Presentation on “Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation” By J. Hanley et al.

Ben Rich

BIOS612 : Advanced Generalized Linear Models

February 9, 2009

Objectives

- Explain the underlying principles of GEE in a way that a non-statistician can understand
- Small worked example to illustrate the calculations that go on "behind the scenes"
- Do the calculations by hand
- Focus on clustered data rather than longitudinal data $\longrightarrow$ exchangeable covariance structure

Example

- Data on standardized heights (z-scores) of 144 children in 54 households in Mexico, randomly selected
- Covariates: gender, SES
- We want to estimate the mean height μ
- Standard error of $\bar{y}$ depends on sample size $\rightarrow \sigma/\sqrt{n}$ in the independent case
- How many observations do we have? 144? 54?
- Simplest possible data set: 3 children, 2 households

A quote

“We show how GEE uses weighted combinations of observations to extract the appropriate amount of information from correlated data.”

Why does correlation imply weighting?

- 2 correlated observations contain less information than 2 independent ones
- The variance of $\bar{y}$ is increased
- Downweight correlated observations, by how much?

An example

3 observations in 2 clusters, same variance σ^2, y_2 and y_3 are correlated with correlation coefficient R

$$\bar{y}_w = \frac{1}{1+2w}y_1 + \frac{w}{1+2w}y_2 + \frac{w}{1+2w}y_3$$

$$\mathrm{Var}[\bar{y}_w] = ???$$

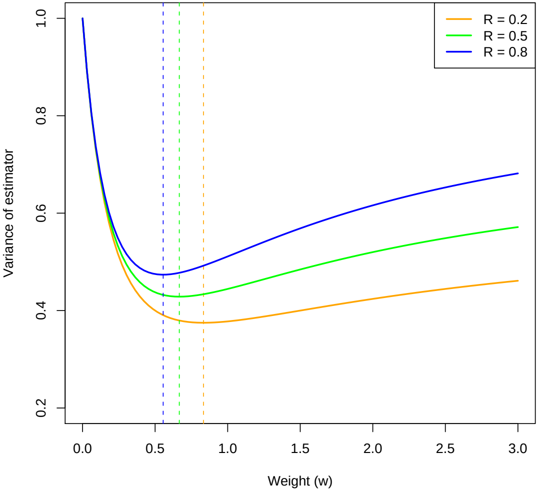
An example

3 observations in 2 clusters, same variance σ^2, y_2 and y_3 are correlated with correlation coefficient R

$$\bar{y}_w = \frac{1}{1+2w} y_1 + \frac{w}{1+2w} y_2 + \frac{w}{1+2w} y_3$$

$$\mathrm{Var}[\bar{y}_w] = ???$$

$$\sigma^2 \left[\left(\frac{1}{1+2w} \right)^2 + \left(\frac{w}{1+2w} \right)^2 (2+2R) \right]$$


R = 0.2
R = 0.5
R = 0.8
Variance of estimator
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Weight (w)

Effective sample size

- Independent case, $\text{Var}[\bar{y}] = \sigma^2/3$
- For $R \neq 0$, with $w = 1/(1+R)$, we have

$$\text{Var}[\bar{y}_w] = ???$$

Effective sample size

- Independent case, $\mathrm{Var}[\bar{y}] = \sigma^2/3$
- For $R \neq 0$, with $w = 1/(1+R)$, we have

$$\mathrm{Var}[\bar{y}_w] = ???$$
$$\sigma^2/(1+2w)$$

- In general, $\mathrm{Var}[\bar{y}_w] = \sigma^2/\sum_i w_i$, so the effective sample size is $\sum_i w_i$

Estimating the nuisance parameter R

- We use the residuals to estimate the (assumed common) variance and covariances

$$\hat{\sigma}^2 = \sum_i \sum_j (y_{ij} - \hat{\mu}_{ij})^2 / (n - p)$$

$$\hat{R} = \frac{\sum_i \sum_{j \neq k} (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik}) / (n_{\text{sum}} - p)}{\hat{\sigma}^2}$$

- Alternate between estimating μ and estimating $R \rightarrow$ convergence

GEE method

- Extension of the GLM framework
- Account for correlation
- Quasi-likelihood approach
  - Correct specification of mean, variance function and covariance structure is sufficient
- Marginal model
  - Recall non-collapsibility in logistic regression
- Model-based or empirical standard errors
- Cluster size should not be related to outcome

“GEE” = “G” + “EE”

- “Estimating Equation”
  - An idea for combining estimates that predates least-squares[1]

$$w_1(y_1 - \hat{\mu}) + w_2(y_2 - \hat{\mu}) + w_3(y_3 - \hat{\mu}) = 0$$

- “Generalized”
  - Can estimate risk difference, risk ratio, odds ratio, etc. by specifying link and variance functions
  - Another level of weights $\rightarrow$ think iteratively reweighted least squares in GLM

[1] Stigler SM. Least squares and the combination of observations. In: *The history of statistics: the measurement of uncertainty before 1900*.

Comparing GEE to mixed-models

- GEE is a marginal model that aims uniquely for more efficient estimates of β, as well as accurate standard errors in the presence of correlation
- Mixed-models explicitly model between-cluster variation
- GEE models within-cluster similarity of residuals instead
- GEE cannot handle
  - multiple levels of clustering
  - both cluster-specific intercepts and slopes (longitudinal setting)

The takehome message

- Don't ignore correlation! Estimates like $\bar{y}$ may be unbiased, but are less efficient (the usual standard error $\sigma/\sqrt{n}$ is wrong)
- Downweighting correlated observations plays an essential role in increasing efficiency
- Quasi-likelihood approach, can fit generalized exponential families like GLM
- Marginal model $\longrightarrow$ interpretability
- Model based or empirical standard errors