Methods Based on Generalized Linear Model Methodology

- Normal-theory methods may be inappropriate
- WLS and randomization model (CMH) methods have shortcomings
  - WLS allows only categorical covariates
  - CMH useful only in one-sample problems
  - Neither can be used in the general repeated measures setting
- In the case of one response per subject:
  - Classical linear models useful for normally-distributed outcomes with constant variance
  - Generalized linear models useful for both categorical and continuous response variables
- Extensions of GLM methodology to the repeated measurements setting are now available

Univariate Generalized Linear Models

- Generalized linear models extend classical linear models for independent normally-distributed random variables with constant variance

- The term "generalized linear model" was first introduced in a landmark paper by Nelder and Wedderburn (1972, *JRSS A*)

- Wedderburn (1974, *Biometrika*) extended the applicability by introducing quasi-likelihood

- A wide range of different problems of statistical modeling and inference were put in an elegant unifying framework:
  - Analysis of variance
  - Analysis of covariance
  - Regression models for normal, binary, Poisson outcomes, etc.

Generalized Linear Models

- The unifying theory of generalized linear models has impacted the way such statistical methods are taught
  - has provided greater insight into connections between various statistical procedures
  - has led to considerable further research
- McCullagh and Nelder (1989) provide a comprehensive account of the theory and applications of generalized linear models
- Dobson (1990) serves as an excellent introduction to the subject

References

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. London: Chapman and Hall.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

A Simple Example

- Let Y_i be a random response variable and let x_i denote an explanatory variable

- In the Gaussian linear model, we assume that

$$Y_i = \beta_0 + \beta_1 x_i + \sigma \epsilon_i,$$

  where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, 1)$

- An equivalent way of writing the model is as $Y_i \sim N(\mu_i, \sigma^2)$, where $Y_1, \ldots, Y_n$ are independent and $\mu_i = \beta_0 + \beta_1 x_i$

- The objectives of this model are to:
  - use the explanatory variable to characterize the variation in the mean of the response distribution across observational units
  - learn about the relationship between the explanatory variable and the response variable

A Simple Example (continued)

- Frequently, interest lies in formulating regression models for responses that have other continuous or discrete distributions

- While the objective is to model the mean, it often must be modeled indirectly via the use of a transformation

- In the case of a single explanatory variable, the model might be of the form $g(\mu_i) = \beta_0 + \beta_1 x_i$

- The error distribution must also be generalized, usually in a way which complements the choice of the transformation g

- This leads to a very broad class of regression models

Components of a GLM

Generalized linear models have three components:

1. *random component*

   identifies the response variable Y

   assumes a specific probability distribution for Y

2. *systematic component*

   specifies the explanatory variables used as predictors in the model

3. *link function*

   describes the functional relationship between the systematic component and the expected value (mean) of the random component

The GLM relates a function of the mean to the explanatory variables through a prediction equation having linear form

The Random Component

- Let $Y_1, \ldots, Y_n$ be independent random variables from the distribution

$$f(y; \theta, \phi) = \exp\Big\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \Big\},$$

for some specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$

If ϕ is known, this is an exponential-family model with canonical parameter θ

It may or may not be a two-parameter exponential family if ϕ is unknown

- Many common discrete and continuous distributions are members of this general family e.g., normal, gamma, binomial, Poisson

- Let $l(\theta, \phi; y)$ denote the log-likelihood function considered as a function of θ and ϕ:

$$l(\theta, \phi; y) = \log\big(f(y; \theta, \phi)\big) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

The Score Function

- It is convenient to find the mean and variance of Y using properties of the score function

$$U = \frac{\partial}{\partial \theta}\Big[l(\theta, \phi; y)\Big]$$

- To find the moments of U, we use the fact that

$$\begin{aligned} &\frac{\partial}{\partial \theta}\Big[\log\big(f(y; \theta, \phi)\big)\Big] \\ &\qquad = \frac{1}{f(y; \theta, \phi)} \frac{\partial}{\partial \theta}\Big[f(y; \theta, \phi)\Big] \end{aligned} \tag{1}$$

- Taking the expectation of both sides of (1) yields

$$\begin{aligned} &\int \frac{\partial}{\partial \theta}\Big[\log\big(f(y; \theta, \phi)\big)\Big] f(y; \theta, \phi) dy \\ &\qquad = \int \frac{\partial}{\partial \theta}\Big[f(y; \theta, \phi)\Big] dy \end{aligned} \tag{2}$$

The Score Function

- Under certain regularity conditions, the right-hand side of (2) is

$$\int \frac{\partial}{\partial\theta}\Big[f(y;\theta,\phi)\Big]dy = \frac{\partial}{\partial\theta}\left[\int f(y;\theta,\phi)dy\right]$$
$$= \frac{\partial}{\partial\theta}\big[1\big] = 0,$$

since $\int f(y;\theta,\phi)dy = 1$

- Therefore, $\mathrm{E}(U) = 0$

- Differentiating both sides of (2) with respect to θ gives

$$\frac{\partial}{\partial\theta}\left[\int \frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]f(y;\theta,\phi)dy\right]$$
$$= \frac{\partial}{\partial\theta}\left[\int \frac{\partial}{\partial\theta}\Big[f(y;\theta,\phi)\Big]dy\right] \tag{3}$$

The Score Function

- Provided that the order of differentiation and integration can be interchanged, the right-hand side of (3) is

$$\frac{\partial^2}{\partial\theta^2}\left[\int f(y;\theta,\phi)dy\right] = 0$$

and the left-hand side is

$$\begin{aligned}&\int \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]f(y;\theta,\phi)\right]dy \\ &= \int\left\{\frac{\partial^2}{\partial\theta^2}\Big[\log\big(f(y;\theta,\phi)\big)\Big]f(y;\theta,\phi)\right. \\ &\qquad \left. + \frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]\frac{\partial}{\partial\theta}\Big[f(y;\theta,\phi)\Big]\right\}dy\end{aligned} \tag{4}$$

- From (1),

$$\frac{\partial}{\partial\theta}\Big[f(y;\theta,\phi)\Big] = f(y;\theta,\phi)\,\frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]$$

The Score Function

- The second term of (4) then simplifies to

$$\frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]\; f(y;\theta,\phi)\; \frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]$$

$$= \left(\frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]\right)^2 f(y;\theta,\phi)$$

- Therefore, equation (3) becomes

$$\int \frac{\partial^2}{\partial\theta^2}\Big[\log\big(f(y;\theta,\phi)\big)\Big] f(y;\theta,\phi)dy$$

$$+ \int \left(\frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]\right)^2 f(y;\theta,\phi)dy = 0$$

or

$$\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\Big[\log\big(f(y;\theta,\phi)\big)\Big]\right]$$

$$+ \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\Big[\log\big(f(y;\theta,\phi)\big)\Big]\right)^2\right] = 0$$

The Score Function

- In terms of the score function

$$U = \frac{\partial}{\partial \theta}\Big[l(\theta, \phi; y)\Big],$$

we have $\mathrm{E}(U') + \mathrm{E}(U^2) = 0$, where $'$ denotes differentiation with respect to θ

- Thus,

$$\mathrm{E}(U) = 0$$

$$\mathrm{Var}(U) = \mathrm{E}(U^2) - [\mathrm{E}(U)]^2 = \mathrm{E}(U^2)$$

$$= -\mathrm{E}(U')$$

- The variance of U is called the *information*

Mean and Variance of Y

- $l(\theta, \phi; y) = \dfrac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$

- $U = \dfrac{\partial}{\partial\theta}\Big[l(\theta, \phi; y)\Big] = \dfrac{y - b'(\theta)}{a(\phi)}$

- $\mathrm{E}(Y) = a(\phi)\mathrm{E}(U) + b'(\theta) = b'(\theta)$

  (since $\mathrm{E}(U) = 0$)

- $U' = \dfrac{\partial}{\partial\theta}\Big[\dfrac{y - b'(\theta)}{a(\phi)}\Big] = \dfrac{-b''(\theta)}{a(\phi)}$

- Since $\mathrm{E}(U^2) = -\mathrm{E}(U')$,

$$\mathrm{E}\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right] = \frac{b''(\theta)}{a(\phi)}$$

  and $\mathrm{Var}(Y) = b''(\theta)a(\phi)$

- Note that the variance of Y is a product of two functions

Example: The Normal Distribution

- If $Y \sim N(\mu, \sigma^2)$,

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y-\mu)^2/(2\sigma^2)\} \\ &= \exp\Big\{-\frac{1}{2\sigma^2}\left(y^2 - 2y\mu + \mu^2\right) - \frac{1}{2}\log(2\pi\sigma^2)\Big\} \\ &= \exp\Big\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\Big(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big\} \end{aligned}$$

- In this case, $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \theta^2/2$, and $a(\phi) = \phi$

- $\mathrm{E}(Y) = b'(\theta) = \theta = \mu$

- $\mathrm{Var}(Y) = b''(\theta)a(\phi) = 1 \times \phi = \sigma^2$

- The variance function is $V(\mu) = 1$ and the dispersion parameter is $\phi = \sigma^2$

Example: The Poisson Distribution

- If $Y \sim P(\mu)$,

$$f(y) = \mu^y \exp(-\mu)/y!$$

$$= \exp\{y \log(\mu) - \mu - \log(y!)\}$$

$$= \exp\{y \log(\mu) - \exp(\log(\mu)) - \log(y!)\}$$

- In this case, $\theta = \log(\mu)$, $a(\phi) \equiv 1$, and $b(\theta) = e^\theta$

- $\mathrm{E}(Y) = b'(\theta) = e^\theta = \mu$

- $\mathrm{Var}(Y) = b''(\theta)a(\phi) = e^\theta = \mu$

- The variance function is $V(\mu) = \mu$ and the dispersion parameter is $\phi = 1$

Example: The Binomial Distribution

- If $Y \sim B(n,p)$, then $f(y) = \binom{n}{y} p^y (1-p)^{n-y}$

$$= \exp\left\{ \log\binom{n}{y} + y\log(p) + (n-y)\log(1-p) \right\}$$

$$= \exp\left\{ y\log\Big(\frac{p}{1-p}\Big) + n\log(1-p) + \log\binom{n}{y} \right\}$$

- In this case, $\theta = \log\big(p/(1-p)\big)$ and $a(\phi) \equiv 1$

- Since $n\log(1-p) = -n\log\Big(\frac{1}{1-p}\Big)$

$$= -n\log\Big(1 + \frac{p}{1-p}\Big),$$

$b(\theta) = n\log\big(1+\exp(\theta)\big)$

- $\mathrm{E}(Y) = b'(\theta) = ne^\theta/(1+e^\theta) = np$

- $\mathrm{Var}(Y) = b''(\theta)a(\phi) = ne^\theta/(1+e^\theta)^2 = np(1-p)$

Systematic Component

- The systematic component of a GLM specifies the explanatory variables

- These enter linearly as predictors on the right hand side of the model equation

- Suppose that each Y_i has an associated $p \times 1$ vector of covariates $x_i = (x_{i1}, \ldots, x_{ip})'$

- The linear combination $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is called the linear predictor

- Some $\{x_j\}$ may be based on others in the model, e.g.,
  - $x_3 = x_1 x_2$ allows for interaction between x_1 and x_2 in their effects on Y
  - $x_3 = x_1^2$ allows for a curvilinear effect of x_1

Link Function

- The link between the random and systematic components specifies how $\mu = \mathrm{E}(Y)$ relates to the explanatory variables in the linear predictor
- One can model the mean μ directly, or model a function $g(\mu)$ of the mean
- The model formula specifies that

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

- The function $g(\cdot)$ is called the link function
  - $g(\cdot)$ is a monotonic differentiable function
- The link function $g(\cdot)$ relates the linear predictor η_i to the expected value μ_i of Y_i
- Link functions that map the parameter space for the mean to the real line are preferred in order to avoid numerical difficulties in estimation

Types of Link Functions

Identity link:

- The simplest link function has the form $g(\mu) = \mu$
- This models the mean directly
- The identity link specifies a linear model for the mean response: $\mu = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$
- This is the form of an ordinary regression model for a continuous response

Other links permit the mean to be nonlinearly related to the predictors:

- $g(\mu) = \log(\mu)$ models the log of the mean

  Appropriate when μ cannot be negative

  A GLM with this link is called a loglinear model
- $g(\mu) = \log(\mu/(1-\mu))$ is called the logit link

  Appropriate when μ is between 0 and 1

  A GLM using this link is called a logit model

Natural Parameters and Canonical Links

- Each probability distribution for the random component has one special function of the mean that is called its *natural parameter*

Normal:	the mean itself
Poisson:	the log of the mean
Bernoulli:	logit of the success probability

- The link function that uses the natural parameter as $g(\mu)$ is called the *canonical link*

Normal:	$g(\mu) = \mu$
Poisson:	$g(\mu) = \log(\mu)$
Bernoulli:	$g(\mu) = \log\big(\mu/(1-\mu)\big)$

- Although other links are possible, the canonical links are most common in practice

- Use of the canonical link function leads to inference for β based solely on sufficient statistics

Sufficient Statistics and Canonical Links

- Let $Y_1, \ldots, Y_n$ be indep. random variables with

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

- The log-likelihood for $Y_1, \ldots, Y_n$ is

$$l = \sum_{i=1}^{n} l(\theta_i, \phi; y_i)$$

$$= \frac{1}{a(\phi)} \sum_{i=1}^{n} y_i\theta_i - \frac{1}{a(\phi)} \sum_{i=1}^{n} b(\theta_i) + \sum_{i=1}^{n} c(y_i, \phi)$$

- If $\theta_i = \eta_i = g(\mu_i) = x_i'\beta$, the first term of l is

$$\frac{1}{a(\phi)} \sum_{i=1}^{n} y_i x_i'\beta$$

- Let $X = (x_1, \ldots, x_n)'$ denote the $n \times p$ matrix of covariate values from all n subjects and let $Y = (Y_1, \ldots, Y_n)'$

Sufficient Statistics and Canonical Links

- The $p \times 1$ vector $X'Y$ with jth component $\sum_{i=1}^{n} x_{ij} Y_i$ is a sufficient statistic for β

- $\eta = \theta$ is called the canonical link function

- The canonical links lead to desirable statistical properties, particularly in small samples

- However, there is usually no a priori reason why the systematic effects in a model should be additive on the scale given by that link

- While it is convenient if effects are additive on the canonical link scale, quality of fit should be the primary model selection criterion.

- Fortunately, the canonical links are usually quite sensible on scientific grounds

Justification of Canonical Links

Normally-distributed responses:

- The identity link is plausible since both η and μ can take any value on the real line

Poisson counts:

- Since $\mu > 0$, the identity link is less attractive (since $\eta = x_i'\beta$ may be negative)
- Models for counts based on independence lead naturally to multiplicative effects
- This is expressed by the log link $\eta = \log(\mu)$

Binary responses:

- Since $0 < \mu < 1$, the link should map the interval (0,1) to the real line
- The logit function satisfies this requirement and also leads to odds ratio interpretations

Overview of Parameter Estimation

- The maximum likelihood estimates of the parameter vector β can be obtained by iterative weighted least squares

- The dependent variable is z rather than y, where z is a linearized form of the link function applied to y

- The weights are functions of the fitted values $\widehat{\mu}$

- The process is iterative because both the adjusted dependent variable z and the weight W depend on the fitted values, for which only current estimates are available

Parameter Estimation

- The log-likelihood for independent responses $Y_1, \ldots, Y_n$ is

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right]$$

- Under certain regularity conditions, the global maximum of l is the solution of $\dfrac{\partial l}{\partial \beta_j} = 0$

- By the chain rule, $\dfrac{\partial l_i}{\partial \beta_j} = \dfrac{\partial l_i}{\partial \theta_i} \dfrac{\partial \theta_i}{\partial \mu_i} \dfrac{\partial \mu_i}{\partial \eta_i} \dfrac{\partial \eta_i}{\partial \beta_j}$

- First, $\dfrac{\partial l_i}{\partial \theta_i} = \dfrac{y_i - b'(\theta_i)}{a(\phi)} = \dfrac{y_i - \mu_i}{a(\phi)}$

- Since $\mu_i = b'(\theta_i)$,

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{Var}(Y_i)}{a(\phi)} = V(\mu_i)$$

Parameter Estimation

- Since $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j, \quad \dfrac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

- Therefore,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

$$= \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

- Thus, the ML estimate of $\beta = (\beta_1, \ldots, \beta_p)'$ is the solution of the equations

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0,$$

  for $j = 1, \ldots, p$

- In general, these equations are nonlinear and must be solved numerically using iterative methods

ML Estimation using the Newton-Raphson Method

- The multidimensional analog of Newton's method requires the $p \times p$ matrix of second derivatives $\dfrac{\partial^2 l}{\partial \beta_j \partial \beta_k}$

- The mth approximation to $\widehat{\beta}$ is then given by

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]^{-1}_{\beta = b^{(m-1)}} \times U^{(m-1)}$$

- $\left[\dfrac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]_{\beta = b^{(m-1)}}$ is the matrix of second derivatives of l evaluated at the estimate of β from the $(m-1)$st iteration

- $U^{(m-1)}$ is the vector of first derivatives of l evaluated at the estimate of β from the $(m-1)$st iteration

Score Function and Information Matrix

- Let $Y_1, \ldots, Y_n$ be independent random variables whose probability distributions depend on parameters $\theta_1, \ldots, \theta_p$, where $p \leq n$

- Let $l_i(\theta; y_i)$ denote the log-likelihood function of Y_i, where $\theta = (\theta_1, \ldots, \theta_p)'$

- The log-likelihood function of $Y_1, \ldots, Y_n$ is $l(\theta, y) = \sum_{i=1}^{n} l_i(\theta; y_i)$, where $y = (y_1, \ldots, y_n)'$

- The total score with respect to θ_j is defined as

$$U_j = \frac{\partial l(\theta; y)}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\theta; y_i)}{\partial \theta_j}$$

- By the same argument as for the univariate case, $\mathrm{E}\Big[\dfrac{\partial l_i(\theta; y_i)}{\partial \theta_j}\Big] = 0$ and so $\mathrm{E}(U_j) = 0$

Score Function and Information Matrix

- The information matrix $\mathcal{I}$ is defined as the variance-covariance matrix of $U = (U_1, \ldots, U_p)'$

- $\mathcal{I} = \mathrm{E}\big[(U - \mathrm{E}(U))(U - \mathrm{E}(U))'\big] = \mathrm{E}[UU']$ has elements

$$\mathcal{I}_{jk} = \mathrm{E}[U_j U_k] = \mathrm{E}\Big[\frac{\partial l_i}{\partial \theta_j}\,\frac{\partial l_i}{\partial \theta_k}\Big]$$

- By an argument analogous to that used in the univariate case

$$\mathrm{E}\Big[\frac{\partial l_i}{\partial \theta_j}\,\frac{\partial l_i}{\partial \theta_k}\Big] = \mathrm{E}\Big[-\frac{\partial^2 l_i}{\partial \theta_j \partial \theta_k}\Big]$$

- Thus, the elements of the information matrix are also given by

$$\mathcal{I}_{jk} = \mathrm{E}\Big[-\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\Big]$$

ML Estimation using the Method of Scoring

- An alternative to Newton-Raphson involves replacing the matrix of second derivatives by the matrix of expected values $\mathrm{E}\Big[\dfrac{\partial^2 l}{\partial\beta_j\partial\beta_k}\Big]$

- Since $\mathrm{E}\Big[\dfrac{\partial^2 l}{\partial\beta_j\partial\beta_k}\Big] = -\mathrm{E}\Big[\dfrac{\partial l}{\partial\beta_j}\,\dfrac{\partial l}{\partial\beta_k}\Big] = -\mathcal{I}$, an alternative iterative procedure is given by

$$b^{(m)} = b^{(m-1)} + \big[\mathcal{I}^{(m-1)}\big]^{-1} U^{(m-1)},$$

where $\mathcal{I}^{(m-1)}$ denotes the information matrix evaluated at $b^{(m-1)}$

- Multiplication of both sides of the above equation by $\mathcal{I}^{(m-1)}$ gives

$$\mathcal{I}^{(m-1)} b^{(m)} = \mathcal{I}^{(m-1)} b^{(m-1)} + U^{(m-1)}$$

ML Estimation using the Method of Scoring

- For generalized linear models, the (j,k)th element of $\mathcal{I}$ is

$$\mathcal{I}_{jk} = \mathrm{E}\left[\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right]$$

$$= \mathrm{E}\left[\frac{(Y_i-\mu_i)\,x_{ij}}{\mathrm{Var}(Y_i)}\,\frac{\partial \mu_i}{\partial \eta_i}\,\frac{(Y_i-\mu_i)\,x_{ik}}{\mathrm{Var}(Y_i)}\,\frac{\partial \mu_i}{\partial \eta_i}\right]$$

$$= \mathrm{E}\left[\frac{(Y_i-\mu_i)^2\,x_{ij}\,x_{ik}}{[\mathrm{Var}(Y_i)]^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]$$

$$= \frac{x_{ij}\,x_{ik}}{\mathrm{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

- Thus, $\mathcal{I} = X'WX$, where W is the $n \times n$ diagonal matrix with elements

$$w_{ii} = \frac{1}{\mathrm{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

ML Estimation using the Method of Scoring

- The iterative procedure can now be written as

$$X'WXb^{(m)} = X'WXb^{(m-1)} + U^{(m-1)}$$

- The jth row of the $p \times n$ matrix $X'W$ is

$$(x_{1j}w_{11}, \ldots, x_{nj}w_{nn}) =$$

$$\left(\frac{x_{1j}}{\mathrm{Var}(Y_1)} \left(\frac{\partial \mu_1}{\partial \eta_1} \right)^2, \ldots, \frac{x_{nj}}{\mathrm{Var}(Y_n)} \left(\frac{\partial \mu_n}{\partial \eta_n} \right)^2 \right)$$

and the jth component of U is

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)} \, \frac{\partial \mu_i}{\partial \eta_i}$$

- Now let v denote the $n \times 1$ vector with ith component $(y_i - \mu_i)\dfrac{\partial \eta_i}{\partial \mu_i}$

ML Estimation using the Method of Scoring

- $U^{(m-1)}$ can now be written as $X'Wv^{(m-1)}$ and the iterative procedure becomes

$$X'WXb^{(m)} = X'WXb^{(m-1)} + X'Wv^{(m-1)}$$
$$= X'Wz$$

- The $n \times 1$ vector z has elements

$$z_i = x_i'b^{(m-1)} + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i},$$

where μ_i and $\dfrac{\partial \eta_i}{\partial \mu_i}$ are evaluated at $b^{(m-1)}$

- Provided that $X'WX$ has rank p, the vector of parameter estimates is given by

$$b^{(m)} = (X'WX)^{-1}X'Wz$$

Comments

- Normal equations are of the same form as for a linear model fitted using weighted least squares

- However, since z and W depend on b, the solution must be obtained iteratively

- The adjusted dependent variable z_i can be written as

$$\widehat{\eta}_i + (y_i - \widehat{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i},$$

where the derivative of the link is evaluated at $\widehat{\mu}$

- The first-order approximation to $g(y)$ is

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu) = \eta + (y - \mu)\frac{\partial \eta}{\partial \mu}$$

- Thus z_i is a linearized form of the link function applied to the data

ML Estimation for Canonical Links

- When the canonical link

$$\eta_i = \theta_i = \sum_{j=1}^{p} x_{ij}\beta_j = x_i'\beta$$

is used, then

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i)$$

- In this case,

$$\begin{aligned}
\frac{\partial l_i}{\partial \beta_j} &= \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) \\
&= \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)}\, b''(\theta_i) \\
&= \frac{(y_i - \mu_i)\, x_{ij}}{a(\phi)},
\end{aligned}$$

since $\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi)$

ML Estimation for Canonical Links

- Thus, $U_j = \dfrac{\partial l}{\partial \beta_j} = \displaystyle\sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{a(\phi)}$,

- The (j, k) component of the matrix of second derivatives is

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \frac{x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_k}\right)$$

- Since these components do not depend on the observations $\{Y_i\}$,

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = \mathrm{E}\Big[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\Big]$$

- Thus, the Newton-Raphson and Fisher scoring algorithms are identical

Quasi-Likelihood

- Most statisticians agree on the importance of the likelihood function in statistical inference
- In order to construct a likelihood function, we must know (or postulate) probability distributions for random variables
- In some cases, there may be no theory available on the specific random mechanism by which the data were generated
- In other situations, the appropriate theoretical probability distribution may be inadequate
- Another possibility is that the underlying theoretical model may be too complicated to permit parameter estimation and statistical inference

Quasi-Likelihood

- However, we may still have substantial information about the data, such as:
  - type of response (discrete, continuous, non-negative, symmetric, skewed, etc.)
  - whether or not the observations are statistically independent
  - how the variability of the response changes with the average response
  - the likely nature of the relationship between the mean response and one or more covariates
- In such situations, quasi-likelihood is a method for statistical inference when it is not possible to construct a likelihood function

Quasi-Likelihood

- Let $Y = (Y_1, \ldots, Y_n)'$ be a vector of independent random variables with mean vector $\mu = (\mu_1, \ldots, \mu_n)'$

- Let $\beta = (\beta_1, \ldots, \beta_p)'$ be a vector of unknown parameters ($p \leq n$)

- We will assume that the parameters of interest, β, relate to the dependence of μ on covariates x

- This will be denoted by the notation that Y_i has mean $\mu_i(\beta)$

- We will also assume that $\text{Var}(Y_i) = \phi\, V(\mu_i)$, where $V(\cdot)$ is a known function and ϕ is a possibly unknown scale parameter

Quasi-Likelihood

- Thus, $\text{Var}(Y) = \phi\, V(\mu)$, where

$$V(\mu) = \text{diag}\{V(\mu_1), \ldots, V(\mu_n)\}$$

- It is important to note that:
  - ϕ is assumed constant for all subjects and does not depend on β
  - $\text{Var}(Y_i)$ depends only on μ_i

    (mathematically necessary, but also physically sensible)
  - It would be permissible to have $\text{Var}(Y_i) = \phi\, V_i(\mu_i)$

    i.e., a possibly different functional relationship for each observation

Construction of Quasi-Likelihood Function

- Consider the random variable $U_i = \dfrac{Y_i - \mu_i}{\phi\, V(\mu_i)}$

- U_i has the following properties in common with a log-likelihood derivative:

$$\mathrm{E}(U_i) = 0,$$

$$\mathrm{Var}(U_i) = \mathrm{E}(U_i^2) = \frac{\mathrm{E}[(Y_i - \mu_i)^2]}{[\phi\, V(\mu_i)]^2} = \frac{1}{\phi\, V(\mu_i)},$$

$$\mathrm{E}\Big(\frac{\partial U_i}{\partial \mu_i}\Big) = \mathrm{E}\bigg[\frac{-\phi\, V(\mu_i) - (Y_i - \mu_i)\phi\, V'(\mu_i)}{[\phi\, V(\mu_i)]^2}\bigg]$$

$$= -\frac{1}{\phi\, V(\mu_i)} = -\mathrm{Var}(U_i)$$

- Most first-order asymptotic theory connected with likelihood functions is founded on the above three properties

Construction of Quasi-Likelihood Function

- Thus, it should not be surprising that the integral

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)}\, dt,$$

  if it exists, should behave like a log-likelihood function for μ_i

- We refer to $Q(\mu_i; y_i)$ as the quasi-likelihood for μ_i based on data y_i

  (more correctly, as the log quasi-likelihood)

- Since the components of Y are independent, the quasi-likelihood for the complete data is

$$Q(\mu; y) = \sum_{i=1}^{n} Q(\mu_i; y_i)$$

Example: The Normal Distribution

- If $Y \sim N(\mu, \sigma^2)$, then $V(\mu) = 1$ and $\phi = \sigma^2$
- In this case, $U = \dfrac{Y - \mu}{\sigma^2}$
- The quasi-likelihood function is

$$\begin{aligned} Q(\mu, y) &= \int_y^\mu \frac{y - t}{\sigma^2}\, dt \\ &= \frac{1}{\sigma^2}\left[yt - \frac{t^2}{2}\right]_y^\mu \\ &= \frac{1}{\sigma^2}\left[y\mu - \frac{\mu^2}{2} - y^2 + \frac{y^2}{2}\right] \\ &= \frac{1}{2\sigma^2}\left[2y\mu - \mu^2 - y^2\right] \\ &= -\frac{(y-\mu)^2}{2\sigma^2} \end{aligned}$$

- This is equivalent to the log likelihood for $N(\mu, \sigma^2)$

Example: The Poisson Distribution

- If $Y \sim P(\mu)$, then $V(\mu) = \mu$ and $\phi \equiv 1$

- In this case, $U = \dfrac{Y - \mu}{\mu}$

- The quasi-likelihood function is

$$\begin{aligned} Q(\mu, y) &= \int_y^\mu \frac{y - t}{t}\, dt \\ &= \int_y^\mu \left(\frac{y}{t} - 1\right) dt \\ &= \Big[y \log(t) - t\Big]_y^\mu \\ &= y \log(\mu) - \mu - y \log(y) + y \end{aligned}$$

- In comparison, the log likelihood for $P(\mu)$ is

$$y \log(\mu) - \mu - \log(y!)$$

Example: The Bernoulli Distribution

- If $Y \sim B(1,p)$, then $\mu = p$, $V(\mu) = \mu(1-\mu)$ and $\phi \equiv 1$
- In this case, $U = \dfrac{Y-p}{p(1-p)}$
- The quasi-likelihood function is

$$\begin{aligned} Q(p,y) &= \int_y^p \frac{y-t}{t(1-t)}\,dt \\ &= \int_y^p \left[\frac{y}{t} + \frac{y-1}{1-t}\right] dt \\ &= \Big[y\log(t) - (y-1)\log(1-t)\Big]_y^p \\ &= y\log\left(\frac{p}{1-p}\right) + \log(1-p) - f(y) \end{aligned}$$

- In comparison, the log likelihood for $B(1,p)$ is

$$y\log\left(\frac{p}{1-p}\right) + \log(1-p) + \log\binom{1}{y}$$

QL Estimating Equations

- If we treat the quasi-likelihood function as if it were a "true" log likelihood, the estimate of β_j satisfies the equation

$$0 = \frac{\partial Q(\mu; y)}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{\partial Q(\mu_i; y_i)}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{\partial Q(\mu_i; y_i)}{\partial \mu_i} \left(\frac{\partial \mu_i}{\partial \beta_j} \right)$$

$$= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi \, V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \beta_j} \right)$$

QL Estimating Equations

- In terms of matrices and vectors, let

$$y_{(n\times 1)} = (y_1, \ldots, y_n)'$$

$$\mu_{(n\times 1)} = (\mu_1, \ldots, \mu_n)'$$

$$V_{(n\times n)} = \text{diag}\{V(\mu_1), \ldots, V(\mu_n)\}$$

$$D_{(n\times p)} = \left(\frac{\partial \mu}{\partial \beta}\right),$$

  where the (i, j) component of D is $\dfrac{\partial \mu_i}{\partial \beta_j}$

- The QL estimating equation is $U(\widehat{\beta}) = 0$, where

$$U(\beta) = D'V^{-1}(y - \mu)/\phi$$

- $U(\beta)$ is called the quasi-score function

QL Estimating Equations

- The covariance matrix of $U(\beta)$, which is also the negative expected value of $\dfrac{\partial U(\beta)}{\partial \beta}$, is $\mathcal{I} = D'V^{-1}D/\phi$

- For QL functions, the matrix $\mathcal{I}$ plays the same role as the Fisher information for ordinary likelihood functions

- In particular, the asymptotic covariance matrix of $\widehat{\beta}$ is

$$\text{Var}(\widehat{\beta}) = \mathcal{I}^{-1} = \phi(D'V^{-1}D)^{-1}$$

- Consistency, asymptotic normality, and optimality are discussed by McCullagh (1983)

QL Estimation of β

- Beginning with an arbitrary estimate $b^{(0)}$ sufficiently close to β, the sequence of parameter estimates generated by the Newton-Raphson method with Fisher scoring is

$$\begin{aligned} b^{(m)} &= b^{(m-1)} + \left[\mathcal{I}^{(m-1)}\right]^{-1} U^{(m-1)} \\ &= b^{(m-1)} + \Big\{ \left[\phi (D' V^{-1} D)^{-1}\right] \\ &\qquad \times \left[D' V^{-1} (y - \mu)/\phi\right] \Big\} \\ &= b^{(m-1)} + (D' V^{-1} D)^{-1} D' V^{-1} (y - \mu), \end{aligned}$$

  where μ, D and V are evaluated at $\mu^{(m-1)}$

- An important property of the estimation procedure is that it does not depend on the value of ϕ

QL Estimation of ϕ

- In the above respects, the quasi-likelihood behaves just like an ordinary log likelihood

- The one exception is in the estimation of ϕ

- The conventional estimator of ϕ is a moment estimator based on the residual vector $y - \widehat{\mu}$, namely

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} = \frac{X^2}{n-p}$$

- X^2 is the generalized Pearson statistic

Comparison Between Quasi-Likelihood and Generalized Linear Models

- The random component of a GLM assumes a specific distribution for the response Y_i

- Quasi-likelihood assumes only a form for the functional relationship between the mean and the variance

- The QL estimating equations for β are

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi\, V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \beta_j} \right) = 0, \qquad j = 1, \ldots, p$$

- The likelihood equations for generalized linear models are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0, \qquad j = 1, \ldots, p$$

Comparison Between Quasi-Likelihood and Generalized Linear Models

- Since

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij},$$

  and

$$\mathrm{Var}(Y_i) = \phi \, V(\mu_i),$$

  the QL estimating equations have the same form as the GLM likelihood equations

- However, QL estimators make only second-moment assumptions about the distribution of $\{Y_i\}$, rather than full distributional assumptions

- Quasi-likelihood can also be motivated in terms of least squares

Characteristics of Methods Based on Generalized Linear Model Methodology

- Useful for discrete and continuous outcomes
  - normal, Poisson, binomial & gamma responses
  - generalizations for ordered categorical data
- No. of repeated measurements per experimental unit need not be constant
- Measurement times need not be the same across subjects
- Covariates may be discrete or continuous, time-independent or time-dependent
- Missing data (MCAR) can be accommodated
- Three types of extensions:
  - Marginal models
  - Random effects models
  - Transition models

Marginal Models

- The marginal expectation $\mu_{ij} = \mathrm{E}(y_{ij})$ is modelled as a function of explanatory variables
  - marginal expectation: the average response over the subpopulation that shares a common value of the covariate vector
- Associations among repeated observations are modelled separately
- The assumptions are as follows:
  a. $g(\mu_{ij}) = x'_{ij}\beta$, where $x'_{ij} = (x_{ij1}, \ldots, x_{ijp})$
  b. $\mathrm{Var}(y_{ij}) = \phi\, V(\mu_{ij})$
     - V is a known variance function
     - ϕ is a possibly unknown scale parameter
  c. The covariance between y_{ij} and $y_{ij'}$ is a known function of μ_{ij}, $\mu_{ij'}$, and a vector of unknown parameters α

Random Effects Models

- Heterogeneity between individuals is accounted for by subject-specific random effects

- These are assumed to account for all of the within-subject correlation present in the data

- Conditional on the values of the random effects, the responses are assumed to be independent

- The assumptions are as follows:

  a. $g\big(\mathrm{E}(y_{ij} \mid b_i)\big) = x'_{ij}\beta + z'_{ij}b_i$

     - b_i is a vector of subject-specific effects
     - z_{ij} is a vector of covariates

  b. $y_{i1}, \ldots, y_{it_i}$ are independent given b_i, for each $i = 1, \ldots, n$

  c. $b_1, \ldots, b_n$ are i.i.d. with density f

Transition Models

- $y_{i1}, \ldots, y_{it_i}$ are correlated because y_{ij} is explicitly influenced by past values $y_{i1}, \ldots, y_{i,j-1}$
- The past outcomes are treated as additional predictor variables
- Conditional expectation of current response, given past responses, is assumed to follow a GLM
- The linear predictor includes:
  - original covariates
  - additional covariates which are known functions of past responses
- The model is

$$g\big(\mathrm{E}(y_{ij} \mid y_{i1}, \ldots, y_{i,j-1})\big) = x'_{ij}\beta + \sum_{r=1}^{s} f_r(y_{i1}, \ldots, y_{i,j-1}; \alpha_1, \ldots, \alpha_s)$$

Comparison of the Three Approaches

- In the linear model case, the three approaches can be formulated to have regression coefficients with the same interpretation
  (coefficients from random effects and transition models can have marginal interpretations)

- Categorical outcome variables, however, require nonlinear link functions

- In this case, the three approaches give different interpretations for the regression coefficients

Transition Model:

- Expresses the conditional mean of y_{ij} as a function of covariates and of past responses

- Difficult to formulate models so that β has the same meaning for different assumptions about the time dependence

Comparison of the Three Approaches

Random Effects Model:

- A "subject-specific" ("cluster-specific") approach
- Heterogeneity among individuals is explicitly modelled using individual-specific effects
- Regression coefficients have interpretations in terms of the influence of covariates on both:
  - an individual's response
  - the average response of the population

Marginal Model:

- A "population-averaged" approach
- Appropriate when inferences about the population average are the focus
- Scientific objectives are to characterize and contrast populations of subjects

Comparison of the Three Approaches

- Marginal models model the effects of covariates on the marginal expectations

  (A model for the association among observations from each subject must also be specified)

- Random effects and transition models model the covariate effects and within-subject associations through a single equation

- In a clinical trial, marginal models are likely to be most appropriate

  (since the average difference between control and treatment is generally most important)

- In addition, software for fitting marginal models is more widely available

The GEE Method

- GEE: generalized estimating equations
  (Liang & Zeger, 1986; Zeger & Liang, 1986)
- An extension of quasi-likelihood to longitudinal data analysis
- The method is semi-parametric in that the estimating equations are derived without full specification of the joint distribution of a subject's observations
- Instead, we specify only the:
  - likelihood for the (univariate) marginal distributions
  - "working" covariance matrix for the vector of repeated measurements from each subject
- Often referred to now as GEE1
  (to distinguish it from more recent extensions)

The GEE Method

- The GEEs have consistent and asymptotically normal solutions, even with misspecification of the time dependence

- The method avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time point

- The covariance structure is treated as a nuisance

- It relies on the independence across subjects to estimate consistently the variance of the regression coefficients (even when the assumed correlation is incorrect)

Advantages of GEE

- Feasible in many situations where maximum likelihood approaches are not, since the full multivariate distribution of the response vector is not required
  - For example, five binary responses per subject gives a multinomial distribution with $2^5 - 1 = 31$ independent parameters
  - With GEE, only the five marginal probabilities and at most $5 \times 4/2 = 10$ correlations are estimated
- Efficiency loss relative to maximum likelihood is often minimal
- Continuous and categorical independent variables can be handled (unlike WLS)

Outline of the GEE Method

a. Relate the marginal response $\mu_{ij} = \mathrm{E}(y_{ij})$ to a linear combination of the covariates:

$$g(\mu_{ij}) = x'_{ij}\,\beta$$

- y_{ij} is the response for subject i at time j
- $x_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is the corresponding $p \times 1$ vector of covariates
- $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown parameters
- $g(\cdot)$ is the link function

b. Describe the variance of y_{ij} as a function of the mean:

$$\mathrm{Var}(y_{ij}) = V(\mu_{ij})\,\phi$$

- $V(\cdot)$ is the variance function
- ϕ is a possibly unknown scale parameter

Link and Variance Functions

- Normally-distributed response:

$$g(\mu_{ij}) = \mu_{ij},$$

$$V(\mu_{ij}) = 1,$$

$$\text{Var}(y_{ij}) = \phi$$

- Binary response:

$$g(\mu_{ij}) = \log\big(\mu_{ij}/(1-\mu_{ij})\big),$$

$$V(\mu_{ij}) = \mu_{ij}(1-\mu_{ij}),$$

$$\phi \equiv 1$$

- Poisson response:

$$g(\mu_{ij}) = \log(\mu_{ij}),$$

$$V(\mu_{ij}) = \mu_{ij},$$

$$\phi \equiv 1$$

Outline of the GEE Method

c. Choose the form of a $t_i \times t_i$ 'working' correlation matrix $R_i(\alpha)$ for each $y_i = (y_{i1}, \ldots, y_{it_i})'$

- The (j, j') element of $R_i(\alpha)$ is the known, hypothesized, or estimated correlation between y_{ij} and $y_{ij'}$

- This working correlation matrix may depend on a vector of unknown parameters α, which is the same for all subjects

- Although this correlation matrix can differ from subject to subject, we commonly use a working correlation matrix $R(\alpha)$ that approximates the average dependence among repeated observations over subjects

Comments on "Working" Correlation Models

- We should choose the form of R to be consistent with the empirical correlations

- R is called a working correlation matrix because with non-normal responses, the actual correlation among a subject's outcomes may depend on the mean values, and hence on $x_{ij}'\beta$

- The GEE method yields consistent estimates of the regression coefficients and their variances, even with misspecification of the structure of the covariance matrix

- In addition, the loss of efficiency from an incorrect choice of R is inconsequential when the number of subjects is large

"Working" Correlation Models

Independence: $R = I$

- When $n >> t$, the correlation influence is often small enough so that ordinary least-squares regression coefficients are nearly efficient

- However, correlation may have a substantial effect on the estimated variances

- These considerations suggest the independence working model with $R = I$

- Solving the GEE is the same as fitting the usual regression models for independent data

- Hence, one can use available software to obtain parameter estimates

"Working" Correlation Models

Completely-specified: $R = R_0$

- Choosing R_0 close to the true (unknown) correlation gives increased efficiency
- Unfortunately, the choice is usually not obvious

Exchangeable: $R_{jj'} = \alpha$

- This is the correlation structure assumed in a random effects model

AR-1: $R_{jj'} = \alpha^{|j-j'|}$

- for normally-distributed y_{ij}, the correlation structure of the continuous time analogue of the first-order autoregressive process

"Working" Correlation Models

Stationary m-dependent:

$$R_{jj'} = \begin{cases} \alpha^{|t_j - t_{j'}|} & \text{if } |t_j - t_{j'}| \leq m \\ 0 & \text{if } |t_j - t_{j'}| > m \end{cases},$$

where t_j is the jth observation time

Unspecified: $R_{jj'} = \alpha_{jj'}$

- In this case, there are $t(t-1)/2$ parameters to be estimated
- Most efficient, but useful only when there are relatively few observation times
- In addition, the occurrence of missing data complicates estimation of R
- The estimate obtained using nonmissing data is not guaranteed to be positive definite

Choosing a Working Correlation Matrix

- Nature of the problem may suggest a structure:
  - Repeated measurements over time
    - Autoregressive, unstructured
  - Individuals within families (clustered data)
    - Exchangeable
- When the number of experimental units is large and the cluster sizes are small, the choice of R often has little impact on the estimation of β
  - Independence model may suffice
- When there are many repeated measurements per experimental unit, modeling the correlation structure may result in increased efficiency
- Consideration of alternative working correlation structures may be useful

Generalized Estimating Equation

- A_i is a $t_i \times t_i$ diagonal matrix with $V(\mu_{ij})$ as the jth diagonal element

- $R_i(\alpha)$ is the $t_i \times t_i$ "working" correlation matrix for the ith subject

- The working covariance matrix for $y_i = (y_{i1}, \ldots, y_{it_i})'$ is

$$V_i(\alpha) = \phi \, A_i^{1/2} \, R_i(\alpha) \, A_i^{1/2}$$

- The GEE estimate of β is the solution of

$$U(\beta) = \sum_{i=1}^{n} \Big(\frac{\partial \mu_i}{\partial \beta} \Big)' \big[V_i(\widehat{\alpha}) \big]^{-1} (y_i - \mu_i) = 0_p,$$

where $\widehat{\alpha}$ is a consistent estimate of α and 0_p is the $p \times 1$ vector $(0, \ldots, 0)'$

Solving the GEE

- Iterate between quasi-likelihood methods for estimating β and a robust method for estimating α as a function of β

  1. Given current estimates of $R_i(\alpha)$ and ϕ, calculate an updated estimate of β using iteratively reweighted least squares

  2. Given the estimate of β, calculate standardized residuals

$$r_{ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{\sqrt{[V_i]_{jj}}}$$

  3. Use the residuals r_{ij} to consistently estimate α

  4. Repeat steps 1.–3. until convergence

Robust Variance Estimate

- One approach to estimating the variance-covariance matrix of $\widehat{\beta}$ would be to use the inverse of the Fisher information matrix:

$$\text{Var}(\widehat{\beta}) = M_0^{-1},$$

where

$$M_0 = \sum_{i=1}^{n} \Big(\frac{\partial \widehat{\mu}_i}{\partial \beta} \Big)' V_i^{-1} \Big(\frac{\partial \widehat{\mu}_i}{\partial \beta} \Big)$$

and $V_i = V_i(\widehat{\alpha})$

- This is called the "model-based" estimator of $\text{Var}(\widehat{\beta})$

- Will not provide a consistent estimator of $\text{Var}(\widehat{\beta})$ unless the underlying model is correct
  - Royall (1986)

Robust Variance Estimate

- Liang and Zeger (1986) recommend the estimator

$$\mathrm{Var}(\widehat{\beta}) = M_0^{-1}\, M_1\, M_0^{-1},$$

  where M_1 is given by

$$\sum_{i=1}^{n} \Big(\frac{\partial \widehat{\mu}_i}{\partial \beta}\Big)' \, V_i^{-1} \, (y_i - \widehat{\mu}_i) \, (y_i - \widehat{\mu}_i)' \, V_i^{-1} \Big(\frac{\partial \widehat{\mu}_i}{\partial \beta}\Big)$$

- This estimator was defined by Royall (1986)
  - known as the "robust" or "information sandwich" estimator
  - a consistent estimator of $\mathrm{Var}(\widehat{\beta})$ even if $R_i(\alpha)$ is not the true correlation matrix of y_i

- If the true correlation structure is correctly modeled, then the robust variance estimator reduces to the model-based estimator

Example

- A randomized, double-blind clinical trial of a new source of botulinum toxin Type A in 75 patients with spasmodic torticollis
  - sponsored by an English company
  - conducted and first analyzed in Germany
  - considered for purchase by a U.S. company

- Patients previously untreated with botulinum toxin were randomized to one of four groups:
  - placebo ($n = 20$)
  - 250 units of botulinum toxin A ($n = 19$)
  - 500 units of botulinum toxin A ($n = 18$)
  - 1000 units of botulinum toxin A ($n = 18$)

- Following a single injection, patients were evaluated at weeks 2, 4, and 8

Example (continued)

- One of the primary outcome variables was a clinical global rating (CGR)
  - 1=symptom free or mild symptoms
  - 0=moderate or severe symptoms

- Covariates of interest include:
  - treatment group (0, 250, 500, 1000 units)
  - age (range: 26-82 years, mean: 47 years)
  - sex (39 males, 36 females)
  - week (2, 4, 8)

- With six exceptions, the data are complete:
  - two patients (both in the 500 unit group) have no follow-up data
  - one patient (1000 unit) missing at week 2
  - one patient (1000 unit) missing at week 4
  - two patients (both placebo) missing at week 8

Clinical Global Ratings

				(0=poor, 1=good)		
ID	Group	Age	Sex	Wk. 2	Wk. 4	Wk. 8
1	Plac.	82	F	0	0	0
2	500	41	F	0	0	0
3	250	62	F	0	0	1
4	1000	63	M	0	0	1
5	500	40	M	1	1	1
6	250	43	F	1	1	1
7	1000	56	F	0	0	0
8	Plac.	48	F	0	0	0
9	1000	34	F	0	1	1
10	500	35	M	0	0	0
11	Plac.	27	M	0	0	0
12	250	39	F	1	1	1
13	1000	54	M	0	0	0
14	500	52	F	.	.	.
15	Plac.	48	M	0	0	0
16	250	55	M	0	0	0
17	1000	79	M	1	0	0
18	250	42	M	0	0	0
19	Plac.	36	M	0	0	.

Clinical Global Ratings (continued)

				(0=poor, 1=good)		
ID	Group	Age	Sex	Wk. 2	Wk. 4	Wk. 8
20	500	26	F	1	1	1
21	1000	60	F	1	1	1
22	Plac.	48	F	0	0	0
23	250	50	F	0	0	0
24	500	29	M	1	0	1
25	1000	44	F	0	1	0
26	500	41	F	0	0	0
27	Plac.	50	M	0	0	0
28	250	53	M	0	1	0
29	250	45	M	0	1	0
30	Plac.	42	M	0	0	0
31	1000	63	F	1	1	1
32	500	47	M	0	0	0
33	1000	36	F	1	1	0
34	250	29	F	1	0	0
35	Plac.	54	M	1	1	1
36	250	44	F	0	0	0
37	1000	55	M	0	0	0
38	500	34	F	1	1	1

Clinical Global Ratings (continued)

				(0=poor, 1=good)		
ID	Group	Age	Sex	Wk. 2	Wk. 4	Wk. 8
39	Plac.	52	M	0	0	0
40	Plac.	48	M	0	0	0
41	250	58	M	0	0	0
42	500	57	F	0	0	0
43	1000	43	M	1	1	1
44	250	46	F	0	0	0
45	500	33	M	0	0	1
46	Plac.	39	F	0	0	0
47	1000	53	F	0	0	1
48	250	51	M	0	0	0
49	500	72	F	0	1	0
50	1000	41	F	0	.	1
51	Plac.	36	M	0	0	1
52	500	53	F	0	1	1
53	250	50	M	0	0	0
54	1000	64	M	0	0	0
55	Plac.	49	M	1	1	.
56	250	29	M	0	0	0
57	Plac.	51	M	0	0	0

Clinical Global Ratings (continued)

				(0=poor, 1=good)		
ID	Group	Age	Sex	Wk. 2	Wk. 4	Wk. 8
58	1000	46	M	.	1	1
59	500	53	M	.	.	.
60	250	42	F	1	1	1
61	Plac.	30	F	0	0	0
62	250	46	M	1	1	0
63	1000	49	M	0	0	0
64	Plac.	33	M	0	0	0
65	500	66	F	1	1	1
66	500	37	M	1	1	1
67	1000	36	F	1	1	0
68	Plac.	49	F	0	0	0
69	500	35	F	1	1	1
70	1000	37	F	1	1	1
71	250	39	F	1	0	0
72	Plac.	46	M	0	0	0
73	Plac.	53	F	1	0	0
74	250	59	M	0	0	0
75	500	55	M	0	1	0

Analysis Issues

Type of model (marginal, transitional, etc.):

- Marginal models appropriate when inferences about average response in subpopulation sharing common covariate vector value are the focus
- Reasonable considering the goals of a clinical trial (since average difference between treatments is generally most important)

Type of working correlation structure:

- Unspecified model is appropriate
  - only three time points, so only 3 parameters
  - data are nearly complete
- Independence and exchangeable working correlation structures can also be considered

Response variable:

- Logit of the probability of a good response
- Binomial variance function

Notation

- $n = 75$ subjects (clusters)
- $t_i = 3$ observations per subject
- Y_{ij} is the response from the ith subject at the jth time point, for $i = 1, \ldots, 75$, $j = 1, \ldots, 3$

$$Y_{ij} = \begin{cases} 1 & \text{if CGR is “good”} \\ 0 & \text{if CGR is “poor”} \end{cases}$$

- $x_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ is a $p \times 1$ vector of covariates for subject i at time j
- The regression model is

$$\log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = x_{ij}'\beta,$$

where $\mu_{ij} = \mathrm{E}(Y_{ij})$ and $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown parameters

Covariates

- $x_{ij1} = 1$ $\quad x_{ij2} = \text{age}$ $\quad x_{ij3} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$
- $x_{ij4} = \begin{cases} 1 & 250 \text{ units} \\ 0 & \text{otherwise} \end{cases}$ $\quad x_{ij5} = \begin{cases} 1 & 500 \text{ units} \\ 0 & \text{otherwise} \end{cases}$
- $x_{ij6} = \begin{cases} 1 & 1000 \text{ units} \\ 0 & \text{otherwise} \end{cases}$ $\quad x_{ij7} = \begin{cases} 0 & \text{placebo} \\ 1 & 250 \text{ units} \\ 2 & 500 \text{ units} \\ 4 & 1000 \text{ units} \end{cases}$
- $x_{ij8} = \begin{cases} 1 & \text{if } j = 2 \\ 0 & \text{otherwise} \end{cases}$ $\quad x_{ij9} = \begin{cases} 1 & \text{if } j = 3 \\ 0 & \text{otherwise} \end{cases}$
- $x_{ij,10} = x_{ij2}\, x_{ij8}$ $\quad x_{ij,11} = x_{ij2}\, x_{ij9}$
- $x_{ij,12} = x_{ij3}\, x_{ij8}$ $\quad x_{ij,13} = x_{ij3}\, x_{ij9}$
- $x_{ij,14} = x_{ij4}\, x_{ij8}$ $\quad x_{ij,15} = x_{ij5}\, x_{ij8}$ $\quad x_{ij,16} = x_{ij6}\, x_{ij8}$
- $x_{ij,17} = x_{ij4}\, x_{ij9}$ $\quad x_{ij,18} = x_{ij5}\, x_{ij9}$ $\quad x_{ij,19} = x_{ij6}\, x_{ij9}$

SAS Data Step Statements

```
data a; infile 'example.dat';
input subject dose age sex wk2 wk4 wk8;
dose=int(dose/250+0.001);
dosesq=dose*dose;
dosecu=dose*dosesq;
dose2=(dose=1);
dose3=(dose=2);
dose4=(dose=4);
male=(sex=1);
week=2; rating=wk2; week4=0; week8=0;
output;
week=4; rating=wk4; week4=1; week8=0;
output;
week=8; rating=wk8; week4=0; week8=1;
output;
data a; set a;
age4=age*(week=4);   age8=age*(week=8);
male4=male*(week=4); male8=male*(week=8);
dose24=dose2*(week=4);
dose28=dose2*(week=8);
dose34=dose3*(week=4);
dose38=dose3*(week=8);
dose44=dose4*(week=4);
dose48=dose4*(week=8);
```

Model 1

Age, sex, dose (3 parameters), week (2 parameters), wk. 4 & 8 incremental effects of age (2 parameters), sex (2 parameters), and dose (6 parameters)

- 18 regression parameters (including intercept)
- Focus is on assessing differential effects of age, sex, and dose at weeks 2, 4, and 8
- The SAS statements are:

```
proc genmod data=a;
class subject;
model rating=age age4 age8 male male4 male8
 dose2 dose3 dose4 dose24 dose28 dose34
 dose38 dose44 dose48 week4 week8 / dist=bin;
repeated subject=subject
 / type=unstr corrw covb sorted;
make 'geeemppest' out=estimate;
make 'geercov' out=cov noprint;
make 'classlevels' out=junk1 noprint;
make 'parminfo' out=junk2 noprint;
make 'modfit' out=junk3 noprint;
make 'geencov' out=junk4 noprint;
```

Model 1 (continued)

- Wald tests of the joint significance of sets of parameters can be computed using PROC IML

```
proc iml; use estimate;
read all var{estimate}
  where(parm ne "Scale") into estimate;
use cov; read all into cov; n=nrow(cov);
do i=2 to n; im1=i-1; do j=1 to im1;
cov[i,j]=cov[j,i]; end; end;
col={"Chi-square" "df" "p-value"};
print "Age X Week Interaction";
x=estimate[3:4,1]; print x;
var=cov[3:4,3:4]; print var;
df=nrow(x); q=x`*inv(var)*x;
pvalue=1-probchi(q,df);
result=q :: df :: pvalue;
print result [colname=col format=7.3];
print "Sex X Week Interaction";
x=estimate[6:7,1]; print x;
var=cov[6:7,6:7]; print var;
df=nrow(x); q=x`*inv(var)*x;
pvalue=1-probchi(q,df);
result=q :: df :: pvalue;
print result [colname=col format=7.3];
   ...
quit;
```

Results from Model 1

- Wald tests of the interaction effects are:

Effect	Chi-square	df	p-value
Age × Week	0.36	2	0.8
Sex × Week	0.12	2	0.9
Dose × Week	5.00	6	0.5
All Interactions	5.98	10	0.8

- The results obtained from the independence and exchangeable working correlation structures were similar

- Since model 1 has a large number of parameters relative to the number of observations, separate models with main effects and only one of the interaction effects were also considered

- In each of these models, there was also no evidence of interactions with week

Model 2

Age, sex, dose (3 parameters), week (2 parameters)

- Eight regression parameters (with intercept)
- Two parameterizations are used:
  - Indicator variables for dose
  - Linear, quadratic, and cubic dose
- The SAS statements are:

```
proc genmod data=a;
class subject;
model rating=age male dose2 dose3 dose4
             week4 week8 / dist=bin;
repeated subject=subject
       / type=unstr corrw covb sorted;

proc genmod data=a;
class subject;
model rating=age male dose dosesq dosecu
             week4 week8 / dist=bin;
repeated subject=subject
       / type=unstr corrw covb sorted;
```

Model 2 (continued)

- Multiple degree of freedom contrasts were also tested using PROC IML
- Wald tests of interest:

Effect	Chi-square	df	p-value
Age	3.45	1	0.06
Sex	2.31	1	0.13
Age and Sex	6.58	2	0.04
Dose	9.88	3	0.02
250 vs placebo	1.62	1	0.20
500 vs placebo	6.60	1	0.01
1000 vs placebo	6.86	1	0.01
Nonlinear dose	3.05	2	0.22
Week	0.56	2	0.76

- Similar results were obtained from independence and exchangeable working correlation structures
- Since the week effect is nonsignificant, these two terms will first be omitted

Model 3

Age, sex, three dose parameters

- Six regression parameters (with intercept)
- Two parameterizations are used:
  - Indicator variables for dose
  - Linear, quadratic, and cubic dose
- The SAS statements are:

```
proc genmod data=a;
class subject;
model rating=age male
       dose2 dose3 dose4 / dist=bin;
repeated subject=subject
         / type=unstr corrw covb sorted;

proc genmod data=a;
class subject;
model rating=age male
      dose dosesq dosecu / dist=bin;
repeated subject=subject
      / type=unstr corrw covb sorted;
```

Model 3 (continued)

- Multiple df contrasts were also tested
- The results from this model are:

Wald Tests			
Effect	Chi-square	df	p-value
Age	3.42	1	0.06
Sex	2.44	1	0.12
Age and Sex	6.69	2	0.03
Dose	9.80	3	0.02
250 vs placebo	1.48	1	0.22
500 vs placebo	6.36	1	0.01
1000 vs placebo	6.80	1	0.01
Nonlinear dose	2.82	2	0.24

	Regression Coefficient		
Covariate	Estimate	S.E.	Odds Ratio
Age	−0.03	0.02	0.97
Male gender	−0.66	0.42	0.52
250 units	0.88	0.73	2.42
500 units	1.88	0.74	6.53
1000 units	1.91	0.73	6.77

Model 4

Age, sex, linear dose

```
proc genmod data=a; class subject;
model rating=age male dose / dist=bin;
repeated subject=subject
   / type=unstr corrw covb sorted;
make 'geeemppest' out=estimate;
make 'geercov' out=cov noprint;
make 'classlevels' out=junk1 noprint;
make 'parminfo' out=junk2 noprint;
make 'modfit' out=junk3 noprint;
make 'geencov' out=junk4 noprint;
proc iml;
use estimate; read all var{estimate}
  where(parm ^= "Scale") into estimate;
use cov; read all into cov; n=nrow(cov);
do i=2 to n; im1=i-1; do j=1 to im1;
cov[i,j]=cov[j,i]; end; end;
col={"Chi-square" "df" "p-value"};
print "Age and Sex";
x=estimate[2:3,1]; print x;
var=cov[2:3,2:3]; print var;
df=nrow(x); q=x`*inv(var)*x;
pvalue=1-probchi(q,df);
result=q :: df :: pvalue;
print result [colname=col format=7.3];
quit;
```

Results from Model 4

Wald Tests			
Effect	Chi-square	df	p-value
Age	3.83	1	0.050
Sex	2.62	1	0.105
Age and Sex	7.22	2	0.027
Linear dose*	9.30	1	0.002

	Regression Coefficient		
Covariate	Estimate	S.E.	Odds Ratio
Age	−0.03	0.02	0.97
Male gender	−0.70	0.43	0.50
Linear dose*	0.44	0.14	1.55

*0=placebo, 1=250 units, 2=500 units, 4=1000 units

- Odds of a good response:
  - decrease as age increases
  - lower for males than females
  - increase as dose increases

Comments on the Analysis

- In model 3, the test of nonlinearity of the dose effect is not significant
  - chi-square=2.82 with 2 df, p=0.24
- However, parameter estimates of the effects for the two highest doses (500, 1000 units) are:
  - nearly identical
  - twice as large as those for the 250 unit dose (1.88 and 1.91, respectively, versus 0.88)
- Thus, the model with indicator effects for dosage may be most appropriate
- The results that follow are based on model 3 parameterized with indicator dosage effects

Effect of Working Correlation Structure

Covariate	Working Correlation	Regression Coefficient Est.	S.E.	z
Age	Unspecified	−0.0285	0.0154	−1.85
	Exchangeable	−0.0281	0.0152	−1.85
	Independence	−0.0285	0.0153	−1.86
Male sex	Unspecified	−0.6627	0.4239	−1.56
	Exchangeable	−0.6707	0.4244	−1.58
	Independence	−0.7221	0.4262	−1.69
250 units	Unspecified	0.8819	0.7254	1.22
	Exchangeable	0.9022	0.7322	1.23
	Independence	0.9850	0.7325	1.34
500 units	Unspecified	1.8757	0.7440	2.52
	Exchangeable	1.8465	0.7492	2.46
	Independence	1.9294	0.7508	2.57
1000 units	Unspecified	1.9122	0.7334	2.61
	Exchangeable	1.9026	0.7402	2.57
	Independence	1.9614	0.7393	2.65

Working Correlation Matrices from Model with Age, Sex, and Dose Indicators

- Unspecified working correlation structure:

$$\begin{pmatrix} 1.00 & 0.61 & 0.39 \\ 0.61 & 1.00 & 0.41 \\ 0.39 & 0.41 & 1.00 \end{pmatrix}$$

- Exchangeable working correlation structure:

$$\begin{pmatrix} 1.00 & 0.54 & 0.54 \\ 0.54 & 1.00 & 0.54 \\ 0.54 & 0.54 & 1.00 \end{pmatrix}$$

- Independence working correlation structure:

$$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$$

Comparison with Univariate Analyses

		Regression Coefficient		
Covariate	Model	Est.	S.E.	z
Age	GEE	−0.029	0.015	−1.85
	Week 2	−0.034	0.025	−1.37
	Week 4	−0.020	0.024	−0.81
	Week 8	−0.034	0.025	−1.33
Male sex	GEE	−0.663	0.425	−1.56
	Week 2	−0.824	0.551	−1.50
	Week 4	−0.737	0.547	−1.35
	Week 8	−0.650	0.567	−1.15
250 units	GEE	0.882	0.725	1.22
	Week 2	0.893	0.819	1.09
	Week 4	1.359	0.906	1.50
	Week 8	0.708	0.953	0.74
500 units	GEE	1.876	0.744	2.52
	Week 2	1.300	0.824	1.58
	Week 4	2.314	0.910	2.54
	Week 8	2.220	0.918	2.42
1000 units	GEE	1.912	0.733	2.61
	Week 2	1.379	0.842	1.64
	Week 4	2.348	0.918	2.56
	Week 8	2.187	0.921	2.37

Comment

- Disadvantages of Wald statistics for testing hypotheses about individual parameters or sets of parameters
  - Dependent on the measurement scale
  - Require estimation of the covariance matrix (unstable if number of clusters is small and cluster size is large)

- With n=75 and t_i=3, their performance may be satisfactory

- Rotnitzky and Jewell (1990, *Biometrika*) discuss the use of 'working' score and likelihood ratio tests
  - these are not yet implemented in standard software packages

Assessing the Adequacy of the Working Correlation Matrix

- Inferences regarding the regression coefficients β can be made using the:

  1. robust variance estimator $M_0^{-1} M_1 M_0^{-1}$
     - consistent even if $R(\alpha)$ is misspecified
     - may be inefficient
  2. naive (model-based) variance estimator M_0^{-1}
     - assumes that $R(\alpha)$ is correctly specified

- Consider testing the hypothesis that the first q components of β are equal to specified values

- Rotnitzky and Jewell (1990, *Biometrika*) show that if variance estimation is based on M_0^{-1}, the Wald statistic is asymptotically equal to

$$c_1 X_1 + c_2 X_2 + \cdots c_q X_q$$

Assessing the Adequacy of the Working Correlation Matrix

- $c_1 \geq c_2 \geq \cdots \geq c_q \geq 0$ are the eigenvalues of a matrix Q
- Q is a function of $\left(\frac{\partial \mu_i}{\partial \beta}\right)$, V_i, and A_i
- $X_1, \ldots, X_q$ are independent χ^2_1 random variables
- Examination of the weights c_j provides information on:
  - how close the working correlation matrix $R(\alpha)$ is to the true correlation structure
  - the effect of a particular choice of V_i on inference about the components of β
- The asymptotic mean and variance of the Wald statistic are $\sum c_j$ and $2\sum c_j^2$, respectively

Assessing the Adequacy of the Working Correlation Matrix

- If V_i is close to $\text{Cov}(y_i)$, then $\overline{c}_1 = \sum c_j/q$ and $\overline{c}_2 = \sum c_j^2/q$ will both approximately equal 1

- Points close to $(1, 1)$ in a plot of $\overline{c}_1$ versus $\overline{c}_2$ for different choices of $R(\alpha)$ indicate reasonable choices of the working correlation structure

- Note that $\overline{c}_1$ and $\overline{c}_2$ can be computed without computation of the individual eigenvalues
  - $q\overline{c}_1 = \text{trace}(Q), \quad q\overline{c}_2 = \text{trace}(Q^2)$

- Probability statements about $\overline{c}_1$ and $\overline{c}_2$ would, however, require the null distribution of $\widehat{Q}$

- Hadgu et al. (1997, *Statistics in Medicine*) and Hadgu (1998, *J Biopharm Statist*) demonstrate the use of this approach

Studies of the Properties of GEE

Emrich LJ and Piemonte MR (1992). *J Statist Comput Simul* 41:19–29

Gunsolley JC, Getchell C & Chinchilli VM (1995). *Commun Statist Simul Comput* 24:869–878

Li N (1994). Unpublished Ph.D. dissertation

Lipsitz SR, Laird NM, and Harrington DP (1991). *Biometrika* 78:153–160

Lipsitz SR, Fitzmaurice GM, Orav EJ, and Laird NM (1994). *Biometrics* 50:270–278

Paik M (1988). *Commun Statist Simul Comput* 17:1155–1171

Park T (1993). *Statist Med* 12:1723–1732

Park T, Davis CS, and Li N (1998). *Comput Statist Data Analysis* 28:243–256

Sharples S and Breslow N (1992). *J Statist Comput Simul* 42:1–20

Properties of GEE (for Categorical Data)

- Lipsitz, Laird, and Harrington (1991) simulated binary data with $n = 100$, $t = 2$, $p = 1$ and seven correlation structures
  - Parameter estimates biased slightly upward
  - Bias increased as the correlation increased
  - Confidence interval coverage probabilities were close to nominal 95%
  - Additional simulations with $n = 40$ led to convergence problems

- Emrich and Piedmonte (1992) simulated binary data with $n = 20$, $t = 64$, $p = 4$, and four correlation structures
  - Parameter estimates were unbiased
  - Type I error rates were inflated (from 5%)
    - to as high as 8% for individual parameters
    - to as much as 17% for joint tests

Properties of GEE (for Categorical Data)

- Lipsitz et al. (1994) simulated binary data with $n = 15, 30, 45$, $t = 3$, $p = 4$ and three exchangeable correlation structures
  - Type I error rates were close to nominal 5%
  - Confidence interval coverage probabilities were close to nominal 95%

- Li (1994) simulated binary data with $n = 25$, 50, 100, 200, $t = 3$, $p = 1, 2, 3$ and four correlation structures
  - Test sizes and confidence interval coverages were close to nominal levels
  - GEE with unspecified correlation structure had convergence problems with $n = 25$
  - Properties of WLS estimates and confidence intervals were similar to those from GEE (even when $n = 25$)

Properties of GEE (for Continuous Data)

- Paik (1988) investigated the small sample properties for correlated gamma data in a limited study
  - $t = 4$, $p = 1$: point estimates and confidence intervals perform satisfactorily if $n \geq 30$
  - $t = 4$, $p = 4$: point estimates and confidence intervals perform satisfactorily if $n \geq 50$

- Park (1993) simulated multivariate normal data ($t = 4$) with $p = 3$, $n = 30, 50$, and missing data probabilities of 0.1, 0.2, and 0.3
  - For $n = 30$, confidence interval coverage probabilities are less than nominal levels
  - For $n = 50$, coverage probabilities are close to nominal levels
  - GEE estimators are more sensitive to missing data than the MLEs

Computer Software

Programs for GEE1:

- Karim and Zeger (1988) SAS macro
  - requires PROC IML
- Lipsitz and Harrington (1990)
- Davis (1993) FORTRAN program
  - runs on any type of computer
  - not as user-friendly
- Carey (1994) S-PLUS program
  - available from STATLIB
- SUDAAN Release 7 (MULTILOG procedure)
- SAS (version 6.12) GENMOD procedure

Cautions Concerning the Use of GEE

- GEE is semiparametric (not nonparametric)
  - correct specification of marginal mean and variance are required
- Missing data cannot depend upon observed or unobserved responses
- A moderate to large number of independent experimental units (n) is required
- Bias & efficiency for finite samples may depend on
  - Number of experimental units (n)
  - Distribution of cluster sizes
  - Magnitudes of the correlations among repeated measurements
  - Number and type of covariates
    e.g., cluster level (time-independent) and/or observation level (time-dependent)

A Caution with Time-Dependent Covariates

- Pepe and Anderson (1994). *Commun Statist Simul Comput*, 23, 939–951

- When there are time-dependent covariates, $\widehat{\beta}$ may not always be a consistent estimator of β

- In this case, one must either:

  1. use a diagonal working covariance matrix

  2. verify that the marginal expectation $\mathrm{E}(y_{ij}|x_{ij})$ is equal to the partly-conditional expectation $\mathrm{E}(y_{ij}|x_{i1}, \ldots, x_{it_i})$

- Note that when covariates are time-independent, the second condition is trivially satisfied

- Pepe and Anderson (1994) describe some general classes of correlation structures for which condition 2. does and does not hold

Alternative GEE Estimation Procedures

- The second step of the GEE iteration procedure uses the Pearson residuals

$$r_{ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{\sqrt{[V_i]_{jj}}}$$

- Although this choice may be most appropriate for continuous, normally-distributed outcomes, it may not be best for categorical responses

- In univariate generalized linear models, other types of residuals have been considered:
  - Anscombe residual
  - Deviance residual

- Modifying the GEE estimation procedure to use a type of residual more appropriate to the response variable might lead to better properties

Anscombe Residuals

- Anscombe (1953) proposed defining a residual using a function $A(y)$ instead of y

- The function A is chosen to make the distribution of $A(y)$ more normal, and is given by

$$\int \frac{d\mu}{V^{1/3}(\mu)}$$

- Then for Poisson outcomes,

$$r_{ij}^{A} = \frac{\frac{3}{2}(y_{ij}^{2/3} - \widehat{\mu}_{ij}^{2/3})}{\widehat{\mu}_{ij}^{1/6}}$$

- For binary outcomes, $A(y) = \int_0^y t^{-1/3}(1-t)^{-1/3}dt$

- This can be computed using algorithms for the incomplete beta function $I(\frac{2}{3}, \frac{2}{3})$

Deviance Residuals

- In univariate generalized linear models, the deviance is often used as a measure of discrepancy

- The deviance residual is the signed square root of the contribution of each observation to the likelihood ratio statistic

- For Poisson outcomes, the deviance residual is

$$r_{ij}^D = \text{sign}(y_{ij} - \widehat{\mu}_{ij})\sqrt{2(y_{ij}\log(y_{ij}/\widehat{\mu}_{ij}) - y_{ij} + \widehat{\mu}_{ij})}$$

- For binary outcomes,

$$r_{ij}^D = \begin{cases} -\sqrt{2\,|\log(1-\widehat{\pi}_{ij})|} & \text{if } y_{ij} = 0 \\ \sqrt{2\,|\ln(\widehat{\pi}_{ij})|} & \text{if } y_{ij} = 1 \end{cases}$$

where $\widehat{\pi}_{ij}$ is the predicted probability at the current value of β

Comparisons of GEE Estimation Procedures

- The three approaches were compared in a model for generating correlated categorical responses with arbitrary covariance structure

- The specific case considered was:
  - Two groups ($p = 1$ dichotomous covariate)
  - Three time points ($t = 3$)
  - y_{hij} is the dichotomous (0,1) response at time j for subject i in group h, for $h = 1, 2$, $i = 1, \ldots, n_h$, and $j = 1, 2, 3$

- The model was

$$\text{logit}(y_{1ij}) = \beta_1 + \beta_3\, j,$$

$$\text{logit}(y_{2ij}) = \beta_1 + \beta_2 + (\beta_3 + \beta_4)j,$$

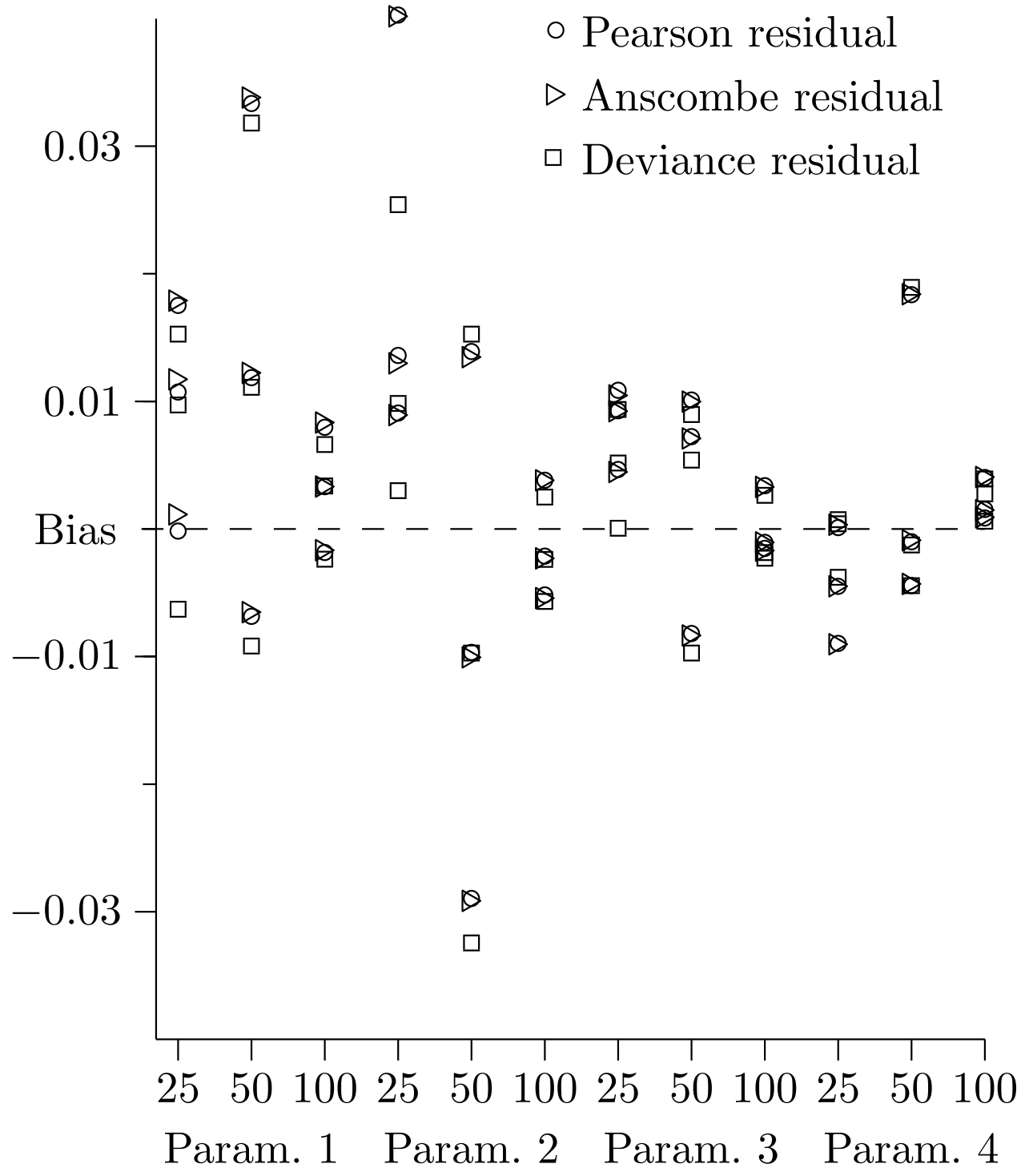
  with $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.1, 0.2, 0.2, 0.0)$

  (A linear logistic model with separate intercepts and a common slope)

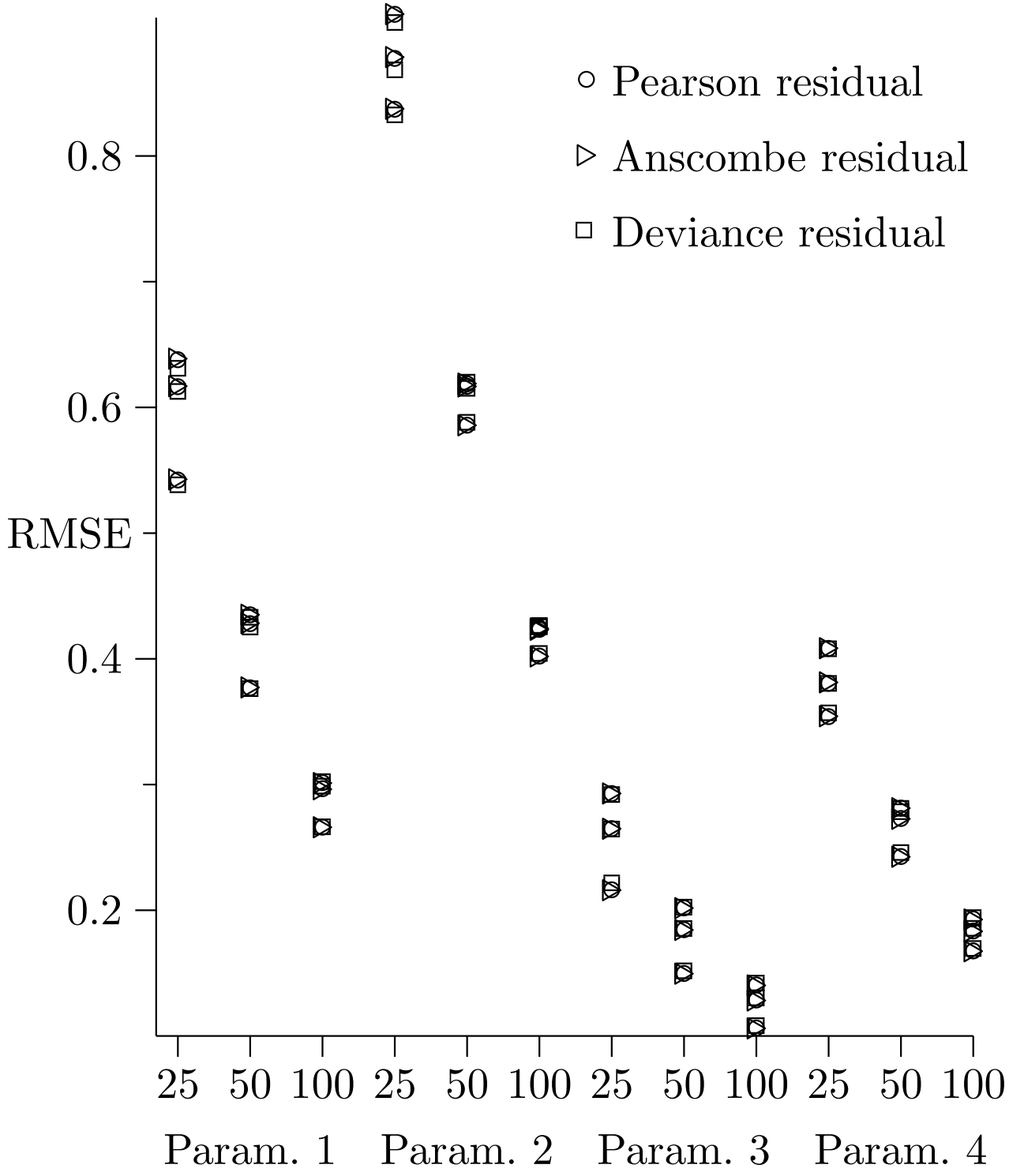
Comparisons of GEE Estimation Procedures

- Three correlation structures were considered:
  - common AR-1 with $\rho = 0.5$
  - exchangeable ($\rho = 0.5$) in group 1, AR-1 ($\rho = 0.5$) in group 2
  - exchangeable ($\rho = 0.1$) in group 1, AR-1 ($\rho = 0.5$) in group 2
- Model parameters were estimated using GEE with the unstructured working correlation model
- Sample sizes of 25, 50, and 100 observations per group were studied
- 2000 replications were carried out for each combination of the model parameters
- The results were summarized in terms of bias, root mean square error, confidence interval coverage, and test size (for $H_0: \beta_4 = 0$)

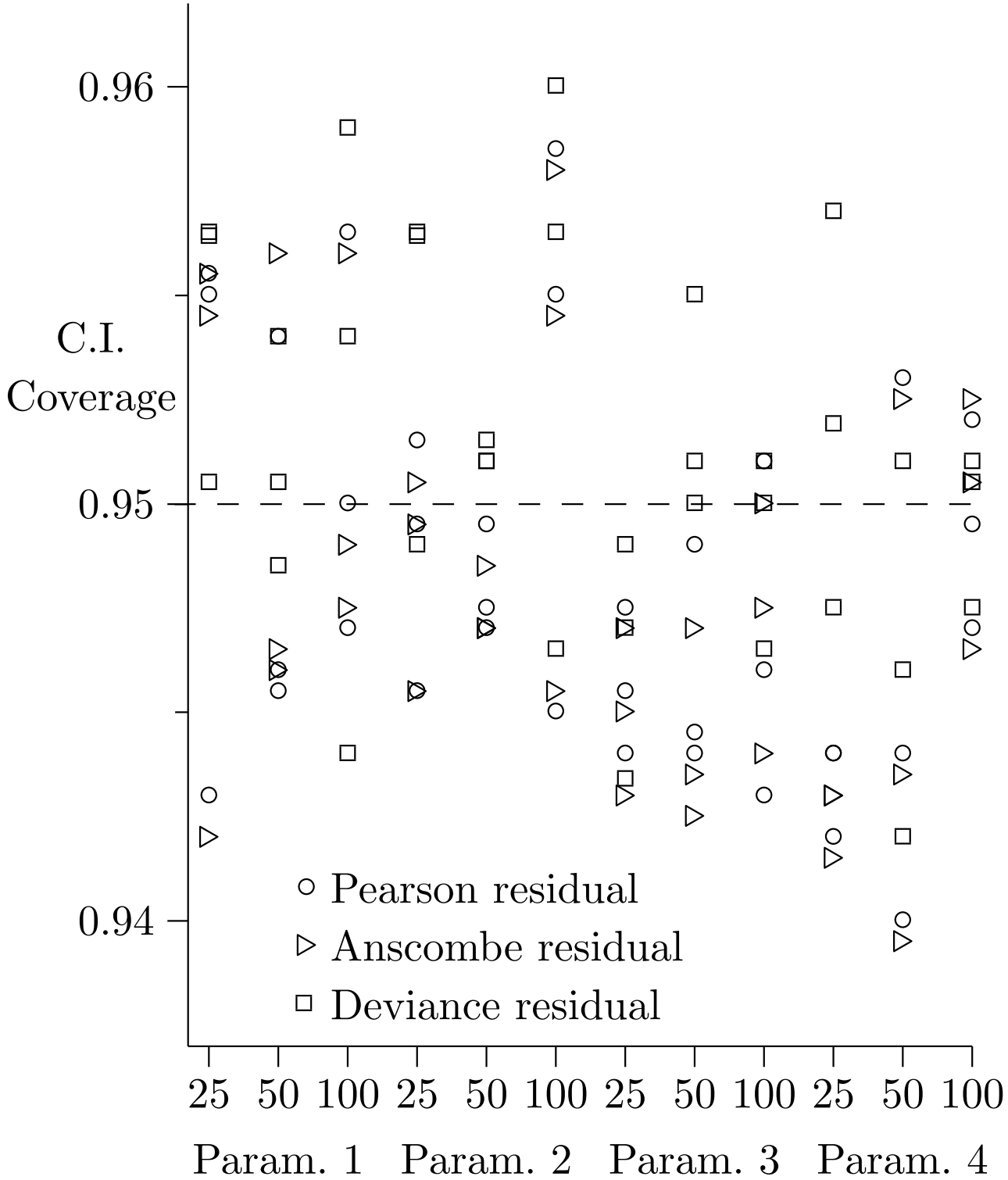
Alternative GEE Estimation: Binary Outcome
Bias versus Sample Size



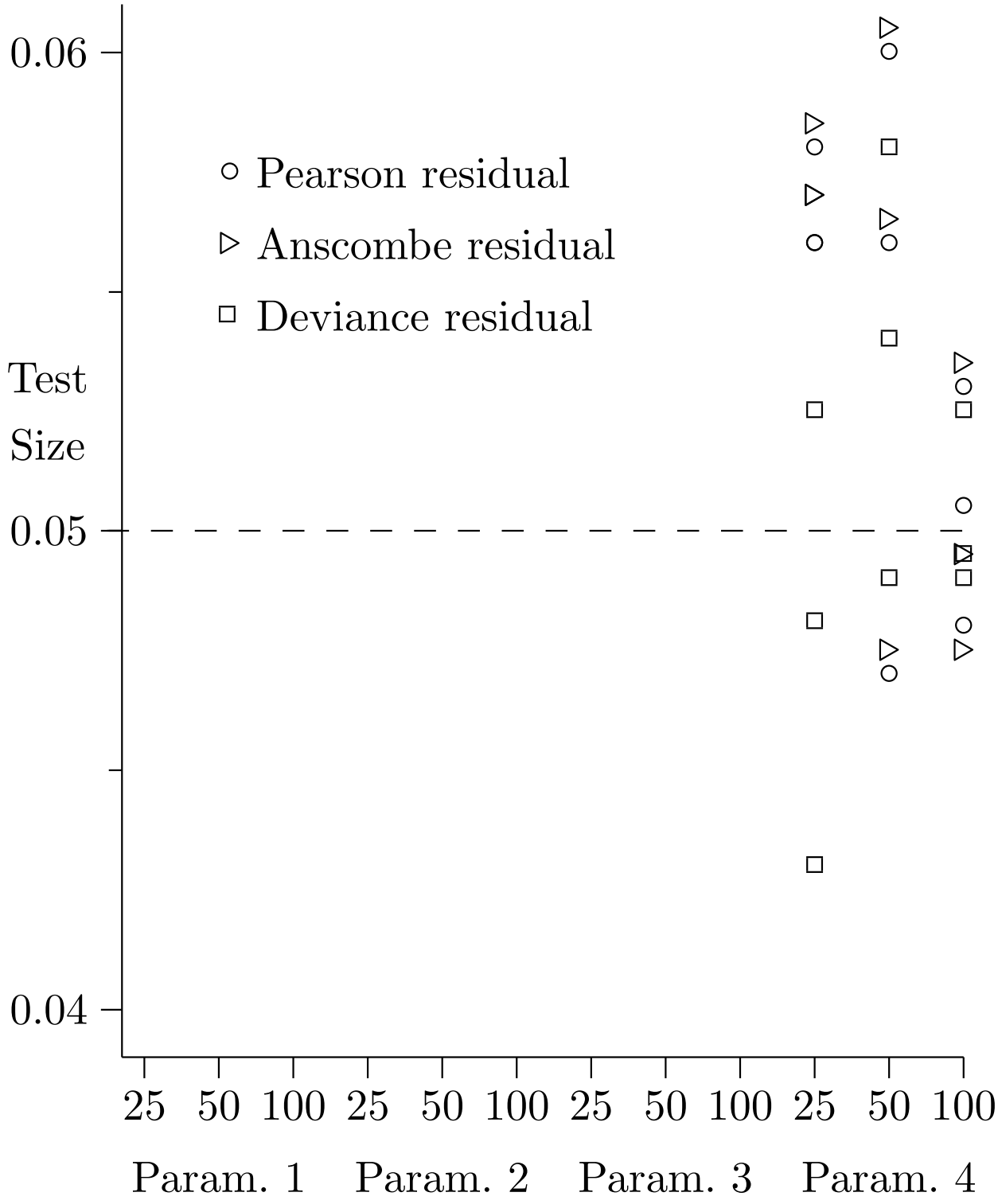
Alternative GEE Estimation: Binary Outcome
Root Mean Square Error versus Sample Size



Alternative GEE Estimation: Binary Outcome
C.I. Coverage versus Sample Size



Alternative GEE Estimation: Binary Outcome
Test Size versus Sample Size



Power Comparisons

- The powers of the three estimation methods were also compared

- The hypothesis of parallelism (H_0: $\beta_4 = 0$) was tested at the alternatives $\beta_4 = 0.1, 0.3, 0.5, 0.7$

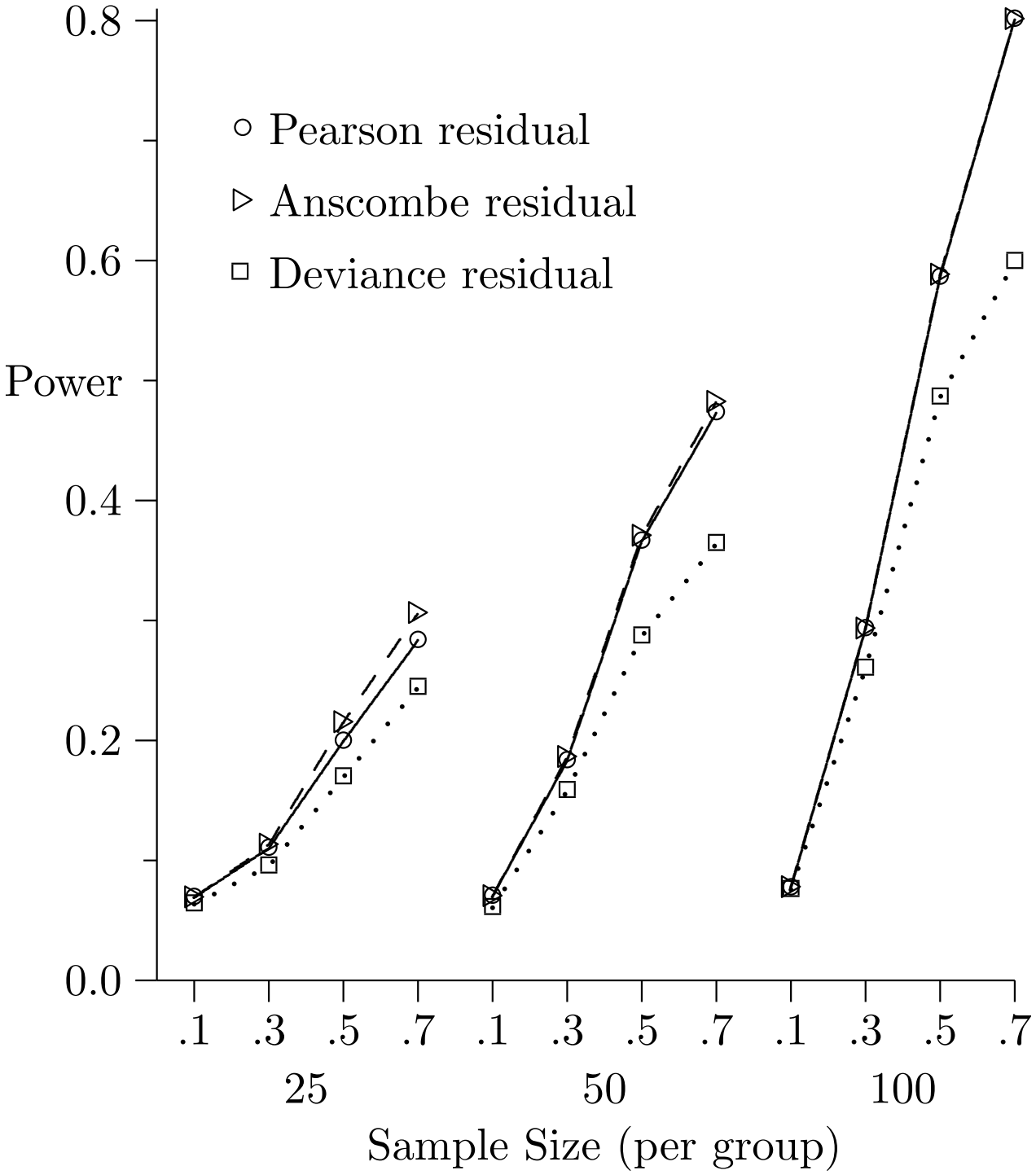
- The true correlation model was:
  - exchangeable ($\rho = 0.1$) in group 1
  - AR-1 ($\rho = 0.5$) in group 2

- Model parameters were estimated using the unstructured working correlation model

- Sample sizes of 25, 50, and 100 observations per group were studied

- 2000 replications were carried out for each combination of the factors studied

Alternative GEE Estimation: Binary Outcome
Power of Test for Parallelism versus
Treatment Difference



Summary of Results

- There are no clear distinctions among methods
- The properties of the GEE estimates, confidence intervals, and test sizes are satisfactory even when correlation structures differ among groups
- In particular, test sizes were between 0.04 and 0.06 for all sample sizes considered
- Estimation using deviance residuals gives lower power than Pearson or Anscombe residuals
- The conclusions based on simulations of Poisson outcomes are similar
- There is no compelling reason to consider use of alternatives to Pearson residuals

Other Developments and Extensions

- Lipsitz, Laird, and Harrington (1991) study using the odds ratio as the measure of association (instead of the Pearson correlation coefficient)
  - may be easier to interpret
  - pairwise odds ratios are not constrained by the marginal probabilities
  - not constrained to be in the interval $(-1, 1)$
  - approach applies only to binary outcomes
  - in a simulation study with $n = 100$, $t = 2$, and $p = 1$, the parameter estimates from the odds ratio association model appeared to be slightly more efficient

- Carey et al. (1993) ALR method

- Chaganty (1997) QLS method

Other Developments and Extensions

- Lipsitz, Fitzmaurice, Orav, and Laird (1994). *Biometrics*, 50, 270–278

- A one-step estimator to circumvent convergence problems associated with the GEE estimation algorithm was proposed

- In a simulation study with a binary response, $n = 15, 30, 45$, $t = 3$, and $p = 4$, the performance of the one-step estimator was similar to that of the fully iterated estimator

- They recommend the one-step approach when the sample size is small and the association between binary responses is high

  (In this case, the fully iterated GEE algorithm often has convergence problems)

Other Developments and Extensions

- Robins, Rotnitzky, and Zhao (1995, *JASA*) propose an extension of GEE that allows for data to be MAR, rather than MCAR
- Thus, the probability that y_{ij} is missing may depend on past values of the outcome and covariates
- However, correct specification of a model for the probability of nonresponse is required
- Rotnitzky and Wypij (1994, *Biometrics*) propose a general approach for calculating the asymptotic bias of GEE estimators calculated from incomplete data
- In an example, they show that use of the exchangeable working correlation structure can result in larger bias than the independence working correlation model

Subsequent Developments

Prentice (1988)

- Considered the special case of binary data
- Proposed GEE estimator of the vector α of correlation parameters
- Improved efficiency vs. original GEE formulation

Zhao & Prentice (1990), Prentice & Zhao (1991), Liang, Zeger & Qaqish (1992)

- Proposed alternative equation for simultaneous estimation of regression parameters β and covariance parameters α
- Requires modeling the third and fourth moments of y_{ij} (instead of just the mean and variance)
- This extension is now called GEE2 and the original formulation is GEE1

Distinctions Between GEE1 and GEE2

- In GEE1, the regression parameters β are considered to be orthogonal to association parameters α (even though they are not)

- GEE1 thus gives consistent estimates of β even when association parameters are modeled incorrectly

- GEE2 gives consistent estimates of β and α only when the marginal means and associations are modeled correctly

- In this case, GEE2 provides parameter estimates which have high efficiency relative to maximum likelihood

Distinctions Between GEE1 and GEE2

- GEE1 gives slightly less efficient estimates of β, but may give inefficient estimates of α
- GEE2 sacrifices the appeal of requiring only first and second moment assumptions

Conclusion:

- Use GEE1 if regression parameters are the primary focus
- Use GEE2 if:
  - efficient estimation of association parameters is of interest
  - model for covariance structure is known to be correctly specified

Other Developments and Extensions

- Hall (1994) and Hall & Severini (1998) propose extended generalized estimating equations (EGEE)

- Uses ideas from extended quasi-likelihood
  - Nelder and Pregibon (1987, *Biometrika*)
  - McCullagh and Nelder (1989)

- Provides estimating equations for regression and association parameters simultaneously

- Makes only first and second moment assumptions

- Estimates α efficiently (like GEE2)
  - consistency of $\widehat{\alpha}$ requires correct covariance specification

- Does not require a correct covariance specification for consistency of regression parameter estimates

Summary

- Recent extensions of generalized linear model methodology are especially useful in the analysis of repeated categorical and continuous outcomes

- In many types of applications, marginal models may be more appropriate than random effects and transition models
  - and software is more widely available

- GEE1 (and EGEE) require weaker assumptions than GEE2
  - and GEE1 software is more widely available

- GEE1 estimators and test statistics generally have satisfactory properties

Comments on Random Effects Models

- More difficult to fit, since evaluation of the likelihood (or even the first two moments) requires numerical methods in most cases

- Mauritsen (1984) proposed a mixed effects model known as the logistic binomial
  - eases the computational burden
  - available in the EGRET software package

- Conaway (1990, *Biometrics*) proposed a random effects model for binary data based on the complementary log-log link and a log-gamma random effects distribution
  - yields a closed form expression for the full likelihood, thus simplifying likelihood analysis
  - regression parameters, however, do not have log odds ratio interpretations

Comments on Random Effects Models

- One approach to avoiding numerical integration is to approximate the integrands with simple expansions whose integrals have closed forms
  - Stiratelli, Laird, and Ware (1984, *Biometrics*)
  - Breslow and Clayton (1993, *JASA*)

- These approximate techniques give effective estimates of the fixed effects, but are biased for estimating random effects and the random effects variance matrix

- Waclawiw and Liang (1993, *JASA*) propose an alternative strategy based on optimal estimating equations

- Zeger and Karim (1991, *JASA*) describe a Bayesian approach using Gibbs sampling

Comments on Random Effects Models

- The SAS macro GLIMMIX fits generalized linear mixed models using restricted pseudo likelihood (REPL)
  - Wolfinger and O'Connell (1993). *J Statist Comput Simul*, 48, 233–243

- For the mixed effects logistic model, estimates of fixed effects and variance components are biased under some common conditions:
  - moderate to large variance components, i.e., moderate to large within-cluster correlation
  - small to moderate cluster sizes

- This was shown by Kuk (1995, *JRSS B*) and Breslow and Lin (1995, *Biometrika*)

- These authors provide methods that reduce the bias (but not yet implemented in GLIMMIX)

Additional Developments

- Version 7 of SAS contains an experimental procedure, PROC NLMIXED, for fitting nonlinear models with fixed and random effects
- Estimation techniques are not the same as those used in the NLINMIX and GLIMMIX macros
- Parameters are estimated by maximizing an approximation to the likelihood integrated over the random effects
- Different integral approximations are available, including:
  - adaptive Gaussian quadrature
  - a first-order Taylor series approximation
- A variety of alternative optimization techniques are available to carry out the maximization
  - the default is a dual quasi-Newton algorithm

Methods for the Analysis of Ordered Categorical Repeated Measurements

Three general approaches to the analysis of ordered categorical repeated measurements:

1. CMH mean score and correlation tests
   - Applicable only in the one-sample setting
   - Landis et al. (1988, *Statistics in Medicine*)

2. Weighted least squares
   - Polytomous logit, cumulative logit, and mean score response functions for the one-sample and multi-sample repeated measures settings
   - Unless sample sizes are quite large, only mean score models may be feasible

3. Methods based on extensions of generalized linear model methodology

Generalized Linear Model Methodology

- Extensions of GEE1 studied by:
  - Stram, Wei, and Ware (1988, *JASA*)
  - Liang, Zeger, and Qaqish (1992, *JRSS B*)
  - Agresti, Lipsitz, and Lang (1992, *JSCS*)
  - Kenward & Jones (1992, *J Biopharm Statist*)
  - Miller, Davis and Landis (1993, *Biometrics*)
  - Lipsitz, Kim, and Zhao (1994, *Statist Med*)

- Software:
  - Shaw, Kenward et al. (1994) SAS macro
  - Lipsitz, Kim, and Zhao (1994) SAS macro
  - SUDAAN Release 7 (MULTILOG procedure)
  - FORTRAN program for Stram-Wei-Ware (Davis and Hall, 1996)

Ordinal Response Considerations

- Polytomous response variables are often ordinal
- Advantageous to construct logits that:
  - account for category ordering
  - are less affected by the number or choice of response categories

    i.e., if a new category is formed by combining adjacent categories of the old scale, the form of the conclusions should be unaffected
- Unnecessary to restrict consideration to only two response categories at a time
- Instead, logits can be formed by grouping categories that are contiguous
- These considerations lead to models based on cumulative response probabilities

Cumulative Logits

- Cumulative response probabilities are

$$\gamma_j = \Pr(Y \leq j), \qquad j = 0, 1, \ldots, c$$

- Thus, $\gamma_0 = \pi_0$, $\gamma_1 = \pi_0 + \pi_1, \ldots, \gamma_c = 1$

- Cumulative logits are

$$\lambda_j = \log\Big(\frac{\gamma_{j-1}}{1 - \gamma_{j-1}}\Big) \quad j = 1, \ldots, c$$

- Each cumulative logit uses all $c + 1$ response categories

- A model for λ_j is similar to the ordinary logit model for a binary response

  (categories 0 to $j - 1$ form the first category and categories j to c form the second category)

The Proportional Odds Model

- $\lambda_j(x) = \alpha_j + x'\beta, \quad \text{for } j = 1, \ldots, c$
- $x' = (x_1, \ldots, x_p)$ is a vector of explanatory variables
- $\beta' = (\beta_1, \ldots, \beta_p)$ is a vector of unknown parameters
- Relationship between x_k and a dichotomized response Y does not depend on j, the point of dichotomization
- Ordinality is an integral feature
- Unnecessary to assign scores to the categories
- Some authors consider: $\lambda_j(x) = \alpha_j - x'\beta$ (negative sign ensures that large values of $x'\beta$ lead to an increase in the probability of higher numbered categories)

Parameter Interpretation

- For individuals with covariate vectors x^* and x, the odds ratio for response below category j is

$$\Psi_j(x^*, x) = \frac{\dfrac{\Pr(Y < j \mid x^*)}{\Pr(Y \geq j \mid x^*)}}{\dfrac{\Pr(Y < j \mid x)}{\Pr(Y \geq j \mid x)}}$$

$$= \frac{\exp\{\lambda_j(x^*)\}}{\exp\{\lambda_j(x)\}}$$

$$= \exp\{\lambda_j(x^*) - \lambda_j(x)\}$$

$$= \exp\{(\alpha_j + x^{*\prime}\beta) - (\alpha_j + x'\beta)\}$$

$$= \exp\{x^{*\prime}\beta - x'\beta\}$$

$$= \exp\{(x^* - x)'\beta\}$$

- Note that $\Psi_j(x^*, x)$ does not depend on j

Motivation

- Suppose that the underlying continuous (and perhaps unobservable) response variable is Z

- The ordinal response Y is produced via cut-off points $\alpha_1, \ldots, \alpha_c$

- The categories of Y are envisaged as contiguous intervals on the continuous scale

- Points of division α_j are assumed unknown

- Therefore,

$$Y = \begin{cases} 0 & \text{if } Z \leq \alpha_1 \\ 1 & \text{if } \alpha_1 < Z \leq \alpha_2 \\ \vdots & \vdots \\ c-1 & \text{if } \alpha_{c-1} < Z \leq \alpha_c \\ c & \text{if } Z > \alpha_c \end{cases}$$

Motivation

- Suppose that Z has the logistic distribution under some set of standard baseline conditions

$$\Pr(Y \leq j) = \Pr(Z \leq \alpha_{j+1})$$

$$= \frac{e^{\alpha_{j+1}}}{1 + e^{\alpha_{j+1}}}, \quad \text{for } j = 0, \ldots, c-1$$

- Suppose that the effect of explanatory variables is represented by a simple location shift of the distribution of Z

  i.e., $Z + x'\beta$ has the standard logistic distribution

- The common effect β for different j in the proportional odds model can be motivated by assuming that a regression model holds when the response is measured more finely

  Anderson and Philips (1981)

Motivation

- Under these assumptions,

$$\begin{aligned}\Pr(Y \leq j-1) &= \Pr(Z \leq \alpha_j) \\ &= \Pr(Z + x'\beta \leq \alpha_j + x'\beta) \\ &= \frac{\exp(\alpha_j + x'\beta)}{1+\exp(\alpha_j + x'\beta)},\end{aligned}$$

  for $j = 1, \ldots, c$

- Therefore,

$$\begin{aligned}\lambda_j(x) &= \log\left(\frac{\Pr(Y \leq j-1)}{1-\Pr(Y \leq j-1)}\right) \\ &= \log\left(\frac{\frac{\exp(\alpha_j + x'\beta)}{1+\exp(\alpha_j + x'\beta)}}{1-\frac{\exp(\alpha_j + x'\beta)}{1+\exp(\alpha_j + x'\beta)}}\right) \\ &= \alpha_j + x'\beta,\end{aligned}$$

  for $j = 1, \ldots, c$

Comments on the Proportional Odds Model

- Since the c response curves are constrained to have the same shape, the model cannot be fit using separate logit models for each cutpoint

- Not equivalent to a log-linear model

  (unlike other logit models)

- Walker & Duncan (1967) and McCullagh (1980) give Fisher scoring algorithms for iterative calculation of MLEs of parameters

  Similar to Newton-Raphson, except expected (rather than observed) values are used in the second derivative matrix

- It is not difficult to find examples of non-proportional odds (Peterson & Harrell, 1990)

  Therefore, the model may not be applicable

Stram, Wei, and Ware (1988)

- Applicable when ordered categorical responses are obtained at a common set of time points

- At each time point, the marginal distribution of the response variable is modeled using the proportional odds regression model

- The model parameters are assumed to be specific to each occasion and are estimated by maximizing the occasion-specific likelihoods

- The joint asymptotic distribution of the estimates of the occasion-specific regression coefficients is obtained without imposing any parametric model of dependence on the repeated observations

Stram, Wei, and Ware (1988)

- The vector of estimated regression coefficients is asymptotically multivariate normal

- Covariance matrix can be estimated consistently

- Provides procedures to test hypotheses about:
  - covariates at a single time point (occasion-specific)
  - a single covariate across time points (parameter-specific)

  and to estimate pooled effects of covariates across time points

- The approach allows for both time-dependent covariates and missing data
  - missing values are assumed to be MCAR (missing completely at random)

Example

- A comparison of the effects of varying dosages of an anesthetic on post-surgical recovery

- Sixty young children undergoing outpatient surgery were randomized to one of four dosages (15, 20, 25 and 30 mg/kg)
  - 15 children per group

- Recovery scores assigned upon admission to recovery room and at minutes 5, 15, and 30

- The response at each of the four time points was an ordinal categorical variable ranging from 0 (least favorable) to 6 (most favorable)

- Two covariates in addition to dosage:
  - patient age (months)
  - duration of surgery (minutes)

Model 1

- Covariate vector for subject i at time j is:

$$x_{ij1} = \begin{cases} 1 & \text{20 mg/kg dose} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij2} = \begin{cases} 1 & \text{25 mg/kg dose} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij3} = \begin{cases} 1 & \text{30 mg/kg dose} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij4} = \text{age (months)}$$

$$x_{ij5} = \text{duration of surgery (minutes)}$$

- Note that all covariates are time-independent

- Since Stram et al. use the parameterization

$$\lambda_k(x) = \alpha_k - x'\beta$$

at each time point, parameter estimates with positive signs are associated with increased probability of higher (more favorable) responses

Results from Model 1

Covariate	Time Point	Regression Coefficient		
		Estimate	Standard Error	Est./S.E.
20 mg/kg	1	−0.105	0.799	−0.13
vs.	2	−0.249	0.758	−0.33
15 mg/kg	3	−0.558	0.724	−0.77
	4	0.194	0.897	0.22
25 mg/kg	1	−0.634	0.770	−0.82
vs.	2	−0.441	0.771	−0.57
15 mg/kg	3	−0.072	0.688	−0.10
	4	−0.371	0.837	−0.44
30 mg/kg	1	−1.010	0.751	−1.34
vs.	2	−0.675	0.735	−0.92
15 mg/kg	3	−0.701	0.708	−0.99
	4	−0.465	0.884	−0.53
Age	1	−0.011	0.018	−0.61
(months)	2	−0.011	0.018	−0.61
	3	−0.028	0.020	−1.45
	4	−0.014	0.020	−0.70
Duration	1	−0.012	0.008	−1.40
of	2	−0.003	0.007	−0.41
Surgery	3	−0.008	0.007	−1.14
(minutes)	4	−0.018	0.009	−1.92

Results from Model 1

- Nearly all of the estimated regression coefficients are negative

  (indicating that the probability of a more favorable outcome decreases as the dosage, age of the patient, or duration of the surgical procedure increases)

- There is no consistent evidence (across time) of statistically significant effects due to dosage, age, or duration of surgery
  - The test statistics "Estimate/S.E." are approximately standard normal
  - None are individually significant based on a two-sided 5%-level test

Time-Specific Hypothesis Tests

- The joint effect of all covariates is not significantly different from zero

  i.e., $H_0: \beta_{j1} = \cdots = \beta_{j5} = 0$

  (p-values at times 1–4 are 0.44, 0.91, 0.46, and 0.31, respectively)

- The overall dosage effect is not significantly different from zero

  i.e., $H_0: \beta_{j1} = \beta_{j2} = \beta_{j3} = 0$

  (p-values at times 1–4 are 0.55, 0.82, 0.68, and 0.86, respectively)

- The nonlinear components of the dosage effect are not significantly different from zero

  i.e., $H_0: \beta_{j1} = \beta_{j2} - \beta_{j1}, \quad \beta_{j1} = \beta_{j3} - \beta_{j2}$

  (p-values at times 1–4 are 0.95, 0.99, 0.63, and 0.88, respectively)

Model 2

- Dosage is used as a quantitative variable:

$$x_{ij1} = \text{dosage (mg/kg)}$$

$$x_{ij2} = \text{age (months)}$$

$$x_{ij3} = \text{duration of surgery (minutes)}$$

- The parameter estimates are:

		Regression Coefficient		
Covariate	Time	Estimate	S.E.	Est./S.E.
Dosage	1	−0.070	0.049	−1.43
	2	−0.044	0.047	−0.95
	3	−0.033	0.046	−0.72
	4	−0.037	0.056	−0.66
Age	1	−0.013	0.016	−0.81
	2	−0.011	0.017	−0.62
	3	−0.025	0.019	−1.32
	4	−0.017	0.019	−0.93
Duration of Surgery	1	−0.012	0.007	−1.57
	2	−0.003	0.007	−0.45
	3	−0.008	0.007	−1.12
	4	−0.017	0.009	−1.94

Parameter-Specific Tests and Estimators

- No dosage effect

  $H_0: \beta_{11} = \beta_{21} = \beta_{31} = \beta_{41} = 0$ $\quad p = 0.72$

- Equality of dosage effects

  $H_0: \beta_{11} = \beta_{21} = \beta_{31} = \beta_{41}$ $\quad p = 0.83$

- No age effect

  $H_0: \beta_{12} = \beta_{22} = \beta_{32} = \beta_{42} = 0$ $\quad p = 0.58$

- Equality of age effects

  $H_0: \beta_{12} = \beta_{22} = \beta_{32} = \beta_{42}$ $\quad p = 0.61$

- No surgery duration effect

  $H_0: \beta_{13} = \beta_{23} = \beta_{33} = \beta_{43} = 0$ $\quad p = 0.09$

- Equality of surgery duration effects

  $H_0: \beta_{13} = \beta_{23} = \beta_{33} = \beta_{43}$ $\quad p = 0.12$

Pooled Estimators of Effects

Variable	Estimate	S.E.	Est./S.E.
Dosage	−0.0460	0.0424	−1.09
Age	−0.0143	0.0162	−0.88
Surgery Duration	−0.0091	0.0065	−1.40

- The odds of having a recovery score higher than a given cutpoint are:
  - $e^{-0.0460} = 0.955$ times as high per 1 mg/kg increase in dosage
  - $e^{-0.0143} = 0.986$ times as high per 1 month increase in age
  - $e^{-0.0091} = 0.991$ times as high per 1 minute increase in surgery duration

- Although there is modest evidence of an effect due to surgery duration, there is essentially no evidence that dosage or age influence recovery

GEE Approach

- The Stram-Wei-Ware methodology:
  1. models the data separately at each time point
  2. combines the resulting estimates

  This approach requires a common set of time points for each experimental unit

- The SAS GENMOD procedure now allows for the analysis of repeated ordered categorical outcome variables using the GEE approach

- Using the GEE approach:
  - The number of repeated measurements per experimental unit need not be constant
  - Measurement times need not be the same across experimental units

- The proportional odds model is used for the marginal distribution

- The "working" correlation matrix is the independence model

SAS Statements

- The statements on the following page:
  - read in the original data set (one observation for each subject)
  - restructure the data set to have one observation per time point
  - fit a model with effects for:
    - dosage (mg/kg)
    - age (months)
    - duration of surgery (minutes)
- Note that dosage is used both as a:
  - class variable (in order to distinguish duplicate subject identifiers across dosages)
  - numeric variable (in the model statement)
- Also note that this model is analogous to the Stram-Wei-Ware model 2

SAS Statements

```
data a;
input dosage id age durat min0 min5
      min15 min30;
dose=dosage;
cards;
15  1 36 128 3 5 6 6
15  2 35  70 3 4 6 6
   ...
30 14 27  61 3 5 5 6
30 15 56 106 0 1 1 3
;

data b; set a;
time=0;   score=min0;   output;
time=5;   score=min5;   output;
time=15;  score=min15;  output;
time=30;  score=min30;  output;

proc genmod;
class dosage id;
model score=dose age durat
      / dist=multinomial;
repeated subject=id(dosage)
      / type=ind;
```

Comments

- The parameter estimates are similar to the pooled estimators from the Stram-Wei-Ware model (but are of opposite sign)

- The GEE approach would also allow one to include time as a factor in the model:

```
proc genmod;
class dosage id;
model score=dose age durat time
      / dist=multinomial;
repeated subject=id(dosage)
      / type=ind;
```

- The odds of having a recovery score higher than a given cutpoint are $e^{0.0946} = 1.1$ times as high per minute in the recovery room
  - This effect is highly significant ($p < 0.0001$)

- Interactions between covariates and time could also be investigated using this approach