Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

Lecture 14: Adjusting for between- and within-cluster covariates in the analysis of clustered data

May 14, 2009

XH Andrew Zhou

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Adjustment for cluster-level covariates

- If cluster-level covariates are available, we can directly adjust for their effects in a multi-level model.
- If cluster-level covariates are not available, we may still adjust for cluster-level effects by including cluster means in the model because variability in cluster means can confound the estimated association between the individual level covariate measurement and outcome.
- It has been shown that inference on the individual-level covariate can be misleading without adjusting for cluster-level means.

An example

- Let us consider the impact of birth weight on childhood intelligence (or IQ) as measured by the Wechsler Intelligence Scale for children.
- Most studies demonstrate that heavier babies tend to have higher IQs.
- However, birth weight is known to be associated with family socio-economic status (SES), with families of higher status having larger babies.
- SES of a family is also related to the IQ measurements of children in that family. Thus, SES can act as a potential confounder of the relationship between birth weight and IQ (Begg and Parides, 2003).

Adjustment for cluster-level covariate

- If we could measure SES by a covariate, we can include that cluster-level covariate into a regression model.
- However, SES is difficult to measure accurately.
- Alternatively, if we had measures of birth weight and IQ for multiple siblings within the same family, we could use these data to make ‘within-family’ comparisons that would be tightly controlled for SES.
- We can also evaluate individual-level birth weight as a predictor of individual IQ as well as the effect of ‘family-averaged’ birth weight on IQ.

Adjustment, continued

- This leads to separation, or partitioning, of the effect of birth weight that conforms to interesting research hypothesis.
- Hence by breaking down the birth weight effect into individual-level and family-level components, we are able to obtain better estimates of these effects and draw more accurate conclusions.

Notations

- Y_{ij}: the outcome for subject j in cluster i.
- X_{ij}: a corresponding continuous covariate, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$.
- The mean covariate measurement for the group can be computed as $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$.

Generalized linear models models

- To study the relationship of Y and X, we consider the following five models:

$$h(E[Y_{ij}|X_{ij}]) = \beta_{01} + \beta_1 X_{ij}, \quad (1)$$

$$h(E[Y_{ij}|a_i, X_{ij}]) = \beta_{02} + \beta_2 X_{ij} + \gamma_2 \bar{X}_i, \quad (2)$$

$$h(E[Y_{ij}|X_{ij}]) = \beta_{03} + \beta_3 (X_{ij} - \bar{X}_i) + \gamma_3 \bar{X}_i, \quad (3)$$

$$h(E[Y_{ij}|X_{ij}]) = \beta_{04} + \beta_4 (X_{ij} - \bar{X}_i), \quad (4)$$

$$h(E[Y_{ij}|X_{ij}]) = \beta_{05} + \gamma_5 \bar{X}_i. \quad (5)$$

(6)

Model interpretation

- Model (2) and Model (3) are mathematical equivalent.

Interpretation of the coefficients

- For Model (1), β_1 measures the change in the expectation corresponding to one unit increase in the covariate, X_{ij}, which is distorted due to confounding by cluster-level effect.
- For Model (2), β_2 measures the change in the expectation corresponding to one unit increase in the covariate, given the same cluster-averaged $\bar{X}$.
- For Model (2), γ_2 can be interpreted as the effect of one unit increase in cluster-averaged $\bar{X}$ on Y, holding the fixed individual-level covariate, X_{ij}.
- That is, given two subjects with the same individual covariate, the subject whose cluster has a 1 unit higher average X can be expected to have Y increase about γ_2.

Interpretation of the coefficients, continued

- For Model (3), β_3 measures the change in the expectation corresponding to a one unit increase in the covariate, given the same cluster-averaged $\bar{X}$.
- For Model (3), γ_3 represents the mean difference in Y associated with one unit increase in cluster-averaged $\bar{X}_i$ simultaneous with one unit increase in individual X_{ij}.

Model comparison

- Model (1) may lead to biased results, and should not be used.
- Model (2) is the model of choice for most purpose for adjusting for cluster-level confounders and easy interpretation.
- Model (3) can adjust for cluster-level confounders, but interpretation of the γ_3 is not easy.

Remarks

1. When covariate values vary within a cluster, the cluster-level mean covariate is often the most possible confounder of the association between individual-level covariate and response.
2. At least, it may be served as a proxy for some relevant cluster-level characteristics.
3. When the clusters are relative large (as in clinical center-based), the cluster mean can be estimated much more precisely than in data sets where cluster sizes are small (as in family studies).

Generalized linear mixed effects models (GLMM)

- We can also consider the following five mixed effects models:

$$h(E[Y_{ij}|X_{ij}]) = a_{1i} + \beta_1 X_{ij}, \quad (7)$$

$$h(E[Y_{ij}|a_i, X_{ij}]) = a_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_i, \quad (8)$$

$$h(E[Y_{ij}|X_{ij}]) = a_{3i} + \beta_3 (X_{ij} - \bar{X}_i) + \gamma_3 \bar{X}_i, \quad (9)$$

$$h(E[Y_{ij}|X_{ij}]) = a_{4i} + \beta_4 (X_{ij} - \bar{X}_i), \quad (10)$$

$$h(E[Y_{ij}|X_{ij}]) = a_{5i} + \gamma_5 \bar{X}_i, \quad (11)$$

(12)

where $a_{ki} \sim N(\beta_k, \sigma^2_{ak})$, and a_{ki} and X_{ij} are independent.

SAS PROC GLIMMIX for GLMM

- METHOD (=RSPL, MSPL, RMPL, MMPL) specifies the estimation method in a generalized linear mixed model (GLMM).
- The default is METHOD=RSPL.
- Estimation methods ending in "PL" are pseudo-likelihood techniques.
- The first letter identifier determines whether estimation is based on a residual likelihood ("R") or a maximum likelihood ("M").
- The second letter identifies the expansion locus for the underlying approximation. The expansion locus of the expansion is either the vector of random effects solutions ("S") or the mean of the random effects ("M").

NACC UDS

- The National Alzheimer's Coordinating Center's (NACC) Uniform Data Set (UDS) is an ongoing longitudinal database of subjects seen at one of the National Institute on Aging's 29 funded Alzheimer's Disease Centers (ADC) located throughout the USA.
- Subjects seen at the ADCs represent a clinical sample of individuals who are either referred to the clinic for evaluation of dementia, self-referred to the clinic, or are recruited by clinics to participate in dementia research.
- Longitudinal follow-up began in 2005. As of December 2008, 16225 subjects aged 50 or older had at least one clinic visit. Up to four observations per subject were available, with 8101 subjects having at least two observations.

Generalized linear models with GEE

- We use the baseline of the NACC UDS data set.
- Response (Y): Demented– Does the subjects have dementia? (1–Yes, 0–No)
- Covariate (X): Mini-Mental State Examination (MMSE) score (0-30).

MMSE

- Any MMSE score over 27 (out of 30) is effectively normal.
- Below this, 20-26 indicates some cognitive impairment.
- Any MMSE between 10 and 19 moderate to severe cognitive impairment.
- Below 10 indicates very severe cognitive impairment

Estimation for GLM Models 1 and 2

	Model 1			Model 2		
Parameters	Estimate	Std.err	p-value	Estimate	Std.err	p-value
Intercept	1.0412	0.1551	$<$0.0001	1.5411	0.8654	0.0749
$\overline{\mathrm{MMSE}}$	–	–	–	-0.0207	0.0323	0.5228
MMSE	-0.0223	0.0069	$<$0.0013	-0.0217	0.0066	0.0009

Model 1:$logit(P(Y_{ij} = 1)) = \beta_{01} + \beta_{11} MMSE_{ij}.$

Model 2:$logit(P(Y_{ij} = 1)) = \beta_{02} + \beta_{12} MMSE_{ij} + \beta_{22} \overline{MMSE}_i.$

Estimation for GLM Models 3 and 4

	Model 3			Model 4		
Parameters	Estimate	Std.err	p-value	Estimate	Std.err	p-value
Intercept	1.5411	0.8654	0.0749	0.4855	0.1196	$<$0.0001
$\overline{\mathrm{MMSE}}$	-0.0424	0.0350	0.2254	–	–	–
$\mathrm{MMSE} - \overline{\mathrm{MMSE}}$	-0217	0.0066	0.0009	-0.0215	0.0062	0.0005

Model 3:$logit(P(Y_{ij} = 1)) = \beta_{03} + \beta_{13}(MMSE_{ij} - \overline{MMSE_i}) + \beta_{23}\overline{MMSE_i}.$

Model 4:$logit(P(Y_{ij} = 1)) = \beta_{04} + +\beta_{14}(MMSE_{ij} - \overline{MMSE_i}).$

Estimation for GLM Model 5

Parameters	Estimate	Std.err	p-value
Intercept	1.5079	0.8235	0.0671
$\overline{\text{MMSE}}$	-0.0411	0.0330	0.2128

Model 5:$logit(P(Y_{ij} = 1)) = \beta_{05} + \beta_{15}\overline{MMSE}_i$.

GLMM

We consider generalized linear mixed effects models

Estimation for GLMM Models 1 and 2

	Model 1			Model 2		
Parameters	Estimate	Std.err	p-value	Estimate	Std.err	p-value
Intercept	1.1001	0.1418	$<$0.0001	1.2721	1.2921	0.3327
$\overline{\text{MMSE}}$	–	–	–	-0.0068	0.05092	0.8937
MMSE	-0.02335	0.00149	$<$0.0001	-0.02335	0.00149	$<$.0001

Model 1:$logit(P(Y_{ij} = 1)) = a_{i1} + \beta_{01} + \beta_{11} MMSE_{ij}.$

Model 2:$logit(P(Y_{ij} = 1)) = a_{i2} + \beta_{02} + \beta_{12} MMSE_{ij} + \beta_{22} \overline{MMSE_i}.$

Estimation for GLMM Models 3 and 4

	Model 1			Model 2		
Parameters	Estimate	Std.err	p-value	Estimate	Std.err	p-value
Intercept	1.2721	1.2921	0.3327	0.5107	0.1373	0.0008
$\overline{\mathrm{MMSE}}$	-0.03016	0.0509	0.5535	–	–	–
$\mathrm{MMSE} - \overline{\mathrm{MMSE}}$	-0.02335	0.001487	$<$0.0001	-0.02335	0.001487	$<$0.0001

Model 3:$logit(P(Y_{ij} = 1)) = a_{i3} + \beta_{03} + \beta_{13}(MMSE_{ij} - \overline{MMSE_i}) + \beta_{23}\overline{MMSE_i}.$

Model 4:$logit(P(Y_{ij} = 1)) = a_{i4} + \beta_{04} + +\beta_{14}(MMSE_{ij} - \overline{MMSE_i}).$

Estimation for GLMM Model 5

Table 1: Estimation for Model 5

Parameters	Estimate	Std.err	p-value
Intercept	1.3003	1.2431	0.3039
$\overline{\text{MMSE}}$	-0.03182	0.04896	0.5158

Model 5:$logit(P(Y_{ij} = 1)) = a_{i5} + \beta_{05} + \beta_{15}\overline{MMSE}_i.$

Research questions

- We are interested in assessing the effect of individual MMSE on the probability of having dementia, free of confounding by center-level influences.
- We are also interested in assessing whether center-average MMSE has an independent effect on the probability of having dementia.

Violation of independence between random effects and covariates

- Let us consider the following linear mixed effects model:

$$Y_{ij} = a_i + \beta X_{ij} + \epsilon_{ij},$$

where $X_{ij} \sim (0, \sigma_X^2)$, $a_i \sim (0, \sigma_b^2)$, $\epsilon_{ij} \sim (0, \sigma_\epsilon^2)$, a_i and ϵ_{ij} are independent, X_{ij} and ϵ_{ij} are independent.

- The ordinary least squared estimator (OLS) for β,

$$\widehat{\beta}_{ols} = (\sum_{i,j} X_{ij}^2)^{-1} \sum_{i,j} X_{ij} Y_{ij},$$

is unbiased and consistent if the covariates and random effects are uncorrelated.

Violation of independence between random effects and covariates

- However, it is easy to show that when $n_i = n$,

$$\widehat{\beta}_{ols} \rightarrow \beta + \frac{\sigma_{xb}}{\sigma_x^2}.$$

- This is a standard econometrics result; that is that the error term correlated with a predictor can introduce bias in regression coefficients.

A shared random-effects model with normal distributions

- $Y_i = (Y_{i1}, \ldots, Y_{in_i})$ are conditionally independent given $X_i = (X_{i1}, \ldots, X_{in_i})$ and a_i.
- $$Y_{ij} = \beta_0 + a_i + \beta_B \bar{X}_i + \beta_W (X_{ij} - \bar{X}_i) + \epsilon_{ij},$$

  where $\epsilon_{ij} \sim N(0, \sigma_w^2)$, $a_i \sim N(0, \sigma_b^2)$, and $X_{ij} \mid a_i \sim N(\delta_0 + \delta_1 a_i, \sigma_X^2)$.

A shared random-effects model, continued

- Neuhaus and McCulloch (2006) showed that the likelihood controbution from the ith cluster is

$$\int_a f(y_i \mid x_i, a) f_X(x_i \mid a) dG(a) =$$

$$f(y_i - \bar{y}_i \mid x_i - \bar{x}) f(x_i - \bar{x}) \int_a f(\bar{y}_i \mid \bar{x}_i, a) f(\bar{x}_i \mid a) dG(a).$$

- They further showed that $f(y_i - \bar{y}_i \mid x_i - \bar{x})$ involves only β_W, but not β_B.
- But $f(\bar{y}_i \mid \bar{x}_i, a)$ involves β_B but not β_W.

Implications

- We may still get a consistent estimator of β_W, based on $f(y_i - \bar{y}_i \mid x_i - \bar{x})$, which is the contribution to the conditional likelihood, proposed by Neuhaus and McCulloch (2006).
- However, we cannot separate estimation of β_B from the specification of the distribution for the random effects, a.
- Misspecification of the distribution may bias the estimator of β_B.

Implication, continued

- Models with common between- and within-cluster covariate effects do not distinguish information from the two sources.
- Poor estimation of between-cluster covariate effects due to features such as correlations between covariates and random effects may yield inconsistent estimates of overall covariate effects.