Methods for a Longitudinal Quantitative Outcome With a Multivariate Gaussian Mixture Distribution Multi-dimensionally Censored by Therapeutic Intervention

By

Wanjie Sun, Michael D. Larsen, John M. Lachin

July 21, 2012

Motivation: DCCT/EDIC Example

DCCT (1982-1993)

Multicenter, randomized clinical trial. Intensive vs. Conventional treatment.

EDIC (1994-2016)

Observational follow-up study of the DCCT cohort.

Objective

- What is the prolonged treatment effect of the former intensive therapy on blood pressure (BP) and albuminuria excretion rate (AER) sixteen years after the end of DCCT.
- What are the genetic determinants for elevated BP or AER in the DCCT/EDIC cohort?

Motivation: DCCT/EDIC Example

Complications: Non-randomized, non-trial medication use (ACEI/ARB)

- Prohibited in DCCT, but allowed in EDIC.
- 50% by EDIC Year 16,
- Intervened for multiple reasons: hypertension, albuminuria or prophylactic reason.

Motivation: DCCT/EDIC Example

Result: Y is altered!

Observed $Y_{obs} \leq$ true underling Y

Bias:

- Clinical Trial: explanatory treatment effect
- Epidemiology: exposure or risk factor effect
- Genetic: genetic association

Objective

Objective:

- to develop methods for a more general longitudinal data structure where the quantitative outcomes are altered by non-randomized non-trial intervention
- applied in clinical trials and epidemiologic and genetic studies.
- to combine the advantages of the current statistical methods while minimizing their restrictive assumptions.

In particular:

Objective:

- to extend Cook's (1997) MI for a restrictive longitudinal data structure to one for a more general longitudinal data set up,
- to extend Tobin *et al.*'s (2005) censored normal regression model for cross-sectional data to one for longitudinal data

Model Specification

Assumptions:

- $Y_n \mid X_n \sim N(\mu_n, \Sigma_n)$
- Y is right censored at y_{obs}, $Y \geq y_{obs}$
- Once intervened, a subject will continue to be intervened.
- Ignorable dropout: $P(M \mid Y, \phi) = P(M \mid Y_{obs}, \phi))$

Model Specification

Model:

Partition Y into $Y_n = \begin{bmatrix} Y_{o1,n} \\ Y_{m,n} \end{bmatrix}$,where $Y_{m,n} \geq y_{o2,n}$.

The joint distribution of the observed underlying Y prior to intervention and the missing underlying Y post intervention is

$$\begin{bmatrix} Y_{o1,n} \\ Y_{m,n} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{o1,n}(\beta) \\ \mu_{m,n}(\beta) \end{bmatrix}, \begin{bmatrix} \Sigma_{o1,n}(R_n, D) & \Sigma_{(o1,m),n}(R_n, D) \\ \Sigma_{(m,o1),n}(R_n, D) & \Sigma_{m,n}(R_n, D) \end{bmatrix} \right)$$

Model Specification

Model:

$\Theta = (\beta, R_n, D)$ is based on a general linear mixed effect model (LMM),

$$y_{nj} = u + b_{0n} + (\beta_0 + \beta_1 G_n + b_{1n}) T_j + W^T \gamma + \varepsilon_{nj}$$

where

$$\begin{bmatrix} b_{0n} \\ b_{1n} \end{bmatrix} \sim N(0, D)$$

and

$$\varepsilon_n = \begin{bmatrix} \varepsilon_{n1} \\ \dots \\ \varepsilon_{n,t_n} \end{bmatrix} \sim N(0, R_n).$$

Extensions to parameter estimation $\Theta = (\beta, b, R_n, D)$ in a N-component MVN mixture distribution with multi-dimensional censored data

(1) Proper imputation using Monte Carlo in EM to simplify the integration and optimization of Q

randomly draw m values from the posterior distribution $f(y_m \mid y_{obs}, \theta^{(t)})$ of missing data to get imputed $y_m^{(j)}$

$Q_{t+1}(\Theta \mid \Theta^{(t)}) = \frac{1}{m} \sum_{j=1}^{m} l(\theta \mid y_{obs}, y_m^{(j)}, \Theta^{(t)})$

when m=1 and $y_m^{(j)}$ is imputed by a good statistics of the missing data: posterior mode or expected value.

Extensions to parameter estimation $\Theta = (\beta, b, R_n, D)$ in a N-component MVN mixture distribution with multi-dimensional censored data

(2) A simplified calculation for K-dimensionally censored MVN posterior distribution $f(y_m \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, ..., Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)})$:

(i) MVN?

- Posterior predictive distribution of the underlying Y_m given a multi-dimensionally truncated distribution: may not be normal, likely to be skewed.
- Schafer (1997): MVN is robust in imputation to distributions that are manifestly not normal when the amount of missing data is not large.

Continue

(2) A simplified calculation for K-dimensionally censored MVN posterior distribution $f(y_m \mid y_{o1}, Y_{m_1} \geq y_{o2_1}, \ldots, Y_{m_{q_n}} \geq y_{o2_{q_n}}, \Theta^{(t)})$:

(ii) Posterior expectation / variance / covariance of a missing underlying Y at post-intervention visit i:

take only the most informative subset

Proposed Method I

A modified EM-type algorithm with a single imputation

Starting data: observed Y_{obs}
Starting model: the Multi-level model (White *et al.* 2001)
E-step: a deterministic draw from the simplified posterior expectation: $\widetilde{y_{m,n}}^{(t)} = E(Y_{m_i} \mid y_{o1}, Y_{m_i} \geq y_{o2_i}, \Theta^{(t-1)})$
M-step: Q function reduces to a regular loglikelihood of LMM to get $\Theta^{(t)}$.
Iterate to get the final MLE estimate $\Theta = (\beta, b, R_n, D)$

Proposed Method II

A modified MCEM-MI algorithm with a multiple imputation

For the ith imputation,

E-step: a random draw from the simplied posterior distribution, $MVN(E(Y_m \mid y_{obs}, \Theta^{(t-1)}), V(Y_m \mid y_{obs}, \Theta^{(t-1)}))$

M-step: Q function reduces to a regular loglikelihood of LMM to get $\Theta^{(t)}$.

Iterate to get one set of MLE estimate $\Theta_i = (\beta, b, R_n, D)_i$

Repeat the whole process m times to get m sets of parameters

Combine the m sets of parameters to get the final parameter estimate Θ by Rubin's rule (1987).

Simulation Results

Compared the bias and efficiency of six class of methods

- reference: the true model or no medication model;
- the "ignore" model;
- the "exclude" model;
- the Multilevel Model (White. *et al.* 2001);
- the Modified EM-type algorithms including a full iteration, one-step and two-step interation algorithms;
- the Modified MCEM-MI algorithms including a full iteration, one-step and two-step iteration algorithms;
- adding a constant (5, 10, 15, 20) to the treated value y_{obs} (Tobin *et al.* 2005, Cui *et al.* 2003),

Table 3.2	Scenario 2 Fixed effect Model							
Description	100% onMed if Y >90, slope:5vs1, 5vsts, MedCutoff 90,AR$\rho = 0.9$							
% Med/Grp	27.7% in Group 0, and 8.0% in Group 1							
Model	NoMed	Ignore	Excld	MedAdj	EMtp	EM2stp	MI	MI2stp
True Slopes								
Slope G0=5	5.0|.1	3.71|.1	3.72|.2	4.15|.1	5.06|.2	4.98|.1	5.09|.2	5.00|.1
Slope G1=1	1.0|.1	0.57|.1	1.17|.2	0.71|.1	1.04|.1	1.03|.1	1.04|.1	1.02|.1
Slope Diff=4	4.0|.2	3.14|.1	2.56|.2	3.44|.1	4.02|.2	3.95|.2	4.05|.2	3.97|.2
95%CI Cover								
Slope G0=5	93%	0%	0%	0%	86%	90%	91%	96%
Slope G1=1	95%	7%	75%	37%	89%	92%	93%	93%
Slope Diff=4	95%	0%	0%	0%	88%	89%	95%	95%
Bias								
Slope G0:	0.00	1.29	1.28	0.85	0.06	0.02	0.09	0.00
Slope G1:	0.00	0.43	0.17	0.29	0.04	0.03	0.05	0.02
Slope Diff:	0.00	0.86	1.44	0.56	0.02	0.05	0.05	0.03
% Bias								
Slope G0:	0.0	25.8	25.6	17.0	1.2	0.4	1.8	0.1
Slope G1:	0.0	43.3	16.6	28.9	4.4	2.5	4.5	2.5
Slope Diff:	0.0	21.4	36.1	14.0	0.5	1.2	1.2	0.7

Improvements over Current Methods

- Improvements over current available methods for a quantitative outcome censored by a non-trial, non-randomized intervention.
- Improvements over Censored Normal Regression (Tobin *et al.* 2005)
- Improvements over Cook's 1997 MI method
- Improvements over Monte Carlo EM (Wei and Tanner 1990)
- Improvements over Lee and Scout's EM algorithms for censored multivariate Gaussian mixture model (working paper)
- Improvements over current conclusions about medication adjustment model (Tobin *et al.* 2005, White *et al.* 2001)