6.2 Nonparametric fractional imputation

Jae-Kwang Kim

Iowa State University

Basic Setup

- $\mathbf{x}_i$: auxiliary variable, completely observed
- y_i: study variable, subject to missingness.
- Assume MAR in the sense that $P(\delta = 1 \mid x, y)$ does not depend on y.
- Without loss of generality, assume that $\delta_i = 1$ for $i = 1, \cdots, r$ and $\delta_i = 0$ for $i = r + 1, \cdots, n$.
- May use a regression model (either parametric model or nonparametric model) to predict y_i using $\mathbf{x}_i$ for $\delta_i = 0$.

Table: Data structure for regression imputation

Sample Partition	ID	X	Y
Respondents	1	$\mathbf{x}_1$	y_1
	2	$\mathbf{x}_2$	y_2
		$\vdots$	
	r	$\mathbf{x}_r$	y_r
Nonrespondents	$r+1$	$\mathbf{x}_{r+1}$	y^*_{r+1}
	$r+2$	$\mathbf{x}_{r+2}$	y^*_{r+2}
		$\vdots$	$\vdots$
	n	$\mathbf{x}_n$	y^*_n

Regression imputation

- Regression model can be used to construct regression imputation for $\theta = E(Y)$:

$$\hat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\hat{y}_i\right\},$$

where

$$\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}, \tag{1}$$

and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{r}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{r}\mathbf{x}_i y_i.$$

Fractional Imputation interpretation of regression imputation

- Note that we can express the regression imputation (1) as

$$\hat{y}_i = \sum_{j=1}^{r} w_{ij}^* y_j \tag{2}$$

where

$$w_{ij}^* = \mathbf{x}_i' \left(\sum_{k=1}^{r} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_j \tag{3}$$

which takes the form of fractional imputation (FI), where w_{ij}^* is the fractional weight assigned to donor $j \in \{1, \cdots, r\}$ to $\hat{y}_i$ in (2).

- The fractional weight w_{ij}^* satisfies

$$\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_j = \mathbf{x}_i. \tag{4}$$

- Thus, if $y_i = \mathbf{x}_i' \boldsymbol{\beta}$ for some $\boldsymbol{\beta}$, then the prediction for y_i using (2) is accurate.

Remark 1

- Under the regression model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i$$

where $e_i \mid \mathbf{x}_i \sim (0, \sigma^2)$.

- The MSPE of $\hat{y}_i = \sum_{j=1}^r w_{ij}^* y_j$ is

$$E\left\{(\hat{y}_i - y_i)^2\right\} = \left\{ \left(\sum_{j=1}^r w_{ij}^* \mathbf{x}_j - \mathbf{x}_i \right)' \boldsymbol{\beta} \right\}^2 + \sum_{j=1}^r \left(w_{ij}^*\right)^2 \sigma^2 + \sigma^2$$

- The regression fractional weights in (3) is obtained by minimizing $\sum_{j=1}^r \left(w_{ij}^*\right)^2$ subject to $\sum_{j=1}^r w_{ij}^* \mathbf{x}_j = \mathbf{x}_i$.

Justification

- Method 1: Use Lagrange multiplier method (check).
- Method 2: Use the GLS method by writing $\mathbf{w}_i = (w_{i1}^*, \ldots, w_{ir}^*)'$ as a regression parameter and X_r as a $r \times p$ design matrix in the GLS. That is, minimize

$$Q(\mathbf{w}_i) = \begin{pmatrix} \mathbf{0} - I_r \mathbf{w}_i \\ \mathbf{x}_i - X_r' \mathbf{w}_i \end{pmatrix}^T \begin{pmatrix} \sigma^2 I_r & 0 \\ 0 & I_p \end{pmatrix} \begin{pmatrix} \mathbf{0} - I_r \mathbf{w}_i \\ \mathbf{x}_i - X_r' \mathbf{w}_i \end{pmatrix}$$

  wrt $\mathbf{w}_i$ and then let $\sigma^2 \to 0$.
- The solution to GLS is

$$\begin{aligned} \hat{\mathbf{w}}_i &= \left(\sigma^2 I_r + X_r X_r'\right)^{-1} \left(\sigma^2 I_r \cdot \mathbf{0} + X_r I_p \mathbf{x}_i\right) \\ &= X_r \left(X_r' X_r + \sigma^2 I_p\right)^{-1} \mathbf{x}_i \end{aligned}$$

Example

- For $\mathbf{x}_i = (1, x_i)'$, the fractional weights in (3) can be written as

$$w_{ij}^* = \frac{1}{r} + \frac{(x_i - \bar{x}_r)(x_j - \bar{x}_r)}{\sum_{k=1}^r (x_k - \bar{x}_r)^2},$$

where $\bar{x}_r = r^{-1} \sum_{j=1}^r x_j$.

- Toy example: A table of w_{ij}^* with $r = 5$

	x_j					
x_i	1	2	3	4	5	$\hat{x}_i = \sum_{j=1}^{5} w_{ij}^* x_j$
3.0	0.200	0.200	0.200	0.200	0.200	3.0
4.5	-0.100	0.050	0.200	0.035	0.500	4.5
6.0	-0.400	-0.100	0.200	0.500	0.800	6.0

Remark 2

- For each $i = r+1, \cdots, n$, w_{ij}^* takes the maximum at $j = j^*$ if $\mathbf{x}_i \cong \mathbf{x}_{j^*}$.
- Thus, we can treat w_{ij}^* as a kernel function constructed from $\mathbf{x}_i$:

$$w_{ij}^* = K(\mathbf{x}_i, \mathbf{x}_j)$$

- Property (4) is essentially the reproducing property of the kernel function.
- Imposing the reproducing property for each $i = r+1, \cdots, n$ can lead to negative fractional weights, which is not desirable.

Ridge regression approach

- Let $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$, where the intercept term is included in $\mathbf{x}_1$.
- For given λ, find the minimizer of

$$Q_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{r} \left(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i}'\boldsymbol{\beta}_2\right)^2 + \lambda \left\|\boldsymbol{\beta}_2\right\|^2 \tag{5}$$

where $\left\|\boldsymbol{\beta}_2\right\|^2 = \boldsymbol{\beta}_2'\boldsymbol{\beta}_2$ and λ is the tuning parameter.

- We can use $\hat{y}_i = \mathbf{x}_{1i}'\hat{\beta}_1 + \mathbf{x}_{2i}'\hat{\boldsymbol{\beta}}_2$ as the imputed value for y_i, where $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ is the solution to the mixed model equation:

$$\begin{pmatrix} \sum_{i=1}^r \mathbf{x}_{1i}\mathbf{x}_{1i}' & \sum_{i=1}^r \mathbf{x}_{1i}\mathbf{x}_{2i}' \\ \sum_{i=1}^r \mathbf{x}_{2i}\mathbf{x}_{1i}' & \sum_{i=1}^r \mathbf{x}_{2i}\mathbf{x}_{2i}' + \lambda I_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^r \mathbf{x}_{1i} y_i \\ \sum_{i=1}^r \mathbf{x}_{2i} y_i \end{pmatrix}$$

- We can express

$$\hat{y}_i = \sum_{j=1}^{r} w_{ij}^* y_j \tag{6}$$

where

$$w_{ij}^* = \left(\mathbf{x}_{1i}', \mathbf{x}_{2i}'\right) \begin{pmatrix} \sum_{i=1}^r \mathbf{x}_{1i}\mathbf{x}_{1i}' & \sum_{i=1}^r \mathbf{x}_{1i}\mathbf{x}_{2i}' \\ \sum_{i=1}^r \mathbf{x}_{2i}\mathbf{x}_{1i}' & \sum_{i=1}^r \mathbf{x}_{2i}\mathbf{x}_{2i}' + \lambda I_q \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_{1j} \\ \mathbf{x}_{2j} \end{pmatrix} \tag{7}$$

- The reproducing property holds for $\mathbf{x}_1$, but not for $\mathbf{x}_2$. That is,

$$\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_{1j} = \mathbf{x}_{1i}$$

but

$$\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_{2j} \cong \mathbf{x}_{2i}$$

Remark

- The ridge regression can be justified under linear mixed model

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + e_i$$

where $\boldsymbol{\beta}_2 \sim (0, \lambda\sigma^2 I_q)$ and $e_i \sim (0, \sigma^2)$.

- The MSPE of $\hat{y}_i = \sum_{j=1}^{r} w_{ij}^* y_j$ is

$$\begin{aligned} E\left\{(\hat{y}_i - y_i)^2\right\} &= \left\{(\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_{1j} - \mathbf{x}_{1i})'\boldsymbol{\beta}_1\right\}^2 \\ &+ \left\{\sum_{j=1}^{r}\left(w_{ij}^*\right)^2 + \lambda(\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_{2j} - \mathbf{x}_{2i})^{\otimes 2}\right\}\sigma^2 + \sigma^2 \end{aligned}$$

- The fractional weights in (7) is obtained by minimizing $\sum_{j=1}^{r}(w_{ij}^*)^2 + \lambda(\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_{2j} - \mathbf{x}_{2i})^{\otimes 2}$ subject to $\sum_{j=1}^{r} w_{ij}^* \mathbf{x}_{1j} = \mathbf{x}_{1i}$.

Justification

- Let $\mathbf{w}_i = (w_{i1}^*, \cdots, w_{ir})'$.
- The optimization problem can be formulated as minimizing

$$Q(\mathbf{w}_i^*) = \begin{pmatrix} \mathbf{0} - I_r\mathbf{w}_i \\ \mathbf{x}_{1i} - X_{1r}'\mathbf{w}_i \\ \mathbf{x}_{2i} - X_{2r}'\mathbf{w}_i \end{pmatrix}^T \begin{pmatrix} \sigma^2 I_r & 0 & 0 \\ 0 & I_p & 0 \\ 0 & 0 & \sigma^2\lambda I_q \end{pmatrix} \begin{pmatrix} \mathbf{0} - I_r\mathbf{w}_i \\ \mathbf{x}_{1i} - X_{1r}'\mathbf{w}_i \\ \mathbf{x}_{2i} - X_{2r}'\mathbf{w}_i \end{pmatrix}$$

wrt $\mathbf{w}_i$ and then let $\sigma^2 \to 0$.

Penalized regression imputation

- More generally, the penalized regression imputation estimator can be expressed as

$$\hat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\hat{\boldsymbol{\beta}}_\lambda$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the minimizer of

$$Q_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{r}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 + p_\lambda(\boldsymbol{\beta}).$$

- Note that we have a bias-correction term in $\boldsymbol{\beta}$ such that $\sum_{i=1}^{r}(y_i - \hat{y}_i) = 0$.
- We can still express $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\lambda = \sum_{j=1}^{r} w_{ij}^* y_j$ where

$$w_{ij}^* = \mathbf{x}_i'\left(\sum_{i=1}^{r}\mathbf{x}_i\mathbf{x}_i' + \Omega(\hat{\boldsymbol{\beta}}_\lambda;\lambda)\right)^{-1}\mathbf{x}_j.$$

- Let $\boldsymbol{\beta}^*$ be the probability limit of $\hat{\boldsymbol{\beta}}$.
- We can express

$$
\begin{aligned}
\hat{\theta}_I &= \bar{\mathbf{x}}_n'\boldsymbol{\beta}^* + \bar{\mathbf{x}}_n'\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\right) \\
&= \bar{\mathbf{x}}_n'\boldsymbol{\beta}^* + \bar{\mathbf{x}}_n'\left(\sum_{i=1}^{r}\mathbf{x}_i\mathbf{x}_i' + \Omega(\hat{\boldsymbol{\beta}}_\lambda;\lambda)\right)^{-1}\sum_{i=1}^{r}\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}^*\right) \\
&= \bar{\mathbf{x}}_n'\boldsymbol{\beta}^* + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r} w_{ij}^*\left(y_j - \mathbf{x}_j'\boldsymbol{\beta}^*\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{x}_i'\boldsymbol{\beta}^* + \delta_i\omega_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}^*\right)\right\}
\end{aligned}
$$

where $\omega_j = \sum_{i=1}^{n} w_{ij}^*$.

- Under some conditions, the penalized regression estimator is asymptotically equivalent to

$$\hat{\theta}_\ell = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{y}_i + \delta_i\omega_i(y_i - \hat{y}_i)\right\}.$$

- The linearization variance estimator is obtained by

$$\hat{V} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{d}_i - \bar{d}_n\right)^2,$$

  where $\hat{d}_i = \hat{y}_i + \delta_i\omega_i(y_i - \hat{y}_i)$.

Nonparametric regression imputation using kernel function

- Let $K_h(x_i, x_j) = K((x_i - x_j)/h)$ be the Kernel function with bandwidth h such that $K(x) \geq 0$ and

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \sigma_K^2 \equiv \int x^2 K(x)dx > 0.$$

- Examples include the following:
  - Boxcar kernel: $K(x) = \frac{1}{2} I(x)$
  - Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$
  - Epanechnikov kernel: $K(x) = \frac{3}{4}(1 - x^2) I(x)$
  - Tricube Kernel: $K(x) = \frac{70}{81}(1 - |x|^3)^3 I(x)$

where

$$I(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1. \end{cases}$$

- Define
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)$$
to be the Kernel-based estimator of the marginal density of X, where $K_h(x) = h^{-1}K(x/h)$, h is the bandwidth, and $K(\cdot)$ is the Kernel function. For simplicity, assume $\dim(x) = 1$.
- Note that
$$\int \hat{f}(x)dx = 1.$$

- It is well known that

$$E\{\hat{f}(x)\} = f(x) + O(h^2)$$

and

$$V\{\hat{f}(x)\} = O((nh)^{-1}),$$

for each x, where $f(x)$ is the true density function. Thus,

$$MSE\{\hat{f}(x)\} = O(h^4 + (nh)^{-1}).$$

- The optimal choice of the bandwidth is $h^* = c(x)n^{-1/5}$ and the MSE is $O(n^{-4/5})$.

Nonparametric regression

- Nonparametric regression estimator of $m(x) = E(Y \mid x)$:

$$\hat{m}(x) = \sum_{i=1}^{r} l_i(x) y_i \tag{8}$$

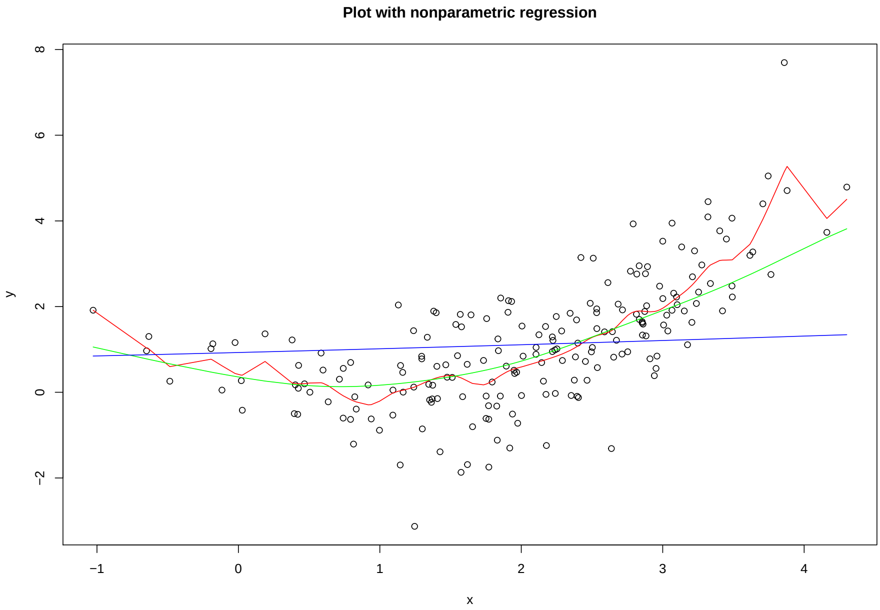
where

$$l_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_j K\left(\frac{x-x_j}{h}\right)}.$$

Estimator in (8) is often called Nadaraya-Watson kernel estimator.

Listing 1: R-code for nonparametric regression

```r
library(np)

x <- rnorm(200, 2,1)
e <- rnorm(200,0,1)
y <- 0.5*(x-1)^2 + e

plot(x,y)
title(main = "Plot with nonparametric regression")

pred1 <- npreg(y~x, bws = 0.1)
pred2 <- npreg(y~x, bws = 0.5)
pred3 <- npreg(y~x, bws = 3.0)

xgrid1 <- pred1$eval[[1]]
xorder1 <- order(xgrid1)
yval1 <- pred1$mean

xgrid2 <- pred2$eval[[1]]
xorder2 <- order(xgrid2)
yval2 <- pred2$mean

xgrid3 <- pred3$eval[[1]]
xorder3 <- order(xgrid3)
yval3 <- pred3$mean


lines(xgrid1[xorder1], yval1[xorder1], col="red")
lines(xgrid2[xorder2], yval2[xorder2], col="green")
lines(xgrid3[xorder3], yval3[xorder3], col="blue")
```


Plot with nonparametric regression
y
x
8
6
4
2
0
−2
−1
0
1
2
3
4

Bandwidth selection

- Use Leave-one-out cross validation

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{m}_h^{(-i)}(x_i)\right\}^2$$

where $\hat{m}_h^{(-i)}(x)$ is the nonparametric regression estimator by omitting the i-th pair (x_i, y_i).

- Generalized cross validation:

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{m}_h(x_i)}{1 - L_{ii}}\right\}^2$$

where $L_{ii} = l_i(x_i)$ and $l_i(x)$ is defined in (8).

Theorem 6.2 (Cheng, 1994)

Theorem

Under some regularity conditions, the imputed estimator of θ using (8) can achieve the $\sqrt{n}$-consistency. That is,

$$\hat{\theta}_{NP} = \frac{1}{n}\left\{\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{m}(x_i)\right\} \tag{9}$$

achieves

$$\sqrt{n}\left(\hat{\theta}_{NP} - \theta\right) \longrightarrow N(0, \sigma^2) \tag{10}$$

where $\sigma^2 = E\{v(x)/\pi(x)\} + V\{m(x)\}$, $m(x) = E(y \mid x)$, $v(x) = V(y \mid x)$ *and* $\pi(x) = E(\delta \mid x)$.

Remark

- Theorem 6.2 essentially states that $\hat{\theta}_{NP}$ is asymptotically equivalent to $\tilde{\theta}_{NP} = n^{-1}\sum_{i=1}^{n} d(x_i, y_i, \delta_i)$ with influence function

$$d(x_i, y_i, \delta_i) = m(x_i) + \delta_i \frac{1}{\pi(x_i)} \{y_i - m(x_i)\}. \tag{11}$$

  The variance of $\tilde{\theta}_{NP}$ is equal to $n^{-1}\sigma^2$, where σ^2 is defined after (10).

- We can express $\hat{\theta}_{NP}$ in (9) as a nonparametric fractional imputation (NFI) estimator of the form

$$\hat{\theta}_{NFI} = \frac{1}{n}\left\{\sum_{i=1}^{r} y_i + \sum_{j=r+1}^{n}\sum_{i=1}^{r} w_{ij}^{*} y_i^{*(j)}\right\}$$

  where $w_{ij}^{*} = l_i(x_j)$, which is defined after (8), and $y_i^{*(j)} = y_i$.

Variance estimation

- For variance estimation of the NFI estimator $\hat{\theta}_{NFI}$, we need to estimate the influence function in (11).
- We can use

$$\hat{\omega}(x) = \sum_{j=1}^{n} \left\{ \frac{K_h(x_j, x)}{\sum_{k=1}^{n} \delta_k K_h(x_j, x_k)} \right\}$$

as an estimator of $1/\pi(x)$. Note that $\hat{\omega}(x)$ satisfies

$$\sum_{i=1}^{n} \delta_i \hat{\omega}(x_i) y_i = \sum_{i=1}^{n} \hat{m}(x_i).$$

- Thus, we can use

$$\hat{d}_i = \hat{m}(x_i) + \delta_i \hat{\omega}(x_i) \{y_i - \hat{m}(x_i)\}$$

Using the above $\hat{d}_i$, we can apply the standard variance estimation formula to estimate the asymptotic variance of the NFI estimator.