Introduction to Analysis Methods for Longitudinal/Clustered Data, Part 3: Generalized Estimating Equations

Mark A. Weaver, PhD
Family Health International

Office of AIDS Research, NIH
ICSSC, FHI
Goa, India, September 2009

Objectives

1. To develop a basic conceptual understanding of what generalized estimating equation (GEE) methods are and
2. When they might be applicable
3. With a focus on interpretation of results

2

What are GEE?

- GEE (Liang and Zeger, 1986) provide a method of inference for a wide variety of models when responses are correlated
  - Linear regression (for continuous outcomes)
  - Logistic regression (for binary outcomes)
  - Poisson regression (for outcomes that are counts)
  - Proportional odds (for ordinal categorical outcomes)

What are GEE?

- GEE are an analysis method, not models in and of themselves
  - You specify a model that you'd like to fit using GEE
  - Model is specified through
    1. A link function that relates the mean response to the regression equation
       - "link = logit" for logistic regression
       - "link = log" for Poisson regression
    2. An assumed distribution for the response, although distributional assumptions not really strong
       - "binomial" for logistic regression
       - "poisson" for Poisson regression
    3. A working correlation matrix (more on this later)

4

GEE Benefits

- GEE provides many of the same benefits as mixed models.
  - Accounts for within-subject/within-cluster correlations
  - Allows for missing data (but requires stronger assumption in this regard than mixed models)
  - Allows for time-varying covariates
  - Allows for irregularly-timed or mistimed measurements
  - Range of correlation structures

5

Differences Between GEE and Mixed Models

- Mixed models can fit multiple levels of correlations
  - Ex., longitudinal data from children clustered within schools
- GEE, as implemented in software, is generally restricted to one level of correlation
- Mixed models fit subject-specific models – GEE fit marginal models (population average).
- Mixed models require normality assumptions – GEE allow for weaker distributional assumptions.

Differences Between GEE and Mixed Models

- Mixed models require assumption that you have correctly specified the correlation structure – GEE do not.
  - GEE provide consistent (i.e., asymptotically unbiased) parameter and standard error estimates even if you do not pick the correct correlation structure.
    - Requires number of clusters (or subjects in RCT) to be large (e.g., >50 is probably ok, >100 better)
    - As long as “robust” standard errors are used, not “model-based” standard errors
    - However, choosing a correlation structure that is closer to the truth improves efficiency of estimates

7

Example

- Intervention: American Heart Association 8-week School Program (among 3rd and 4th graders)
- Is the intervention effective to reduce proportion of kids with high systolic BP?
  - Y =HiBP
  - T = Treatment (0=Control, 1=Intervention)

- New: 3 study visits
  1. Visit 1 – baseline, pre-randomization
  2. Visit 2 – 8 weeks following start of intervention
  3. Visit 3 – 1 Year post-intervention

8

“Long” Data Structure

- Reminder, GEE requires “long” data structure
  - One record per observation per subject
  - True for all standard statistical packages

STID	VISIT	HiBP	T	SEX	Other VARs
105	1	0	1	F	…
105	2	0	1	F	…
105	3	1	1	F	…
202	1	0	0	M	…
202	2	0	0	M	…
202	3	0	0	M	…
…					

9

Summary Stats

T = 0 (Control)			
	HiBP		
Visit	0	1	Total
1	291	16 (5.2%)	307
2	261	46 (15.0%)	307
3	256	51 (16.6%)	307

T = 1 (Intervention)			
	HiBP		
Visit	0	1	Total
1	247	12 (4.6%)	259
2	243	16 (6.2%)	259
3	237	22 (8.5%)	259

No missing data here (unrealistic!), but GEE would have still worked fine even if there had been.

Questions of Interest

- Primary question: Is intervention effective at reducing proportion with high BP over 8 weeks?
  1. Re-expressed: Do the groups differ wrt change in proportion with high BP from baseline to post-intervention (V1 to V2), controlling for other important variables?

Questions of Interest

- Secondary questions:
  2. Do the groups differ wrt change from baseline to 1 year following intervention (V1 to V3), controlling for other important variables?

  3. Do the groups differ wrt change from immediately post-intervention to 1 year later (V2 to V3), controlling for other important variables?

Review of Logistic Regression Model

- We'll start by reviewing Mario's model:

$$\text{logit}(\theta) = \beta_0 + \beta_1 * T + \beta_2 * B\_HiBP$$
$$+ \beta_3 * METSUM + \beta_4 * B\_AGE + \beta_5 * MALE + \beta_6 * URBAN$$

where θ = Pr{ HiBP = 1 | T, B_HiBP, … }

logit(θ) = log[θ / (1- θ)] = log odds of high BP

$$\text{odds} = \exp\{ \beta_0 + \beta_1 * T + \beta_2 * B\_HiBP + \dots \}$$

13

Longitudinal Logistic Regression Model

- Now consider adding terms for additional visits:

$$\text{logit}(\theta) = \beta_0 + \beta_1 * T + \beta_2 * V2 + \beta_3 * V3 + \beta_4 * T * V2 + \beta_5 * T * V3$$

$$+ \beta_6 * METSUM + \beta_7 * B\_AGE + \beta_8 * MALE + \beta_9 * URBAN$$

14

Interpretation of Logistic Parameters

T	Visit	Odds
0	1	$\exp\{ \beta_0 \}$
0	2	$\exp\{ \beta_0 + \beta_2 \}$
0	3	$\exp\{ \beta_0 + \beta_3 \}$
1	1	$\exp\{ \beta_0 + \beta_1 \}$
1	2	$\exp\{ \beta_0 + \beta_1 + \beta_2 + \beta_4 \}$
1	3	$\exp\{ \beta_0 + \beta_1 + \beta_3 + \beta_5 \}$

- $\exp\{ \beta_2 \}$ is the OR comparing V2 to V1 for T = 0
- $\exp\{ \beta_2 + \beta_4 \}$ is the OR comparing V2 to V1 for T = 1
- $\exp\{ \beta_4 \}$ is the ratio of these ORs, which compares change from V1 to V2 for intervention group to that for control

15

Relating Model Parameters to Questions

1. Do the groups differ wrt change in proportion with high BP from baseline to post-intervention (V1 to V2)?

   H_0: $\exp\{ \beta_4 \} = 1$ $\longleftrightarrow$ H_0: $\beta_4 = 0$

2. … from V1 to V3?

   H_0: $\exp\{ \beta_5 \} = 1$ $\longleftrightarrow$ H_0: $\beta_5 = 0$

3. … from V2 to V3?

   H_0: $\exp\{ \beta_5 - \beta_4 \} = 1$ $\longleftrightarrow$ H_0: $\beta_5 - \beta_4 = 0$

16

Where Does GEE Come In?

- So far, everything we've talked about has been a review of standard logistic regression
- So, where does GEE come in?
- Keep in mind that the goal of GEE is to account for within-subject correlations…

GEE: Working Correlation Structure

- In GEE, you choose a "working" correlation structure:
  - Available choices: compound symmetry (exchangeable), unstructured, few more
  - In a sense, choice doesn't matter in large samples (e.g., > 100 clusters) because parameter estimates will be unbiased regardless
  - Also provides unbiased estimates of parameters' standard errors

Let's Fit the Model!

- Stata: XTGEE command

```
xtgee   dependent var   independent vars ,
                i( stid )
              link( logit )
            family( binomial )
          corr( exchangeable )
                robust
```

19

Let's Fit the Model!

- SPSS: GENLIN command (or use point and click)

```
link=logit
distribution=binomial
/ repeated subject=stid
corrtype=exchangeable
```

Analyze → Generalized Linear Models → GEE

20

Model with Compound Symmetric Working Correlation

Working Correlation Matrix			
	Col1 (Visit 1)	**Col2 (Visit 2)**	**Col3 (Visit 3)**
Row1 (Visit 1)	1.00	0.28	0.28
Row2 (Visit 2)	0.28	1.00	0.28
Row3 (Visit 3)	0.28	0.28	1.00

21

Model Parameter Estimates

Parameter	Estimate	SE	Pr > |Z|
Intercept	β_0 = -6.20	1.27	<.001
trt	β_1 = -0.11	0.39	0.773
Vis2	β_2 = 1.18	0.24	<.001
Vis3	β_3 = 1.30	0.27	<.001
trt*Vis2	β_4 = -0.87	0.39	0.027
trt*Vis3	β_5 = -0.65	0.41	0.116
metsum	β_6 = 0.01	0.01	0.160
b_age	β_7 = 0.36	0.14	0.012
male	β_8 = -0.10	0.22	0.647
urban	β_9 = -0.50	0.21	0.018

Answering Our Questions of Interest

1. Do the groups differ wrt change in proportion with high BP from baseline to post-intervention (V1 to V2)?
   - Recall: OR comparing V2 to V1 for T=0

     $\exp\{ \beta_2 \} = \exp\{ 1.18 \} = 3.25$
   - OR comparing V2 to V1 for T=1

     $\exp\{ \beta 2 + \beta 4 \} = \exp\{ 1.18 - 0.87 \} = 1.36$
   - Ratio of these two ORs compares change btw groups

     $\exp\{ \beta 4 \} = \exp\{ - 0.87 \} = 0.42$

23

Answering Our Questions of Interest

Question	Estimated ratio of ORs	95% CI	p-value
1. V1 to V2	$\exp\{ \beta_4 \} = 0.42$	(0.19, 0.90)	0.027
2. V1 to V3	$\exp\{ \beta_5 \} = 0.52$	(0.23, 1.17)	0.116
3. V2 to V3	$\exp\{ \beta_5 - \beta_4 \} = 1.25$	(0.64, 2.44)	0.511

Model with Unstructured Working Correlation

Working Correlation Matrix			
	Col1 (Visit 1)	**Col2 (Visit 2)**	**Col3 (Visit 3)**
Row1 (Visit 1)	1.00	0.35	0.22
Row2 (Visit 2)	0.35	1.00	0.27
Row3 (Visit 3)	0.22	0.27	1.00

25

Answering Our Questions of Interest

Question	Estimated ratio of ORs	95% CI	p-value
1. V1 to V2	$\exp\{ \beta_4 \} = 0.42$	(0.19, 0.90)	0.026
2. V1 to V3	$\exp\{ \beta_5 \} = 0.52$	(0.23, 1.17)	0.114
3. V2 to V3	$\exp\{\beta_5 - \beta_4 \} = 1.25$	(0.64, 2.44)	0.512

Virtually identical to those for the model with the compound symmetric working correlation matrix

Concluding Remarks

- For our example, final conclusions did not depend on the chosen working correlation matrix
- But, sometimes they will!
- Suggested approach for GEE: prior to analysis, specify a reasonable form for the working correlation matrix in the analysis plan and stick with it
  - To avoid biasing (or appearance of biasing) results

27