Nonparametric regression approach to fractional imputation and propensity score estimation

Jae-Kwang Kim

Iowa State University

Basic Setup

- Bivariate data: (x_i, y_i)
- x_i are completely observed but y_i is subject to missingness.
- Joint distribution of (x, y) completely unspecified.
- 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 $i = r + 1, \cdots, n$.
- We are only interested in estimating $\theta = E(Y)$.

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 $O(n^{-4/5})$.

Lemma

Under some regularity conditions,

$$n^{-1} \sum_{i=1}^{n} K_h(x_i - x) y_i = f(x) \cdot E(Y \mid x) + O_p(h^2 + (nh)^{-1/2}). \tag{1}$$

Li and Racine (2007) presents a rigorous proof.

Nonparametric regression

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

$$\hat{m}(x) = \sum_{i=1}^{r} l_i(x) y_i \tag{2}$$

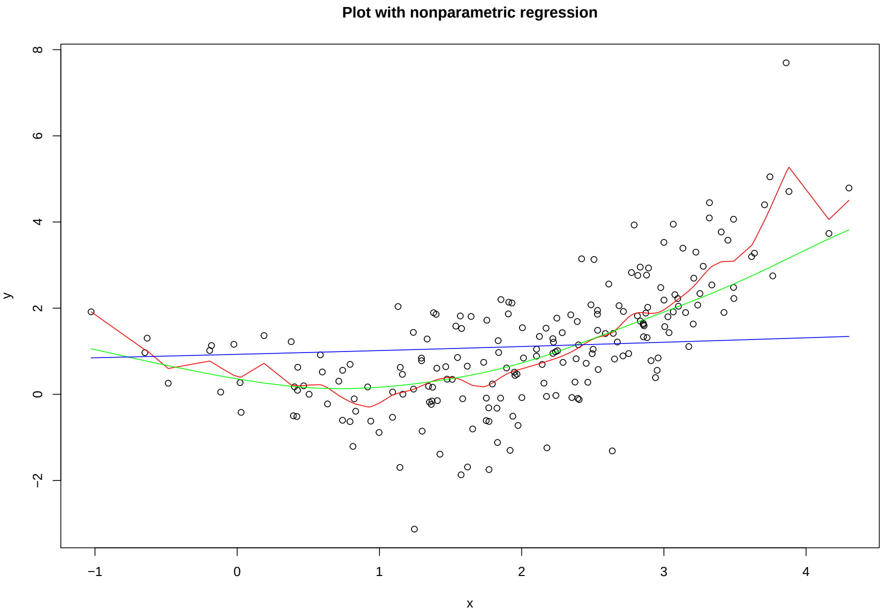
where

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

Estimator in (2) is often called Nadaraya-Watson kernel estimator.

Listing 1: R-code for nonparametric regression

```
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
8
6
4
2
0
−2
x
−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 (2).

Theorem 6.2 (Cheng, 1994)

Theorem

Under some regularity conditions, the imputed estimator of θ using (2) can achieve the $\sqrt{n}$-consistency. That is,

$$\hat{\theta}_{NPI} = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\hat{m}(x_i)\right\} \tag{3}$$

achieves

$$\sqrt{n}\left(\hat{\theta}_{NPI} - \theta\right) \longrightarrow N(0, \sigma^2) \tag{4}$$

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

$$d(x_i, y_i, \delta_i) = m(x_i) + \delta_i \frac{1}{\pi(x_i)} \{y_i - m(x_i)\}. \quad (5)$$

  The variance of $\tilde{\theta}_{NPI}$ is equal to $n^{-1}\sigma^2$, where σ^2 is defined after (4).
- We can express $\hat{\theta}_{NP}$ in (3) 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 (2), and $y_i^{*(j)} = y_i$.

Sketched Proof (by Wang and Chen (2009))

- The nonparametric regression imputation estimator under MAR can be written as

$$\begin{aligned}\hat{\theta}_{NRI} &= \frac{1}{n}\sum_{i=1}^{n}\{\delta_i y_i + (1-\delta_i)\hat{m}(x_i)\} \\ &:= S_n + T_n + R_n,\end{aligned}$$

where

$$\begin{aligned}S_n &= n^{-1}\sum_{i=1}^{n}\delta_i\{y_i - m(x_i)\} \\ T_n &= n^{-1}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}(x_i) - m(x_i)\} \\ R_n &= n^{-1}\sum_{i=1}^{n} m(x_i)\end{aligned}$$

- To prove (4), we have only to show that

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left\{ \frac{1}{\pi(x_i)} - 1 \right\} (y_i - m(x_i)) + o_p(n^{-1/2})$$

- Also, we can express

$$\hat{m}(x) = \frac{C_n(x)}{D_n(x)}$$

where

$$C_n(x) = n^{-1}\sum_{j=1}^{n} \delta_j K_h(x_j - x) y_j$$

$$D_n(x) = n^{-1}\sum_{j=1}^{n} \delta_j K_h(x_j - x).$$

- Using (1), we can obtain

$$\begin{aligned} C_n(x) &= E(\delta \mid x)E(Y \mid x)f(x) + O_p(a_n) \\ D_n(x) &= E(\delta \mid x)f(x) + O_p(a_n) \end{aligned}$$

  where $a_n = h^2 + (nh)^{-1/2}$
- Now, by Taylor expansion,

$$\frac{C_n(x)}{D_n(x)} - m(x) = -\frac{1}{E(\delta \mid x)f(x)} \{C_n(x) - m(x)D_n(x)\} + O_p(a_n).$$

- Thus, writing $\pi(x) = E(\delta \mid x)$,

$$\begin{aligned} T_n &= \frac{1}{n^2}\sum_{i=1}^{n}(1-\delta_i)\frac{1}{\pi(x_i)f(x_i)}\sum_{j=1}^{n}\delta_j K_h(x_j - x_i)\{y_j - m(x_i)\} \\ &\quad + O_p(h^2 + n^{-1}h^{-1/2}). \end{aligned}$$

- If h satisfies $nh \to \infty$ and $nh^4 \to 0$, then we have

$$h^2 + n^{-1}h^{-1/2} = o(n^{-1/2}).$$

Thus,

$$T_n = \frac{1}{n^2}\sum_{i=1}^{n}(1-\delta_i)\frac{1}{\pi(x_i)f(x_i)}\sum_{j=1}^{n}\delta_j K_h(x_j - x_i)\{y_j - m(x_i)\} + o_p(n^{-1/2}).$$

- Ignoring the smaller order terms, we can express

$$T_n = \frac{1}{n(n-1)}\sum_i \sum_{j\neq i} h(z_i, z_j)$$

where $z_i = (x_i, \delta_i, y_i)$ and

$$\begin{aligned} h(z_i, z_j) &= \frac{1}{2}\Bigg[(1-\delta_i)\delta_j \frac{1}{\pi(x_i)f(x_i)}K_h(x_i, x_j)\{y_j - m(x_i)\} \\ &\quad + (1-\delta_j)\delta_i \frac{1}{\pi(x_j)f(x_j)}K_h(x_j, x_i)\{y_i - m(x_j)\}\Bigg] \\ &:= \frac{1}{2}(\zeta_{ij} + \zeta_{ji}). \end{aligned}$$

- Now, it can be shown that

$$\begin{aligned} E(\zeta_{ij} \mid z_i) &= (1-\delta_i)\frac{1}{\pi(x_i)f(x_i)}\frac{1}{h}\int K\left(\frac{x_i - x_j}{h}\right)\pi(x_j)\left\{m(x_j) - m(x_i)\right\} f(x_j \\ &= O(h^2) \end{aligned}$$

and

$$\begin{aligned} E(\zeta_{ji} \mid z_i) &= \delta_i \frac{1}{h}\int \frac{1-\pi(x_j)}{\pi(x_j)f(x_j)} K\left(\frac{x_j - x_i}{h}\right)\{y_i - m(x_j)\} f(x_j) dx_j \\ &= \delta_i \left\{\frac{1}{\pi(x_i)} - 1\right\}\{y_i - m(x_i)\} + O(h^2). \end{aligned}$$

- Thus, using the theory of U-statistics (van der Vaart, 1998; Ch. 12),

$$T_n = \frac{2}{n}\sum_{i=1}^{n} E\{h(z_i, z_j) \mid z_i\} + o_p(n^{-1/2})$$

we can obtain

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left\{\frac{1}{\pi(x_i)} - 1\right\}\{y_i - m(x_i)\} + o_p(n^{-1/2})$$

Motivation

- So far, we have assumed a parametric model for $\pi(x) = Pr(\delta = 1 \mid x)$.
- Using the nonparametric regression technique, we can use a nonparametric estimator of $\pi(x)$ given by a nonparametric regression estimator of $\pi(x) = E(\delta \mid x)$ can be obtained by

$$\hat{\pi}_h(x) = \frac{\sum_{i=1}^n \delta_i K_h(x_i, x)}{\sum_{i=1}^n K_h(x_i, x)}, \tag{6}$$

  where K_h is the kernel function which satisfies certain regularity conditions and h is the bandwidth.
- Once a nonparametric estimator of $\pi(x)$ is obtained, the nonparametric PS estimator $\hat{\theta}_{NPS}$ of $\theta_0 = E(Y)$ is given by

$$\hat{\theta}_{NPS} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_h(x_i)} y_i. \tag{7}$$

Theorem 7.5

Under some regularity conditions, we have

$$\hat{\theta}_{NPS} = \frac{1}{n}\sum_{i=1}^{n}\left[m(x_i) + \frac{\delta_i}{\pi(x_i)}\{y_i - m(x_i)\}\right] + o_p(n^{-1/2}), \tag{8}$$

where $m(x) = E(Y \mid x)$ and $\pi(x) = P(\delta = 1 \mid x)$. Furthermore, we have

$$\sqrt{n}\left(\hat{\theta}_{NPS} - \theta\right) \to N\left(0, \sigma_1^2\right),$$

where $\sigma_1^2 = V\{m(X)\} + E\left[\{\pi(X)\}^{-1} V(Y \mid X)\right]$.

Originally proved by Hirano et al. (2003).

Remark

- Unlike the usual asymptotic for nonparametric regression, $\sqrt{n}$-consistency is established.
- The nonparametric PS estimator achieves the lower bound of the asymptotic variance (Robins et al., 1994), which was discussed in Theorem 5.1.
- The asymptotic variance of nonparameteric PS estimator is equal to the that of the nonparametric fractional imputation estimator of Cheng (1994), presented in Theorem 6.2.

REFERENCES

Cheng, P. E. (1994), 'Nonparametric estimation of mean functionals with data missing at random', *Journal of the American Statistical Association* **89**, 81–87.

Hirano, K., G. Imbens and G. Ridder (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**, 1161–1189.

Li, Q. and J. S. Racine (2007), *Nonparametric Econometrics*, Princeton University Press.

Robins, James M, Andrea Rotnitzky and Lue Ping Zhao (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American statistical Association* **89**(427), 846–866.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Wang, D. and S. X. Chen (2009), 'Empirical likelihood for estimating equations with missing values', *The Annals of Statistics* **37**, 490–517.