Using Generalized Estimating Equations for Longitudinal Data Analysis

Presented by Ziyi Li

Department of Biostatistics and Bioinformatics
Ziyi.li@emory.edu

March 29, 2016

Author

Using Generalized Estimating Equations for Longitudinal Data Analysis

Published on *Organizational Research Methods*, April 2004

Gary A. Ballinger Ph.D, Organizational Behavior, Purdue University
Mcintire school of commerce at University of Virginia

Content

1. **Why we need Generalized Estimating Equations (GEE)?**
   - Motivating example
   - Limited-range dependent variables
   - Correlation of response
2. **What is GEE?**
3. **How to use GEE?**
   - Step 1: Specify link function.
   - Step 2: Specify the distribution of the outcome variables.
   - Step 3: Specify the form of correlation of responses within subjects or nested within group in the the sample.
4. **Two examples of using GEE to analyze data.**
   - Longitudinal data with counted responses
   - Normally distributed responses & Correlated within branch offices

Why we need Generalized Estimating Equations (GEE)? A motivating example

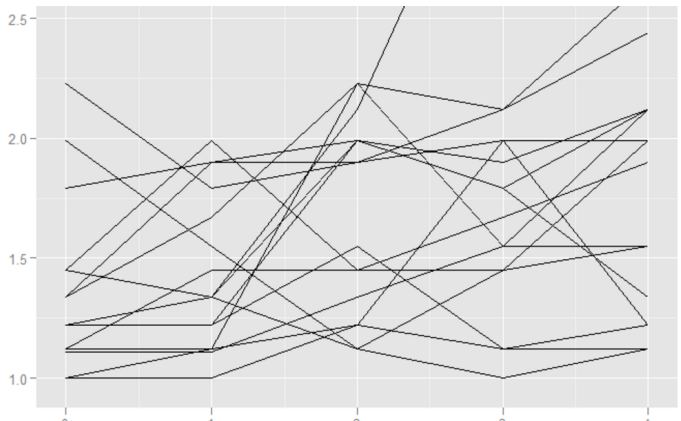
A laboratory experiment involving groups assembling Lego

- 52 groups through four consecutive session ;
- 1 min to view four objects and select one object to assemble;
- One person is allowed to go out of the room and view object during assembly task;
- Response of interest: the number of trips out of the room (Counted data);
- Covariates: Object(old or new), Time, Group satisfaction, Group size

Why we need Generalized Estimating Equations (GEE)? A motivating example

A laboratory experiment involving groups assembling Lego

- 52 groups through four consecutive session ;
- 1 min to view four objects and select one object to assemble;
- One person is allowed to go out of the room and view object during assembly task;
- Response of interest: the number of trips out of the room (Counted data);
- Covariates: Object(old or new), Time, Group satisfaction, Group size



Why we need Generalized Estimating Equations (GEE)? Limited-range dependent variables

Different kinds of response variables appear in organizational research:

- Binary variable: absenteeism (do I show up to work today or not?)
- Counted variable: number of patents received by different firms
- Multi-categorized data:

Why we need Generalized Estimating Equations (GEE)? Limited-range dependent variables

Different kinds of response variables appear in organizational research:

- Binary variable: absenteeism (do I show up to work today or not?)
- Counted variable: number of patents received by different firms
- Multi-categorized data:

!! Ordinary Least Square (OLS) regression cannot be used if normality assumption is not met.

Why we need Generalized Estimating Equations (GEE)? Correlation of responses

- Repeated responses may be correlated within a subject over repeated measures or within a cluster of observations in a particular group.

	Jan	Feb	Mar	Apr	May
Jan	1				
Feb	0.45	1			
Mar	0.25	0.48	1		
Apr	0.2	0.25	0.43	1	
May	0.1	0.15	0.25	0.45	1

- Consequences of ignoring correlations between responses in analysis:
  1. Making ncorrect inferences about the regression coefficients;
  2. Having inefficient or biased estimates of the regression coefficients.
  3. Efficiency losses were large as correlation increased, as the asymptotic relative efficiency of parameter estimates assuming independence fell to approximately 40% for within cluster correlations of .5 or more.

What is GEE?

- Response $Y = (Y_{ij})$ for each subject i, measured at different occasions $j = 1, 2, \ldots, n_i$.
- $X = (X_1, X_2, \ldots, X_k)$ be a set of explanatory variables which can be discrete, continuous, or a combination. X_i is $n_i \times k$ matrix of covariates.
- GEE models the expected value of the marginal response for the population $\mu_i = E(y_i)$.
- A transformation function that allows the response to be expressed as a vector of parameter estimates in the form of an additive model : $g(\mu_i) = X_i^T \beta$

How to use GEE?
Step 1: Specify link function.

- Selection of link function depends on the distribution of the underlying dependent variable and how the user wishes to interpret the coefficients.
- Most commonly used link function:
  1. **Normal Distribution** Identity link
  2. **Binomial Distribution** Logit link and Probit link
  3. **Possion Distribution**(Counted Data) Log link
  4. **Negative Binomial Distribution** Power link

  See a complete list of choosing link functions for different distributions in Appendix.

How to use GEE?
Step 2: Specify the distribution of the outcome variables.

- GEEs permit specification of distributions from the exponential family of distributions, which includes normal, inverse normal, binomial, Poisson, negative binomial, and Gamma distributions.
- User should make every reasonable effort to correctly specify the distribution of the response variable so that the variance can be efficiently calculated as a function of the mean and the regression coefficients can be properly interpreted.
- Example: Poisson distribution / Negative bionomial distribution

How to use GEE?
Step 3: Specify the form of correlation of responses within subjects or nested within group in the the sample.

- The goal of selecting a working correlation structure is to estimate β more efficiently .
- Incorrect specification of the correlation structure can affect the efficiency of the parameter estimates.

Independence - (correlation between time points is independent)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Exchangable (or **Compund Symmetry**)

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

AutoRegressive Order 1 (AR 1)

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Unstructured

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

Two examples of using GEE to analyze data
Longitudinal data with counted responses

- The responses are not normally distributed because they consist of a count of the number of trips out of the room. (Possible distributions used: normal distribution, binomial distribution, poisson distribution.)
- There are four responses within subject, and they are not independent (they are correlated with each other). (Possible working correlation matrix used: independent correlation, one-period autoregressive correlation, unstructured correlation.)
- There are time-dependent covariates. Group satisfaction changed over the course of the study, and the groups were allowed to select a new object at each trial.

Two examples of using GEE to analyze data
Longitudinal data with counted responses

Table 2
Comparisons of Generalized Estimating Equation (GEE) Regressions for Example 1 ($N = 52$ Groups)

	Method									
	Ordinary Least Squares, Normal Distribution, Independent Correlation		*Logistic, Binomial Distribution, Independent Correlation*		*GEE, Poisson Distribution, Independent Correlation*		*GEE, Poisson Distribution, Unstructured Correlation*		*GEE, Poisson Distribution, One-Period Autoregressive Correlation*	
	Unstandardized Coefficient	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE
Object	–2.01	.29***	–.32	.14*	–.29	.10**	–.61	.22**	–.29	.10**
New choice	–1.49	.26***	–.10	.08	1.29	.17***	.97	.17***	1.17	.17***
Trial	3.52	.36***	.36	.16*	–.20	.14	–.35	.15*	–.20	.13
Size	.26	.28	.04	.09	.05	.09	.18	.07*	.04	.09
Group satisfaction	–.54	.23*	–.00	.11	–.02	.09	–.09	.10	–.06	.08
Trial × Object	.31	.10**	.05	.04	–.04	.06	–.00	.08	–.05	.06
Constant	9.41	2.19	–.32	.99	1.42	.74	1.97	.75	1.72	.72
R^2_{Marg}					.69		.69		.62	
Wald χ^2 (6 *df*)	788.07		26.32		408.61		522.37		388.54	
Dispersion	4.25		1.22		1.84					
Pan statistic					0.00		18.29		24.85	
Mean fitted	2.82				2.82					
Minimum fitted	–1.06				0.22					

Note. Binomial model failed to converge after 100 iterations.
*$p < .05$. **$p < .01$. ***$p < .001$.

Two examples of using GEE to analyze data
Longitudinal data with counted responses

Table 4
Comparison of Within-Group Correlation Estimates for Example 1

	Trial 1	*Trial 2*	*Trial 3*	*Trial 4*	*Trial 5*
Unstructured correlation					
Trial 1	1				
Trial 2	.17	1			
Trial 3	.24	.06	1		
Trial 4	.20	.35	.28	1	
Trial 5	.38	.66	.13	.22	1
One-period autoregressive correlation					
Trial 1	1				
Trial 2	.15	1			
Trial 3	.02	.15	1		
Trial 4	.00	.02	.15	1	
Trial 5	.00	.00	.02	.15	1

Two examples of using GEE to analyze data
Normally distributed responses & Correlated within branch offices

The correlation within clusters is estimated to be quite small: In this case, it is estimated by the GEE model as .0244.

Table 5
Comparison of Regression Results for Example 2

	Parameter							
	OLS, Naïve SE		*OLS, Robust* SE		*GEE, Independence*		*GEE, Exchangeable*	
	Unstandardized Coefficient	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE	*Unstandardized Coefficient*	SE
Supervision	.042	.051	.042	.055	.042	.054	.040	.055
Pay	.235	.040***	.235	.044***	.235	.039***	.228	.039***
Growth	.133	.063*	.133	.073	.133	.068*	.138	.067*
Security	.146	.048**	.146	.054**	.146	.059*	.152	.058*
Constant	1.212	.201	1.212	.239	1.212	.281	1.212	.282
R^2 (R^2_{Marg})	.2454		.2454		.2454		.2451	
$F(4, 406)$	33.01		30.51					
Wald χ^2 (4 *df*)					75.14		72.82	

Note. OLS = ordinary least squares. $N = 50$ hospitals; 411 employees.
*$p < .05$. **$p < .01$. ***$p < .001$.

Cautions regarding GEE

1. The estimate of the variance produced under GEE models could be highly biased when the number of subjects within which observations are nested is small.
2. Goodness-of-fit statistics for GEEs that would function as the equivalent to measures such as the magnitude of the squared differences of observed versus predicted values or dispersion measures are not widely accepted for most classes of dependent variables beyond binary data or for different correlation structures. (Comment: QIC proposed by Pan(2001) is implemented in SAS and R, and is used frequently for GEE model selection now.)
3. Comparing with generalized linear mixed model (GLMM).

Thank you!