Chapter 2: Likelihood-based approach (Part 4)

Discuss some statistical properties of the observed score function in the missing data setup.

Definition

1. Observed score function: $S_{\rm obs}(\eta) = \frac{\partial}{\partial \eta} \log L_{\rm obs}(\eta)$
2. Fisher information from observed likelihood: $\mathrm{I}_{\rm obs}(\eta) = -\frac{\partial^2}{\partial \eta \partial \eta^T} \log L_{obs}(\eta)$
3. Expected (Fisher) information from observed likelihood: $\mathcal{I}_{\rm obs}(\eta) = \mathrm{E}_{\eta}\left\{\mathrm{I}_{\rm obs}(\eta)\right\}$.

Theorem 2.6

Under regularity conditions,

$$\mathrm{E}_{\eta}\{S_{\mathrm{obs}}(\eta)\} = 0, \quad \text{and} \quad \mathrm{V}_{\eta}\{S_{\mathrm{obs}}(\eta)\} = \mathcal{I}_{\mathrm{obs}}(\eta),$$

where $\mathcal{I}_{\mathrm{obs}}(\eta) = \mathrm{E}_{\eta}\{\mathrm{I}_{\mathrm{obs}}(\eta)\}$ is the expected information from the observed likelihood.

4 Observed information

- Under missing data, the MLE $\hat{\eta}$ is the solution to $S_{\rm obs}(\eta) = 0$.
- Under some regularity conditions, $\hat{\eta}$ converges in probability to η_0 and has the asymptotic variance $\{\mathcal{I}_{\rm obs}(\eta_0)\}^{-1}$ with

$$\mathcal{I}_{\rm obs}(\eta) = \mathrm{E}_{\eta}\left\{-\frac{\partial}{\partial \eta^T} S_{\rm obs}(\eta)\right\} = \mathrm{E}_{\eta}\left\{S_{\rm obs}^{\otimes 2}(\eta)\right\},$$

where $B^{\otimes 2} = BB^T$.

Decomposition of information matrix

- $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta)$: prediction model (or imputation model) for $\mathbf{y}_{\text{mis}}$.
- $S_{\text{mis}}(\eta)$: the score function with $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta)$

$$S_{\text{mis}}(\eta) = S_{\text{com}}(\eta) - S_{\text{obs}}(\eta)$$

- Bartlett identify

$$E_{\eta}\{S_{\text{mis}}(\eta) \mid \mathbf{y}_{obs}, \boldsymbol{\delta}\} = 0$$

- Orthogonality

$$Cov\{S_{\text{obs}}(\eta), S_{\text{mis}}(\eta)\} = 0. \tag{1}$$

- Pythagorean theorem

$$\mathrm{V}\{S_{\text{com}}(\eta)\} = \mathrm{V}\{S_{\text{obs}}(\eta)\} + \mathrm{V}\{S_{\text{mis}}(\eta)\}$$

Proof of (1)

- Missing information principle (Orchard and Woodbury, 1972):

$$\mathcal{I}_{\mathrm{mis}}(\eta) = \mathcal{I}_{\mathrm{com}}(\eta) - \mathcal{I}_{\mathrm{obs}}(\eta),$$

  where $\mathcal{I}_{\mathrm{com}}(\eta) = \mathrm{E}_{\eta}\left\{-\partial S_{\mathrm{com}}(\eta)/\partial \eta^T\right\}$ is the expected information with complete-sample likelihood .

- An alternative expression of the missing information principle is

$$\mathrm{V}\{S_{\mathrm{mis}}(\eta)\} = \mathrm{V}\{S_{\mathrm{com}}(\eta)\} - \mathrm{V}\{S_{\mathrm{obs}}(\eta)\}. \tag{2}$$

  Note that $\mathrm{V}\{S_{\mathrm{com}}(\eta)\} = \mathcal{I}_{\mathrm{com}}(\eta)$ and $\mathrm{V}\{S_{\mathrm{obs}}(\eta)\} = \mathcal{I}_{\mathrm{obs}}(\eta)$.

Example 2.7

1. Consider the following bivariate normal distribution:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right],$$

   for $i = 1, 2, \cdots, n$. Assume for simplicity that σ_{11}, σ_{12} and σ_{22} are known constants and $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ be the parameter of interest.

2. The complete sample score function for $\boldsymbol{\mu}$ is

$$S_{\text{com}}(\boldsymbol{\mu}) = \sum_{i=1}^{n} S_{\text{com}}^{(i)}(\boldsymbol{\mu}) = \sum_{i=1}^{n} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_{1i} - \mu_1 \\ y_{2i} - \mu_2 \end{pmatrix}.$$

   The information matrix of $\boldsymbol{\mu}$ based on the complete sample is

$$\mathcal{I}_{\text{com}}(\boldsymbol{\mu}) = n \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1}.$$

Example 2.7 (Cont'd)

3. Suppose that there are some missing values in y_{1i} and y_{2i} and the original sample is partitioned into four sets:

$$\begin{aligned} H &= \text{both } y_1 \text{ and } y_2 \text{ respond} \\ K &= \text{only } y_1 \text{ is observed} \\ L &= \text{only } y_2 \text{ is observed} \\ M &= \text{both } y_1 \text{ and } y_2 \text{ are missing.} \end{aligned}$$

   Let n_H, n_K, n_L, n_M represent the size of H, K, L, M, respectively.

4. Assume that the response mechanism does not depend on the value of (y_1, y_2) and so it is MAR. In this case, the observed score function of $\boldsymbol{\mu}$ based on a single observation in set K is

$$\begin{aligned} E\left\{S_{\text{com}}^{(i)}(\boldsymbol{\mu}) \mid y_{1i}, i \in K\right\} &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_{1i} - \mu_1 \\ E(y_{2i} \mid y_{1i}) - \mu_2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11}^{-1}(y_{1i} - \mu_1) \\ 0 \end{pmatrix}. \end{aligned}$$

5. Similarly, we have

$$E\left\{S_{\text{com}}^{(i)}(\boldsymbol{\mu}) \mid y_{2i}, i \in L\right\} = \begin{pmatrix} 0 \\ \sigma_{22}^{-1}(y_{2i} - \mu_2) \end{pmatrix}.$$

6. Therefore, the observed information matrix of $\boldsymbol{\mu}$ is

$$\mathcal{I}_{\text{obs}}(\boldsymbol{\mu}) = n_H \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} + n_K \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + n_L \begin{pmatrix} 0 & 0 \\ 0 & \sigma_{22}^{-1} \end{pmatrix}$$

and the asymptotic variance of the MLE of $\boldsymbol{\mu}$ can be obtained by the inverse of $\mathcal{I}_{\text{obs}}(\boldsymbol{\mu})$.

Remark 1

- In the special case of $n_L = n_M = 0$,

$$\{\mathcal{I}_{\rm obs}(\boldsymbol{\mu})\}^{-1} = \left\{ n_H \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} + n_K \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right\}^{-1}.$$

- Using the following Woodbury matrix identity

$$(A + c\mathbf{b}\mathbf{b}')^{-1} = A^{-1} - A^{-1}\mathbf{b}\left(c^{-1} + \mathbf{b}'A^{-1}\mathbf{b}\right)^{-1}\mathbf{b}'A^{-1}$$

with

$$A = n_H \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1},$$

$\mathbf{b} = (1, 0)'$ and $c = n_K \sigma_{11}^{-1}$, we have $c^{-1} + \mathbf{b}'A^{-1}\mathbf{b} = (1/n_H + 1/n_K)\,\sigma_{11}$ and

$$\{\mathcal{I}_{\rm obs}(\boldsymbol{\mu})\}^{-1} = \frac{1}{n_H} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} + \left(\frac{1}{n} - \frac{1}{n_H}\right) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{12}^2/\sigma_{11} \end{pmatrix}.$$

- Thus, the asymptotic variance of the MLE of μ_1 is equal to σ_{11}/n and the asymptotic variance of the MLE of μ_2 is equal to

$$V(\hat{\mu}_2) \doteq \frac{1}{n}\sigma_{22}\rho^2 + \frac{1}{n_H}\left(1-\rho^2\right)\sigma_{22} = \frac{1}{n_H}\sigma_{22} - \rho^2\sigma_{22}\left(\frac{1}{n_H} - \frac{1}{n}\right), \qquad (3)$$

  where $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$.

- Note that we obtain

$$V(\hat{\mu}_2) = V(\hat{\mu}_{2,H}) - \rho^2\sigma_{22}\left(\frac{1}{n_H} - \frac{1}{n}\right),$$

  where $\hat{\mu}_{2,H}$ is the MLE of μ_2 using sample H only.

- Thus, by incorporating the partial response, the asymptotic variance is reduced by $\rho^2\sigma_{22}(1/n_H - 1/n)$.

Return to Example 2.3

- Observed log-likelihood

$$\ln L_{obs}(\theta) = \sum_{i=1}^{n} \delta_i \log(\theta) - \theta \sum_{i=1}^{n} y_i$$

- MLE for θ:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \delta_i}$$

- Fisher information: $I_{obs}(\theta) = \sum_{i=1}^{n} \delta_i / \theta^2$
- Expected information: $\mathcal{I}_{obs}(\theta) = \sum_{i=1}^{n} (1 - e^{-\theta c_i})/\theta^2$.

Which one do you prefer ?

Motivation

- $L_{com}(\eta) = f(\mathbf{y}, \boldsymbol{\delta}; \eta)$: complete-sample likelihood with no missing data
- Fisher information associated with $L_{com}(\eta)$:

$$\mathrm{I}_{com}(\eta) = -\frac{\partial}{\partial \eta^T} S_{com}(\eta) = -\frac{\partial^2}{\partial \eta \partial \eta^T} \log L_{com}(\eta)$$

- $L_{obs}(\eta)$: the observed likelihood
- Fisher information associated with $L_{obs}(\eta)$:

$$\mathrm{I}_{obs}(\eta) = -\frac{\partial}{\partial \eta^T} S_{obs}(\eta) = -\frac{\partial^2}{\partial \eta \partial \eta^T} \log L_{obs}(\eta)$$

- How to express $\mathrm{I}_{\mathrm{obs}}(\eta)$ in terms of $\mathrm{I}_{\mathrm{com}}(\eta)$ and $S_{\mathrm{com}}(\eta)$?

Theorem 2.7 (Louis, 1982; Oakes, 1999)

Under regularity conditions allowing the exchange of the order of integration and differentiation,

$$\begin{aligned} \mathrm{I}_{\mathrm{obs}}(\eta) &= \mathrm{E}\{I_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - \left[\mathrm{E}\{S_{com}^{\otimes 2}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - \bar{S}(\eta)^{\otimes 2}\right] \\ &= \mathrm{E}\{I_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - V\{S_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\}, \end{aligned} \tag{4}$$

where $\bar{S}(\eta) = \mathrm{E}_{\eta}\{S_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\}$.

Check (using Example 2.3)

- Recall that

$$S_{\rm com}(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} t_i$$

and

$$I_{\rm com}(\theta) = \frac{n}{\theta^2}$$

- Now,

$$\begin{aligned} V\{S_{com}(\eta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} &= \sum_{i=1}^{n}(1-\delta_i)V(t_i \mid y_i, t_i > c_i) \\ &= \sum_{i=1}^{n}(1-\delta_i)\frac{1}{\theta^2}. \end{aligned}$$

- Thus,

$$\begin{aligned} \mathrm{I}_{\rm obs}(\theta) &= \mathrm{E}\{I_{com}(\theta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} - V\{S_{com}(\theta)|\mathbf{y}_{obs}, \boldsymbol{\delta}\} \\ &= \sum_{i=1}^{n}\delta_i \frac{1}{\theta^2} \end{aligned}$$

Proof of Theorem 2.7

Remark

- Result (4) is closely related to the missing information principle in (2). Note that

$$
\begin{aligned}
V\left\{S_{\text{mis}}(\eta)\right\} &= E\left[V\left\{S_{\text{mis}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\}\right] + V\left[E\left\{S_{\text{mis}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\}\right] \\
&= E\left[V\left\{S_{\text{mis}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\}\right] \\
&= E\left[V\left\{S_{\text{com}}(\eta) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}\right\}\right]
\end{aligned}
$$

  and we can see that $I_{\text{obs}}(\eta)$ in (4) is unbiased for $V\{S_{\text{obs}}(\eta)\} = V\{S_{\text{com}}(\eta)\} - V\{S_{\text{mis}}(\eta)\}$.
- To evaluate the conditional expectation in (4), we use the prediction model $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \eta)$.

REFERENCES

Louis, T. A. (1982), 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society: Series B* **44**, 226–233.

Oakes, D. (1999), 'Direct calculation of the information matrix via the em algorithm', *Journal of the Royal Statistical Society: Series B* **61**, 479–482.

Orchard, T. and M.A. Woodbury (1972), A missing information principle: theory and applications, *in* 'Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, California, pp. 695–715.