling: survey respondents are not simple random samples.
- Stratification, clustering (PSU), and weighting influence the size of SEs for estimates.
- Design effect measures the net effect of the combined influences of stratification, clustering, and weighting.

$$D^2(\hat{\theta}) = \frac{Var(\hat{\theta})_{complex}}{Var(\hat{\theta})_{SRS}}$$

Often measures efficiency losses relative to a SRS.

(Heeringa, West, & Berglund, 2010)

Weighting

- Weights = 1/sample inclusion probability

  = $1/\pi_i$ for a respondent i

- Reflect unequal sample inclusion probabilities
- Compensate for non-response and under-coverage
- Used to get unbiased estimates or to attenuate selection bias esp. for non-ignorable sampling designs

Variance estimation

- Stratification, clustering, and weighting complicate variance estimation
- Dependent observations affect sampling variances of model parameters and reference distribution for test statistics
- Taylor series linearization

$$\bar{y}_w = \frac{\sum_h \sum_\alpha \sum_i w_{h\alpha i} y_{h\alpha i}}{\sum_h \sum_\alpha \sum_i w_{h\alpha i}} = \frac{u}{v}$$

$y_{h\alpha i}$ = response of a person i in cluster α in stratum h

By Taylor series expansion

$$Var\left(\frac{u}{v}\right) \cong \frac{var(u) + \bar{y}_w^2\, var(v) - 2\bar{y}_w\, cov(u, v)}{v^2}$$

- Resampling variance estimation

  Nonparametric methods:

  Balances repeated replication (BRR), jackknife repeated replication (JRR), bootstrap

  Use replicated subsampling of the sample database to develop sampling variance estimates

  Involves replicate weights

(Skinner & Holmes, Ch 14 in Chambers & Skinner, 2003

Heeringa et al., 2010)

Regression models for complex survey data

Pfeffermann, et al. (1998)

Normally distributed outcome, Ground work for methods of incorporating sampling weights in multilevel models for complex survey data

Rabe-Hesketh & Skrondal, (2006); Rabe-Hesketh, Skrondal, & Pickles (2004)

Generalized linear mixed models with an arbitrary number of levels using adaptive quadrature

Stata GLLAMM

Multilevel modeling for non-normal response

Rabe-Hesketh & Skrondal (2006)

- Longitudinal survey / multi-stage design

Level 3	PSU	Geo Areas
Level 2	Individuals	Schools
Level 1	Time points	Students

- Model (two-stage sampling):

$$g\{E(y_{it}|\boldsymbol{x}_{it}, b_i)\} = x_{it}\beta + b_i$$

y_{it} = response from a respondent i at wave t

$g\{\ \}$ = link function

Random intercept $b_i \sim N(0, \tau^2)$

Parameter Estimation

- Likelihood function is constructed at each level

$$L(y) = \sum_i w_i^{(2)} L_i^{(2)}(y_i)$$

Need separate weights at each level!

- Pseudo-maximum likelihood (PML)

Weights enter as if they were freq weights at each level, representing the # times that each unit should be replicated

Adaptive quadrature for approximation to the integrals

Weights

- Define

  π_i = Individual's probability of inclusion / School's inclusion probability

  $\pi_{t|i}$ = Probability that individual i responds at time t / Probability that a student t responds at school i

- The weights w_i and w_{it} are usually given.

$$w_i = {}^{1}/_{\pi_i} \text{ and } w_{it} = {}^{1}/_{\pi_i\, \pi_{t|i}}$$

  Therefore,

$$w_{t|i} = \frac{w_{it}}{w_i}$$

- Usually base weight $w_i = w_{i1} = 1/\pi_{i1}$ (level 2)

$w_{1|i} = w_{i1}/w_i = 1, w_{2|i} = w_{i2}/w_i,\ w_{3|i} = w_{i3}/w_i, \ldots$ (level1)

- Scaling the lower level weights affects parameter estimates

$$w^*_{t|i} = \frac{t_i^*}{\sum_{t=1}^{t_i^*} w_{t|i}} w_{t|i}$$

t_i^* = the last wave at which individual i responds

Average rescaled weights for a person $i = 1$

- For more details for scaled weights

Longford (1996)

Pfeffermann et al. (1998)

Graubard & Korn (1996)

<u>Clustering & variance</u>

Q: PSU as the level 3?

Need the PSU selection probability

Not as a level in the model, but should be accounted for variance estimation $\rightarrow$ Sandwich-type estimator of SE

More regression models and variance estimation

Skinner & Holmes (2003)

Extended the work of Pfeffermann et al. to allow for serially correlated responses

Covariance structure approach

- Ignores the clustering of repeated obs within clusters
- Uses the linearization, and weight = w_{it}

Skinner & Vieira (2007)

Studied the impacts of cluster sampling on variance estimation

Including random cluster effect may seriously underestimate the effects of clustering on the SEs

Recommended the linearization approach for "robust" SEs

Vieira & Skinner (2008)

PML combined with the linearization for variance estimation

Limitation: used "complete" respondents

Reference

Chambers, R. L., & Skinner, C. J. (Eds.). (2003). *Analysis of survey data*. Chichester, West Sussex, England ; Hoboken, NJ: Wiley.

Graubard, B. I., & Korn, E. L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, *5*(3), 263–281.

Heeringa, S., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Longford, N. t. (1996). Model-Based Variance Estimation in Surveys with Stratified Clustered Design. *Australian Journal of Statistics*, *38*(3), 333–352. doi:10.1111/j.1467-842X.1996.tb00687.x

Lynn, P. (2009). Methods for longitudinal surveys. *Methodology of Longitudinal Surveys*, 1–19.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *60*(1), 23–40.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805–827. doi:10.1111/j.1467-985X.2006.00426.x

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM Manual. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Retrieved from http://biostats.bepress.com/ucbbiostat/paper160

Skinner, C., & Vieira, M. de T. (2007). Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology*, *33*(1), 3–12.

Vieira, M. D. T., & Skinner, C. (2008). Estimating Models for Panel Survey Data under Complex Sampling. *Journal of Official Statistics*, *24*(3), 343–364.

Thank you!

Email ykim96@uic.edu for questions