Debiased two-phase regression estimation

Caleb Leedy Jae Kwang Kim

May 7, 2024

1 Introduction

- Introduce two-phase sampling
- Introduce classical two-phase regression estimation under parametric regression model

  $$\hat{Y}_{\text{reg}} = \sum_{i \in A_1} w_{1i} m(\mathbf{x}_i; \hat{\beta}) + \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} \left(y_i - m(\mathbf{x}_i; \hat{\beta}) \right)$$

  Properties: If $\dim(\beta) = p$ is fixed in the asymptotic sense, then $\hat{Y}_{\text{reg}}$ is asymptotically equivalent to

  $$\hat{Y}_{\text{reg},\ell} = \sum_{i \in A_1} w_{1i} m(\mathbf{x}_i; \beta^*) + \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} \left(y_i - m(\mathbf{x}_i; \beta^*) \right).$$

  Note that $\hat{Y}_{\text{reg},\ell}$ is design unbiased and achieves the optimality if $E(Y \mid \mathbf{x}_i) = m(\mathbf{x}_i; \beta_0)$ for some β_0 and $\hat{\beta}$ is consistent for β_0.

- Two-phase regression estimation under non-parametric regression

  $$\hat{Y}_{\text{np,reg}} = \sum_{i \in A_1} w_{1i} \hat{m}(\mathbf{x}_i) + \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} \left(y_i - \hat{m}(\mathbf{x}_i) \right)$$

  where $\hat{m}(\mathbf{x})$ is a consistent estimator of $m(\mathbf{x}) = E(Y \mid \mathbf{x})$, but the convergence rate is slow.

- Generally speaking, $\hat{Y}_{\text{rep,np}}$ is not asymptotically equivalent to the oracle regression estimator given by

  $$\hat{Y}_{\text{oracle}} = \sum_{i \in A_1} w_{1i} m(\mathbf{x}_i) + \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} \left(y_i - m(\mathbf{x}_i) \right). \tag{1}$$

Note that

$$
\begin{aligned}
\hat{Y}_{\text{np,reg}} - \hat{Y}_{\text{oracle}} &= \sum_{i \in A_1} w_{1i} \left\{ \hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} - \sum_{i \in A_2} w_{1i} \pi_{2i|1}^{-1} \left\{ \hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} \\
&= \sum_{i \in A_1} w_{1i} \left\{ \hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} \left(1 - \delta_i \pi_{2i|1}^{-1} \right) \\
&:= D_n .
\end{aligned}
$$

In general, the estimation error term D_n is not asymptotically negligible.

2 Proposed estimator

One possible idea is to develop a debiased estimation under two-phase sampling. If the covariates are high dimensional or the regression employs nonparametric regression (such as spline or random forest, etc), then the resulting two-phase regression estimator can be biased. To correct for the bias, we can use the following approach.

1. Split the sample A_2 into two parts: $A_2 = A_2^{(a)} \cup A_2^{(b)}$. We can use SRS to split the sample, but other sampling designs can be used.

2. Use the observations in $A_2^{(a)}$ only to obtain a predictor of y_i, $\hat{m}^{(a)}(\mathbf{x}_i)$. Also, use the observations in $A_2^{(b)}$ only to obtain a predictor of y_i, $\hat{m}^{(b)}(\mathbf{x}_i)$.

3. Let
$$
\hat{m}(\mathbf{x}_i) = \left(\hat{m}^{(a)}(\mathbf{x}_i) + \hat{m}^{(b)}(\mathbf{x}_i) \right) / 2
$$
be the predictor combining two samples.

4. The final debiased two-phase regression estimator is given by
$$
\hat{Y}_{\text{d,reg}} = \sum_{i \in A_1} w_{1i} \hat{m}(\mathbf{x}_i) + \sum_{i \in A_2^{(a)}} w_{1i} \pi_{2i|1}^{-1} \left\{ y_i - \hat{m}^{(b)}(\mathbf{x}_i) \right\} + \sum_{i \in A_2^{(b)}} w_{1i} \pi_{2i|1}^{-1} \left\{ y_i - \hat{m}^{(a)}(\mathbf{x}_i) \right\} \tag{2}
$$

Note that $\hat{Y}_{\text{d,reg}}$ can be expresssed as

$$
\hat{Y}_{\text{d,reg}} = \left(\hat{Y}_{\text{d,reg}}^{(a)} + \hat{Y}_{\text{d,reg}}^{(b)} \right) / 2
$$

where

$$\hat{Y}_{\text{d,reg}}^{(a)} = \sum_{i \in A_1} w_{1i} \hat{m}^{(b)}(\mathbf{x}_i) + \sum_{i \in A_2^{(a)}} w_{1i} 2\pi_{2i|1}^{-1} \left(y_i - \hat{m}^{(b)}(\mathbf{x}_i) \right)$$

and

$$\hat{Y}_{\text{d,reg}}^{(b)} = \sum_{i \in A_1} w_{1i} \hat{m}^{(a)}(\mathbf{x}_i) + \sum_{i \in A_2^{(b)}} w_{1i} 2\pi_{2i|1}^{-1} \left(y_i - \hat{m}^{(a)}(\mathbf{x}_i) \right).$$

Now, we can establish the following result.

Theorem 1. *Under some regularity conditions (to be explained later), we obtain*

$$N^{-1} \left(\hat{Y}_{\text{d,reg}} - \hat{Y}_{\text{oracle}} \right) = o_p(n^{-1/2}), \tag{3}$$

where $\hat{Y}_{\text{oracle}}$ is defined in (1).

Proof. Define

$$\hat{Y}_{\text{oracle}}^{(a)} = \sum_{i \in A_1} w_{1i} m(\mathbf{x}_i) + \sum_{i \in A_2^{(a)}} w_{1i} 2\pi_{2i|1}^{-1} \left(y_i - m(\mathbf{x}_i) \right)$$

and

$$\hat{Y}_{\text{oracle}}^{(b)} = \sum_{i \in A_1} w_{1i} m(\mathbf{x}_i) + \sum_{i \in A_2^{(b)}} w_{1i} 2\pi_{2i|1}^{-1} \left(y_i - m(\mathbf{x}_i) \right).$$

Note that

$$\hat{Y}_{\text{oracle}} = \left(\hat{Y}_{\text{oracle}}^{(a)} + \hat{Y}_{\text{oracle}}^{(b)} \right) / 2$$

where $\hat{Y}_{\text{oracle}}$ is defined in (1).

Note that

$$\begin{aligned} \hat{Y}_{\text{d,reg}}^{(a)} - \hat{Y}_{\text{oracle}}^{(a)} &= \sum_{i \in A_1} w_{1i} \left\{ \hat{m}^{(b)}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} - \sum_{i \in A_2^{(a)}} w_{1i} 2\pi_{2i|1}^{-1} \left\{ \hat{m}^{(b)}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} \\ &= \sum_{i \in A_1} w_{1i} \left\{ \hat{m}^{(b)}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} \left(1 - 2\delta_i I_i^{(a)} \pi_{2i|1}^{-1} \right) \\ &:= D_n^{(a)}. \end{aligned}$$

It can be shown that, under some regularity conditions, we obtain

$$N^{-1} D_n^{(a)} = o_p(n^{-1/2}). \tag{4}$$

Also, writing

$$D_n^{(b)} = \hat{Y}_{\text{d,reg}}^{(b)} - \hat{Y}_{\text{oracle}}^{(b)},$$

we can establish

$$N^{-1} D_n^{(b)} = o_p(n^{-1/2}). \tag{5}$$

Combining (4) and (5), we can establish that

$$N^{-1}\left(\hat{Y}_{\text{d,reg}} - \hat{Y}_{\text{oracle}}\right) = o_p(n^{-1/2}). \tag{6}$$

Therefore, asymptotic unbiasedness of the $\hat{Y}_{\text{d,reg}}$ in (2) is established. □

Theorem 1 means that the estimation error of $\hat{m}(\mathbf{x})$ can be safely ignored in the asymptotic sense. There are several advantages of the debiased two-phase regression estimator in (2).

1. Unlike the classical two-phase regression estimator using nonparametric regression, we can establish asymptotic unbiasedness and $\sqrt{n}$-consistency.

2. Even if we use the sample split, there is no efficiency loss. That is, the asymptotic variance is equal to

$$V\left(\hat{Y}_{\text{d,reg}}\right) = V\left(\hat{Y}_1\right) + E\left[V\left\{\sum_{i\in A_2} w_{1i}\pi_{2i|1}^{-1}\left(y_i - m(\mathbf{x}_i)\right) \mid A_1\right\}\right]$$

   where $m(\mathbf{x}_i)$ is the probability limit of $\hat{m}(\mathbf{x}_i)$.

3. Variance estimation is also straightforward. We can compute

$$\hat{\eta}_i = \hat{m}(\mathbf{x}_i) + \delta_i\pi_{2i|1}^{-1}I_i^{(a)}\left\{y_i - \hat{m}^{(b)}(\mathbf{x}_i)\right\} + \delta_i\pi_{2i|1}^{-1}I_i^{(b)}\left\{y_i - \hat{m}^{(a)}(\mathbf{x}_i)\right\}$$

   and apply to the usual variance estimation formula for the first-phase sample, where $I_i^{(a)}$ is the indicator function for $A_2^{(a)}$ such that $I_i^{(a)} = 1$ if $i \in A_2^{(a)}$ and $I_i^{(a)} = 0$ otherwise. Also, $I_i^{(b)} = 1 - I_i^{(a)}$.

My variance estimator is

$$\hat{V} = \frac{1}{n_1(n_1-1)} \sum_{i \in A_1} (\hat{\eta}_i - \bar{\eta}_n)^2$$

I also have an idea on how to implement the above debiased regression estimator using calibration. I will give more details once we are confident in the proposed idea.