HOME

MedicalBiostatistics.com

Generalized linear models

General linear models

GENERALIZED ESTIMATING EQUATIONS

The method of generalized estimating equations (GEE) is used to estimate the parameters of a model where there are several response (**dependent) variables** that are correlated and there may be several explanatory (**independent) variables**. Thus this is an extension of the **generalized linear models** to the setup where the responses are correlated. The **correlation** can be because same subjects are measured at several points of time (as in **longitudinal data**), each subject measured at several sites (such as a particular brain function at several locations in the brain), subjects sharing a common environment (such as those living together in a family), or any other such setup. In short, such data can be called clustered in the sense that they are more similar within the cluster than outside the cluster. The GEE method provides a framework for analyzing such correlated data. The method is comprehensive since the correlated data can be almost of any type—continuous (**Gaussian** or nonGaussian) or discrete (proportions or counts).What about rates and ratios? It can be adapted to estimate **fixed effects**, **random effects**, and mixed effects. The difference between general linear models, generalized linear models and generalized estimating equations is shown in the following table.

Table: Difference between general linear models, generalized linear models and generalized estimating equations (in all, the explanatory variables can be of almost any type—continuous or discrete—and may define fixed or random effects should they be linear combination?)

Type of response variable	Condition on responses	Method
Continuous, nearly Gaussian pattern	Uncorrelated	General linear models
Continuous (Gaussian or nonGaussian) or discrete (proportions or counts)	Uncorrelated	Generalized linear models
Continuous (Gaussian or nonGaussian) or discrete (proportions or counts)	Correlated	Generalized estimating equations

Consider cholesterol level of 200 members of 70 randomly selected families. Note first that the number of persons in different families would be different and second that *n* actually is 70 and not 200. You cannot calculate the **standard error (SE)** of the estimates with *n* = 200 in this case. Family is a cluster in this example. Because of common heritance and similar diet, the cholesterol values of members of the same family will be correlated. While the families are statistically **independent**, the individuals within the families are not. You can also have information on the sex of each person, his or her age, blood group, body mass index, physical exercise, etc. These could be the explanatory variables in this setup. These could be **discrete** or **continuous** or combination—that does not matter. The objective is to find the extent and form of linear relationship between the explanatory variables and the response variable. The explanatory variables could be like x^2 and log$x$ but the regression coefficients have to be linear. The response variable is the cholesterol level of the members of the families in this example. When the response variable is continuous as in this example so far, you would hope that clustered values jointly follow a multivariate Gaussian pattern so that the usual method of **general linear models** can be used to estimate the **regression coefficients** and to test **hypothesis** on them for their statistical **significance**. Beside the Gaussian pattern, this will also require an estimate of the **correlations** between the members of the same family and that these correlations are similar across families. The estimates of correlations will come from the sample values in this setup. Unequal clusters and the requirement of multivariate Gaussianity along with the estimate of the correlation structure and their homogeneity can be a challenge. In addition, in the generalized linear model setup, in place of actual cholesterol level, you may only have the information that it is within normal limit or is high. This will make the dependent variable **dichotomous**. This will require that a **logistic** kind of relationship be studied and not usual regression. However because of correlation among the family members, the usual logistic is not applicable. If there is no correlation, both these setups (and many others) can indeed be woven into a unifying method of generalized linear models. When correlations are present, these various setups can be studied by the GEE method.

Same setup arises when a characteristic is measured repeatedly over a period of time in a **longitudinal study**. For example, you may like to measure the pain score before and 1 minute, 2 minutes, 5 minutes, and 10 minutes after administering an anesthesia in a patient being prepared for a surgery. Some patients can be measured only 1 or 2 times and some others 3 or 4 times depending on how they respond to the anesthesia. The number of observations available on each subject may or may not be the same. But these pain scores on the same patients at different points in time will be correlated. The explanatory variables in this setup could be hemoglobin (Hb) level as a marker of nutrition, blood pressure, body mass index, etc. GEE method is the most commonly used method for analyzing such longitudinal data. Multiple observations on each patient at different points in time form a cluster in this example.

Since the mathematical details are too complex for medical professionals, we try to explain GEE method heuristically so that you have at least the elementary knowledge about this method. The strength of the method stems from (i) using only the mean (and to some extent variance) of the values but not any particular **distribution**—this makes it a semiparametric method, (ii) not requiring any joint or multivariate distribution of the clustered values, and (iii) not worrying much about the specific correlation structure. In fact, the correlation structure is considered kind of nuisance and it has been shown that the estimates obtained by GEE method are statistically *consistent* in the sense that as n increases they tend to become the actual value in the population with certainty, mostly even when the correlation structure is misspecified. This method is able to produce reasonably valid **standard errors** (SEs) of the estimates of the regression coefficients—thus believable **confidence intervals** can be obtained when the sample is truly representative. Unlike the usual methods of generalized linear models, the GEE method does not explicitly model between-cluster or within-cluster variation but directly models the mean response. Note that within-cluster variation is a kind of correlation that we have stated as a nuisance under this method.

The GEE method can be implemented by using an appropriate statistical package. But specifying it correctly for commands in the software is difficult. Also deciphering the output provided by the software can be a challenge. Thus, this method should not be used by inexperts.

The method requires that a working correlation structure for responses within clusters be specified, although the actual values of the correlations are not required (???). If you consider appropriate, you have the option to consider that there is no correlation. This would mean that the values within clusters are independent and reduces the GEE method to the usual generalized linear modeling. The other option is that the correlation between the first and second values within a cluster is the same as between the first and third values, and between the second and the fourth values, etc. This is called **exchangeable** or compound symmetry of the correlations. Third option is that the longitudinal values are **autocorrelated**—that is if the correlation between the first and second values is ρ, the correlation between the first and the third values is ρ^2, etc. Fourth option is to consider that different values within the cluster have different correlations, called unstructured. One of these structures is required to solve, what are called, generalized estimating equations. These equations are obtained by maximizing the likelihood without using the Gaussian distribution, called *quasi-likelihood*. As mentioned earlier, the choice of the correlation structure is not terribly important under GEE method but correct specification does help in getting more reliable estimates of the SEs.

Tang et al. [1] used GEE method to identify the determinants of quality of life during the dying process of terminally ill cancer patients who were longitudinally followed till death, and concluded that optimal quality of life during the dying process may be achieved by interventions designed to adequately manage physical and psychological symptoms, enhance social support, lighten perceived sense of burden to others, and facilitate experiences of posttraumatic growth. Van Rijn et al. [2] studied the effects of single or multiple concordant HPV infections at various anatomical sites (anal canal, penile shaft, and oral cavity) on type-specific HPV seropositivity by using logit link in GEE.

The GEE method was developed by Liang and Zeger [3]. For further details, see Hanley et al. [4]. For technical details, see Agresti [5].

[1] Tang ST, Chang WC, Chen JS, Su PJ, Hsieh CH, Chou WC. Trajectory and predictors of quality of life during the dying process: roles of perceived sense of burden to others and posttraumatic growth.*Support Care Cancer* 2014 May 28. [Epub ahead of print]. http://link.springer.com/article/10.1007%2Fs00520-014-2288-y

[2] van Rijn VM, Mooij SH, Mollers M, Snijders PJ, Speksnijder AG, King AJ, de Vries HJ, van Eeden A, van der Klis FR, de Melker HE, van der Sande MA, van der Loeff MF. Anal, penile, and oral high-risk HPV infections and HPV seropositivity in HIV-positive and HIV-negative men who have sex with men. *PLoS One* 2014 Mar 20;9(3):e92208. doi: 10.1371/journal.pone.0092208. eCollection 2014. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3961332/

[3] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986 (April);73:13-22. http://www.biostat.jhsph.edu/~fdominic/teaching/bio655/references/extra/liang.bka.1986.pdf

[4] Hanley JA, Abdissa Negassa A, deB. Edwardes MD, Forrester JE. Statistical Analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol* 2003;157:364-375. http://aje.oxfordjournals.org/content/157/4/364.full.pdf

[5] Agresti A. *Categorical Data Analysis*, Third Edition. Wiley 2013.