Generalized Estimating Equations for Multilevel Data

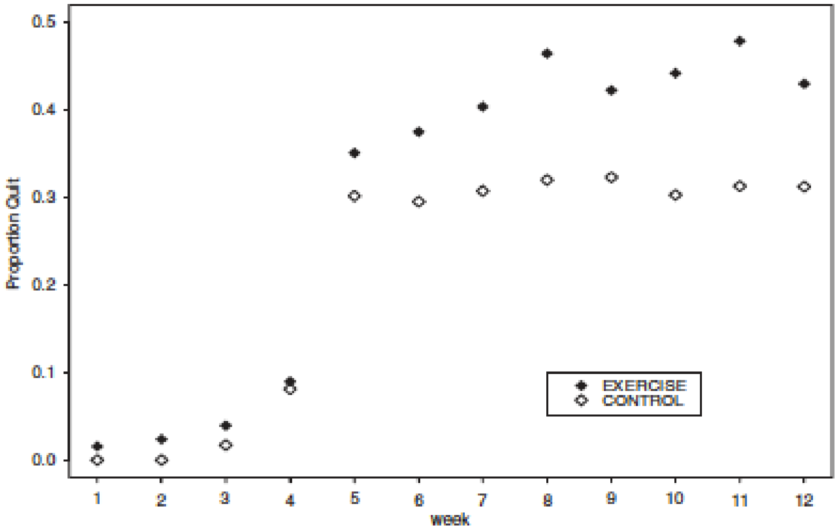
Nambari Short Course

17 July 2019

Overview

- Continue to use CTQ data
- GEE as a method to fit generalized linear models for correlated data
- Compare conditional (subject-specific) versus marginal (population-averaged) treatment effect

Smoking Cessation Study: Summaries



Motivation for GEE

- Generalized estimating equations (GEE) is a method to fit generalized linear models to correlated data
  - Clustered data
  - Multilevel data
  - Longitudinal data

- Multilevel models motivated by hierarchical sampling structure
  - Explicit modeling of variation at each level
  - Interpretation of coefficients as *conditional* effects
  - Usually relies on parametric assumptions (e.g., normality)

- GEE motivated by correlated data structure
  - Explicit modeling of the *marginal* distribution
  - Specify mean, variance, correlation structure
  - Typically does not need parametric modeling assumptions

Generalized linear model

A generalized linear model (GLM) is used to model the mean of a response variable Y as a function of covariates $\boldsymbol{X}$, namely $\mu = E(Y \mid \boldsymbol{X})$.

Requires the user to specify two things:

- Link function $g(\cdot)$ linking the mean to a linear predictor

$$g(\mu) = \boldsymbol{X}\beta$$

  - Examples: logit, log, identity.

- Variance function characterizing $\text{var}(Y \mid \boldsymbol{X})$. This usually depends on the type of outcome.
  - Binary data: $\text{var}(Y \mid \boldsymbol{X}) = \mu(1-\mu)$
  - Count data: $\text{var}(Y \mid \boldsymbol{X}) = \mu$
  - Continuous data: $\text{var}(Y \mid \boldsymbol{X}) = \sigma^2$

Connection between marginal and multilevel model specifications

Consider normal distribution with random intercept

Multilevel model

$$\begin{aligned} Y_{ij} &\sim \mathcal{N}(\alpha_i + X_{ij}\beta, \sigma^2) \\ \alpha_i &\sim \mathcal{N}(\theta, \tau^2) \end{aligned}$$

Writing this as a marginal model

$$\begin{aligned} E(Y_{ij} \mid X_{ij}) &= E(\alpha_i + X_{ij}\beta) \\ &= \theta + X_{ij}\beta \\ &= \mu_{ij} \end{aligned}$$

$$\begin{aligned} \mathrm{var}(Y_{ij} \mid X_{ij}) &= \sigma^2 + \tau^2 \\ &= v_{ij} \end{aligned}$$

What's missing from this is a specification of *correlation*

Connection between marginal and multilevel model specifications

For the normal model with random intercept, can show that

$$\begin{aligned} \mathrm{cov}(Y_{ij}, Y_{ik}) &= \tau^2 \\ \mathrm{corr}(Y_{ij}, Y_{ik}) &= \frac{\tau^2}{\tau^2 + \sigma^2} \\ &= \rho_{ijk} \end{aligned}$$

Hence the *marginal* distribution can be described using three features: mean, variance, correlation.

Representation of this model as a marginal model

Mean (link function is identity)

$$\mu_{ij} = \theta + X_{ij}\beta$$

Variance

$$\mathrm{var}(Y_{ij} \mid X_{ij}) = v$$

Correlation

$$\mathrm{corr}(Y_{ij}, Y_{ik}) = \rho$$

Representation in matrix form

$$\boldsymbol{\mu}_i = \begin{pmatrix} \theta + X_{i1}\beta \\ \theta + X_{i2}\beta \\ \vdots \\ \theta + X_{iJ}\beta \end{pmatrix}$$

$$\operatorname{var}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) = \begin{pmatrix} v & & & \\ 0 & v & & \\ \vdots & & & \\ 0 & 0 & \cdots & v_J \end{pmatrix}$$

$$\operatorname{corr}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) = \begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \vdots & & & \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

General representation

For a generalized linear model, need link function g.

$$
\begin{aligned}
\mathbf{g}(\boldsymbol{\mu}_i) &= \begin{pmatrix} \theta + X_{i1}\beta \\ \theta + X_{i2}\beta \\ \vdots \\ \theta + X_{iJ}\beta \end{pmatrix} \\
\text{var}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) &= \begin{pmatrix} v_1 & & & \\ 0 & v_2 & & \\ \vdots & & & \\ 0 & 0 & \cdots & v_J \end{pmatrix} \\
\text{corr}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) &= \begin{pmatrix} 1 & & & \\ \rho_{21} & 1 & & \\ \vdots & & & \\ \rho_{J1} & \rho_{J2} & \cdots & 1 \end{pmatrix}
\end{aligned}
$$

Notes about specification

Variance

- Variance usually determined by type of outcome (continuous, count, binary)
- For count and binary, can add scale parameter to capture extra variation

$$\mathrm{var}(Y_{ij} \mid X_{ij}) = \phi v_j$$

Correlation

- Typically specify correlation *structure*
  - Independence, exchangeable, AR-1, unstructured, etc.

Example using CTQ data

In these specifications, $\mu_{ij} = P(Y_{ij} = 1 \mid X_{ij})$

Model 1: Effect of nicotine dependence

$$\begin{aligned} \text{logit}(\mu_{ij}) &= \theta + \beta X_i \\ \text{var}(Y_{ij} \mid X_i) &= \phi \mu_{ij}(1 - \mu_{ij}) \end{aligned}$$

Model 2: Effect of treatment

$$\text{logit}(\mu_{ij}) = \alpha + \beta T_j + \theta T_j Z_i$$

Correlation structures: independence, exchangeable, unstructured

```
> GO.indep = geeglm(Y ~ totfager, family=binomial("logit"), data=ctq,
+                   id=id, corstr="independence", waves=week)
> summary(GO.indep)

Call:
geeglm(formula = Y ~ totfager, family = binomial("logit"), data = ctq,
    id = id, waves = week, corstr = "independence")

 Coefficients:
            Estimate Std.err Wald Pr(>|W|)
(Intercept)   0.3984  0.3938 1.02   0.3117
totfager     -0.1867  0.0623 8.98   0.0027 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)    0.999  0.0592
```

```
> GO.exch = geeglm(Y ~ totfager, family=binomial("logit"), data=ctq,
+                id=id, corstr="exchangeable", waves=week)
> summary(GO.exch)

Call:
geeglm(formula = Y ~ totfager, family = binomial("logit"), data = ctq,
    id = id, waves = week, corstr = "exchangeable")

 Coefficients:
            Estimate Std.err Wald Pr(>|W|)
(Intercept)   0.1280  0.3855 0.11    0.740
totfager     -0.1757  0.0611 8.27    0.004 **
---

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.09   0.104

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.643  0.0565
```

```
> G0.unst = geeglm(Y ~ totfager, family=binomial("logit"), data=ctq,
+                    id=id, corstr="unstructured", waves=week)
> summary(G0.unst)

Call:
geeglm(formula = Y ~ totfager, family = binomial("logit"), data = ctq,
    id = id, waves = week, corstr = "unstructured")

 Coefficients:
            Estimate Std.err Wald Pr(>|W|)
(Intercept)  -0.1609  0.3896 0.17   0.6797
totfager     -0.1961  0.0642 9.33   0.0023 **
---

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.43   0.296
```

```
Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2  0.11650  0.0578
alpha.1:3  0.10869  0.0588
alpha.1:4  0.09500  0.0585
alpha.1:5  0.02332  0.0512
alpha.1:6  0.02857  0.0508
alpha.1:7  0.01258  0.0504
alpha.1:8  0.01526  0.0514
alpha.1:9  0.00438  0.0515
alpha.2:3  0.81458  0.1446
alpha.2:4  0.79954  0.1428
alpha.2:5  0.75740  0.1379
alpha.2:6  0.70171  0.1332
alpha.2:7  0.67125  0.1255
alpha.2:8  0.72796  0.1361
alpha.2:9  0.60318  0.1196
alpha.3:4  0.94377  0.1593
alpha.3:5  0.91568  0.1548
alpha.3:6  0.78175  0.1435
alpha.3:7  0.77467  0.1432
```