Chapter 6

Longitudinal Data Analysis for Counts and Binary Outcomes: Generalized Estimating Equations (GEE)

In many settings, the outcomes recorded on individuals are counts or binary responses. In this chapter we extend the theory in the preceding chapters to permit a regression analysis which does not require the mean responses to be linear in X_i. In the univariate setting, the generalized linear model (GLM, McCullagh and Nelder, 1989) offers an approach which unifies linear, log and logistic regression analysis. It was extended to the distribution free multivariate setting by Liang and Zeger (1988) and Zeger and Liang (1988). Extensions of the likelihood approach and the random effects models to the nonlinear setting is more complex; we will review some of the suggested approaches in subsequent chapters. In this chapter, we begin by reviewing the basic ideas for GLM's in the univariate setting, and then discuss the GEE extension to correlated data.

6.1 The Generalized Linear Model (GLM) for Univariate Outcomes.

Suppose now that Y_i is a scalar outcome, X_i is a $1 \times p$ row vector of covariates, β is a $p \times 1$ vector of regression coefficients and

$$\mu_i = E(Y_i) = g(X_i\beta), \tag{6.1}$$

where

$$g^{-1}(\mu_i) \equiv \ell(\mu_i) = X_i\beta.$$

Here $g(\cdot)$ and $\ell(\cdot)$ are known functions; $\ell(\cdot)$ is called the link function and $g(\cdot)$ is the inverse link function.

Examples. For the linear model, both $\ell(\cdot)$ and $g(\cdot)$ are the identity functions: $\ell(\mu_i) = \mu_i$, and $\ell(\cdot)$ is called the identity link. If Y_i is a count, so that $\mu_i > 0$, a natural link function is the log:

$$\log\ \mu_i = X_i\beta \Rightarrow \mu_i = e^{X_i\beta}.$$

Here $\ell(\cdot)$ is the log link. With binary data, $E(Y_i) = P(Y_i = 1)$, hence $0 < \mu_i < 1$ and a popular link function is

$$\text{logit}\mu_i = \log\left(\mu_i/(1-\mu_i)\right) = X_i\beta,$$

or

$$\mu_i = e^{X_i\beta}/(1 + e^{X_i\beta}).$$

Here $\ell(\cdot)$ is the logit link.

As in the LMCD setting, it is possible to implement a distribution free analysis using only the assumption of the mean model (6.1), or we may fully specify the distribution of Y_i (possibly as a function of other parameters) and use a fully parametric analysis. The distribution free approach estimates β by minimizing the objective function

$$Q(\beta) = \sum_{i=1}^{N} W_i(Y_i - \mu_i)^2$$

for some arbitrary choice of weights, W_i. Straightforward differentiation of $Q(\beta)$ with respect to β gives a $p \times 1$ vector of derivatives $\partial Q(\beta)/\partial\beta_j, j = 1, \ldots, p$:

$$\frac{\partial Q(\beta)}{\partial\beta} = 2\sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right) W_i\left(Y_i - \mu_i\right); \tag{6.2}$$

setting (6.2) equal to zero gives $\widehat{\beta}$, since given β,

$$\widehat{\mu}_i = g(X_i\widehat{\beta}),$$

and the weights are assumed known. This can be viewed as a semiparametric approach because:

i) Any estimator of β that is consistent and asymptotically normal, assuming *only* (6.1) is true, is asymptotically equivalent to $\widehat{\beta}(W)$ for some choice of W.

ii) The choice of weights which gives $\widehat{\beta}(W)$ the smallest variance among estimators in this class is $W_i^{-1} = \text{var}(Y_i|X_i) = V_i$.

iii) The asymptotic distribution of $\widehat{\beta}(W)$ satisfies

$$\sqrt{N}\left(\widehat{\beta}(W) - \beta\right) \to N(0, C)$$

where

$$C = \lim_{N\to\infty} I_0^{-1} I_1 I_0^{-1}, \tag{6.3}$$

$$I_0 = \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) W_i \left(\frac{\partial \mu_i}{\partial \beta}\right)^T\right] \Big/ N,$$

and

$$I_1 = \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) W_i V_i W_i \left(\frac{\partial \mu_i}{\partial \beta}\right)^T\right] \Big/ N.$$

A consistent estimator of C is obtained by evaluating $(\partial\mu_i/\partial\beta)$ at $\widehat{\beta}$, and substituting $(Y_i - \widehat{\mu}_i)^2$ for V_i. The same asymptotic limiting distribution will obtain when the W_i are replaced by estimated $\widehat{W}_i$.

In the GLM, we additionally assume that

$$V_i = \text{var}(Y_i|X_i) = V(\mu_i)\phi,$$

where $V(\mu_i)$ is a known function depending upon the mean and ϕ is a known or unknown scalar factor. This implies that V_i depends upon the covariates X_i only through the mean μ_i.

Examples. With Y_i binary, $\text{var}(Y_i) = \mu_i(1-\mu_i) = V(\mu_i)$ and $\phi = 1$. If we assume a Poisson variance for count data Y_i, var $(Y_i) = \mu_i\phi$, where

ϕ is a dispersion parameter. In the linear case, we usually assume the variance does not depend upon μ_i, and take $V(\mu_i) = 1$ and $\sigma^2 = \phi$.

Notice that if

$$W_i = (V(\mu_i)\phi)^{-1},$$

ϕ drops out of the estimating equations, so we may equivalently take

$$W_i = (V(\mu_i))^{-1},$$

so that the estimating equations become

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)_{|\widehat{\beta}} V(\widehat{\mu}_i)^{-1}(Y_i - \widehat{\mu}_i) = 0. \tag{6.4}$$

Under this assumption, the same limiting distribution holds, with now

$$\text{var}\sqrt{N}(\widehat{\beta} - \beta) = C$$

for

$$C = \lim_{N\to\infty} \left[\frac{1}{N}\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T\right]^{-1}, \tag{6.5}$$

and

$$\widehat{C} = \left[\frac{1}{N}\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)_{|\widehat{\beta}} \widehat{V}_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T_{|\widehat{\beta}}\right]^{-1}.$$

REMARKS.

(i) The GLM assumes that the model $\text{var}(Y_i) = V(\mu_i)\phi$ is correct for $\text{var}(Y_i) = V_i$; once we estimate β, we have an estimate for V_i (up to a proportionality constant).

(ii) Equation (6.5) for the $\widehat{\text{var}}(\sqrt{N}(\widehat{\beta} - \beta))$ is only consistent when $\text{var}(Y_i) = V(\mu_i)\phi$ is correct.

(iii) The estimating equations in (6.4) are called quasi-likelihood score equations (Weddeburn, 1974; McCullagh and Nelder, 1989).

(iv) Suppose further that Y_i has an exponential family distribution with canonical parameter θ_i, so that

$$f(Y_i) = e^{(Y_i\theta_i + \tilde{a}(\theta_i) + C(Y_i))/\phi},$$

where μ_i is some function of θ_i, and

$$\tilde{a}(\theta_i)/\phi = -\ln \int e^{(t\theta_i + C(t))/\phi} dt.$$

For this family of distributions, it is easily shown that

$$E(Y_i) = -\partial\tilde{a}(\theta_i)/\partial\theta_i$$

and

$$\text{var}\,(Y_i) = -\partial^2\tilde{a}(\theta_i)/\partial^2\theta_i.$$

We now show that the quasi-likelihood equations correspond exactly to the likelihood score equations. Here we treat ϕ as a fixed scale parameter. Note that (iv) implies

$$\mathcal{L}_\phi\,(\beta)) = \prod_{i=1}^{N} f\,(Y_i) \propto e^{\left(\sum_{i=1}^{N} Y_i\theta_i + \sum_{i=1}^{N}\tilde{a}(\theta_i)\right)/\phi}$$

so that

$$\frac{\partial \ln \mathcal{L}_\phi(\beta)}{\partial\beta} = \left[\sum_{i=1}^{N} Y_i\left(\frac{\partial\theta_i}{\partial\beta}\right) + \sum_{i=1}^{N}\left(\frac{\partial\tilde{a}(\theta_i)}{\partial\beta}\right)\right]\frac{1}{\phi}.$$

Using the chain rule we have that

$$\frac{\partial\tilde{a}(\theta_i)}{\partial\beta} = \frac{\partial\tilde{a}(\theta_i)}{\partial\theta_i}\;\frac{\partial\theta_i}{\partial\beta} = -\mu_i\frac{\partial\theta_i}{\partial\beta}.$$

But

$$\frac{\partial\mu_i}{\partial\beta} = \frac{\partial\mu_i}{\partial\theta_i}\;\frac{\partial\theta_i}{\partial\beta} = -\frac{\partial^2\tilde{a}(\theta_i)}{\partial\theta_i^2}\left(\frac{\partial\theta_i}{\partial\beta}\right) = \text{var}(Y_i)\left(\frac{\partial\theta_i}{\partial\beta}\right),$$

so that

$$\frac{\partial\theta_i}{\partial\beta} = \text{var}(Y_i)^{-1}\frac{\partial\mu_i}{\partial\beta}$$

and

$$\frac{\partial\tilde{a}(\theta_i)}{\partial\beta} = -\mu_i\text{var}(Y_i)^{-1}\frac{\partial\mu_i}{\partial\beta}.$$

Now using the fact $\text{var}(Y_i) = V(\mu_i)\phi$, we have that

$$\frac{\partial \ln \mathcal{L}_\phi(\beta)}{\partial\beta} \propto \sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right) V(\mu_i)^{-1}(Y_i - \mu_i). \tag{6.6}$$

This shows that the likelihood equations are equal to the quasi-likelihood score equations when Y_i has an exponential family density with specified mean and variance, and $W_i = V(\mu_i)^{-1}$.

With exponential families, θ_i is the canonical parameter. We can use it to define the canonical link. If

$$\ell\,(\mu_i) = \theta_i = X_i\beta$$

then ℓ is said to be the canonical link. In this case, the likelihood can be further simplified by noting that

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \theta_i} X_i^T$$

and

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{\partial^2 \tilde{a}(\theta_i)}{\partial \theta_i^2} = \text{var}(Y_i)$$

hence the likelihood equations become simply

$$\sum_{i=1}^{N} X_i^T (Y_i - \widehat{\mu}_i) = 0.$$

To continue with the likelihood approach assuming the exponential family density (iv), the asymptotic variance of $\widehat{\beta}$ is given by the expected value of $-\partial^2 \ln \mathcal{L}_\phi(\beta)/\partial\beta\partial\beta^T$. Differentiating (6.6) with respect to β^T, we see that only one term has nonzero expectation:

$$-E\left(\frac{\partial \ln \mathcal{L}_\phi(\beta)}{\partial\beta\partial\beta^T}\right) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta} (\text{var}(Y_i))^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T,$$

hence

$$A \text{ var } \widehat{\beta} = \left\{\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) (\text{var}(Y_i))^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T\right\}^{-1}$$

in agreement with (6.5). As we will see, the estimating equation and likelihood based approaches generally do not coincide in the multivariate setting with generalized linear models.

6.2 Generalized Linear Models for Longitudinal Data

As before, we will assume an $n_i \times 1$ vector of outcomes, Y_i, where any missingness in the data are MCAR. In addition, each observation is assumed to have a $p \times 1$ vector of covariates X_{ij} so that

$$E(Y_{ij}) = \mu_{ij} = g\left(X_{ij}^T \beta\right)$$

and

$$\ell(\mu_{ij}) = X_{ij}^T \beta$$

for some suitable link function $\ell(\cdot)$. Thus we may write

$$E(Y_i) = \mu_i$$

where

$$\ell(\mu_i) = X_i\beta$$

and $\ell(\mu_i)$ denotes the vector: $(\ell(\mu_{i1}), \ldots, \ell(\mu_{in_i}))^T$. This model is the natural extension of the longitudinal data model considered in the linear setting, with the only difference being that we allow for a generalized link-function linking the mean response vector μ_i to the covariates. In all other respects, the two are similar; it permits unbalanced designs, unequal clusters, etc. It is sometimes referred to as a marginal model to emphasize the point that the means, μ_{ij}, are marginal for each Y_i:

$$E(Y_{ij}|X_i\beta) = \mu_{ij}.$$

In this respect it does not differ from the linear model case.

Example 1. The Harvard Six-Cities Study of Air Pollution and Health gathered data annually on school children in six cities. One outcome studied was the presence or absence of respiratory illness in the preceding year. The relationship between maternal smoking status and rates of respiratory illness was one feature of interest in the study. Here each child has four annual indicators

$$\begin{aligned} Y_{ij} &= 1 \quad \text{if illness is past year} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and we assume

$$\text{logit}\,(\mu_{ij}) = \beta_0 + \beta_1 MS_i + \beta_2 \text{ age}_{ij} + \beta_3 MS_i \times \text{ age}_{ij}$$

where $MS_i = 1$ if mother smoked at the beginning of the study, 0 otherwise and age_{ij} is the age of the ith child at the jth occasion.

Example 2. Diggle *et al.* (1994) describe a clinical trial of progabide in the treatment of epileptic seizures. Patients were randomized to progabide (31 patients) or placebo (28 patients), and measured at baseline, and every two weeks until week 8. Responses were number of seizures in each period. Covariates include baseline seizure rate, period and treatment group. Here Y_{ij} is a count of the number of seizures for the ith subject in the jth period, $j = 1, \ldots, 4$. We assume

$$\log \mu_{ij} = X_{ij}^T\beta$$

where X_{ij} can include baseline seizure counts (perhaps transformed), treatment, period and treatment $\times$ period.

It is natural in this setting to further assume that

$$\text{var}(Y_{ij}) = V(\mu_i)\phi$$

for suitable $V(\cdot)$ because with count and binary data, the variance does typically depend upon the mean. For example, if Y_{ij} is binary, then by definition $V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$ and $\phi = 1$. With count data, $\text{var}(Y_{ij}) = \mu_i$ can be a rather strong assumption derived from Poisson theory. Over dispersion, $\phi > 1$, implies $\text{var}(Y_{ij}) > \mu_i$, so this can be a more reasonable model. In the multivariate setting

$$\text{var}(Y_i) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

where $A_i = \text{diag}\{V(\mu_{ij})\}$. Now however, there is no natural set of assumptions as to how $R_i(\alpha)$ should depend upon μ_i. Thus we will leave $R_i(\alpha)$ as unspecified. As before, we let the true variance of Y_i be denoted by Σ_i, and let

$$V_i \propto W_i^{-1} = A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

denote a "working variance" assumption. Some authors refer to $R_i(\alpha)$ as a "working correlation matrix," implicitly assuming the variance assumption is correct, but not necessarily $R_i(\alpha)$.

6.3 Estimation via GEE.

The basic GEE strategy is to simply generalize the quasi-likelihood equations to the multivariate setting by replacing Y_i and μ_i by their vector counterparts, and using a weight matrix W_i. This yields

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T_{\Big|\widehat{\beta}} \widehat{W}_i \left(Y_i - \widehat{\mu}_i\right) = 0.$$

Here we define $\partial\mu_i/\partial\beta$ as an $n_i \times p$ matrix whose jth row is $\partial\mu_{ij}/\partial\beta^T$. Although optimally we would take $W_i = V_i^{-1}$, in fact any positive definite and symmetric matrix can by used for W_i. If $W_i = V_i^{-1}$, it now depends upon β, but $R_i(\alpha)$ can be specified arbitrarily provided that each W_i remains symmetric and positive definite. In fact, if $R_i(\alpha) = I$, then the GEE just reduces to the GLM analysis treating all Y_{ij} as independent observations, e.g., if $V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$ and $\ell(\cdot)$ is the logit (or probit) link, then GEE reduces to ordinary logistic (or probit) regression, treating all Y_{ij} as independent. Notice also, that for the identity link and $W_i = \Sigma_i^{-1}$, the GEE reduces to the multivariate normal likelihood equations.

We have the following property for $\widehat{\beta}$: $\sqrt{N}(\widehat{\beta}(W) - \beta) \longrightarrow N(0, C)$, where

$$C = \lim_{N\to\infty} I_0^{-1} I_1 I_0^{-1},$$

$$I_0 = \lim_{N\to\infty} \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T W_i \left(\frac{\partial \mu_i}{\partial \beta}\right) \right] \Big/ N, \qquad (6.7)$$

$$I_1 = \lim_{N-\infty} \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T W_i \Sigma_i W_i \left(\frac{\partial \mu_i}{\partial \beta}\right) \right] \Big/ N.$$

As before, asymptotic efficiency of estimation is best with $W_i = \Sigma_i^{-1}$.

The GEE equations can be simplified as follows. First define

$$\frac{\partial \mu_i}{\partial \beta} = \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \beta_1} & \cdots & \frac{\partial \mu_{i1}}{\partial \beta_p} \\ \vdots & & \vdots \\ \frac{\partial \mu_{in_i}}{\partial \beta_1} & \cdots & \frac{\partial \mu_{in_i}}{\partial \beta_p} \end{pmatrix}_{n_i \times p} .$$

Recall that $\ell(\mu_{ij}) = X_{ij}^T \beta = \ell_{ij}$, thus

$$\frac{\partial \mu_{ij}}{\partial \beta} = \frac{\partial \mu_{ij}}{\partial \ell_{ij}} \; \frac{\partial \ell_{ij}}{\partial \beta} = \frac{\partial \mu_{ij}}{\partial \ell_{ij}} X_{ij},$$

so that

$$\left(\frac{\partial \mu_i}{\partial \beta}\right)^T = X_i^T \Delta_i \text{ with } \Delta_i = \text{ diag } \left\{\frac{\partial \mu_{ij}}{\partial \ell_{ij}}\right\},$$

and the GEE equations become

$$\sum_{i=1}^{N} X_i^T \widehat{\Delta}_i \widehat{W}_i (Y_i - \mu_i) = 0.$$

Examples. If ℓ_{ij} is the identity link, then $\ell_{ij} = \mu_{ij}$ and $\Delta_i = I$ and we have generalized least squares. If $\ell_{ij} = \log(\mu_{ij}/(1-\mu_{ij}))$, then $\partial \mu_{ij}/\partial \ell_{ij} = \mu_{ij}(1-\mu_{ij}) = V(\mu_{ij})$, and $\Delta_i =$ diag $(\mu_{ij}(1-\mu_{ij}))$.

If, in addition, we assume that marginally, each Y_{ij} follows the exponential family density, with canonical parameter θ_i then

$$\begin{aligned} \frac{\partial \mu_{ij}}{\partial \ell_{ij}} &= \frac{\partial \mu_{ij}}{\partial \theta_{ij}} \frac{\partial \theta_{ij}}{\partial \ell_{ij}} \\ &= b_{ij} \text{ var}(Y_{ij}) \qquad \text{for } b_{ij} = \frac{\partial \theta_{ij}}{\partial \ell_{ij}}, \end{aligned}$$

and we can write $\Delta_i = B_i A_i$, where $B_i =$ diag $\{b_{ij}\}$, and A_i is diag$\{Y_i\}$, so that

$$\left(\frac{\partial \mu_i}{\partial \beta}\right)^T = \underset{p\times n_i}{X_i^T} \quad \underset{n_i\times n_i}{B_i} \quad \underset{n_i\times n_i}{A_i}\,.$$

Note that if the canonical link is used $\ell_{ij} = \theta_{ij}$ and $B_i = I$, and further, if $R(\alpha) = I$, so that

$$\sum_{i=1}^{N} X_i^T (Y_i - \mu_i) = 0.$$

6.4 Estimating the Correlation Matrix.

Assuming that $\text{var}(Y_{ij}) = V(\mu_{ij})\phi$ where V is known, the parameters in A_i will be determined by $\widehat{\beta}$, thus to estimate W_i, it remains only to model and estimate α. Models for the correlation are not different from those considered in the linear setting (except for random effects models to be considered later), i.e., we may choose unstructured (in the balanced setting), compound symmetry, serial correlation models, etc. Zeger and Liang (1984) proposed the following procedure to estimate α:

i) Estimate β by setting $R_i(\alpha) = I$ to get $\widehat{\beta}_I$ (independence working assumption).

ii) Obtain an estimate of α using the normalized residuals $A_i^{1/2}(Y_i - \widehat{\mu}_i)$, with $\widehat{\mu}_i$ evaluated at $\widehat{\beta}_I$. The details of this step depend upon the model for α and degree of balance in the data. Call this $\widehat{\alpha}_1$.

iii) Set $\widehat{W}_i^{-1} = \widehat{A}_i^{1/2} R_i(\widehat{\alpha}_1)\widehat{A}_i^{1/2}$ and use GEE to get $\widehat{\beta}_1$, holding $R_i(\widehat{\alpha}_1)$ fixed. Here $\widehat{A}_i$ depends upon $\widehat{\mu}_i$.

Iterate ii) and iii) to convergence. In practice, one step is often used, and may, in fact, be preferable if estimates of $R_i(\alpha)$ are unstable due to sparse data or small sample sizes. To compute $\widehat{\beta}$ given a fixed α we can use Fisher Scoring.

Estimating the α parameter can be done using the same method-of-moment approach used in the semi-parametric linear model setting. First consider the balanced and complete case with $n_i = n$ and unstructured $R(\alpha)$. As before, $A_i =$ diag $V(\mu_i)$, so that A_i depends only on β. Given $\widetilde{\beta}$, we may estimate α and ϕ as follows. Let $\widetilde{\mu}_i$ denote μ_i evaluated at $\widetilde{\beta}$, and

$$\widetilde{V}_i = (Y_i - \widetilde{\mu}_i)\,(Y_i - \widetilde{\mu}_i)^T\,,$$

so in large samples we have (assuming $\text{var}(Y_i) \doteq \phi A_i^{1/2} R(\alpha) A_i^{1/2}$):

$$E\left(\widetilde{V}_i\right) \doteq \phi \widetilde{A}_i^{1/2} R(\alpha) \widetilde{A}_i^{1/2}.$$

Thus we take

$$\widehat{R}(\widehat{\alpha}) = \sum_{i=1}^{N} \left\{ \widetilde{A}_i^{-1/2} \widetilde{V}_i \widetilde{A}_i^{-1/2} \right\} \Big/ \widehat{\phi} N^*$$

and (6.8)

$$\widehat{\phi} = \sum_{i=1}^{N} \frac{\widetilde{\psi}_i^T \widetilde{\psi}_i}{N^*} \qquad \text{for } \widetilde{\psi}_{ij} = (Y_{ij} - \widetilde{\mu}_{ij}) / A_{ij}^{1/2},$$

where $N^* = \sum_{i=1}^{N} n_i$. Notice that ϕ is a constant variance inflation function for $V(\mu_{ij})$. In practice, using (6.8) yields an $\widehat{R}$ such that $\text{diag}(\widehat{R})$ will not be all ones, so they are usually forced to one at each iteration. This is because the variance terms are estimated from the model, but the covariance parameters are not.

We can formalize this method-of-moments estimation of α by using a similar set of estimating equations for an arbitrary $R_i(\alpha)$. This is convenient for the setting where the occasions of measurement may vary from subject to subject, but the correlations can be modeled with a limited number of parameters. Let

$$\rho_{ijk} = \rho_{ijk}(\alpha)$$

where

$$\rho_{ijk} = \text{corr}\ (Y_{ij}, Y_{ik})\ ,$$

and

$$r_{ijk} = \phi (Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik}) / A_{ij}^{1/2} A_{ik}^{1/2}$$

so that

$$E(r_{ijk}) = \rho_{ijk}\ .$$

Then estimating equations for α are given by

$$\sum_{i=1}^{N} C_i^T U_i^{-1} (r_i - \rho_i) = 0$$

where

$$\begin{aligned} \rho_i^T &= \left(\rho_{i12}, \ldots, \rho_{in_i(n_i-1)}\right)^T, \\ r_i^T &= \left(r_{i12}, \ldots, r_{in_i(n_i-1)}\right), \\ C_i &= \frac{\partial \rho_i}{\partial \alpha} \quad \text{and} \quad U_i = \text{var}(r_i)\ . \end{aligned}$$

Note that the dimension of r_i and ρ_i is $n_i(n_i-1)/2 = m_i$. Specifying U_i for optimal estimation requires both the third and fourth moments of Y_{ij}, so usually we set $U_i = I_{m_i}$, to give

$$\Sigma C_i^T(r_i - \rho_i) = 0 \ .$$

When the two sets of estimating equations are used to estimate β and α, solving them iteratively but separately we have:

Given (α^k, β^k):

i) Fix α^k, solve GEE equations to get β^{k+1}, where

$$W_i^k = \left(A_i^{1/2} R_i(\alpha^k) A_i^{1/2}\right)^{-1} \ .$$

ii) Fix β^{k+1}, solve for $\alpha^{(k+1)}$ using the α estimating equations, where $\mu_i, V(\mu_i)$ are evaluated at β^{k+1}; ϕ is estimated as before.

Comments.

1. In many cases, especially with $U_i = I$, the α estimating equations can be solved non-iteratively.
2. When the data are highly unbalanced, these estimating equations for α may not be so attractive.
3. One difficulty encountered with binary data is that the correlations are not a natural measure of association as they are in the linear model setting. In particular, the correlation is restricted by the range of the data; all values between -1 and $+1$ are generally not possible.

To elaborate on point 3, consider two binary variables Y_1, Y_2, with means μ_1, μ_2. Then

$$\text{corr} \ = \frac{E(Y_1Y_2) - \mu_1\mu_2}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}}$$

. But $E(Y_1Y_2) = \Pr(Y_1 = Y_2 = 1) = \mu_{11}$ where

		Y_2		
		1	0	
Y_1	1	μ_{11}	$\mu_1 - \mu_{11}$	μ_1
	0	$\mu_2 - \mu_{11}$	$1 - \mu_2 - \mu_1 + \mu_{11}$	$(1 - \mu_1)$
		μ_2	$(1 - \mu_2)$	1

The maximum value of μ_{11} is $\min(\mu_1, \mu_2)$. Assume $\mu_1 < \mu_2$, then the max of $\mu_{11} = \mu_1$ and

$$\begin{aligned}\text{corr} &= \frac{\mu_1 - \mu_1\mu_2}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}} = \frac{\mu_1(1-\mu_2)}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}} \\ &= \frac{\sqrt{\mu_1(1-\mu_2)}}{\sqrt{(1-\mu_1)\mu_2}} < 1\end{aligned}$$

because $\mu_1 < \mu_2 \Rightarrow (1-\mu_2) < (1-\mu_1)$. The correlation can attain one only if $\mu_1 = \mu_2$.

With binary response, and sometimes with count data as well, we often use odds ratios to describe association:

$$\text{OR} = \frac{P(Y_1 = Y_2 = 1)(P(Y_1 = Y_2 = 0)}{P(Y_1 = 1, Y_2 = 0)P(Y_1 = 0, Y_2 = 1)}.$$

This is an desirable measure of association for a variety of reasons:

1. OR = 1 or ln OR = 0 implies (Y_1, Y_2) are independent.
2. ln (OR) is symmetric about 0 and unbounded; it is not constrained by the marginal moments of (Y_1, Y_2).
3. It is invariant to marginal specification of μ_1 and μ_2. That is, any (μ_1, μ_2) pair is compatible with any value of OR; this explains its appeal in case-control studies.

Various authors (Prentice, 1988; Lipsitz *et al.*, 1990, 1991; Liang *et al.*, 1992) have suggested replacing the α estimating equations by a set of odds-ratio estimating equations. The idea is that in the 2×2 table, if the margins are fixed (μ_1, μ_2), there is one remaining degree-of-freedom for determining association. We can use it to estimate the odds ratio, then calculate the correlation needed for the β equations as a function of the odds ratio. This approach has some attractive features, but has limitations when there are more than two responses $(n_i > 2)$. In this case, the set of $n_i(n_i - 1)/2$ odds ratios are given by

$$\Omega_{ijk} = \frac{P(Y_{ij} = 1,\ Y_{ik} = 1)\ P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0)\ P(Y_{ij} = 0, Y_{ik} = 1)}.$$

In particular, the parameter space of the Ω_{ijk}'s, like that of the ρ_{ijk}'s depends upon the μ_i.

Liang *et al.* (1992) proposed an extension to GEE termed GEE-1 which can be used to estimate β and α for an arbitrary parameterization of the association, say η_{ijk}, that permits a unique transformation from η_i to ρ_i. This permits one to obtain an expression for W_i in terms of $R_i(\alpha)$. Let $S_{ijk} = (Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})$ and $E(S_{ijk}) = \eta_{ijk} = E(Y_{ij}Y_{ik}) - \mu_{ij}\mu_{ik}$. Then the association parameters indexed by α can be defined in terms of η_{ijk} and μ_i.

For example,

$$\rho_{ijk} = \eta_{ijk} / \left(\mu_{ij}(1-\mu_{ij})\mu_{ik}(1-\mu_{ik})\right)^{1/2}$$

and the odds ratio Ω_{ijk} can be expressed as:

$$\Omega_{ijk} = \frac{(\eta_{ijk} + \mu_{ij}\mu_{ik})\left((1-\mu_{ij})(1-\mu_{ik}) + \eta_{ijk}\right)}{(\mu_{ij}(1-\mu_{ik}) - \eta_{ijk})\left(\mu_{ik}(1-\mu_{ij}) - \eta_{ijk}\right)}.$$

We define the α estimating equations by

$$\sum_{i=1}^{N} C_i^T U_i^{-1}(S_i - \eta_i) = 0$$

where

$$\begin{aligned} C_i &= \partial\eta_i/\partial\alpha, \\ S_i^T &= \left(S_{i12}, \ldots, S_{in_i(n_i-1)}\right), \\ \eta_i^T &= \left(\eta_{i12}, \ldots, \eta_{in_i(n_i-1)}\right). \end{aligned}$$

Again, U_i is often taken to be I. Notice that we could also use an appropriate link function, i.e., we might assume $\ln \Omega_{ijk} = Z_{ijk}^T \alpha$ for some covariates Z_{ijk}. Putting these two sets of equations together we get

$$\sum_{i=1}^{N} \begin{pmatrix} \left(\frac{\partial\mu_i}{\partial\beta}\right)^T & 0 \\ 0 & \left(\frac{\partial\eta_i}{\partial\alpha}\right)^T \end{pmatrix} \begin{pmatrix} V_i & 0 \\ 0 & U_i \end{pmatrix} \begin{pmatrix} Y_i - \mu_i \\ S_i - \eta_i \end{pmatrix}$$

These estimating equations are called GEE-1 by Liang *et al.* (1992).

COMMENT. If our primary interest is in estimating β, then asymptotic theory tells us that it matters little how we estimate α, since the same asymptotic distribution for $\widehat{\beta}$ applies for any consistent estimate of α. With finite samples, less is known about the impact of the estimate for α.