Generalized Estimating Equations (GEE)
An Introduction
BIOS 667

Bahjat F. Qaqish
Department of Biostatistics
CB 7420, McGavran Greenberg Hall
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420
email: qaqish@bios.unc.edu
www.bios.unc.edu/~qaqish

Outline

- The 6-City study (correlated binary responses)
- GEE using proc GENMOD in SAS
- Some details on GEE
- GEE - strengths and weaknesses
- Study design considerations and sample size calculations
- GEE2 (or GEE-II) for variance and correlation parameters
- Other issues.
- Summary

The 6-City Study (correlated binary responses)

- Data on 537 children from Stubenville, Ohio, examined from age 7 to age 10.
  Response: Respiratory infection in the year prior to the exam reported by the mother.
  Covariates: Mother smoking, age.
- One child = one cluster
  Number of clusters = $K = 537$.
  Cluster size = number of observations in a cluster, n_i = size of the i-th cluster.
  In this example $n_i = 4$ for all clusters.

- Does MS increase risk of respiratory infection? How much?
- Does risk of respiratory infection vary with age? How much?
- Does the effect of MS on risk of respiratory infection vary with age? How much?

- % with respiratoty infection:

	N	age 7	age 8	age 9	age 10
MS=0 (nonsmoker)	350	16	15	14	11
MS=1 (smoker)	187	17	21	19	14

- Pairwise odds-ratios (below diagonal) and correlations (above diagonal):

  MS=0 (nonsmoker):

	age 7	age 8	age 9	age 10
age 7		0.34	0.29	0.31
age 8	7.1		0.43	0.33
age 9	5.5	11.4		0.39
age 10	6.9	7.8	11.1	

  MS=1 (smoker):

	age 7	age 8	age 9	age 10
age 7		0.37	0.34	0.36
age 8	7.4		0.46	0.33
age 9	6.4	11.2		0.36
age 10	7.9	6.3	7.8	

Generalized Estimating Equations

– Model: Suppose $Y_1, \cdots, Y_K$ are independent vectors with means $\mu_1, \cdots, \mu_K$, functions of β.

$$E[Y_{ij}] = \mu_{ij},$$

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^T \beta,$$
$$\mathrm{Var}(Y_{ij}) = \phi h(\mu ij),$$

g is the link function, h the variance function, η_{ij} the linear predictor, ϕ the scale parameter.

– Consider elementary estimating functions $Y_i - \mu_i$. Then the matrix $\mathrm{Var}(Y)$ is block-diagonal and the optimal linear combination is

$$\sum_{i=1}^{K} D_i^T (\mathrm{Var}(Y_i))^{-1} (Y_i - \mu_i),$$

where $D_i := \partial \mu_i / \partial \beta$.

– The diagonals of $\mathrm{Var}(Y_i)$ are determined by

$$\mathrm{Var}(Y_{ij}) = \phi h(\mu_{ij}).$$

The off-diagonals involve correlations that so far have not been defined. First write

$$\mathrm{Var}(Y_i) = \Sigma_i = \phi A_i C_i A_i,$$

where $A_i = \mathrm{diag}(\sqrt{h(\mu_{ij})})$ and $C_i = \mathrm{Corr}(Y_i)$.

Assumptions are made about the correlation matrix C_i. Specifically, it is parametrized by an $s \times 1$ parameter vector ρ. An estimate of ρ is plugged-in and estimation proceeds. The assumed structure is called the **working correlation matrix**, denoted R_i, may not be identcal to the true correlation, thus the different notation.

Define

$$V_i = A_i R_i A_i.$$

– Then the GEE is

$$\Sigma_{i=1}^{K} D_i^T V_i^{-1}(Y_i - \mu_i) = 0.$$

The GEE is solved for β. The solution is the estimator $\hat{\beta}$.

– Intuitively, the GEE is

$$\Sigma \text{ derivative (variance })^{-1}(O - E) = 0,$$

essentailly, a weighted sum of **observed - expected**. The weights are chosen according to certain “optimality” criteria.

– Requirements:

A $\sqrt{K}-$consistent estimator of ϕ at the true β.

A $\sqrt{K}-$consistent estimator of ρ at the true β and ϕ.

Regularity conditions.

– The estimator $\hat{\beta}$ is consistent and asymptotically Gaussian:

As $K \to \infty$

$$\sqrt{K}(\hat{\beta}_K - \beta) \xrightarrow{d} N(0, S),$$

$$S := \lim_{K \to \infty} K H_1^{-1} H_2 H_1^{-1},$$

where

$$H_1 := \sum_{i=1}^{K} D_i^T V_i^{-1} D_i,$$

$$H_2 := \sum_{i=1}^{K} D_i^T V_i^{-1} \Sigma_i V_i^{-1} D_i.$$

– Variance estimation (S):

The matrices H_1 and H_2 are evaluated at the estimates and Σ_i in H_2 is replaced by

$$(Y_i - \mu_i)(Y_i - \mu_i)^T.$$

The estimator thus obtained is known as the **sandwich**, **robust**, or **empirical** variance estimator. This is becuase it is generally a consistent estimator of the true variance of $\hat{\beta}$ even when the correlation is mispecified ($R_i \neq C_i$).

– If the assumed correlation structure is correct, i.e. ($R_i = C_i$), then $H_1 = H_2$ and the asymptotic variance simplifies

$$S = \lim_{K \to \infty} K H_1^{-1}.$$

The estimator thus obtained is known as the **naive** or **model-based** variance estimator becuase it is generally an inconsistent estimator of the true variance of $\hat{\beta}$ unless $R_i = C_i$.

– Some choices for R_i:

  * Exchangeable: For $j \neq k$

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho.$$

  * Autoregressive: e.g. AR(1):

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|j-k|}.$$

  * M-dependent: For $|j - k| \leq m$

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{|j-k|}.$$

  * Unstructured:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk}.$$

  * Independence: For $j \neq k$

$$\text{Corr}(Y_{ij}, Y_{ik}) = 0.$$

– As long as interest is mainly in the mean structure, there is no need to worry too much about getting the correaltion structure exactly right.

Estimating Equations - Strengths

- Can generate consistent estimates even in cases where the MLE fails.
- Even when MLE is theoretically available, EE can be much easier computationally.
- Can be used to estimate parameters of interest (e.g. β) with minimal concern about nuisance parameters (e.g. correlations).
- Can generate reasonable estimates for a wide class of distributions, fewer assumptions.

Estimating Equations - Weaknesses

– Can generate inefficient estimates.

– If the EE has several roots, there is no clear way to choose one as the estimate.
  Suggested approaches:
  1) Li and McCullagh (1994, Ann. Stat.).
  2) McLeish and Small (1992, Bka.).
  3) Li (1993, Bka.), Hanfelt and Liang (1995, Bka.).
  4) Heyde and Morton (1998, Bka.).

– No "likelihood-ratio"-type tests. Some of the above references try to construct a "likelihood-like" function.

Binary Outcomes - Correlation and Odds Ratio

- Suppose $(Y_1, Y_2) \sim$ bivariate Bernoulli, $E[Y_1] = \mu_1, E[Y_2] = \mu_2, E[Y_1 Y_2] = \mu_{12}$. Note that μ_{12} must obey the Frechet bounds

$$\max(0, \mu_1 + \mu_2 - 1) \leq \mu_{12} \leq \min(\mu_1, \mu_2)$$

  Knowing or fixing μ_1 and μ_2 puts some limits on μ_{12}, and hence on the Pearson correlation between Y_1 and Y_2.

- Pearson correlation:

$$\rho_{12} = \frac{\mu_{12} - \mu_1 \mu_2}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}}$$

- The range of Pearson correlation for a given (μ_1, μ_2): Define $\psi_j = \sqrt{\mu_j/(1-\mu_j)}, j = 1, 2$. The range is

$$-\min(\psi_1\psi_2, \frac{1}{\psi_1\psi_2}) \leq \rho_{12} \leq \min(\frac{\psi_1}{\psi_2}, \frac{\psi_2}{\psi_1}).$$

  Example: $\mu_1 = 0.1, \mu_2 = 0.4, -0.272 \leq \rho_{12} \leq 0.408$ (bounds were rounded toward 0).

- Odds ratio ("$ad/(bc)$"):

$$e^{\lambda_{12}} = \frac{\mu_{12}(1 - \mu_1 - \mu_2 + \mu_{12})}{(\mu_1 - \mu_{12})(\mu_2 - \mu_{12})}$$

  The log odds ratio is λ_{12}.

- The range of the log odds ratio for a given (μ_1, μ_2) is $(-\infty, \infty)$, i.e. no restrictions.

- For three or more variables there are restrictions on the set of pairwise odds ratios. e.g. For three variables, the valid parameter space for $(\lambda_{12}, \lambda_{13}, \lambda_{23})$ is a proper subset of R^3, not the whole R^3.

- Model Fitting in SAS:
  The logistic regression model:

  $$\text{logit P}(Y = 1; \text{AGE}, \text{MS}) = \beta_1 + \beta_2 \text{ AGE} + \beta_3 \text{MS} + \beta_4 \text{ AGE} * \text{MS}.$$

  can be fitted to independent responses using the statements:

```
proc genmod data=B;
  model y / one =  ms age msxage
        / d=b;
```

  A binary, or binomial distribution is specified by **d=b**. Default link for binomial is logit ($\log(\mu/(1-\mu))$).

- For dependent responses, estimation by GEE is triggered by the **repeated** statement (file: 6city.sas):

```
proc genmod data=B;
  class id;
  model y / one =  ms age msxage
        / d=b ;
  repeated subject=id / type=exch;
```

- For dependent responses, the **repeated** statement identifes the cluster id, which must be declared in a **class** statement. The working correlation structure is selected by the **type=** option.
- For binary responses, 0 or 1, the mean determines the variance completely and the scale (dispersion parameter) $\phi = 1$.

• Abridged SAS output with [comments]:

```
->                   The GENMOD Procedure
                      Model Information
Description                    Value     Label

Data Set                       WORK.B
Distribution                   BINOMIAL
Link Function                  LOGIT
Dependent Variable             Y         Respiratory illness
Dependent Variable             ONE
Observations Used              2148
Number Of Events               326
Number Of Trials               2148


->    Class Level Information
  Class     Levels  Values
  ID           537  1 2 3 4 5 6 7 8 9 10 11 12 13
                        ...
                    530 531 532 533 534 535 536 537


-> Parameter Information
Parameter     Effect
PRM1          INTERCEPT
PRM2          MS
PRM3          AGE


->       Criteria For Assessing Goodness Of Fit
```

```
...     [ not usable for dependent responses ]

->         Analysis Of Initial Parameter Estimates
              [ assuming independence ]

Parameter     DF     Estimate      Std Err    ChiSquare  Pr>
INTERCEPT      1      -1.8837       0.0838     504.7826  0.0
MS             1       0.2721       0.1235       4.8578  0.0
AGE            1      -0.1134       0.0541       4.3976  0.0
SCALE          0       1.0000       0.0000            .


->            GEE Model Information

Description                      Value
Correlation Structure            Exchangeable
Subject Effect                   ID (537 levels)
Number of Clusters               537
Correlation Matrix Dimension     4
Maximum Cluster Size             4
Minimum Cluster Size             4


->          Working Correlation Matrix
           [ cluster size = 4 ]
                 COL1      COL2      COL3      COL4
ROW1           1.0000    0.3543    0.3543    0.3543
ROW2           0.3543    1.0000    0.3543    0.3543
ROW3           0.3543    0.3543    1.0000    0.3543
ROW4           0.3543    0.3543    0.3543    1.0000
```

-> Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Empirical Std Err	95% Confidence Limits Lower	Uppe
INTERCEPT	-1.8804	0.1139	-2.1037	-1.657
MS	0.2651	0.1777	-0.0833	0.613
AGE	-0.1134	0.0439	-0.1993	-0.027
Scale	1.0000	.	.	

- Estimates and empirical SE's based on exchangeble working correlation:

Parameter	Estimate	Std Err	Lower	Upper	Z	p
INTERCEPT	-1.88	0.114	-2.103	-1.66	-16.5	0.00
MS	0.265	0.178	-0.0833	0.614	1.5	0.13
AGE	-0.113	0.044	-0.199	-0.0274	-2.6	0.01

- $\hat{\beta}$ using different correlation structures:

Parameter	INDEP Estimate	EXCH Estimate	AR(1) Estimate
INTERCEPT	-1.90	-1.90	-1.92
MS	0.314	0.314	0.295
AGE	-0.141	-0.141	-0.147
MSXAGE	0.071	0.071	0.082

- Estimated SE of $\hat{\beta}$:

Parameter	INDEP Std Err	INDEP Empirical Std Err	EXCH Empirical Std Err	AR(1) Empirical Std Err
INTERCEPT	0.089	0.119	0.119	0.120
MS	0.139	0.188	0.188	0.190
AGE	0.070	0.058	0.058	0.059
MSXAGE	0.111	0.088	0.088	0.091

The Model

- There are $K = 537$ independent children. The response of the ith child is the vector $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})^\top$.
- Observations are ordered by age, 7–10.
- The mean vector for the ith child is $\mu_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})^\top$.
- Logit link: $\eta_{ij} = \text{logit}(\mu_{ij})$ or, in vector form $\eta_i = \text{logit}(\mu_i)$, with the logit applied elementwise.
- The covariance matrix for the ith child is Σ_i. The jth diagonal element is $\mu_{ij}(1 - \mu_{ij})$. The correlation matrix is C_i.
- Different *working* correlation matrices were used.

The Linear Predictor

- The model assumes that the marginal mean, on the logit scale, follows a linear trend with age. Each MS group has its own line; there are two different slopes and two different intercepts.

- An unstructured correlaion matrix can be described by:
  The correlation matrix is assumed to be the same in both groups, but otherwise has no special structure.

- Other models are certainly possible, including models that are not linear in age. Even within linear (in age) models, other options include: 1) A model that assumes that the two groups have the same intercept but different slopes. This implies that the groups start with the same mean at time 0, but later diverge. 2) A model that assumes that the two groups have the same slope but different intercepts. This implies that the difference between the two group means (on the logit scale) is constant over time (parallel lines). 2) A model that assumes that the two groups have the same line (same intercept, same slope).

- The 4×1 covariate vector for the ith child at jth occasion is x_{ij},
  $$x_{ij} = (1, t_{ij}, MS_{ij}, t_{ij}MS_{ij}),$$
  where t_{ij} is age-9 and MS_{ij} is the mother's smoking status (0=non-smoker, 1=smoker), $i = 1, \cdots, K, j = 1, 2, 3, 4$.

- That is,
  $x_{ij1} = 1,$
  $x_{ij2} = t_{ij},$
  $x_{ij3} = MS_{ij},$
  $x_{ij4} = t_{ij}MS_{ij}.$

- Since $(t_{i1}, t_{i2}, t_{i3}, t_{i4}) = (-2, -1, 0, 1)$, it holds that $t_{ij} = j - 3$ for all (i, j).
- Since the mother's smoking status does not change over time MS_{ij} can be written as MS_i with no confusion.
- Rewriting,
  $x_{ij1} = 1$,
  $x_{ij2} = j - 3$,
  $x_{ij3} = MS_i$,
  $x_{ij4} = (j - 3)MS_i$.
- $$\eta_{ij} = \beta_1 + \beta_2(j - 3) + \beta_3 MS_i + \beta_4(j - 3)MS_i.$$
- Cell linear predictors

	MS = 0	MS = 1	Contrast
Age 7	$\beta_1 - 2\beta_2$	$\beta_1 - 2\beta_2 + \beta_3 - 2\beta_4$	$\beta_3 - 2\beta_4$
Age 8	$\beta_1 - \beta_2$	$\beta_1 - \beta_2 + \beta_3 - \beta_4$	$\beta_3 - \beta_4$
Age 9	β_1	$\beta_1 + \beta_3$	β_3
Age 10	$\beta_1 + \beta_2$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$	$\beta_3 + \beta_4$

Interpretation

- β_3 is the contrast of smokers versus non-smokers at age 9. So β_3 is the log of the ratio of odds of respiratory infection in children of smokers to the odds of respiratory infection in children of non-smokers at age 9.
- In short, β_3 is the log odds ratio relating RI at age 9 to mother smoking.
- $\hat{\beta}_3 = 0.314, e^{0.314} = 1.37, \hat{se}(\hat{\beta}_3) = 0.188$. A 95% confidence interval (CI) for β_3 is $0.314 \pm 1.96(0.188) = (-0.0545, 0.682)$ and a 95% confidence interval (CI) for e^{β_3} is $(e^{-0.0545}, e^{0.682}) = (0.947, 1.98)$
- At age 9, the estimated ratio of the odds of respiratory infection in children of smokers to the odds of respiratory infection in children of non-smokers is 1.37 (95% CI: 0.947–1.98).
- At age 9, the estimated ratio of the odds of respiratory infection in children of smokers to the odds of respiratory infection in children of non-smokers is 1.37, 95% CI: 0.947–1.98.
- Confidence intervals can also be given using ():
  ... is 1.37, 95% CI: (0.947, 1.98).
- The estimated contrasts at ages 7–10 are:
  (0.172, 0.243,0.314,0.385).
- So the estimated odds ratios at ages 7–10 are:
  (1.19, 1.28, 1.37, 1.47).
- The contrasts will be constant over ages 7–10 if $\beta_4 = 0$. So $H_0 : \beta_4 = 0$ is the hypothesis that mother smoking has a constant effect, as measured by the odds ratio, over ages 7–10. The

Wald-type test is to refer $(\hat{\beta}_4 - 0)/\hat{se}(\hat{\beta}_4) = 0.807$ to the standard normal distribution, or equivalently, $(\hat{\beta}_4 - 0)^2/a\hat{v}ar(\hat{\beta}_4)$ to the χ^2_1 distribution.