Biost 572: Final Talk

Jing Fan

University of Washington

jingfan@uw.edu

May 31, 2012

Paper

Title: A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data.

Authors: Pepe MS, Anderson GL

Publication: Communications in Statistics and Simulation, 1994

A mathematically-easy but conceptually-difficult paper

Background concepts

Marginal or Cross-sectional model:

$$E[Y_{it}|X_{it}] = g(X_{it}\beta)$$

Cross-sectional study:
A *descriptive* study providing data on the entire population at some specific time.
Note: no control on covariates, and so covariates are RANDOM

Longitudinal data:
Each subject is followed over a period of time, and repeated observations of the outcome and relevant covariates are recorded.
Note: CORRELATED outcomes

Generalized Estimating Equations (GEE)

$$S(\beta) = \sum_{i=1}^{N} D_i^T V_i^{-1}(Y_i - g(X_i\beta)) = 0,$$

β is the parameter representing cross-sectional association between outcomes Y and covariates X.

D_i is a partial derivative matrix of $\mu_i \equiv g(X_i\beta)$ with respect to β.

$V_i = A_i^{1/2} R_i A_i^{1/2}$ is a working *covariance* matrix, where A_i is a diagonal matrix with $\mathrm{Var}(Y_{it})$ on its diagonal line and R_i is a working *correlation* matrix.

GEE

As long as the marginal mean model $E[Y_i|X_i] = g(X_i\beta)$ is correctly specified, GEE method gives a consistent and asymptotically Normal estimate for β, for many sorts of correlated data.

However, GEE method fails (e.g. gives a biased estimate for β) for some types of longitudinal data.

Examples in the paper

Three different ways to generate longitudinal data:

1. $Y_{it} = \alpha Y_{i(t-1)} + \beta X_{it} + \epsilon_{it}$, where $Y_{i0} = 0$, X_{it} and ϵ_{it} have mean 0 and are independent of $Y_{i(t-1)}$ (i.e. an autoregressive process).
2. $Y_{it} = \beta X_{it}(\alpha Y_{i(t-1)}) + \epsilon_{it}$, where $\alpha = \beta^{-1}$, $Y_{i0} = 1$ and X_{it} has mean 1.
3. $Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$, where α_i is random effect with mean 0 and independent of X_{it} and ϵ_{it} (i.e. random-effect model).

All models above have the same marginal mean model $E(Y_{it}|X_{it}) = X_{it}\beta$

Examples in the paper

With the true $\beta = 0.5$ and 1,000 simulations, the average GEE estimates for β are:

Model	*Identity*	*Opt*1	*Opt*2	*Opt*3
1	0.501	0.416	0.399	0.428
2	0.498	0.416	0.393	0.413
3	0.500	0.495	0.498	0.500

- The estimates in the first column and last row are unbiased and consistent.
- Other estimates are clearly biased.

For longitudinal data, we can get a consistent and asymptotically Normal estimate for β if either (i) we use independent working correlation matrix, or (ii) we validate the following equality:

$$E(Y_{it}|X_{it}) = E(Y_{it}|X_{is}, s = 1, ..., n_i)$$

An insight look at a simple autoregressive model

$Y_{it} = Y_{i(t-1)} + \beta X_{it} + \epsilon_{it}$, where $Y_{i0} = 0$, X_{it} and ϵ_{it} have mean 0 and are independent of $Y_{i(t-1)}$. For simplicity, we assume all cluster sizes are equal to m.

$$E(Y_{it}|X_{it}) = X_{it}\beta$$

$$\begin{aligned}
\because Y_{it} &= Y_{i(t-1)} + \beta X_{it} + \epsilon_{it} \\
&= Y_{i(t-2)} + \beta(X_{it} + X_{i(t-1)}) + (\epsilon_{i(t-1)} + \epsilon_{it}) \\
&= \cdots \\
&= Y_{i0} + \beta \sum_{s=1}^{t} X_{is} + \sum_{s=1}^{t} \epsilon_{is}
\end{aligned}$$

$$\therefore E(Y_{it}|X_{is}, s = 1, ..., m) = \sum_{s=1}^{t} X_{is}\beta$$

An insight look at a simple autoregressive model

Using diagonal (i.e.independent) working correlation matrix:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{m} X_{it}Y_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{m} X_{it}^2}$$

$$E[\hat{\beta}|X] = [1 + \frac{\sum_{i=1}^{n}\sum_{t=2}^{m} X_{it}\sum_{j=1}^{t-1} X_{ij}}{\sum_{i=1}^{n}\sum_{t=1}^{m} X_{it}^2}]\beta \approx \beta$$

$$E[\hat{\beta}] = E[E[\hat{\beta}|X]] = \beta$$

An insight look at a simple autoregressive model

Using non-diagonal working correlation matrix:
For simplicity, suppose $A = V^{-1} = (a_{ij})$ which is a non-diagonal matrix, for all clusters

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (\sum_{t=1}^{m} a_{tj} X_{it}) Y_{ij}}{\sum_{i=1}^{n} X_i' A X_i}$$

$$E[\hat{\beta}] = E[E[\hat{\beta}|X]] \rightarrow [1 + \frac{\sum_{t=1}^{m-1} \sum_{j=t+1}^{m} a_{tj}}{\sum_{j=1}^{m} \lambda_j}]\beta$$

Proof of sufficient conditions

The unbiasedness and consistency of GEE estimates $\hat{\beta}$ result from

$$E[S(\beta)] = 0$$

Let $W_i = V_i^{-1}$ and $\mu_i = g(X_i\beta)$, and then:

$$
\begin{aligned}
S(\beta) &= \sum_{i=1}^{N} D_i^T W_i (Y_i - \mu_i) \\
&= \sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \beta_0} & \cdots & \frac{\partial \mu_{in_i}}{\partial \beta_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_{i1}}{\partial \beta_{p-1}} & \cdots & \frac{\partial \mu_{in_i}}{\partial \beta_{p-1}} \end{pmatrix} \begin{pmatrix} w_{i11} & \cdots & w_{i1n_i} \\ \vdots & \ddots & \vdots \\ w_{in_i1} & \cdots & w_{i1n_in_i} \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu_{i1} \\ \vdots \\ Y_{in_i} - \mu_{in_i} \end{pmatrix} \\
&= \sum_{i=1}^{N} \begin{pmatrix} \sum_{k=1}^{n_i} w_{ik1} \frac{\partial \mu_{ik}}{\partial \beta_0} & \cdots & \sum_{k=1}^{n_i} w_{ikn_i} \frac{\partial \mu_{ik}}{\partial \beta_0} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n_i} w_{ik1} \frac{\partial \mu_{ik}}{\partial \beta_{p-1}} & \cdots & \sum_{k=1}^{n_i} w_{ikn_i} \frac{\partial \mu_{ik}}{\partial \beta_{p-1}} \end{pmatrix} \begin{pmatrix} Y_{i1} - \mu_{i1} \\ \vdots \\ Y_{in_i} - \mu_{in_i} \end{pmatrix}
\end{aligned}
$$

Proof of sufficient conditions

$$= \sum_{i=1}^{N} \begin{pmatrix} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} w_{ikj} \frac{\partial \mu_{ik}}{\partial \beta_0} (Y_{ij} - \mu_{ij}) \\ \vdots \\ \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} w_{ikj} \frac{\partial \mu_{ik}}{\partial \beta_{p-1}} (Y_{ij} - \mu_{ij}) \end{pmatrix}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} w_{ikj} (\frac{\partial \mu_{ik}}{\partial \beta})^T (Y_{ij} - \mu_{ij})$$

Note that:

$$\frac{\partial \mu_{ik}}{\partial \beta} = \frac{\partial \mu_{ik}}{\partial (X_{ik}\beta)} \frac{\partial X_{ik}\beta}{\partial \beta}$$

$$= \frac{\partial \mu_{ik}}{\partial (X_{ik}\beta)} X_{ik}$$

Proof of sufficient conditions

Therefore,

$$S(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} X_{ik}^T w_{ikj} (\frac{\partial \mu_{ik}}{\partial (X_{ik}\beta)})(Y_{ij} - \mu_{ij})$$

Proof of sufficient conditions

If W_i is diagonal (i.e. working independence), $w_{ikj} = 0$ unless $k = j$, and so,

$$S(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} X_{ij}^T w_{ijj} \left(\frac{\partial \mu_{ij}}{\partial (X_{ij}\beta)}\right)(Y_{ij} - \mu_{ij})$$

$$\begin{aligned}
E[S(\beta)] &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} E[X_{ij}^T w_{ijj} \left(\frac{\partial \mu_{ij}}{\partial (X_{ij}\beta)}\right)(Y_{ij} - \mu_{ij})] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} E[E[X_{ij}^T w_{ijj} \left(\frac{\partial \mu_{ij}}{\partial (X_{ij}\beta)}\right)(Y_{ij} - \mu_{ij})]|X_{ij}] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} E[X_{ij}^T w_{ijj} \left(\frac{\partial \mu_{ij}}{\partial (X_{ij}\beta)}\right)(E[Y_{ij}|X_{ij}] - \mu_{ij})] \\
&= 0
\end{aligned}$$

Proof of sufficient conditions

If W_i is non-diagonal, then,

$$
\begin{aligned}
E[S(\beta)] &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E[X_{ik}^T w_{ikj}(\frac{\partial \mu_{ik}}{\partial(X_{ik}\beta)})(Y_{ij}-\mu_{ij})] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E[E[X_{ik}^T w_{ikj}(\frac{\partial \mu_{ik}}{\partial(X_{ik}\beta)})(Y_{ij}-\mu_{ij})]|X_{is}, s=1,...,n_i] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} E[X_{ik}^T w_{ikj}(\frac{\partial \mu_{ik}}{\partial(X_{ik}\beta)})(E[Y_{ij}|X_{is}, s=1,...,n_i]-\mu_{ij})]
\end{aligned}
$$

Clearly, $E[S(\beta)] = 0$, if $E[Y_{ij}|X_{is}, s=1,...,n_i] = \mu_{ij} \equiv E[Y_{ij}|X_{ij}]$.

An big question

Q: Does the biased estimator issue happen to GEE method ONLY?

A: No, it could happen to ANY estimator whose unbiasedness and consistency rely on $E[S(\beta)] = 0$, where $S(\beta)$ could be written in a form of $X^T V^{-1}(Y - g(X\beta))$.

- Linear Model and LMM in which $\beta : S(\beta) = X^T V^{-1}(Y - X\beta) = 0$
- Some GLMs and GLMMs
- Some general likelihood-based methods

General likelihood-based methods

Approach 1: most common method

$$L(Y,X) = L(Y|X)L(X)$$
$$= (\prod_{i=1}^{N} L(Y_i|X_i))L(X)$$

Approach 2: also called Telescope/Sequencing method

$$L(Y_i|X_i) = \prod_{t=1}^{ni} L(Y_{it}|X_{it}, Y_{is}, X_{is}, s < t)L(X_{it}|Y_{is}, X_{is}, s < t)$$

Summary:
Those two likelihood-based methods do not help make correct inference on cross-sectional effect β, unless $L(Y_{it}|X_{is}, s = 1, ..., n_i) = L(Y_{it}|X_{it})$ which leads to $E[Y_{it}|X_{is}, s = 1, ..., n_i] = E[Y_{it}|X_{it}]$.

A simulation illustration: setup

Model:

- $Y_{it} = \gamma_0 + \gamma_1 X_{it} + \gamma_2 X_{i(t-1)} + b_i + e_{it}$
- $X_{it} = \rho X_{i(t-1)} + \epsilon_{it}$
- $b_i, e_{it}, \epsilon_{it}$ are mutually independent with mean zero

Then,

- $E[Y_{it}|X_{is}, s = 1, ..., n_i] = \gamma_0 + \gamma_1 X_{it} + \gamma_2 X_{i(t-1)}$
- $E[Y_{it}|X_{it}] = \beta_0 + \beta_1 X_{it}$, where $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1 + \rho\gamma_2$

For the simulations, we use

- $b_i \sim N(0, 1), e_{it} \sim N(0, 1), \epsilon_{i0} \sim N(0, 1)$ and $\epsilon_{it} \sim N(0, 1 - \rho^2)$
- $\gamma_0 = 0, \gamma_1 = 1, \gamma_2 = 1$

A simulation illustration: results

For several different values of ρ, the average GEE estimates of β_1 are:

ρ	0.9	0.7	0.5	0.3	0.1
β_1	1.9	1.7	1.5	1.3	1.1
Independence	1.90	1.70	1.50	1.30	1.10
Exchangeable	1.73	1.54	1.37	1.19	1.01
AR(1)	1.70	1.31	1.07	0.89	0.74
LMM	1.73	1.54	1.37	1.19	1.01

Summary:

- Similar to the examples in the paper, the estimates are biased, unless independence working correlation matrix is used.
- In addition, linear mixed effect model could not save us.

Which model to fit?

Unless satisfying the sufficient conditions, we cannot get rid of biased estimate issue when we apply marginal model to longitudinal data. So, why not use other modelling approaches for longitudinal data?

We cannot arbitrarily choose model, since the choice should depend on the question of scientific interest.

- Fully-conditional model: interested in association between outcomes and covariates at ALL times
- Partly-conditional model: interested in association between outcomes and covariates at not all but SOME times
- Marginal model: interested in association between outcomes and covariates at the SAME time.
- Random/mixed effect model: interested in modelling mean AND covariance

Advantages of marginal models

Even when we are free to choose a model (e.g. exploratory study), it still worth applying marginal model to longitudinal.data

- Conceptually and computationally simple
- Easy to deal with missing values
- Simple data display (e.g. scatterplot) is okay

Further considerations and concerns

- When using GEE method to analyze longitudinal data in practice, ALWAYS use independence working correlation matrix, if efficiency is not a problem.
- If efficiency matters, try to validate/assume $E(Y_{it}|X_{it}) = E(Y_{it}|X_{is}, s = 1, ..., n_i)$ based on correlation structures (e.g. observation-driven model v.s. parameter-driven model).
- How about fixed covariates in controlled study? Does the same issue occur?

Questions?