

Generalized Estimating Equations

Paul E. Johnson <pauljohn@ku.edu>

5th April 2004

Liang and Zeger (1986) proposed a modeling strategy that they called GEE, Generalized Estimating Equations. GEE is an extension of the "quasi-likelihood" approach to estimation that is inspired by Generalized Least Squares. Christopher Zorn (American Journal of Political Science, 2001 offers a very understandable survey with examples).

Liang and Zeger proved that the estimates from the GEE are consistent, meaning asymptotically accurate.

1 What is GEE good for?

My inclination is to get lost in details without explaining what all of this work is for. Let's avoid that.

GEE is useful because it can help you to build models of cross-sectional/time-series data in which the dependent variable is a "logit" or a "Poisson" or some other nonNormal thing.

GEE is an extension of FGLS to nonNormal distributions and it is also an extension of quasi-likelihood to deal with inter-correlated observations.

The only key requirements are that the observations are "clustered" in a meaningful, and that the observations within the clusters are intercorrelated in some substantively interesting way.

If the clustered observations form themselves into short time-series, we have longitudinal data.

The key thing is that the clusters, however they are considered, are stochastically independent. If you want to write a model in which observations are interrelated, the data should be conceptualized so that those observations are within a cluster.

That is to say, we are ruling-out the "contemporaneous correlation" across units that is assumed to be important in the panel-corrected standard error (Beck and Katz, 1995).

2 GLM

Consult my GLM1 handout in case this is unfamiliar to you.

The mean of y_i is hypothesized to be $E(y_i) = \mu_i = g^{-1}(X_i b)$. The so-called link function is $g(\mu_i)$ and $g^{-1}(X_i b)$ is the inverse link function. The inverse link function takes input from the linear predictor $\eta_i = X_i b$ and gives back a predicted value.

Assume the observations follow an exponential distribution.

$$f(y_i) = \exp \left\{ \frac{y_i \theta_i - c(\theta_i)}{\phi} + h(y_i, \phi) \right\} \quad (1)$$

The variance of observed y_i is hypothesized to be separable into two parts, one is a function of the μ_i , commonly called the variance function $V(\mu_i)$ and a constant "dispersion" or "scale" parameter ϕ :

$$\text{Var}(y_i) = \phi V(\mu_i)$$

Note that, right from the start, I'm adopting the simplification that a single dispersion parameter is shared among all observations. There's no subscript i on ϕ , in other words. Note that one must be cautious because authors vary in their notation. In some work, authors use a scale parameter $1/\phi$. But they use the symbol ϕ for it. (Liang and Zeger, 1986, use such notation.)

The first order condition for the likelihood function is a score equation. For the noncanonical link (the most general kind of model), as stated in McCullagh & Nelder (p. 41), MM&V (p. 331), or Dobson (p. 63)

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^N \frac{1}{V(\mu_i)} [y_i - \mu_i] \frac{\partial \mu_i}{\partial \eta_i} x_{ik} = 0 \quad (2)$$

which is the same as

$$\frac{\partial l}{\partial b_k} = U_k = \sum_{i=1}^N [y_i - \mu_i] \frac{\partial \theta_i}{\partial \eta_i} x_{ik} = 0 \quad (3)$$

In matrix form:

$$X'W(y - \mu) = 0 \quad (4)$$

The matrix W is $N \times N$ square, but it has only diagonal elements $\frac{1}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \theta_i}{\partial \eta_i}$. The fine points are described in depth in the handout GLM#1. The dispersion parameter ϕ disappears because, since it is a constant, it plays no role in equating the left and right hand sides.

3 Quasi-likelihood

What if the probability model for y_i is unknown? If one is willing to provide a function for the mean and variance of observations of y_i , then quasi-likelihood offers a good strategy to estimate the coefficients b .

Let the predicted mean (as a function of observed input variables X_i and parameters \hat{b}) be:

$$\hat{\mu}_i = g^{-1}(X_i \hat{b})$$

In complete generality, the variance of the observations is an $N \times N$ matrix V .

The proponents of quasi-likelihood estimators (following Wedderburn, 1974; Liang and Zeger 1986) propose that one should estimate b by solving a quasi-score equation:

$$D'V^{-1}(y - \hat{\mu}) = 0 \quad (5)$$

The inspiration for that is said to be found in Generalized Least Squares. See my handout CXTS1 which tries to draw out this comparison.

While the estimate \hat{b} is being calculated, it is interesting because b is an element in the formulas for D , V , and $\hat{\mu}$.

3.1 The D term

D' represents $[\partial \hat{\mu} / \partial \hat{b}]'$.

$$D = \begin{bmatrix} \frac{\partial \hat{\mu}_1}{\partial \hat{b}_1} & \frac{\partial \hat{\mu}_2}{\partial \hat{b}_1} & \dots & \frac{\partial \hat{\mu}_{N-1}}{\partial \hat{b}_1} & \frac{\partial \hat{\mu}_N}{\partial \hat{b}_1} \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{\partial \hat{\mu}_1}{\partial \hat{b}_p} & \frac{\partial \hat{\mu}_2}{\partial \hat{b}_p} & \dots & \dots & \frac{\partial \hat{\mu}_N}{\partial \hat{b}_p} \end{bmatrix}$$

If we have an estimate \hat{b} , we can calculate an estimate $\hat{\mu}_i$ for the i 'th case, and we can also calculate an estimate of D .

3.1.1 Digression: $D' = X' \text{diag}[\partial\hat{\mu}/\partial\hat{\eta}]$

D' can be simplified and rewritten in various ways.

The value of $\partial\hat{\mu}_i/\partial\hat{b}_k$ depends on the link function. Since, by definition

$$\partial\hat{\mu}_i/\partial\hat{b}_k = \partial\hat{\mu}_i/\partial\hat{\eta}_k \cdot \partial\hat{\eta}_i/\partial\hat{b}_k$$

and this is a linear model,

$$\partial\hat{\eta}_i/\partial\hat{b}_k = x_{ik}$$

so

$$\partial\hat{\mu}_i/\partial\hat{b}_k = x_{ik} \cdot \partial\hat{\mu}_i/\partial\hat{\eta}_k$$

This finding can be used to define

$$\Gamma = \text{diag}[\partial\hat{\mu}/\partial\hat{\eta}] = \begin{bmatrix} \partial\hat{\mu}_1/\partial\hat{\eta}_1 & 0 & \dots & 0 & 0 \\ 0 & \partial\hat{\mu}_2/\partial\hat{\eta}_2 & 0 & 0 & 0 \\ \vdots & & \ddots & 0 & 0 \\ 0 & 0 & & \partial\hat{\mu}_{N-1}/\partial\hat{\eta}_{N-1} & 0 \\ 0 & 0 & \dots & & \partial\hat{\mu}_N/\partial\hat{\eta}_N \end{bmatrix}$$

And thus

$$D' = X'\Gamma$$

3.2 Independent observations

Take the simple case of independent–uncorrelated–observations. Then the variance matrix for the observations, would be simple:

$$V = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \sigma_{N-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_N^2 \end{bmatrix}$$

I have in mind that $\sigma_i^2 = \phi V(\mu_i)$, so all the user need supply is $V(\mu_i)$ and we estimate ϕ .

Here's the big result

The quasi-score equation for this case of independent observations reduces to, for the k 'th parameter in b ,

$$U_k = \frac{\partial \ln L}{\partial b_k} = \sum_{i=1}^N \frac{1}{\sigma_i^2} [y_i - \mu_i] \frac{\partial \mu_i}{\partial b_k} = 0$$

$$U_k = \frac{\partial \ln L}{\partial b_k} = \sum_{i=1}^N \frac{1}{\sigma_i^2} [y_i - \mu_i] \frac{\partial \mu_i}{\partial \eta_i} x_{ik} = 0$$

which (we should not be shocked to find) is the same as the GLM score equation in equation 2.

So, when the mean and variance stipulated in a GLM matches the mean and variance stipulated in a quasi-likelihood model, the 2 result in the same parameter estimates.

You can write the matrix equation as:

$$X'GV^{-1}(y - \hat{\mu}) = 0$$

The big difference between GLM and quasi-likelihood, of course, is that the quasi-likelihood is defined for many more situations than the GLM.

Now here's the but...

The equality of the two models is strictly true only when the variance matrix of y is diagonal, since that is the only case in which the GLM is defined. If, instead of a diagonal matrix V , we have the full, complicated looking thing

$$V = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & & & \sigma_{2N} \\ \sigma_{31} & & & & \sigma_{3N} \\ \vdots & & & \sigma_{N-1}^2 & \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{N(N-1)} & \sigma_N^2 \end{bmatrix}$$

then quasi-likelihood is a different beast altogether, more similar in appearance to the GLS score in ??(in my opinion).

In any case, the longitudinal model requires us to deal with intercorrelated observations. Rather than dealing with the problem in complete generality, the GEE approach takes advantage of the separation of observations among clusters.

3.2.1 Digression

In the case of independence, Liang and Zeger have the matrix version of the score equation

$$U = X'\Delta[y - \hat{\mu}] = 0$$

where

$$\Delta = \begin{bmatrix} d\theta_1/d\eta_1 & 0 & 0 & 0 & 0 \\ 0 & d\theta_2/d\eta_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & d\theta_{N-1}/d\eta_{N-1} & 0 \\ 0 & 0 & 0 & 0 & d\theta_N/d\eta_N \end{bmatrix}$$

This is found to be equivalent to the result in the preceding section, recalling that

$$\Delta = \begin{bmatrix} \frac{1}{V(\mu_1)} \frac{\partial \mu_1}{\partial \eta_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{V(\mu_2)} \frac{\partial \mu_2}{\partial \eta_2} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{V(\mu_{N-1})} \frac{\partial \mu_{N-1}}{\partial \eta_{N-1}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{V(\mu_N)} \frac{\partial \mu_N}{\partial \eta_N} \end{bmatrix}$$

The dispersion parameter ϕ “disappeared” because it is a constant.

4 GEE

Turn back to the basic problem of longitudinal data analysis. There are a few observations (T) about each of N subjects. Each “cluster” of observations has its own V_i to describe the inter-correlation of its observed $y'_i = (y_{i1}, y_{i2}, y_{i3}, \dots, y_{iT})'$

$$Var(y) = \begin{bmatrix} V_1 & & & & 0 \\ & V_2 & & & \\ & & V_3 & & \\ & & & \ddots & \\ 0 & & & & V_N \end{bmatrix} \quad (6)$$

4.1 Working Correlation Matrix

Among other things, the seminal paper by Liang and Zeger (1986) contributed a workable system for representing our ideas about the matrix V_i and a coherent estimation and interpretation strategy.

Here are the essentials. The “working correlation” $R(\alpha)$ matrix is:

$$V_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi \quad (7)$$

(Note, here’s a case where ϕ is relabeled as $1/\phi$, if you know what I mean.)

$R(\alpha)$ is a $T \times T$ matrix of correlation coefficients, numbers between -1 and +1. The parameter α is a “tunable parameter” on which $R(\alpha)$ depends.

The symbol $A_i^{1/2}$ is the diagonal matrix that holds the square root of the variances of the observed y_i .

$$A_i = \text{diag}[\sigma_i^2] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \sigma_{T-1}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_T^2 \end{bmatrix}$$

$$A_i^{1/2} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \sigma_{T-1} & 0 \\ 0 & 0 & 0 & 0 & \sigma_T \end{bmatrix}$$

4.2 Aside: Correlation/Covariance

The correlation and covariance concepts fit together. In case you have forgotten, correlation is defined as

$$r = \frac{Cov(X1, X2)}{\sqrt{\sigma_{X1}^2 \cdot \sigma_{X2}^2}}$$

So, if one is given r , one can recover $Cov(X1, X2)$ by multiplying:

$$Cov(X1, X2) = \sqrt{\sigma_{X1}^2 \sigma_{X2}^2} \cdot r = \sigma_{X1} \cdot \sigma_{X2} \cdot r$$

The matrix equivalent of this trick for recovering the Covariance from the Correlation is the formula given in expression 7.

4.3 Examples of correlation matrices

1. Unstructured
2. Exchangeable
3. AR(1)

4.4 General Estimating Equations

The “general estimating equations” (Liang and Zeger, 1986, p. 15) are the result of taking this structure into account within a quasi-likelihood context. The general estimating equations are defined in the matrix form as

$$D'Var(y)^{-1}(y - \hat{\mu}) = 0$$

Because $Var(y)$ has a block-diagonal structure (lots of 0’s surrounded by square matrices), this simplifies to a sum of scores for the N clusters, so the generalized estimating equations are often written as:

$$\sum_{i=1}^N D_i'V_i^{-1}(y_i - \hat{\mu}_i) = 0$$

Liang and Zeger (1986) have $D_i = [d\hat{\mu}/d\hat{b}] = A_i\Delta_iX_i$, where

$$\Delta_i = \text{diag}(d\theta_{it}/d\hat{\eta}_{it}) = \begin{bmatrix} \frac{d\theta_{11}}{d\hat{\eta}_{11}} & 0 & 0 & 0 & 0 \\ 0 & \frac{d\theta_{12}}{d\hat{\eta}_{12}} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{d\theta_{N(T-1)}}{d\hat{\eta}_{N(T-1)}} & 0 \\ 0 & 0 & 0 & 0 & \frac{d\theta_{NT}}{d\hat{\eta}_{NT}} \end{bmatrix}$$

4.5 Iterative Estimation Procedure

A numerical maximization procedure is described in such horrifying detail in the textbooks that I can’t bear to write it down. Maybe next time.

4.6 Statistical Properties

The estimate from the GEE approach are consistent and Normally distributed.

Furthermore, the variance/covariance matrix of the estimate \hat{b} is

$$Var(\hat{b}) = \hat{\phi}(X'\Delta Var(y)\Delta X)^{-1}$$

That assumes $Var(y)$ is specified correctly. Computer output from GEE programs will present these “model standard errors”.

Liang and Zeger proposed the robust estimator (in the Huber-White tradition). Their robust estimator of the variance/covariance matrix of \hat{b} is stated on their p. 15.

$$\text{robust}Var(\hat{b}) = \left(\sum D_i'V_i^{-1}D_i\right)^{-1} \left\{\sum D_i'V_i^{-1} \left(\widehat{Var}(y)\right) V_i^{-1}D_i\right\} \left(\sum D_i'V_i^{-1}D_i\right)^{-1}$$

The information matrix is $\sum(D_i'V_i^{-1}D_i)$. This really is an information sandwich.

The robust estimator is also sometimes called the “empirical estimator”.

$$robustVar(\hat{b}) = \left(\sum D_i' V_i^{-1} D_i \right)^{-1} \left\{ \sum D_i' V_i^{-1} (y - \hat{\mu})(y - \hat{\mu})' V_i^{-1} D_i \right\} \left(\sum D_i' V_i^{-1} D_i \right)^{-1}$$

Focus on the middle part, say, for cluster 1:

$$\begin{aligned} & D_1' V_1^{-1} (y_1 - \hat{\mu}_1)(y_1 - \hat{\mu}_1)' V_1^{-1} D_1 \\ &= D_1' V_1^{-1} \begin{bmatrix} y_{11} - \hat{\mu}_{11} \\ y_{12} - \hat{\mu}_{12} \\ y_{13} - \hat{\mu}_{13} \end{bmatrix} \begin{bmatrix} y_{11} - \hat{\mu}_{11} & y_{12} - \hat{\mu}_{12} & y_{13} - \hat{\mu}_{13} \end{bmatrix} V_1^{-1} D_1 \\ &= D_1' V_1^{-1} \begin{bmatrix} (y_{11} - \hat{\mu}_{11})^2 & (y_{12} - \hat{\mu}_{12})(y_{11} - \hat{\mu}_{11}) & (y_{13} - \hat{\mu}_{13})(y_{11} - \hat{\mu}_{11}) \\ (y_{11} - \hat{\mu}_{11})(y_{12} - \hat{\mu}_{12}) & (y_{12} - \hat{\mu}_{12})^2 & (y_{13} - \hat{\mu}_{13})(y_{12} - \hat{\mu}_{12}) \\ (y_{11} - \hat{\mu}_{11})(y_{13} - \hat{\mu}_{13}) & (y_{12} - \hat{\mu}_{12})(y_{13} - \hat{\mu}_{13}) & (y_{13} - \hat{\mu}_{13})(y_{13} - \hat{\mu}_{13}) \end{bmatrix} V_1^{-1} D_1 \end{aligned}$$