MEDIAN REGRESSION MODELS FOR LONGITUDINAL DATA WITH MISSING OBSERVATIONS

Grace Yi

Department of Statistics and Actuarial Science, University of Waterloo

OUTLINE

$*$ GEE AND WEIGHTED GEE

$*$ MODEL FORMULATION

$*$ ESTIMATION PROCEDURES

GEE AND WEIGHTED GEE

LONGITUDINAL DATA

- Y_{ij}: response for subject i at time point j

$$Y_{i1} \quad Y_{i2} \quad Y_{i3} \quad Y_{i4} \quad Y_{i5} \quad Y_{i6}$$

- $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5}, Y_{i6})'$

MEAN MODEL OF INTEREST

- $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ - mean vector

- GEE (Liang and Zeger 1986):

$$\sum_i \boldsymbol{D}_i \cdot \boldsymbol{V}_i^{-1} \cdot \boldsymbol{\epsilon}_i = \mathbf{0}$$

where $\boldsymbol{\epsilon}_i = (Y_{i1} - \mu_{i1}, ..., Y_{im} - \mu_{im})'$

MEDIAN MODEL OF INTEREST

- $\boldsymbol{\mu}_j$=mdeian of $\boldsymbol{Y}_i$, given $\boldsymbol{x}_i$

- GEE (Jung 1996; Godambe 2001):

INCOMPLETE LONGITUDINAL DATA

- Y_{ij}: response for subject i at time point j

$$Y_{i1} \quad Y_{i2} \quad Y_{i3} \quad Y_{i4} \quad Y_{i5} \quad Y_{i6}$$

- $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \underline{Y_{i3}}, Y_{i4}, \underline{Y_{i5}}, Y_{i6})' = (\boldsymbol{Y}_i^{obs'}, \boldsymbol{Y}_i^{mis'})'$

- $R_{ij} = I(Y_{ij} \text{ is observed})$

SELECTION MODELS (Little and Rubin 1987; Little 1995)

MISSING DATA MECHANISMS (Little and Rubin 2002)

- Missing Completely At Random (MCAR)

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\alpha}) = f(\boldsymbol{r}_i|\boldsymbol{x}_i; \boldsymbol{\alpha})$$

- Missing At Random (MAR)

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\alpha}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^{obs}, \boldsymbol{x}_i; \boldsymbol{\alpha})$$

- Not Missing At Random (NMAR)

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\alpha}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{mis}, \boldsymbol{x}_i; \boldsymbol{\alpha})$$

WEIGHTED GEE FOR INCOMPLETE LONGITUDINAL DATA

- Weighted GEE for Mean Model (Robins, Rotnitzky, and Zhao 1995)
  - GEE: with MCAR: $\Sigma_i \boldsymbol{D}_i^{obs} \cdot (\boldsymbol{V}_i^{obs})^{-1} \cdot (\boldsymbol{Y}_i^{obs} - \boldsymbol{\mu}_i^{obs}) = \mathbf{0}$

    $$E_{(\boldsymbol{Y}_i,\boldsymbol{R}_i)}(U_i(\beta)) = E_{\boldsymbol{Y}_i}\{\sum_{\boldsymbol{r}} \boldsymbol{D}_i^{obs}(\boldsymbol{V}_i^{obs})^{-1}(\boldsymbol{Y}_i^{obs} - \boldsymbol{\mu}_i^{obs}) \cdot P(\boldsymbol{R}_i = \boldsymbol{r}|\boldsymbol{Y}_i)\}$$

  - WGEE: with MAR: $\Sigma_i \boldsymbol{D}_i \cdot \boldsymbol{V}_i^{-1} \cdot \boldsymbol{\Delta}_i \cdot (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$
- Weighted GEE for Median Model

  Lipsitz et al. (1997)

  - serial correlation is not accounted for
  - asymptotic properties are not established

MODEL FORMULATION

MEDIAN REGRESSION MODEL

- Notation: n subjects are followed up longitudinally at m occasions

  Y_{ij}: continuous response; $\quad \boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{im})'$

  $\boldsymbol{x}_{ij}$: covariate vector; $\quad \boldsymbol{x}_i = (\boldsymbol{x}'_{i1}, \boldsymbol{x}'_{i2}, ..., \boldsymbol{x}'_{im})'$

  μ_{ij}: median of Y_{ij}, given $\boldsymbol{x}_i$; $\quad \boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{im})'$

  $f(\mu_{ij})$: pdf of Y_{ij} at μ_{ij}

- Regression Model

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

MODEL FOR THE MISSING DATA PROCESS

- Notation:

  $M_i = \Sigma_{j=1}^{m} R_{ij} + 1$: the drop-out time

  monotone missing data patterns: $R_{ij} = 0 \Rightarrow R_{ik} = 0$ for $k > j$

  conditional probability: $\lambda_{ij} = P(R_{ij} = 1 | R_{i,j-1} = 1, \boldsymbol{y}_i, \boldsymbol{x}_i)$

- Regression Model (MAR):

$$\text{logit}\lambda_{ij} = \boldsymbol{u}'_{ij}\boldsymbol{\alpha}$$

where $\boldsymbol{\alpha}$ = parameters for the missing data process

ESTIMATION PROCEDURES

WEIGHTED GEE FOR $\boldsymbol{\beta}$

$$\boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{D}_i \boldsymbol{\Gamma}_i \boldsymbol{V}_i^{-1} \cdot \boldsymbol{\Delta}_i(\boldsymbol{\alpha}) \cdot \boldsymbol{\epsilon}_i$$

$$\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) = \mathbf{0}$$

where $\quad \boldsymbol{D}_i = \partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}; \quad \boldsymbol{\Gamma}_i = \mathrm{diag}(f(\mu_{ij}), j = 1, 2, ..., m)$

$$\epsilon_{ij} = I(Y_{ij} \geq \mu_{ij}) - 1/2; \quad \boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im})'$$

<u>CONSTRUCTION OF WEIGHTS</u>

- $\boldsymbol{\Delta}_i(\boldsymbol{\alpha}) = \text{diag}(I(R_{ij} = 1)/\pi_{ij}, 1 \leq j \leq m)$

  where $\quad \pi_{ij} = P(R_{ij} = 1|\boldsymbol{y}_i, \boldsymbol{x}_i) = \Pi_{t=2}^{j} \lambda_{it}$

<u>ESTIMATING EQUATIONS FOR $\boldsymbol{\alpha}$</u>

- Likelihood: $L_i(\boldsymbol{\alpha}) = \Pi_{t=1}^{m_i - 1} \lambda_{it} \cdot (1 - \lambda_{im_i})$

- score: $\boldsymbol{S}_i(\boldsymbol{\alpha}) = \partial \ell_i(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$

$^{n}\ \boldsymbol{S}_i(\boldsymbol{\alpha}) = \mathbf{0}$

ASYMPTOTIC PROPERTIES

<u>THEOREM</u>: Under some regularity conditions, we have, as $n \to \infty$,

1. $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$
2. $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{P}^{-1}\boldsymbol{\Sigma}[\boldsymbol{P}^{-1}]')$

where

$$\boldsymbol{P} = E\left[\partial \boldsymbol{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}'\right]$$

$$\boldsymbol{\Sigma} = E\{\boldsymbol{Q}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})\boldsymbol{Q}_i'(\boldsymbol{\beta}, \boldsymbol{\alpha})\}$$

APPLICATION

DATA (Davis 1991)

83 individuals: 43 treated and 40 in placebo group

6 scheduled assessments: 30 minutes apart

amount of pain: measured on a 100mm line

0= no pain; 100= extreme pain

59% of women have missing values: monotone missing data patterns

1	2	3	4	5	6
5.0	1.0	1.0	0	5.0	.

RESPONSE MODEL

$$\mu_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2j} + \beta_3 x_{i1} x_{i2j}$$

where $x_{i1} = 0$ if subject i received treatment, and 1 otherwise

$x_{i2j} = j$ indexes the assessment time for subject i

MISSING DATA MODEL

$$\text{logit } \lambda_{ij} = \alpha_0 + \alpha_1 y_{i,j-1} + \alpha_2 x_{i1}$$

SUMMARY OF RESULTS

- Missing Data Process:
  - MAR mechanism appears reasonable (p-value= 0.046 for $H_o: \ \alpha_1 = 0$).
- Response Process:
  - Patients in the treatment group do not suffer increasing pain as time goes by (p-value= 0.273 for $H_o: \ \beta_2 = 0$).
  - The degree of pain in the control group would increase as time elapses (p-value $\simeq 0$ for $H_o: \ \beta$[illegible] $= 0$).

SIMULATION STUDY

MODELS

- $\boldsymbol{Y}_i \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{V})$; $\quad \boldsymbol{V} = \sigma^2[v_{st}]_{m\times m}$ with $v_{ss} = 1$ and $v_{st} = \rho$ for $s \neq t$
- Response and Missing Data Models: same as before
- Setting: $m = 6$, $n = 1000$, 200 simulations

$$\boldsymbol{\beta} = (6.0, -5.0, 1.0, 15.0)'; \quad \sigma = 1.0$$

$\boldsymbol{\alpha} = (1.0, 0.1, -0.5)'$: about 20% missing values

ANALYSES

SUMMARY

- Finite sample biases for Method 1 are smaller than those of Method 2; As ρ increases, biases for Method 1 tend to reduce, while biases for Method 2 do not change much.

- The standard errors for both Methods 1 and 2 seem to vary on the same scale.

- The coverage rates for Method 1 agree reasonably well with the nominal