Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Bureau of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625901734-5, Fax: +6625918581, Email: osireditor@osirjournal.net, http://www.osirjournal.net

Grammar of Science: Gee Whiz… It's GEE!

Jaranit Kaewkungwal*

Mahidol University, Thailand

* Corresponding author, email address: jaranit.kae@mahidol.ac.th

"I have one to many children in a house!" A public health officer looks so worried.

"What is the problem with that?"

"I did home visits in a community, to the household with a tuberculosis (TB) case; and I want to know whether the TB case will be the source of disease transmission to the children under five years old within his/her house or not. But each house has different number of children: some houses only one child, others vary 2-5 children. I even find a house with 10 children. I performed tuberculin skin test in all children – and some of them are positive while other negative, though they are living in the same house. If the children in a house are called household contacts and the TB patient is called an index case – then, how can I estimate the risk of acquiring infection from the index case among the household contacts?"

"This is called 'clustered data' structure. There are many ways to analyze clustered data. One of the popular statistical methods that can handle this type of data is 'Generalized Estimation Equation', so-called GEE. This clustered data cannot be analyzed by standard statistical models like linear regression, logistic regression, etc. The main reason is that the outcomes measured from each individual are considered "not independent", but potentially "correlated", among the individuals (household contacts) who share the same exposures (index case and other household characteristics). Let's take a look in more detail."

What kinds of data can be used in GEE model?

GEE is a statistical method that can be applied for "clustered data" and "repeated measures data".[1-4] When we talk about these two types of data structure, they are the "multivariate" datasets, meaning that there are more than one outcome observations (Y's) per case/unit, which is different from the "univariate" datasets with only one outcome observation (Y) per case/unit (Figure 1). Repeated measures data structure refers to the sets of data when we have repeated observations of an outcome variable measured from each individual (case) over time on multiple visits (Y's of an individual at different times: Yt1,Yt2,Yt3). Clustered data structure refers to the sets of data when outcome observation of different individuals (Y's) are grouped (or nested) within a certain unit (subgroup/cluster). The study may have either one exposure variable (X) or more than one exposure variables (Xs). The statistical method with >1 Xs is called "multi-variables" analysis.

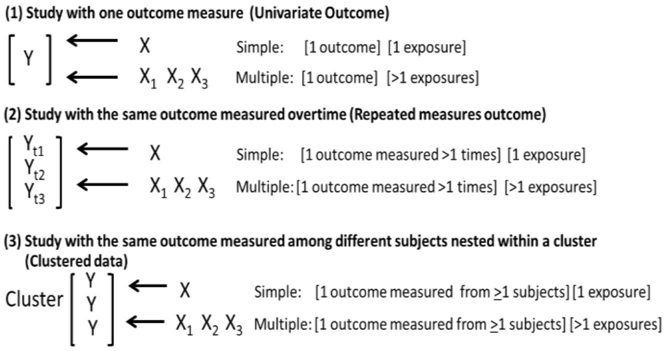
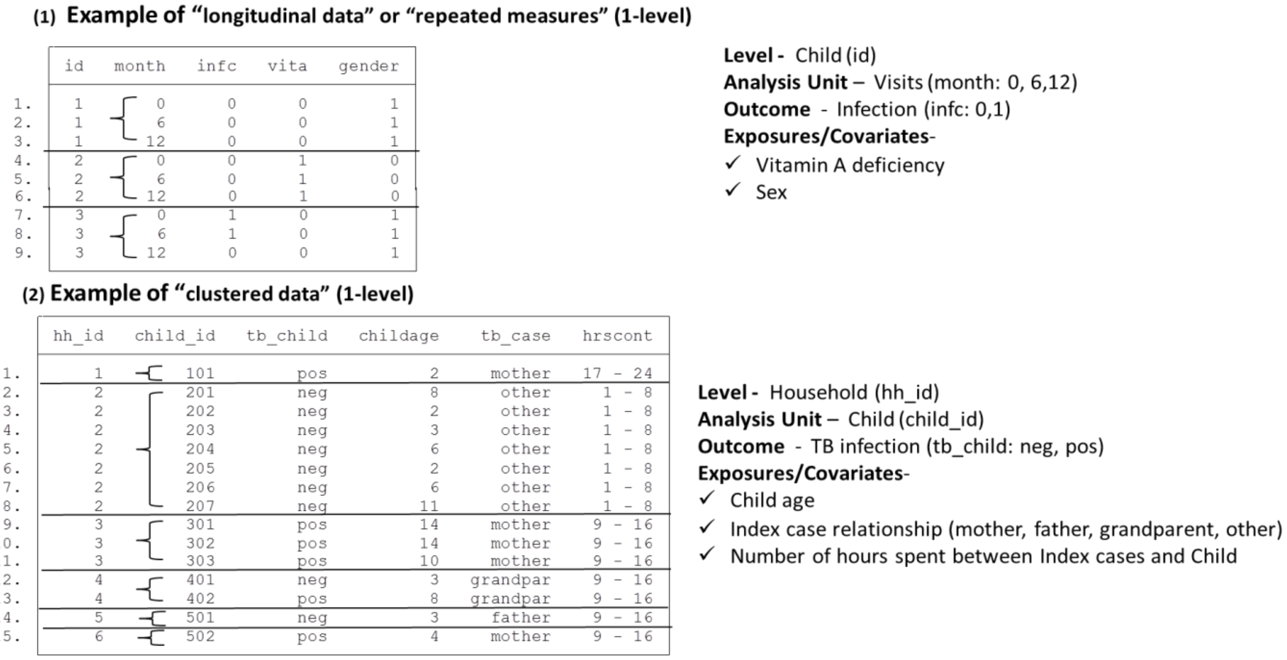

(1) Study with one outcome measure (Univariate Outcome)
Y
X
X1 X2 X3
Simple: [1 outcome] [1 exposure]
Multiple: [1 outcome] [>1 exposures]
(2) Study with the same outcome measured overtime (Repeated measures outcome)
Yt1
Yt2
Yt3
X
X1 X2 X3
Simple: [1 outcome measured >1 times] [1 exposure]
Multiple: [1 outcome measured >1 times] [>1 exposures]
(3) Study with the same outcome measured among different subjects nested within a cluster (Clustered data)
Cluster
Y
Y
Y
X
X1 X2 X3
Simple: [1 outcome measured from ≥1 subjects] [1 exposure]
Multiple: [1 outcome measured from ≥1 subjects] [>1 exposures]

Figure 1. Univariate vs. multivariate data structure

Repeated data structure is shown as an example in figure 2 (a). In the study to determine the association between vitamin A deficiency and respiratory infection in school children, the researchers collected data on respiratory infection from each student at three time points (Months 0, 6, 12). Clustered data structure is shown in figure 2 (b) for the study of the public health officer in which he collected data on TB infection among all household contacts within each house. In fact we can say that, in repeated measures, study outcome data are clustered (repeated) within an individual; and in clustered or nested study, outcome data of individuals are clustered within a certain unit. These are the examples of only one level of clustering. It is also possible to have a multilevel data structure, in which we have multiple levels of grouping units, for example: children are clustered



(1) **Example of "longitudinal data" or "repeated measures" (1-level)**

```
      id   month   infc   vita   gender
  1.   1       0      0      0        1
  2.   1       6      0      0        1
  3.   1      12      0      0        1
  4.   2       0      0      1        0
  5.   2       6      0      1        0
  6.   2      12      0      1        0
  7.   3       0      1      0        1
  8.   3       6      1      0        1
  9.   3      12      0      0        1
```

Level - Child (id)
Analysis Unit – Visits (month: 0, 6,12)
Outcome - Infection (infc: 0,1)
Exposures/Covariates-
✓ Vitamin A deficiency
✓ Sex

(2) **Example of "clustered data" (1-level)**

```
      hh_id   child_id   tb_child   childage   tb_case    hrscont
  1.      1        101        pos          2    mother    17 - 24
  2.      2        201        neg          8     other     1 - 8
  3.      2        202        neg          2     other     1 - 8
  4.      2        203        neg          3     other     1 - 8
  5.      2        204        neg          6     other     1 - 8
  6.      2        205        neg          2     other     1 - 8
  7.      2        206        neg          6     other     1 - 8
  8.      2        207        neg         11     other     1 - 8
  9.      3        301        pos         14    mother     9 - 16
 10.      3        302        pos         14    mother     9 - 16
 11.      3        303        pos         10    mother     9 - 16
 12.      4        401        neg          3  grandpar     9 - 16
 13.      4        402        pos          8  grandpar     9 - 16
 14.      5        501        neg          3    father     9 - 16
 15.      6        502        pos          4    mother     9 - 16
```

Level - Household (hh_id)
Analysis Unit – Child (child_id)
Outcome - TB infection (tb_child: neg, pos)
Exposures/Covariates-
✓ Child age
✓ Index case relationship (mother, father, grandparent, other)
✓ Number of hours spent between Index cases and Child

Figure 2. Examples of repeated measures datasets and clustered datasets

within a classroom (level 1), and classrooms are clustered in a school (level 2), and so on.

What is GEE?

GEE was proposed by Liang K-Y and Zeger SL in 1986[5]. GEE is a generalized model unifying in a single method. The model of GEE can be transformed into three classic generalized linear models (GLM): linear, logistic and poisson depending on the type of the outcome (Y) variable.[1,2,6]

- Linear regression (continuous outcome)
    - Distribution of Y: ~ Normal; mean of Y is μ, average of the outcome
    - Transformation of Y: none (identity link)
    - Equation $\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
- Logistic regression (binary outcome)
    - Distribution of Y:~ Bernoulli; mean of Y is p, probability of having outcome
    - Transformation of Y: logit link
    - Equation: $logit(p) = log\ (Odds) = log\ (p\ /1\text{-}p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
- Poisson regression (incidence or count outcome)
    - Distribution of Y ~ Poisson, mean of Y is λ, rate per time unit, or mean count per unit, of the outcome events
    - Transformation of Y: *log link*
    - Equation: $log(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$

The three classic GLM models are based univariate data. In GLM, an outcome variable (Y) is measured for each individual, and thus, the Y's for all individuals in the study are considered "independent". In contrast, GEE models are based on multivariate data where outcome variables are potentially "correlated" because Y's are measured within the same individual (for repeated measures) or among different individuals within the same exposure variable(s) (for clustered data). GEE thus simply extends such GLM models by taking into account the correlated Ys within a case (of repeated measure) or a grouping unit (cluster). If the researchers did not take into account the correlation among Ys, the estimated regression coefficients (βs) will be less efficient (i.e., widely scattering around the parameters or true population values estimated)[7].

How does GEE model fit the data?

In fitting the extended regression model, GEE uses quasi-likelihood estimation method to estimate the expected (predicted) value of the outcome, [E(Y)], via the consistent estimates of regression coefficients, β of Xs - [$g(\beta_i X_i)$], and its variance-covariance (correlations) among Ys.[5,8]

$E(Y) = g(\beta_i X_i)$, $Var(Y) = Corr(Y_{ij}, Y_{ik})$ for subject i^{th} and $j\text{-}k^{th}$ times/units

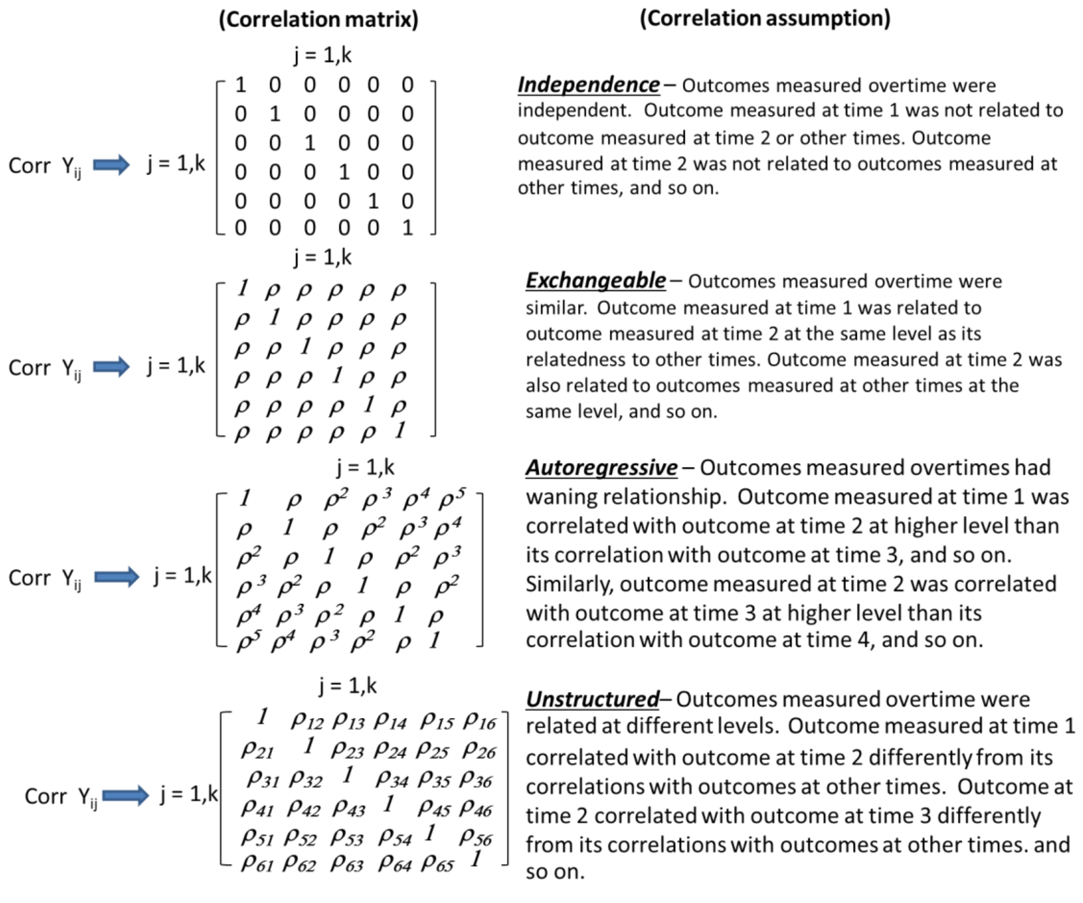
Liang and Zeger (1986) proposed GEE under the asymptotic theory in which they utilize outcome values across study subjects to estimate a "working correlation" matrix, assuming that such correlations are explicitly accounted for the time dependence or the clustering effect, and to achieve greater asymptotic efficiency[5]. To explain asymptotic theory in layman terms, it means "a large sample theory" which is typically used when estimating any parameters or statistical tests based on the assumption that the sample size would grow indefinitely ($n \rightarrow \infty$)[9]. That means GEE fits better when the sample size is getting larger.

GEE is considered as a semi-parametric model as it estimates parameters (β coefficients) in the equation without full specification of the joint distribution of the outcome observations overtime or within clusters. The model derives from the specification of the likelihood for the (univariate) marginal distributions of the outcome variables (Ys) and then incorporates the "working correlation" matrix into the model.[4] In other words, there are three steps in modeling GEE. The three steps are: (1) a naive regression analysis is carried out, assuming the outcome observations within the individual/cluster are independent; (2) the residuals (observed - predicted) are then calculated from the naive model, and used to estimate the working correlation matrix; and (3) the regression coefficients are subsequently refitted using iterative process by treating the within-subject correlation as a nuisance (covariate) variable.[10]

The "working correlation" matrix is based on an important assumption that the outcome observations (Ys) measured over time or within individual/unit are correlated or clustered. That means observations (Y at time 1, 2, 3, ... of each individual; or Y of individuals within each unit) are not independent[3]. There are typically four types of correlation structures that we have to assume prior to fitting the model. Figure 3 presents structure and assumption of each type of correlation matrix[3,6,7].

In analyzing the clustered data, we will typically have an outcome response measured from each study subject within a cluster/unit, and thus, there is usually no problem with missing outcome data. But in the repeated measures situation, there are always study subjects who missed some visits and thus, outcome data are missing.

In analyzing the repeated measures data with missing outcome values at different visits, GEE uses the pairwise method (i.e., "all available pairs"); all non-missing pairs of data are used in estimating the working correlations. That means we do not lose the study subjects that had missing outcome data at certain visit(s)[10]. There is no need to perform imputation for the missing data. However, GEE with robust and optimal option was developed to handle missing data that are either missing at random (MAR) or missing not at random (MNAR)[1].



(Correlation matrix)

(Correlation assumption)

$$\text{Corr } Y_{ij} \Rightarrow \begin{array}{c|cccccc} & & & j = 1,k & & & \\ \hline j = 1,k & 1 & 0 & 0 & 0 & 0 & 0 \\ & 0 & 1 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 1 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 1 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 1 & 0 \\ & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Independence – Outcomes measured overtime were independent. Outcome measured at time 1 was not related to outcome measured at time 2 or other times. Outcome measured at time 2 was not related to outcomes measured at other times, and so on.

$$\text{Corr } Y_{ij} \Rightarrow \begin{array}{c|cccccc} & & & j = 1,k & & & \\ \hline j = 1,k & 1 & \rho & \rho & \rho & \rho & \rho \\ & \rho & 1 & \rho & \rho & \rho & \rho \\ & \rho & \rho & 1 & \rho & \rho & \rho \\ & \rho & \rho & \rho & 1 & \rho & \rho \\ & \rho & \rho & \rho & \rho & 1 & \rho \\ & \rho & \rho & \rho & \rho & \rho & 1 \end{array}$$

Exchangeable – Outcomes measured overtime were similar. Outcome measured at time 1 was related to outcome measured at time 2 at the same level as its relatedness to other times. Outcome measured at time 2 was also related to outcomes measured at other times at the same level, and so on.

$$\text{Corr } Y_{ij} \Rightarrow \begin{array}{c|cccccc} & & & j = 1,k & & & \\ \hline j = 1,k & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{array}$$

Autoregressive – Outcomes measured overtimes had waning relationship. Outcome measured at time 1 was correlated with outcome at time 2 at higher level than its correlation with outcome at time 3, and so on. Similarly, outcome measured at time 2 was correlated with outcome at time 3 at higher level than its correlation with outcome at time 4, and so on.

$$\text{Corr } Y_{ij} \Rightarrow \begin{array}{c|cccccc} & & & j = 1,k & & & \\ \hline j = 1,k & 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} \\ & \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} \\ & \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} & \rho_{36} \\ & \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} & \rho_{46} \\ & \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 & \rho_{56} \\ & \rho_{61} & \rho_{62} & \rho_{63} & \rho_{64} & \rho_{65} & 1 \end{array}$$

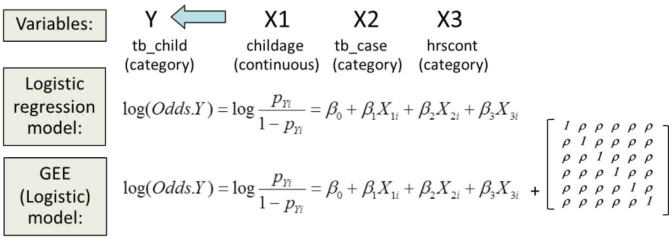
Unstructured– Outcomes measured overtime were related at different levels. Outcome measured at time 1 correlated with outcome at time 2 differently from its correlations with outcomes at other times. Outcome at time 2 correlated with outcome at time 3 differently from its correlations with outcomes at other times. and so on.

Figure 3. "Working correlation" matrix

Case study of GEE

For the case scenario of the public health officer, the GEE model to be fitted is for the clustered categorical outcome (not having or having TB infection). We now consider to fit the logistic regression model with binary outcome data (Y=0,1). Note that, similar to all regression models, the exposures (X's) can be categorical or continuous data. In this study, the exposures are child's age, types of TB index case (father, mother, grandparent, other) and duration of contact/exposure (1-8, 9-16, 17-24 hours per day). As shown in figure 4, the GEE model to be fitted is the extended logistic regression with correlated and clustered data (children residing in each household). While the working correlation matrix for a repeated measures study can be specified as one of the four structures, the appropriate working correlation matrix to be used for clustered data study is only exchangeable.

Based on the analysis of the data collected by the public health officer, the results are shown in figure 5. The goal of GEE is to make inferences about the population parameter(s) when accounting for the within-subject correlation. As GEE is the extended regression model, the interpretation of the model follows the regular regression model such that that for every one-unit increase in a covariate (X) across the population, how much the outcome response (Y) would change[7]. We can say that the odds that a child got infected with TB increases significantly by 2.3 and 4.9 times if the TB-case is the child's father and mother respectively, when compared to the odds of the reference group (TB case whose relationship with the child in "other" category). Compared children whose ages are different by one year, the odds seems to increase by 1.4 times, but is not statistically significant different. Notice the differences of the two logistic regression models, GEE (Figure 5) vs. GLM (Figure 6), the estimates of odds ratios and p-values are different.



Variables: Y ⟸ X1 X2 X3

tb_child (category) childage (continuous) tb_case (category) hrscont (category)

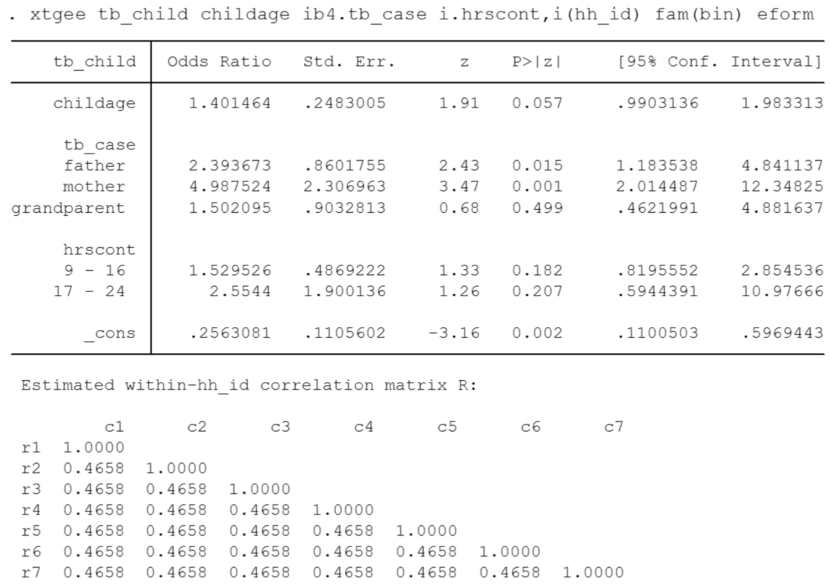
Logistic regression model:

$$\log(Odds.Y) = \log\frac{p_{Yi}}{1-p_{Yi}} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

GEE (Logistic) model:

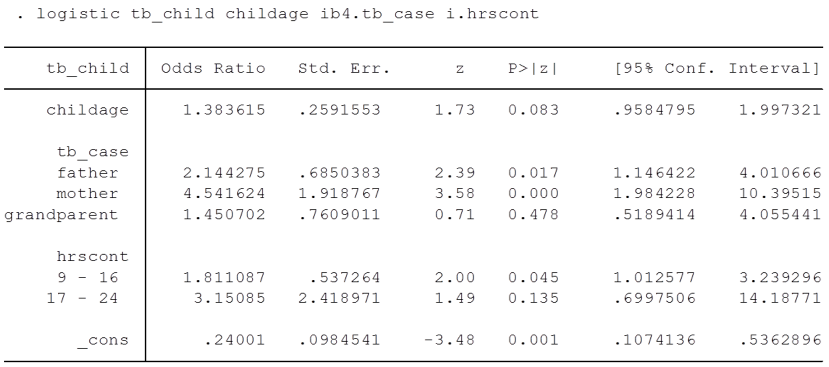
$$\log(Odds.Y) = \log\frac{p_{Yi}}{1-p_{Yi}} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \begin{bmatrix} 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

Figure 4. GEE (extended logistic regression) model



```
. xtgee tb_child childage ib4.tb_case i.hrscont,i(hh_id) fam(bin) eform

    tb_child   Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]

    childage     1.401464    .2483005     1.91   0.057     .9903136    1.983313

     tb_case
      father     2.393673    .8601755     2.43   0.015     1.183538    4.841137
      mother     4.987524    2.306963     3.47   0.001     2.014487    12.34825
 grandparent     1.502095    .9032813     0.68   0.499     .4621991    4.881637

     hrscont
      9 - 16     1.529526    .4869222     1.33   0.182     .8195552    2.854536
     17 - 24       2.5544    1.900136     1.26   0.207     .5944391    10.97666

       _cons     .2563081    .1105602    -3.16   0.002     .1100503    .5969443

 Estimated within-hh_id correlation matrix R:

         c1      c2      c3      c4      c5      c6      c7
 r1  1.0000
 r2  0.4658  1.0000
 r3  0.4658  0.4658  1.0000
 r4  0.4658  0.4658  0.4658  1.0000
 r5  0.4658  0.4658  0.4658  0.4658  1.0000
 r6  0.4658  0.4658  0.4658  0.4658  0.4658  1.0000
 r7  0.4658  0.4658  0.4658  0.4658  0.4658  0.4658  1.0000
```

Figure 5. Analysis of the case scenario with generalized estimation equation (GEE)



```
. logistic tb_child childage ib4.tb_case i.hrscont

    tb_child   Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]

    childage     1.383615    .2591553     1.73   0.083     .9584795    1.997321

     tb_case
      father     2.144275    .6850383     2.39   0.017     1.146422    4.010666
      mother     4.541624    1.918767     3.58   0.000     1.984228    10.39515
 grandparent     1.450702    .7609011     0.71   0.478     .5189414    4.055441

     hrscont
      9 - 16     1.811087     .537264     2.00   0.045     1.012577    3.239296
     17 - 24      3.15085    2.418971     1.49   0.135     .6997506    14.18771

       _cons       .24001    .0984541    -3.48   0.001     .1074136    .5362896
```

Figure 6. Analysis of the case scenario with binary logistic regression (GLM)

The GLM model considers the outcome of each record (household contact case) are independent while the GEE model takes into consideration that the outcomes of household contacts within the same house are correlated. In fact, if GEE model is fitted with working correlation matrix specified as "independent", we will get the same results as shown in GLM model.

How good is GEE?

In fact, another popular method that can be used to analyze repeated measures or clustered data is the "Multilevel Mixed Model" which handles within-subject variation in the regression model with random intercepts/slopes for each individual rather than using the "working correlation" matrix[11]. (We may talk about the Mixed model at other time.) Note that GEE provides the result as a generic equation applied to all in the population of the study, that is why it is called "marginal population average" model; but the Mixed model will provide the result as equations that are subject-specific[2,10,12]. Basically GEE generates the fix effect only. But when the question is to find out the variation of the effect between clusters AND within the clusters, then random effect model like the Mixed effect model could be used. The use of working correlation (or variance-covariance) matrix as a nuisance parameter in the equation has made fitting GEE model easier than Mixed model[1]. Both methods can handle missing data, time-varying covariates (exposures changed overtime or across individuals), irregularly-timed (timing of visits varied across individuals in repeated measures). GEE typically provides consistent estimates even if incorrect correlation structure is specified; but the Mixed model has assumption that the researchers should correctly specified the correlation structure, which is sometimes difficult in practice. GEE is not very strict with the distributional assumptions, but Mixed model requires normality assumptions.[10,12]

GEE is limited that it can handle only one level of correlation or cluster. In the example showed in figure 2, the observations are nested at one level (times/visits within each student, or children within household). However, the Mixed model can handle data nested within more than one level of clusters[10]. For example, malaria patients nested within villages, and villages nested within sub-district, and sub-district nested within district. If the researchers considered different layers of clustering, they need to use the Mixed model.

"Gee Whiz…. It is GEE to handle my clustered data…", the public health officer exclaims.

Suggested Citation

Kaewkungwal J. Grammar of science: Gee Whiz… It's GEE! OSIR. 2018 Dec;11(4):23-28.

References

1. Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments. Advances in Statistics. 2014;2014:11 pages [cited 2018 Nov 20]. <http://dx.doi.org/10.1155/2014/303728>.

2. Rodrıguez G Models for longitudinal and clustered data. 2012 Dec 6 [cited 2018 Nov 20]. <http://data.princeton.edu/wws509/notes/fixedRandom.pdf>.

3. Penn State University, Eberly College of Science. Introduction to generalized estimating equations [cited 2018 Nov 20]. <https://onlinecourses.science.psu.edu/stat504/node/180/>.

4. Hedeker D. GEE for longitudinal data analysis [cited 2018 Nov 20]. <https://bstt513.class.uic.edu/geeLS.pdf>.

5. Liang K-Y, Zeger S. Longitudinal data analysis using generalized linear models. Biomttrika. 1986;73(1):13-22.

6. Hill EG. An introduction to generalized estimating equations. 2008 Oct 16 [cited 2018 Nov 22]. <http://people.musc.edu/~hille/Presentations/GEE_tutorial_Betsy/GEE_Tutorial.pdf>.

7. Columbia University Mailman School of Public Health. Repeated measures analysis [cited 2018 Nov 22]. <https://www.mailman.columbia.edu/research/population-health-methods/repeated-measures-analysis>.

8. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol. 2003 Feb 15;157(4):364-75.

9. Hayashi P. Introduction to large sample theory. 2010 [cited 2018 Nov 22]. <http://froelich.vwl.uni-mannheim.de/fileadmin/user_upload/froelich/teaching/Ch2_Large_Sample_Theory.pdf>.

10. Sainani K. GEE and mixed models for longitudinal data [cited 2018 Nov 22]. <www.pitt.edu/~super4/33011-34001/33151-33161.ppt>.

11. Center for Multilevel Modeling, University of Bristol. Introduction to multilevel modeling [cited 2018 Nov 22]. <http://www.bristol.ac.uk/cmm/software/support/workshops/materials/multilevel-m.html>.

12. Weaver MA. Introduction to analysis methods for longitudinal/clustered data, part 3: generalized estimating equations. 2009 September [cited 2018 Nov 22]. <http://www.icssc.org/Documents/AdvBiosGoa/Tab%2007.00_GEE.pdf>.