# Variance estimation after Kernel Ridge Regression Imputation

Hengfang Wang & Jae Kwang Kim

Center for Survey Statistics and Methodology
Iowa State University

07/17/2020

# Introduction

## Data Structure

- Imputation is a popular technique for handling item nonresponse.
- Suppose we have the dataset $\{(\delta_i, \boldsymbol{x}_i, \delta_i y_i), i = 1, \ldots, n\}$.
- $\boldsymbol{x}_i \in \mathbb{R}^d$ : is fully observed covariate.
- $y_i \in \mathbb{R}$ : response subject to missingness.
- $\delta_i$: response indicator for $y_i$, $\delta_i = 1$ if $y_i$ is observed, for $i = 1, \ldots, n$.

## Target Estimator

Under MAR (missing at random), we can develop a nonparametric estimator $\widehat{m}(\boldsymbol{x})$ of $m(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{x})$ and construct the following imputation estimator:

$$\widehat{\theta}_I = \frac{1}{n} \sum_{i=1}^{n} \{\delta_i y_i + (1 - \delta_i)\widehat{m}(\boldsymbol{x}_i)\}. \tag{1}$$

## KRR Imputation

- We use kernel ridge regression (KRR) to get the corresponding $\hat{m}(\cdot)$ where

$$\widehat{m} = \arg\min_{m \in \mathcal{H}} \left[ \sum_{i=1}^{n} \delta_i \{y_i - m(\boldsymbol{x}_i)\}^2 + \lambda \|m\|_{\mathcal{H}}^2 \right], \tag{2}$$

where $\lambda > 0$ is tuning parameter and $\mathcal{H}$ is a Hilbert space.

# Main Results

**Theorem 1**

*Under regularity conditions, for a Sobolev kernel of order $\ell$, $\lambda \asymp n^{1-\ell}$, we have*

$$\sqrt{n}(\hat{\theta}_I - \tilde{\theta}_I) = o_p(1), \tag{3}$$

*where*

$$\tilde{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\left[m(\boldsymbol{x}_i) + \delta_i \frac{1}{\pi(\boldsymbol{x}_i)}\{y_i - m(\boldsymbol{x}_i)\}\right]. \tag{4}$$

*Furthermore, as $n \to \infty$,*

$$\sqrt{n}\left(\tilde{\theta}_I - \theta\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

*where*

$$\sigma^2 = V\{E(Y \mid \boldsymbol{x})\} + E\{V(Y \mid \boldsymbol{x})/\pi(\boldsymbol{x})\}$$

*and $\pi(\boldsymbol{x}) = E(\delta \mid \boldsymbol{x})$.*

# Main Results

## Methodology

By Theorem 1, we use the following estimator to estimate the variance of $\hat{\theta}_I$ in (1):

$$\hat{\mathrm{V}}(\hat{\theta}_I) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\eta}_i - \bar{\eta})^2 \quad (5)$$

where

$$\hat{\eta}_i = \widehat{m}(\boldsymbol{x}_i) + \delta_i \hat{\omega}_i \{y_i - \widehat{m}(\boldsymbol{x}_i)\},$$

and $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \ldots, \hat{\omega}_n)^{\mathrm{T}} \in \mathbb{R}^n$ is estimated by

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega} \geqslant 1}{\arg\min} \left[ \max_{u \in \widetilde{\mathcal{H}}_n} \left\{ S_n(\boldsymbol{\omega}, u) - \lambda \|u\|_{\mathcal{H}}^2 \right\} + \tau V_n(\boldsymbol{\omega}) \right]. \quad (6)$$

where $S_n(\boldsymbol{\omega}, u) = \left\{ \frac{1}{n} \sum_{i=1}^n (\delta_i \omega_i - 1) u(\boldsymbol{x}_i) \right\}^2$, $V_n(\boldsymbol{\omega}) = \frac{1}{n} \sum_{i=1}^n \delta_i \omega_i^2$, $\widetilde{\mathcal{H}}_n = \{u \in \mathcal{H} : \|u\|_n = 1\}$ and $\|u\|_n^2 = \frac{1}{n} \sum_{i=1}^n u(\boldsymbol{x}_i)^2$. The weight $\hat{w}_i$ estimates $\{\pi(\boldsymbol{x}_i)\}^{-1}$ nonparametrically.