Chapter 4: Imputation

Jae-Kwang Kim

Motivating Example

- Basic Setup: Let $(x, y)'$ be a vector of bivariate random variables. Assume that x_i are always observed and y_i are subject to missingness in the sample.
- In this case, an imputed estimator of $\theta = E(Y)$ can be computed by

$$\hat{\theta}_I = \frac{1}{n} \sum_{i=1}^{n} \{\delta_i y_i + (1 - \delta_i) y_i^*\} \tag{1}$$

  where y_i^* is an imputed value for y_i.
- If y_i^* satisfies

$$E(y_i^* \mid \delta_i = 0) = E(y_i \mid \delta_i = 0), \tag{2}$$

  then the imputation estimator (1) is unbiased.
- A sufficient condition is to assume MAR and generate y_i^* from $f(y_i \mid x_i, \delta_i = 1)$.

Justification

Regression imputation (Example 4.1)

- Imputation model (under MAR) is a linear regression model:

$$y_i \mid x_i \sim \left(\beta_0 + \beta_1 x_i, \sigma_e^2\right),$$

for some $(\beta_0, \beta_1, \sigma_e^2)$.

- Regression imputation: Use $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \left(\bar{y}_r - \hat{\beta}_1 \bar{x}_r, S_{xxr}^{-1} S_{xyr}\right).$$

- The regression imputation estimator can be written as

$$\hat{\theta}_{\mathrm{I,reg}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right\} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

- How to estimate the variance of $\hat{\theta}_{I,reg}$?

Introduction

Basic setup

- $\mathbf{Y}$: a vector of random variables with distribution $F(\mathbf{y};\theta)$.
- $\mathbf{y}_1, \cdots, \mathbf{y}_n$ are n independent realizations of $\mathbf{Y}$.
- We are interested in estimating ψ which is implicitly defined by $E\{U(\psi;\mathbf{Y})\} = 0$.
- Under complete observation, a consistent estimator $\hat{\psi}_n$ of ψ can be obtained by solving estimating equation for ψ:

$$\sum_{i=1}^{n} U(\psi;\mathbf{y}_i) = 0.$$

- A special case of estimating function is the score function. In this case, $\psi = \theta$.
- How to find the asymptotic distribution of $\hat{\psi}_n$?

Asymptotically Linear Estimator

Definition

Let $X_1, \cdots, X_n$ be IID sample from $f(x;\theta_0), \theta_0 \in \Theta$ and we are interested in estimating $\gamma_0 = \gamma(\theta_0)$, where $\gamma(\cdot) : \Theta \to R^k$. An estimator $\hat{\gamma} = \hat{\gamma}_n$ is called asymptotically linear if there exist a random vector $\mathbf{a}(x) = \mathbf{a}(x;\theta_0)$ such that

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{a}(X_i) + o_p(1) \tag{3}$$

with $E_{\theta_0}\{\mathbf{a}(X)\} = 0$ and $E_{\theta_0}\{\mathbf{a}(X)\mathbf{a}(X)'\}$ is finite. Here, $Z_n = o_p(1)$ means that Z_n converges to zero in probability.

Remark

- The function $\mathbf{a}(x)$ is referred to as the influence function for $\hat{\gamma}$. The phrase influence function was used by Hampel (1974) and is motivated by the fact that to the first order $\mathbf{a}(x)$ is the influence of a single observation on the estimator $\hat{\gamma} = \hat{\gamma}(X_1, \cdots, X_n)$.
- The asymptotic properties of an asymptotically linear estimator can be summarized by considering only its influence function.
- Since $\mathbf{a}(X)$ has zero mean, the CLT tells us that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{a}(X_i) \xrightarrow{\mathcal{L}} N\left[0, E_{\theta_0}\{\mathbf{a}(X)\mathbf{a}(X)'\}\right]. \tag{4}$$

Thus, combining (3) with (4) and applying Slutsky's theorem, we have

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma_0\right) \xrightarrow{\mathcal{L}} N\left[0, E_{\theta_0}\{\mathbf{a}(X)\mathbf{a}(X)'\}\right].$$

Example

- Let $X_1, \ldots, X_n$ be IID from $N(\mu, \sigma^2)$.
- The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$$

- What is the influence function of $\hat{\sigma}^2$?

Lemma 4.1

Let $\hat{\psi}$ be the solution to $\hat{U}(\psi) = 0$, where

$$\hat{U}(\psi) = \frac{1}{n}\sum_{i=1}^{n} U(\psi; \mathbf{y}_i).$$

Lemma 4.1

Let ψ_0 be the solution to $E\{U(\psi; \mathbf{Y})\} = 0$. Then, under some regularity conditions,

$$\sqrt{n}\left(\hat{\psi} - \psi_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\tau(\psi_0)\}^{-1}U(\psi_0; \mathbf{y}_i) + o_p(1),$$

where $\tau(\psi) = -E\{\dot{U}(\psi; Y)\}$ and $\dot{U}(\psi; Y) = \partial U(\psi; Y)/\partial \psi'$.

Sketched Proof

- Its proof is based on Taylor linearization:

$$\begin{aligned}\hat{U}(\hat{\psi}) &\cong \hat{U}(\psi_0) + \frac{\partial}{\partial \psi'}\hat{U}(\psi_0)\left(\hat{\psi} - \psi_0\right) \\ &\cong \hat{U}(\psi_0) + E\{\dot{U}(\psi_0)\}\left(\hat{\psi} - \psi_0\right),\end{aligned}$$

  where the second (approximate) equality follows by

$$\frac{\partial}{\partial \psi'}\hat{U}(\psi_0) = E\{\dot{U}(\psi_0)\} + o_p(1)$$

  and $\hat{\psi} = \psi_0 + o_p(1)$.
- Need to assume that $E\left\{\dot{U}(\psi_0)\right\}$ is nonsingular.
- Also, we need conditions for $\hat{\psi} \xrightarrow{p} \psi_0$.

Remark 1 (Sandwich formula)

- Lemma 4.1 can be used to obtain a consistent variance estimator of $\hat{\psi}$:

$$\hat{V}(\hat{\psi}) = \frac{1}{n}\hat{\tau}^{-1}\hat{V}(U)\left(\hat{\tau}^{-1}\right)'$$

where

$$\begin{aligned} \hat{\tau} &= -n^{-1}\sum_{i=1}^{n}\dot{U}(\hat{\psi}; y_i) \\ \hat{V}(U) &= n^{-1}\sum_{i=1}^{n}U(\hat{\psi}; y_i)U(\hat{\psi}; y_i)' \end{aligned}$$

- The above formula is often called the sandwich formula.

Remark 2

- To solve a nonlinear equation $\hat{U}(\psi) = 0$, one might use the Newton method

$$\hat{\psi}^{(t+1)} = \hat{\psi}^{(t)} - \left\{ \dot{U}(\hat{\psi}^{(t)}) \right\}^{-1} \hat{U}(\hat{\psi}^{(t)}), \tag{5}$$

where $\dot{U}(\psi) = \partial \hat{U}(\psi)/\partial \psi'$. However, the partial derivative $\dot{U}(\psi)$ is not symmetric, and the iterative computation in (5) can have numerical problems.

- To deal with the problem, we can use

$$\hat{\psi}^{(t+1)} = \hat{\psi}^{(t)} - \left\{ \dot{U}(\hat{\psi}^{(t)})' \dot{U}(\hat{\psi}^{(t)}) \right\}^{-1} \dot{U}(\hat{\psi}^{(t)})' \hat{U}(\hat{\psi}^{(t)}), \tag{6}$$

which is essentially equivalent to finding $\hat{\psi}$ that minimizes $Q(\psi) = \hat{U}(\psi)' \hat{U}(\psi)$.

Remark 3

- Let $\hat{\theta}_{\mathrm{MLE}}$ be the solution to

$$\hat{S}(\theta) \equiv n^{-1} \sum_{i=1}^{n} S(\theta; y_i) = 0,$$

  where $S(\theta; y) = \partial \log f(y; \theta)/\partial\theta$.

- By Lemma 4.1, we can obtain

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\mathcal{I}(\theta_0)\}^{-1} S(\theta_0; y_i) + o_p(1),$$

  because

$$\mathcal{I}(\theta) = -E_\theta \left\{ \frac{\partial}{\partial\theta'} S(\theta; Y) \right\}.$$

Missing data setup

- Suppose that $\mathbf{y}_i$ is not fully observed.
- $\mathbf{y}_i = (\mathbf{y}_{\mathrm{obs,i}}, \mathbf{y}_{\mathrm{mis,i}})$: (observed, missing) part of $\mathbf{y}_i$
- $\boldsymbol{\delta}_i$: response indicator functions for $\mathbf{y}_i$.
- Under the existence of missing data, we can use the following estimators:

$$\hat{\psi}: \text{ solution to } \sum_{i=1}^{n} E\left\{U\left(\psi; \mathbf{y}_i\right) \mid \mathbf{y}_{obs,i}, \boldsymbol{\delta}_i\right\} = 0. \tag{7}$$

- The equation in (7) is often called the expected estimating equation.
- If $U(\psi; Y)$ is a score function of ψ, then (7) reduces to the mean score equation of R.A. Fisher.

Motivation (for imputation)

Computing the conditional expectation in (7) can be a challenging problem.

1. The conditional expectation depends on unknown parameter values. That is,

$$E\left\{U\left(\psi;\mathbf{y}_i\right)\mid \mathbf{y}_{obs,i},\boldsymbol{\delta}_i\right\}=E\left\{U\left(\psi;\mathbf{y}_i\right)\mid \mathbf{y}_{obs,i},\boldsymbol{\delta}_i;\theta,\phi\right\},$$

   where θ is the parameter in $f(\mathbf{y};\theta)$ and ϕ is the parameter in $p(\boldsymbol{\delta}\mid\mathbf{y};\phi)$.
2. Even if we know $\eta=(\theta,\phi)$, computing the conditional expectation is numerically difficult.

What is imputation?

- Imputation: Monte Carlo approximation of the conditional expectation (given the observed data).

$$E\left\{U\left(\psi; \mathbf{y}_i\right) \mid \mathbf{y}_{obs,i}, \boldsymbol{\delta}_i\right\} \cong \frac{1}{m}\sum_{j=1}^{m} U\left(\psi; \mathbf{y}_{obs,i}, \mathbf{y}_{mis,i}^{*(j)}\right)$$

1. Bayesian approach: generate $\mathbf{y}_{mis,i}^{*}$ from

$$f\left(\mathbf{y}_{mis,i} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}\right) = \int f\left(\mathbf{y}_{mis,i} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta\right) p(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) d\eta$$

2. Frequentist approach: generate $\mathbf{y}_{mis,i}^{*}$ from $f\left(\mathbf{y}_{mis,i} \mid \mathbf{y}_{obs,i}, \boldsymbol{\delta}; \hat{\eta}\right)$, where $\hat{\eta}$ is a consistent estimator.

2. Basic Theory (Frequentist approach)

- Parameter ψ defined by $E\{U(\psi; \mathbf{y})\} = 0$.
- Under complete response, a consistent estimator of ψ can be obtained by solving $\hat{U}(\psi) = 0$, where $\hat{U}(\psi) = n^{-1}\sum_{i=1}^{n} U(\psi; \mathbf{y}_i)$.
- Assume that some part of $\mathbf{y}$, denoted by $\mathbf{y}_{\text{mis}}$, is not observed and m imputed values, say $\mathbf{y}_{\text{mis}}^{*(1)}, \cdots, \mathbf{y}_{\text{mis}}^{*(m)}$, are generated from $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \hat{\eta}_{MLE})$, where $\hat{\eta}_{MLE}$ is the MLE of $\eta_0 = (\theta_0, \phi_0)$.
- The imputed estimating function using m imputed values is computed as

$$\bar{U}_{I,m}(\psi \mid \hat{\eta}_{MLE}) = \frac{1}{m}\sum_{j=1}^{m} \hat{U}(\psi; \mathbf{y}^{*(j)}), \tag{8}$$

  where $\mathbf{y}^{*(j)} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}^{*(j)})$.
- If $m \to \infty$, the imputed estimating function converges to the expected estimating function $\bar{U}_{I,\infty}(\psi \mid \hat{\eta}_{MLE}) = E\{\hat{U}(\psi) \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\delta}; \hat{\eta}_{MLE}\}$

- Let $\hat{\psi}_{I,m}$ be the solution to $\bar{U}_{I,m}(\psi \mid \hat{\eta}_{MLE}) = 0$. We are interested in the asymptotic properties of $\hat{\psi}_{I,m}$.
- Note that $\hat{\psi}_{I,m} = \hat{\psi}_{I,m}(\hat{\eta}_{MLE})$, which emphasize its dependence of $\hat{\eta}_{MLE}$, the MLE of η.
- Here, η is a nuisance parameter in the sense that we are not interested in estimating ψ, but we need an estimator of η in order to estimate ψ.
- Because of the sampling error of $\hat{\eta}_{MLE}$, the asymptotic distribution of $\hat{\psi}_{I,m}(\hat{\eta}_{MLE}) - \psi_0$ is different from that of $\hat{\psi}_{I,m}(\eta_0) - \psi_0$.
- Our goal is to find an inflence function for $\hat{\psi}_{I,m}$.

Example 4.2

- Under the setup of Example 4.1, we are interested in estimating the asymptotic variance of the regression imputation estimator

$$\hat{\theta}_{\mathrm{I,reg}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right),$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ is the solution to

$$\hat{U}(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\delta_i(y_i - \beta_0 - \beta_1 x_i)\begin{pmatrix}1\\ x_i\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}.$$

- How to find the influence function of $\hat{\theta}_{\mathrm{I,reg}}$?

Linearization method

- Write $\hat{\theta}_{\mathrm{I,reg}} = \hat{\theta}_I(\hat{\boldsymbol{\beta}})$ and consider Taylor expansion of $\hat{\theta}_I(\hat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^* = p \lim \hat{\boldsymbol{\beta}}$,

$$
\begin{aligned}
\hat{\theta}_{\mathrm{I,reg}}(\hat{\boldsymbol{\beta}}) &= \hat{\theta}_{\mathrm{I}}(\boldsymbol{\beta}^*) + E\left\{\nabla_\beta \hat{\theta}_I(\boldsymbol{\beta}^*)\right\}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + o_p(n^{-1/2}) \\
&= \hat{\theta}_{\mathrm{I}}(\boldsymbol{\beta}^*) + (1, E(X))\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + o_p(n^{-1/2}) \qquad (9)
\end{aligned}
$$

- Also, Taylor expansion of $\hat{U}(\hat{\boldsymbol{\beta}}) = 0$ around $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ to get

$$
\begin{aligned}
0 &= \hat{U}(\boldsymbol{\beta}^*) + E\left\{\nabla_\beta \hat{U}(\boldsymbol{\beta}^*)\right\}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + o_p(n^{-1/2}) \\
&= \hat{U}(\boldsymbol{\beta}^*) - \begin{pmatrix} E(\delta) & E(\delta X) \\ E(\delta X) & E(\delta X^2) \end{pmatrix}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + o_p(n^{-1/2}) \qquad (10)
\end{aligned}
$$

- Combining (9) with (10), we obtain

$$
\begin{aligned}
\hat{\theta}_{\mathrm{I,reg}} &= \hat{\theta}_{\mathrm{I}}(\boldsymbol{\beta}^*) + (\kappa_1, \kappa_2)\hat{U}(\boldsymbol{\beta}^*) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(\beta_0 + \beta_1 x_i) + \delta_i(y_i - \beta_0 - \beta_1 x_i)(\kappa_1 + \kappa_2 x_i)\} + o_p(n^{-1/2}),
\end{aligned}
$$

where

$$
\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} E(\delta) & E(\delta X) \\ E(\delta X) & E(\delta X^2) \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ E(X) \end{pmatrix}.
$$

- Therefore, the influence function of $\hat{\theta}_{\mathrm{I,reg}}$ is

$$
a(x_i, y_i, \delta_i) = (\beta_0 + \beta_1 x_i) + \delta_i(y_i - \beta_0 - \beta_1 x_i)(\kappa_1 + \kappa_2 x_i) - \theta.
$$

Remark

- If $\hat{\theta} = \hat{\theta}(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ is obtained by solving $\hat{U}(\boldsymbol{\beta}) = 0$, then the linearization takes the following form:

$$\hat{\theta}(\boldsymbol{\beta}, \mathbf{c}) = \hat{\theta}(\boldsymbol{\beta}) + \mathbf{c}'\hat{U}(\boldsymbol{\beta})$$

  For $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, we have $\hat{\theta}(\hat{\boldsymbol{\beta}}, \mathbf{c}) = \hat{\theta}$ regardless of the choice of $\mathbf{c}$.
- Thus, we have only to find $\mathbf{c} = \mathbf{c}^*$ such that no further Taylor expansion is necessary.
- We have only to solve

$$E\left\{\nabla_{\beta}\hat{\theta}(\boldsymbol{\beta}^*, \mathbf{c})\right\} = 0$$

  to get $\mathbf{c}^*$, where $\boldsymbol{\beta}^* = p\lim \hat{\boldsymbol{\beta}}$.
- Originally considered by Randles (1982).

REFERENCES

Hampel, F. R. (1974), 'The influence curve and its role in robust estimation', *Journal of the American Statistical Association* **69**, 383–393.

Randles, Ronald H (1982), 'On the asymptotic normality of statistics with estimated parameters', *The Annals of Statistics* pp. 462–474.