Generalized Linear Models for Longitudinal Data

Consider extensions of GLM to repeated measurements, or generalization of linear models for longitudinal data to non-Gaussian response variables (e.g., count-valued or binary responses).

- Marginal models: the basic premise is to make inferences about population averages.
- Mixed effects models: a subset of the regression coefficients vary from subject to subject.

Marginal Models

- The focus of marginal models is on inferences of population averages.
- Marginal models separately model the mean responses and within-subject association among the repeated responses.
- Marginal models do not require the joint distributional assumptions for the vector of responses, which may be difficult for the discrete data. The avoidance of distributional assumptions leads to a method of estimation of *generalized estimating equations* (GEE).

Features of Marginal Models

- Marginal expectation of Y_{ij} (i.e., μ_{ij}) depends on covariates through a known link function

$$g(\mu_{ij}) = X_{ij}\beta$$

- Marginal variance of Y_{ij} is a function of the marginal mean and a scale parameter

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

- The "within-subject association" among the responses is a function of the means and of additional parameters, say α, that may need to be estimated.

α

$\text{var}(Y_{ij}) \cdot \alpha \rightarrow (\quad)_{\text{cov}}$

Example: Continuous Response

- $\mu_{ij} = X_{ij}\beta$ (i.e., linear regression)
- $\text{var}(Y_{ij}) = \sigma_j^2$ (i.e., heterogenous variance for different visits, but no dependence on mean)
- $\text{corr}(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$ $(0 \leq \alpha \leq 1)$ (i.e., autoregressive correlation. Other correlation structures such as compound symmetry or unstructured are also possible.)

Example: Count Response

- $\log \mu_{ij} = X_{ij}\beta$ (i.e., log linear regression)
- $\text{var}(Y_{ij}) = \phi\mu_{ij}$ (i.e., Poisson variance with potential over dispersion)
- $\text{corr}(Y_{ij}, Y_{ik}) = \alpha$ (i.e., compound symmetry correlation structure)

Example: Binary Response

- $logit(\mu_{ij}) = \log \frac{\mathbb{P}(Y_{ij}=1)}{\mathbb{P}(Y_{ij}=0)} = X_{ij}\beta$ (i.e., logistic regression)
- $\mathrm{var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$ (i.e., binary variance)
- $\mathrm{corr}(Y_{ij}, Y_{ik}) = \alpha_{jk}$ (i.e., unstructured)
- Sometimes, people also use other measures (e.g., log odds ratio) to characterize associations for binary variables

Interpretation of Model Parameters

The regression parameters, β, have "population-averaged" interpretations (where "averaging" is over all individuals within subgroups of the population):

- describe effect of covariates on the average responses
- contrast the means in sub-populations that share common covariate values

age sex

$X_{i1}\beta_1 + X_{i2}\beta_2$

For example, consider the logistic model

$$logit(\mu_{ij}) = logit(\mathbb{E}(Y_{ij}|X_{ij})) = X_{ij}\beta$$

Each element of β measures the change in the log odds of a "positive" response per unit change in the respective covariate, *for sub-populations defined by fixed and known covariate values*.

Parameter Estimation: GEE

GLM: $f(y) = y\cdot\theta - b(\theta)$

$b'(\theta) = \mu$

$g(\mu) = X\beta$

$\max_\beta f(y)$

Solve: $\frac{\partial f(y)}{\partial \beta} = 0$

$$\frac{\partial f(y)}{\partial \beta} = \frac{\partial[y\theta - b(\theta)]}{\partial \beta} = (y - b'(\theta))\cdot\frac{\partial \theta}{\partial \beta} = (y - b'(\theta))\frac{\partial \theta}{\partial \mu}\cdot\frac{\partial \mu}{\partial \beta} = (y - b'(\theta))\frac{1}{b''(\theta)}\cdot\frac{\partial \mu}{\partial \beta}$$

$$\boxed{(y-\mu)\frac{1}{V}\cdot\frac{\partial \mu}{\partial \beta} = 0}$$

- It is difficult to derive a multivariate distribution for discrete response data.
- Thus, no "convenient" likelihood function to maximize.
- Instead, use Generalized Estimating Equations (GEE).
- No need to specify any distribution; just provide the mean function (link) and the association structure.

ind

$$\sum_{i=1}^{m} D^T V_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0$$

where $V_i \approx \text{cov}(\boldsymbol{Y}_i)$ and $D_i = \partial\boldsymbol{\mu}_i/\partial\beta$

Example

Respiratory Illness

- Study how respiratory illness ($\overset{1}{\text{good}}$ or $\overset{0}{\text{poor}}$) is related to several factors (e.g., baseline status, age, sex, treatment, etc)
- 111 subjects with 4 measurements per subject (1 baseline and 3 follow-ups)
- Each response variable is binary
- Fit marginal models with logit link and different correlation structures

GEE with logit link and exchangeable correlation structure

```
> resp_gee2 <- gee(nstat ~ centre + treatment + gender + baseline + age, data = resp, family = "binomial",
+                  id = subject, corstr = "exchangeable", scale.fix = TRUE, scale.value = 1)
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
      (Intercept)           centre2 treatmenttreatment         gendermale       baselinegood                age
      -0.90017133        0.67160098         1.29921589         0.11924365         1.88202860        -0.01816588
> summary(resp_gee2)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Call:
gee(formula = nstat ~ centre + treatment + gender + baseline +
    age, id = subject, data = resp, family = "binomial", corstr = "exchangeable",
    scale.fix = TRUE, scale.value = 1)

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-0.93134415 -0.30623174  0.08973552  0.33018952  0.84307712


Coefficients:
                      Estimate Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept)        -0.90017133  0.4784634 -1.8813796  0.46032700 -1.9555041
centre2             0.67160098  0.3394723  1.9783676  0.35681913  1.8821889
treatmenttreatment  1.29921589  0.3356101  3.8712064  0.35077797  3.7038127
gendermale          0.11924365  0.4175568  0.2855747  0.44320235  0.2690501
baselinegood        1.88202860  0.3419147  5.5043802  0.35005152  5.3764332
age                -0.01816588  0.0125611 -1.4462014  0.01300426 -1.3969169

Estimated Scale Parameter:  1
Number of Iterations:  1

Working Correlation
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.3359883 0.3359883 0.3359883
[2,] 0.3359883 1.0000000 0.3359883 0.3359883
[3,] 0.3359883 0.3359883 1.0000000 0.3359883
[4,] 0.3359883 0.3359883 0.3359883 1.0000000
```

Example

Epileptic Seizure

- Clinical trial of 59 epileptics
- For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks
- Patients were randomized to treatment with the antiepileptic drug progabide or placebo
- Number of seizures was then recorded in 4 consecutive 2-week intervals
- Question: Does seizure rate change over time? Is it related to baseline, age, or treatment group assignment?
- Question: Is the treatment effective?

time × trt