Lecture 22—Wednesday, February 29, 2012

Topics

A marginal approach to generalized linear models

As appealing as mixed effects models are for handling hierarchical clustered/correlated data, the difficulties (discussed last time) in interpreting the parameter estimates from generalized linear mixed effects models with non-identity links suggests that other approaches are needed. One of the most popular of these is generalized estimating equations (GEE). GEE extends generalized linear models to correlated data but differs from mixed effects models in that GEE explicitly fits a marginal model to data. To understand the motivation behind GEE we need to take a closer look at the theory behind generalized linear models.

An abstract formulation of the generalized linear model

The probability distributions used in generalized linear models are related because they are all members of what's called the exponential family of distributions. The density (mass) function of any member of the exponential family takes the following form.

$$f(y;\theta,\phi)=\exp\left[\frac{y\theta-b(\theta)}{a(\phi)}-c(y,\phi)\right] \tag{1}$$

The functions *a*, *b*, and *c* in the formula will vary from distribution to distribution. The parameter θ is called the canonical parameter and is a function of μ. This function of μ is referred to as the canonical link function.

Example: As an illustration we write the Poisson distribution in its exponential form. The formula for the Poisson probability mass function is shown below.

$$f(y) = P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

Through a series of steps we can transform this expression into one that is of the correct exponential form.

$$\frac{e^{-\lambda}\lambda^{y}}{y!} = \exp\left[\log\left(\frac{e^{-\lambda}\lambda^{y}}{y!}\right)\right] = \exp\left[\log e^{-\lambda} + \log\lambda^{y} - \log y!\right]$$

$$= \exp\left[-\lambda + y\log\lambda - \log y!\right] = \exp\left[\frac{y\log\lambda - \lambda}{1} - \log y!\right]$$

Comparing this to the generic exponential form

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

we can make the following identifications.

- $\theta = \log\lambda$
- $b(\theta) = \lambda = \exp(\log\lambda) = \exp\theta$
- $a(\phi) = 1$
- $c(y,\phi) = -\log y!$

* * * * *

The canonical link functions for three members of the exponential family of probability distributions are shown below.

$$\text{Normal: } \theta = \mu$$
$$\text{Poisson: } \theta = \log \mu$$
$$\text{Binomial: } \theta = \log \frac{\mu}{1-\mu}$$

These in turn are the usual link functions used in normal, Poisson, and logistic regression. The mean and variance of a distribution in the exponential family are given generically as follows.

$$\mu = b'(\theta)$$
$$\operatorname{Var}(y) = a(\phi) b''(\theta) = a(\phi) \frac{d\mu}{d\theta} \equiv a(\phi) \operatorname{Var}(\mu)$$

The last step defines Var(μ) as the derivative of the mean with respect to the canonical parameter.

Suppose we have a random sample of size *n* from a member of the exponential family of distributions. The likelihood is given by

$$L(\mathbf{y}) = \prod_{i=1}^{n} \exp\left[\frac{y_i \theta - b(\theta)}{a(\phi)} - c(y_i, \phi) \right]$$

with corresponding log-likelihood

$$\log L(\mathbf{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta - b(\theta)}{a(\phi)} - c(y_i, \phi) \right).$$

To find maximum likelihood estimates analytically we take the derivative of the log-likelihood with respect to the canonical parameter θ, set the result equal to zero, and solve for θ.

$$\frac{d}{d\theta} \log L(\mathbf{y}) = 0$$

Typically we do this in a regression setting in which we express the canonical parameter as a linear combination of predictors.

$$\theta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

So, in this case we would differentiate the log-likelihood with respect to each of the regression parameters separately, set the result equal to zero, and solve for the regression parameters. The shorthand notation for this is to use a vector derivative (gradient).

$$\frac{d}{d\beta} \log L(\mathbf{y}) = 0$$

After some algebraic simplification and using the notation defined previously, we end up with $p + 1$ equations of the following form.

$$\frac{\partial}{\partial \beta_j} \log L(\mathbf{y}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi) \mathrm{Var}(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \qquad \textbf{(2)}$$

These are referred to as estimating equations because we can use them to obtain the maximum likelihood estimate of **β**. The $p + 1$ estimating equations defined by eqn (2) can be written more succinctly using matrix notation. Define the $n \times (p + 1)$ matrix **D**, the $n \times n$ matrix **V**, and the $n \times 1$ vector $\mathbf{y} - \boldsymbol{\mu}$ as follows.

$$\mathbf{D} = \begin{bmatrix} \frac{\partial \mu_1}{\partial \beta_0} & \frac{\partial \mu_1}{\partial \beta_1} & \cdots & \frac{\partial \mu_1}{\partial \beta_p} \\ \frac{\partial \mu_2}{\partial \beta_0} & \frac{\partial \mu_2}{\partial \beta_1} & \cdots & \frac{\partial \mu_2}{\partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_n}{\partial \beta_0} & \frac{\partial \mu_n}{\partial \beta_1} & \cdots & \frac{\partial \mu_n}{\partial \beta_p} \end{bmatrix}, \quad \mathbf{V} = a(\phi) \begin{bmatrix} \mathrm{Var}(\mu_1) & 0 & \cdots & 0 \\ 0 & \mathrm{Var}(\mu_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{Var}(\mu_n) \end{bmatrix},$$

$$\mathbf{y}-\boldsymbol{\mu}=\begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \\ \vdots \\ y_n-\mu_n \end{bmatrix}$$

With these definitions the $p + 1$ estimating equations of eqn (2) can be written as the following single matrix equation.

$$\frac{d}{d\boldsymbol{\beta}}\log L(\mathbf{y})=\begin{bmatrix} \frac{\partial \log L}{\partial \beta_0} \\ \vdots \\ \frac{\partial \log L}{\partial \beta_p} \end{bmatrix}=\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})=\mathbf{0} \qquad \textbf{(3)}$$

From estimating equations to generalized estimating equations (GEE)

The primary problem in dealing with correlated data in a likelihood framework is that the likelihood becomes inordinately complicated. Generalized estimating equations gets around this difficulty by making the following simple observation. Although the likelihood in eqn (1) depends specifically on the form of the probability distribution, (the functions *a*, *b*, and *c*), the estimating equation that we obtain by differentiating the log-likelihood that is shown in eqns (2) and (3) depends on the probability distribution only via the mean **μ** and the variance **V**. The rest of the details about the nature of the original probability distribution are irrelevant. So, when a probability model is a member of the exponential family we don't need the all the details of the probability distribution in order to estimate the parameters of a regression equation, just its mean and variance.

Inspired by this observation, generalized estimating equations performs an end-around the likelihood. Rather than starting with eqn (1), GEE starts with the estimating equations in eqn (2), or equivalently the matrix version eqn (3), and generalizes it in two distinct ways.

1. With independent data the matrix **V** in eqn (3) is a diagonal matrix. A generalization to correlated data is to instead let **V** = **Σ**, an arbitrary variance-covariance matrix. Specifically we let

$$\boldsymbol{\Sigma}=a(\phi)\mathbf{S}^{1/2}\,\mathbf{R}\,\mathbf{S}^{1/2}$$

where **S** is a diagonal matrix and **R** is the proposed correlation matrix for our data. Typically **R** will depend on parameters that need to be estimated from the data when solving eqn (3). If so, an additional estimating equation will be required.

2. GEE allows us to avoid specifying a probability model entirely. Instead we just need a regression model for the mean μ and a variance model Var(μ). The variance model may include a correlation structure or not. For instance we might choose a binomial-like model for the variance, $p(1-p)$, even though the data themselves are not binomial and perhaps not even discrete! The Simpson's diversity index in Assignments 1 and 2 is an example where such an approach might make sense.

Both of these two generalizations lead to what is called a generalized estimating equation. Parameter estimates are obtained by solving the generalized estimating equations numerically typically using an optimization algorithm based on Newton's method.

Quasi-likelihood

The estimating equation of eqn (1) is the derivative of the log-likelihood set equal to zero.

$$\frac{\partial}{\partial \beta_j} \log L(\mathbf{y}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)\mathrm{Var}(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}$$

From calculus we know that the process of differentiation can be reversed by integrating (antidifferentiating) this expression. Because we can add an arbitrary constant to an antiderivative and obtain another antiderivative, antidifferentiation does not yield a unique result. Antidifferentiating the estimating equation yields the following.

$$\begin{aligned} \int \frac{\partial}{\partial \beta_j} \log L(\mathbf{y}) d\beta_j &= \int \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)\mathrm{Var}(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} d\beta_j \\ &= \sum_{i=1}^{n} \int \frac{y_i - \mu_i}{a(\phi)\mathrm{Var}(\mu_i)} d\mu_i \end{aligned}$$

If we start with this last integral, there are two possible scenarios.

1. If the integrand was obtained from an estimating equation that was derived from a log-likelihood then the integral can be written in the form $\log L(\mathbf{y}) + K$ where K contains terms that don't involve μ. So, the log-likelihood will be included in the set of antiderivatives obtained from the estimating equation.
2. If instead we start with a generalized estimating equation by adding a correlation structure to an estimating equation, by postulating a mean-variance relationship Var(μ), or by doing both, then there is no corresponding log-likelihood. Integrating the estimating equation in this case will not yield a true log-likelihood, but instead generates something referred to as a quasi-likelihood (although it might be better to call it a

quasi-loglikelihood). The formal definition of the quasi-likelihood is that it is the antiderivative of the generalized estimating equation evaluated at the parameter estimates, eqn (4).

$$Q(y;\hat{\mu}) = \int^{\hat{\mu}} \frac{y_i - \mu}{a(\phi)\mathrm{Var}(\mu)}\, d\mu \qquad \textbf{(4)}$$

The connection between quasi-likelihood and likelihood theory is actually quite close. It turns out that GEE estimates have the same large sample properties that MLEs do—they're consistent, asymptotically normal, etc. In addition, the quasi-likelihood can be used to generate an AIC-like quantity for use in model selection.

Implementing GEE in practice

When estimating a regression model using maximum likelihood (for instance fitting a generalized linear model) we

1. specify a model for the mean and
2. a probability model (along with a link function) for the response.

When fitting a model using generalized estimating equations we

1. specify a regression model for the mean,
2. a model for the mean-variance relationship Var(μ), and
3. a model for the correlation.

In most GEE software the mean-variance relationship is identified by specifying, using a family argument, a probability model that has the same mean-variance relationship as the one desired. This convention is rather counterintuitive given that there is no probability model in GEE. Table 1 lists some typical choices for the family argument in GEE and the corresponding expression for Var(μ) that results from that choice.

Table 1 Common mean-variance relationships in GEE

Var(μ)	Family
1	gaussian (normal)
μ	poisson
μ(1 – μ)	binomial (Bernoulli)
μ^2	gamma

Depending on the software, a scale parameter φ can also be estimated to account for over- or under-dispersion in Poisson and binomial models.

Table 2 defines some of the common correlation models that are available in GEE software. (The exact name used will depend on the software.) The correlation requested in GEE is referred to as the working correlation and for 2-level hierarchical data it refers to the correlation among observations *j* and *k* coming from the same level-2 unit *i*.

Table 2 Common correlation structures in GEE

Correlation type	Correlation formula
independence	$\text{cor}\left(Y_{ij}, Y_{ik}\right) = 0,\ j \neq k$
exchangeable	$\text{cor}\left(Y_{ij}, Y_{ik}\right) = \rho,\ j \neq k$
AR(1)	$\text{cor}\left(Y_{ij}, Y_{ik}\right) = \rho^{\lvert j-k \rvert},\ j \neq k$
unstructured	$\text{cor}\left(Y_{ij}, Y_{ik}\right) = \rho_{jk},\ j \neq k$
user-defined	Specific values are entered for the correlations

Typically there is also an ID argument that is used to identify the variable that denotes the level-2 units. This is the same as the group variable in mixed effects models. (In **nlme** the group variable appeared to the right of the vertical bar in the random argument: `random = ~1|group_variable`.)

Comparing GEE to GLM

1. Typically, GEE returns parameter estimates that are fairly close to those returned by GLM when the models being compared assume the same mean-variance relationship, i.e., use the same value of the family argument.
2. When there is a hypothesized correlation structure in a GEE model, the estimated variance of the parameter estimates tends to be larger in GEE than it is for GLM.
3. Because GEE is not a likelihood-based method, GEE output does not include a log-likelihood, likelihood ratio tests, or AIC. Reported significance tests are the Wald tests for individual parameter estimates.

The quasi family of models

The family argument of the **glm** function of R permits the use of quasibinomial, quasipoisson, or the more general quasi function. These are not probability distributions per se but are ways to specify a model for Var(μ) in the manner described in Table 1 above. The quasipoisson and quasibinomial choices generalize the expressions given in Table 1 as follows.

Table 3 The quasi values for the family argument

Family	Var(μ)
quasipoisson	φμ
quasibinomial	φμ(1 – μ)

Here $\phi = \frac{\text{Pearson deviance}}{\text{df}}$ is called the dispersion parameter. The Pearson deviance is the sum of the squared Pearson residuals. The quasi function can be used to specify these same variance structures as well as others by giving a formula for the variance along with a link function. For instance, `family=quasi(var="mu(1-mu)", link=logit)` yields the same result as `family=quasibinomial`.

The quasipoisson and quasibinomial families provide a crude way to adjust for overdispersion in data, where overdispersion is defined as deviation from the Poisson or binomial probability models due to clumping. When either family = quasipoisson or family = quasibinomial is specified, the parameter estimates one gets from **glm** are identical to what one gets with family = poisson or family = binomial, but the standard errors are adjusted by multiplying them by the square root of φ.

Full-fledged GEE, on the other hand, assumes the data have a hierarchical structure in which the clusters are identified explicitly with an ID variable. GEE also allows you to model the correlation in a cluster by using one of the correlation structures of Table 2.

Choosing a correlation structure

There are three recommended methods for choosing a correlation structure for GEE (Hardin & Hilbe, 2003).

1. Choose a correlation structure that reflects the manner in which the data were collected. For instance with temporal data a sensible correlation structure is one that includes time dependence, such as AR(1).
2. Choose a correlation structure that minimizes Pan's QIC. QIC is a statistic that generalizes AIC to GEE but is used only for comparing models that are identical except for their different correlation structures. Note: This is not QAIC that is described in Burnham & Anderson (2002).
3. Choose a correlation structure for which the sandwich estimates of the variance most closely approximate the naive estimate of the variance.

I examine the last two options in more detail.

Method 2: Pan's QIC and/or CIC

QIC is due to Pan (2001) and comes in two flavors. One version is used for selecting a correlation structure and the second version is used for choosing models all of which were fit with the same correlation structure. I first discuss the version of QIC that is used for choosing a correlation structure.

In what follows let *R* denote the correlation structure of interest and let *I* denote the independence model. Recall the definition of AIC.

$$\mathrm{AIC} = -2\log L + 2K$$

QIC is defined analogously as follows.

$$\mathrm{QIC} = \underbrace{-2Q\left(\hat{\beta}(R), \mathrm{I}, \mathrm{Data}\right)}_{\text{quasi-likelihood}} + 2\,\mathrm{trace}\left(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_R\right) \tag{5}$$

The first term contains the quasi-likelihood as given in eqn (4) except that it is now extended to the full data set.

$$Q(y;\hat{\mu}) = \sum_{i=1}^{n} \int^{\hat{\mu}_i} \frac{y_i - \mu_i}{a(\phi)\mathrm{Var}(\mu_i)}\, d\mu_i$$

The quasi-likelihood defined in this formula is actually a bit of a hybrid. For $\hat{\mu}$ we use the regression coefficient estimates obtained from a GEE model with correlation model *R*, but for Var(μ) we assume a working correlation structure of independence *I*. Table 4 gives the quasi-likelihood formulas (multiplied by φ) for models in which the mean-variance relationship has the form of a binomial or a Poisson random variable.

Table 4 Quasi-likelihoods

Family	**Var(μ)**	$\phi \cdot Q\left(\hat{\mu}, \mathrm{I}, \mathrm{Data}\right)$
binomial	μ(1 – μ)	$\sum_{i=1}^{n}\left[y_i \log\frac{\hat{\mu}_i}{1-\hat{\mu}_i} + \log\left(1-\hat{\mu}_i\right)\right]$
Poisson	μ	$\sum_{i=1}^{n}\left[y_i \log\hat{\mu}_i - \hat{\mu}_i\right]$

For binary data (binomial with $n = 1$) the estimate of φ described above is not used; instead φ is set to one. The second term of QIC, $2\,\mathrm{trace}\left(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_R\right)$, is a penalty term that is analogous to $2K$ in the formula for AIC. It is defined as follows.

1. Trace refers to "matrix trace", the sum of the diagonal entries of a matrix.

2. $\Omega_I = \mathbf{A}_I^{-1}$ where $\mathbf{A}_I$ is the variance-covariance matrix of the parameter estimates in which an independence model *I* is used for the correlation.
3. $\hat{\mathbf{V}}_R$ is the modified sandwich estimate (explained in the next section) of the variance-covariance matrix of the parameter estimates that is obtained using the hypothesized correlation model *R*. The sandwich estimate of the variance-covariance matrix is part of the standard output from GEE.

So, to calculate QIC we need to fit two models: one that uses the correlation model *R* and the other that uses the independence model *I*. QIC is then used like AIC. We make various choices for *R* and choose the *R* that yields the lowest value of QIC.

Pan (2001) also suggested another version of QIC for comparing models that have the same working correlation matrix *R* and the same quasi-likelihood form (for instance, all Poisson), but involve different predictor sets. He suggested calculating a statistic QIC_u that is defined as follows.

$$QIC_u = -2Q\left(\hat{\beta}(p), \mathrm{I}, \mathrm{Data}\right) + 2p$$

Models with smaller values of QIC_u are to be preferred.

Recently QIC has been criticized as a model selection tool, particularly if the model for the mean does not fit the data very well. Hin and Wang (2009) argue that the two terms in the expression for QIC in eqn (5) are not equally informative for correctly identifying the correlation structure. In particular the quasi-likelihood term corresponds to an apparent error rate that better indicates an inadequacy in the mean model rather than in the correlation model. Through simulations they demonstrate that when the mean model is misspecified, the quasi-likelihood term can mask differences in model quality that are due purely to the different assumptions made about the correlation. As a result they recommend using just the second term of QIC (without the superfluous multiplier of two) when comparing models with different correlation structures. They call this statistic the correlation information criterion, CIC.

$$\mathrm{CIC} = \mathrm{trace}\left(\hat{\Omega}_I \hat{\mathbf{V}}_R\right)$$

Method 3: Comparing sandwich estimates

All GEE packages return something that is variously referred to as the sandwich variance estimate or the robust variance estimate. For many packages this is the default estimate that is displayed in the standard error column of the summary table of the model. The sandwich estimate corrects the variance estimate that is based on the working correlation matrix *R* using a correction that is constructed from the model residuals. It tends to be robust to an incorrectly specified working correlation model *R* and will be nearly correct even if *R* is incorrect. The sandwich estimate is known to behave best for a large sample consisting of many small groups so that there are not too many observations in each group. If the total number of groups is small, the sandwich estimate can be very biased.

In addition to the sandwich variance estimate, GEE packages also return a variance estimate that is based only on the working correlation model R. This is variously called the model-based variance estimate or the naive variance estimate.

A correlation model R can be selected as follows. Fit a series of models that differ only in the choice of correlation matrix R; all of the remaining features of these models are the same. For each model compare the sandwich variance estimates with the model-based variance estimates. The model whose sandwich variance estimates most closely resembles its model-based variance estimates is the one with the best correlation model R.

A problem with GEE and binary data

In addition to invertibility and positive definiteness, correlation matrices for binary data have a further feasibility requirement (Chaganty and Joe 2004). The predicted probabilities of a pair of binary response variables impose a constraint on the legal values that the correlation can take. Let y_i and y_j be two binary observations, let p_i and p_j be their predicted probabilities under a given model, and let r_{ij} be their correlation as estimated from say a GEE model. Prentice (1998) derived the following formula for the joint probability of the binary pair y_i and y_j.

$$P(y_i, y_j) = p_i^{y_i}(1-p_i)^{y_i}\, p_j^{y_j}(1-p_j)^{y_j}\left[1 + r_{ij}\frac{(y_i-p_i)(y_j-p_j)}{\sqrt{p_i p_j (1-p_i)(1-p_j)}}\right]$$

Because probabilities must be non-negative, the expression inside the brackets imposes a constraint on the possible values the correlation r_{ij} can take.

Currently available GEE software, including the **gee** and **geepack** packages of R, do not check for the feasibility of the correlation matrices they estimate for binary data. The general consensus is that if a correlation estimated by GEE is infeasible that is a good reason to reject that correlation model. There has been a lot of discussion of this issue in the biomedical and biostatistical literature. For a recent survey see Ziegler and Vens (2010) as well as the discussion that follows the article (Breitung et al. 2010; Shults 2011).

References

- Breitung, J., N. R. Chaganty, R. M. Daniel, M. G. Kenward, M. Lechner, P. Martus, R. T. Sabo, Y.-G. Wang, and C. Zorn. 2010. Discussion of "Generalized estimating equations: Notes on the choice of the working correlation matrix". *Methods of Information in Medicine* **49**: 426–432.
- Burnham, K. P. and D. R. Anderson. 2002. *Model Selection and Multimodel Inference*. Springer-Verlag: New York.
- Chaganty, N. R. and H. Joe. 2004. Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society B* **66**: 851–860.
- Hardin, James W. and Joseph M. Hilbe. 2003. *Generalized Estimating Equations*. Chapman & Hall/CRC Press: Boca Raton, FL.
- Hin, Lin-Yee and You-Gan Wang. 2009. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* **28**: 642–658.
- Pan, W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrika* **83**: 551–562.
- Prentice R. L. 1988. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**: 1033–1048.

- Shults, J. 2011. Discussion of “Generalized estimating equations: Notes on the choice of the working correlation matrix"—continued. *Methods of Information in Medicine* **50**: 96–99.
- Ziegler, A. and M. Vens. 2010. Generalized estimating equations: Notes on the choice of the working correlation matrix. *Methods of Information in Medicine* **49**: 421–425.

Course Home Page

Jack Weiss
Phone: (919) 962-5930
E-Mail: jack_weiss@unc.edu
Address: Curriculum for the Environment and Ecology, Box 3275, University of North Carolina, Chapel Hill, 27599

Last Revised--February 29, 2012
URL: https://sakai.unc.edu/access/content/group/2842013b-58f5-4453-aa8d-3e01bacbfc3d/public/Ecol562_Spring2012/docs/lectures/lecture22.htm