Cluster Data.

$$\{(X_{i1}, Y_{i1}), \ldots, (X_{imi}, Y_{imi});\ i = 1, \ldots, n\}$$

where Y_{ij} and $X_{ij} = (X_{ij1}, \ldots, X_{ijp})$ are the response variable and the $p \times 1$ coordinate vector of the j-th individual within the i-th cluster, and m_i denotes the random cluster size, $i = 1, \ldots, n;\ j = 1, \ldots, m_i$.

m_i and X_i are related. (informative) A parametric model is considered for each individual Y_{ij},

$$E(Y_{ij}|X_{ij}) = u_c(X_{ij}; \theta)$$

where $\theta = (\theta_1, \ldots, \theta_p)^T$.

Model Assumption

Under the validity of a fitting model, an independent cluster size

$$E[Y_{ij}|X_{ij}, m_i] = E[Y_{ij}|X_{ij}], \quad \forall i, j$$

is usually assumed in estimation procedures, e.g., generalized estimating equation (GEE).

In this lecture, we consider the condition that the cluster size is informative, i.e.,

$$E[E[h_j|X_{ij}, m_i]] \neq E[E[Y_j|X_{ij}]]$$

—

Remark

It can be verified that an estimator derived from the GEE might produce an inconsistent estimator of θ under the assumption of informative cluster size.

—

Let

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i\,m_i} \end{pmatrix}, \quad \mu_i = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{i\,m_i} \end{pmatrix}, \quad V_i = (V_{i\,j_1 j_2}); \qquad i = 1, \ldots, n; \quad j = 1, \ldots, m_i$$

with

$$\begin{aligned} V_{i\,j_1 j_2} &= \mathrm{Cor}(Y_{ij_2}, Y_{ij_2}|X_{ij_1}, X_{ij_2}) \\ &\triangleq h(x_{ij_1}, x_{ij_2}; \theta) \end{aligned}$$

—

In GEE, an estimator is obtained by solving the estimating equations

$$S(\theta) = 0,$$

where

$$S(\theta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \theta} \right) V_i^{-1}(Y_i - \mu_i).$$

Because of the possible dependence between Y_{ij} and m_i, the conditional expectation

$$E\left[\left(\frac{\partial \mu_i}{\partial \theta} \right) V_i^{-1}(Y_{ij} - \mu_i) \mid (X_{i1}, \ldots, X_{im_i}) \right]$$

might not equal zero, and hence, the derived estimator is not consistent.

—

Within Cluster Resampling (WCR) Approach

(Hoffman, Sen, and Weinberg (2001))

Estimation Procedure

Step 1. The qth subsample $\{(X_{1q}, Y_{1q}), \ldots, (X_{nq}, Y_{nq})\}$, $q = 1, \ldots, Q_1 = \prod_{i=1}^{n} m_i$, is drawn with a randomly selected individual (X_{iq}, Y_{iq}) from the ith cluster, $i = 1, \ldots, n$.

Step 2: Based on the qth subsample in Step 1, the qth estimator, say $\hat{\theta}_q$, of θ, is obtained from solving the equation

$$S_1(\theta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \theta} \right) \hat{V}_{i\,qq}^{-1}(Y_{iq} - \mu_{iq}).$$

where $\hat{V}_{i\,qq}$ is a consistent estimator of $V_{i\,qq}$.

Step 3: Repeat Steps 1-2. The final estimator $\bar{\theta}_{wcr}$ takes the average of the Q_i estimators,

$$\bar{\theta}_{wcr} = \frac{1}{Q_1} \sum_{q=1}^{Q_1} \widehat{\theta}_i.$$

—

Remark:

1. The number of possible sub-samples Q_1 is extremely large; a reasonable number of re-sampling is often implemented in the WCR approach.
2. Two merits of the WCR approach:
   (a) The within-cluster correlation does not need to be specified in the estimating equations.
   (b) The estimator $\bar{\theta}_{wcr}$ is a consistent estimator of θ.
3. Drawbacks: The WCR approach is found to be computationally intensive in implementation.

—

Cluster-Weighted generalized estimating equation (CWGEE) approach.

(Williamson, J,M,...(2003)) The estimator, say $\widetilde{\theta}_{cw}$ is obtain by solving the following estimating equations:

$$S_2(\theta) \triangleq \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{\partial \mu_{ij}}{\theta} \hat{V}_{ijj}^{-1} \left(Y_{ij} - \mu_{ij}\right) = 0$$

—

Remark:

$$\begin{aligned}
\mathbb{E}\left[S_2(\theta)\right] &= \sum_{i=1}^{n} E\left[\frac{1}{m_i} \sum_{j=1}^{m_i} E\left[\frac{\partial \mu_{ij}}{\partial \theta} \hat{V}_{ijj}^{-1} \left(Y_{ij} - \mu_{ij}\right) \mid m_i\right]\right] \\
&= \sum_{i=1}^{n} E\left[\frac{\partial \mu_{ij}}{\partial \theta} \hat{V}_{ijj}^{-1} \left(Y_{ij} - \mu_{ij}\right)\right] \\
&= \sum_{i=1}^{n} E[\frac{\partial \mu_{ij}}{\partial \theta} \hat{V}_{ijj}^{-1} \underbrace{E\left[\left(Y_{ij} - \mu_{ij}\right) \mid X_{ij}\right]}_{0}] = 0
\end{aligned}$$

—

Asymptotic Equivalent between the WCR and CWGEE

Under some regularity conditions, one can derive that

$$\hat{\theta}_1 = \theta + \frac{1}{n} H^{-1}(\theta) S_{1q}^*(\theta) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

where

$$S_{1q}^*(\theta) = \sum_{i=1}^{n} \frac{\partial \mu_{iq}}{\partial \theta} V_{iqq} \left(Y_{iq} - \mu_{iq}\right)$$

and

$$H_1(\theta) = E\left[\frac{\partial \mu_{iq}}{\partial \theta} V_{ijj}^{-1} \left(\frac{\partial \mu_{iq}}{\partial \theta}\right)^{\top}\right].$$

As a result,

$$\bar{\theta}_{wcr} = \theta + \frac{1}{n} H^{-1}(\theta) \frac{1}{Q_1} \sum_{q=1}^{Q_1} S_{1q}^*(\theta) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

Similarly

$$\widetilde{\theta} cw = \theta + \frac{1}{n} H^{-1}(\theta) S_2^*(\theta) + o_p(1)$$

—

Theorem. 1

Under some regularity conditions

$$\bar{\theta}_{wct} = \widetilde{\theta}_{cw} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof:

$$
\begin{aligned}
\frac{1}{Q_1} \sum_{q=1}^{Q_1} S_{1q}^*(\theta) & =\sum_{i=1}^n \frac{1}{Q} \cdot \sum_{q=1}^{Q_1}\left(\frac{\partial \mu_{iq}}{\partial \theta}\right) V_{iqq}^{-1}\left(Y_{iq}-\mu_{iq}\right) \\
& =\sum_{i=1}^n \frac{1}{Q}\left(\prod_{l \neq i}^n m_l\right)\left(\sum_{j=1}^{m_i}\left(\frac{\partial \mu_{ij}}{\partial \theta}\right) V_{ijj}^{-1}\left(Y_{ij}-\mu_{ij}\right)\right) \\
& =\sum_{i=1}^n \frac{1}{m_i} \sum_j^{m_i}\left(\frac{\partial \mu_{ij}}{\partial \theta}\right) V_{ijj}^{-1}\left(Y_{ij}-\mu_{ij}\right) \\
& =S_2^*(\theta)
\end{aligned}
$$

Efficient Estimation Method Let $m=\min \left\{m_1, \ldots, m_n\right\} \geq z$ Modified Within Cluster Resampling (MWCR) approach.

Step 1: The q^{th} subsample

$$
\begin{gathered}
\left\{\left(X_i q(m), Y_i q(m)\right), \ldots,\left(X_{nq(m)}, Y_{nq(m)}\right)\right\} . \\
q=1, \ldots, Q_m=\prod_{i=1}^n C^{m_1}
\end{gathered}
$$

is drawn with the q^{th} observation

$$
\begin{aligned}
& X_{ig}(m)=\left(X_{iq}, \ldots, X_{\text{imp}}\right)^{\top}, \\
& Y_{iq}(m)=\left(Y_{iq}, \ldots, Y_{\text{imp}}\right)^{\top}, \\
& \text{ of } X_i \text{ and } Y_i .
\end{aligned}
$$

Step 2: Based on the q^{th} subsample in Step 1, the q-th estimator, Say $\hat{Q}_{q(m)}$ of θ is defined as the solution of

$$
S_{1q}^{(m)}(\theta)=\sum_{i=1}^n \frac{\partial \mu_{iq}(m)}{\partial \theta} \hat{V}_{iq(m)}^{-1}\left(Y_{iq(m)}-\mu_{iq(m)}\right)=0
$$

where $\hat{V}_{iq(m)}=\left(h\left(X_{ijq}, X_{ij2q} ; \alpha\right)\right)$