Regression Analysis with Response-selective Sampling

Kani Chen[1], Yuanyuan Lin[2], Yuan Yao[3] and Chaoxu Zhou[1]

[1] *Hong Kong University of Science and Technology,* [2] *The Chinese University of Hong Kong and* [3] *Hong Kong Baptist University*

Supplementary Material

This supplement contains proof of Theorem 1.

Proof of Theorem 1

Consider the transformation model

$$H(Y^*) = \theta_0' W^* + \epsilon^*, \tag{S1.1}$$

where $H(\cdot)$ is an unknown monotonically increasing function, ϵ^* is the error, independent of W^*, with unspecified distribution, and θ_0 is a $(d+1)$-dimensional vector of regression coefficients. Accordingly, W^* can be decomposed into $W = (Z^*, X^*)$, where Z^* is the covariate corresponding to the fixed regression coefficient and X^* is the other d-dimensional covariate. Hence, the model can be rewritten as

$$H(Y^*) = Z^* + \beta_0' X^* + \epsilon^*.$$

We suppose the covariance decomposition satisfies that $\tilde{Z}^* := Z^* + \beta_0' X^*$ is irrelevant of X^*. Such a decomposition always exists since $\theta_0' W^*$ is a one-dimensional vector in a $(d+1)$-dimensional linear space of random variables with inner product defined as $\langle X, Y\rangle = E(XY)$, so it has a d-dimensional orthogonal compliment which can be defined as X^*. Furthermore, $\tilde{Z}^*$ and X^* are supposed to be independent.

Consistency:

Define $g(\beta) = E[I\{Y_1 < Y_2\}I\{\beta X_1 + Z_1 < \beta X_2 + Z_2\}]$ and $g_n(\beta) = \frac{1}{n^2-n}\sum_{i\neq j} I\{Y_i < Y_j\}I\{\beta X_i + Z_i < \beta X_j + Z_j\}$.

Step 1. We show that $g(\beta)$ has a unique maximum at $\beta = \beta_0$.

In the response-based sampling, the conditional distribution of $(X, Z)|Y$ in the sample is the same as the conditional distribution of $(X^*, Z^*)|Y^*$ in the population. Therefore, for any $t_1 < t_2$,

$$
\begin{aligned}
& E[I\{Y_1 < Y_2\}I\{\beta X_1 + Z_1 < \beta X_2 + Z_2\}|Y_1 = t_1, Y_2 = t_2] \\
= \; & P(\beta X_1 + Z_1 < \beta X_2 + Z_2 | Y_1 = t_1, Y_2 = t_2) \\
= \; & P(\beta X_1^* + Z_1^* < \beta X_2^* + Z_2^* | Y_1^* = t_1, Y_2^* = t_2) \\
= \; & P(Z_1^* - Z_2^* < \beta X_2^* - \beta X_1^* | \beta_0 X_1^* + Z_1^* + \epsilon_1^* = H(t_1), \beta_0 X_2^* + Z_2^* + \epsilon_2^* = H(t_2)) \\
= \; & P(\tilde{Z}_1^* - \tilde{Z}_2^* < (\beta - \beta_0)(X_2^* - X_1^*) | \tilde{Z}_1^* + \epsilon_1^* = \tilde{t}_1, \tilde{Z}_2^* + \epsilon_2^* = \tilde{t}_2) \\
= \; & \frac{\int P(\xi(\beta) > s_1 - s_2) f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2}{\int f_{\tilde{Z}^*}(s) f_{\epsilon^*}(\tilde{t}_1 - s) ds \int f_{\tilde{Z}^*}(s) f_{\epsilon^*}(\tilde{t}_2 - s) ds}, \qquad \text{(S1.2)}
\end{aligned}
$$

where $\tilde{t}_i = H(t_i)$, $i = 1, 2$.

The denominator is irrelevant with β. The numerator will be proved to have a unique maximum at $\beta = \beta_0$. The numerator can be written as

$$\begin{aligned} & \frac{1}{2}\int [1 - sgn(s_1 - s_2)P(|\xi(\beta)| < |s_1 - s_2|)] \\ & \qquad f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 \\ = & \frac{1}{2}\int f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 + \Pi(\beta) \end{aligned}$$

where

$$\begin{aligned} \Pi(\beta) = & -\frac{1}{2}\int sgn(s_1 - s_2) P(|\xi(\beta)| < |s_1 - s_2|) \\ & \qquad f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2. \end{aligned}$$

It then suffices to show that $\Pi(\beta)$ is uniquely maximized at $\beta = \beta_0$. To this end, write

$$\begin{aligned} \Pi(\beta) = & \frac{1}{2}\int_{s_1 < s_2} g^*_\beta(|s_1 - s_2|) f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 \\ & -\frac{1}{2}\int_{s_1 > s_2} g^*_\beta(|s_1 - s_2|) f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 \\ = & \frac{1}{2}\int_{s_1 < s_2} g^*_\beta(|s_1 - s_2|) f_{\tilde{Z}}(s_1) f_{\tilde{Z}}(s_2) \\ & \qquad [f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\epsilon^*}(\tilde{t}_2 - s_2) - f_{\epsilon^*}(\tilde{t}_1 - s_2) f_{\epsilon^*}(\tilde{t}_2 - s_1)] ds_1 ds_2, \end{aligned} \tag{S1.3}$$

where we define $g^*_\beta(t) = P(|\xi(\beta)| < t)$ and then $g^*_{\beta_0} = 1$ since $\xi(\beta_0) \equiv 0$.

Since $g^*(\cdot)$ is only maximized at $\beta = \beta_0$ by assumption, to show that β_0 is the unique maximizer of $g(\beta)$, we only need to prove that the quantity in the square brackets is positive for all $\tilde{t}_1 < \tilde{t}_2$ and $s_1 < s_2$.

Now we show

$$h(\tilde{t}_1 - s_1) + h(\tilde{t}_2 - s_2) > h(\tilde{t}_1 - s_2) + h(\tilde{t}_2 - s_1)$$

for all $\tilde{t}_1 < \tilde{t}_2$ and $s_1 < s_2$, where $h = \log f_\epsilon$.

By the fact that f_{ϵ^*} is log-concave,

$$\frac{\partial}{\partial t}(h(t-s_1) - h(t-s_2)) = \int_{t-s_2}^{t-s_1} \frac{d^2}{ds^2} h(s) ds < 0.$$

Therefore $h(t - s_1) - h(t - s_2)$ is decreasing in t. As a result,

$$h(\tilde{t}_1 - s_1) + h(\tilde{t}_2 - s_2) > h(\tilde{t}_1 - s_2) + h(\tilde{t}_2 - s_1).$$

Step 2. We show that

$$\sup_\beta |g_n(\beta) - g(\beta)| = O_p(\sqrt{\frac{\log n}{n}}). \tag{S1.4}$$

For each $n \in \mathcal{N}$, let $\{\beta_{n_1}, \cdots, \beta_{n_m}\}$ be a $1/n^2$-net of $\mathbf{B}$, which means that

$$\mathbf{B} \subset \cup_{k=1}^m B(\beta_{n_k}, \frac{1}{n^2}).$$

Then $m = O(n^{2d})$.

For $M > 1$, we have

$$\begin{aligned}
& P(\sup_{\beta}[g_n(\beta) - g(\beta)] > M\sqrt{\frac{\log n}{n}}) \\
\leq\ & P(\sup_{k=1,\cdots,m}[g_n(\beta_{n_k}) - g(\beta_{n_k})] > (M-1)\sqrt{\frac{\log n}{n}}) \\
& +P(\sup_{\beta}[g_n(\beta) - g(\beta)] - \sup_{k=1,\cdots,m}[g_n(\beta_{n_k}) - g(\beta_{n_k})] > \sqrt{\frac{\log n}{n}}) \qquad \text{(S1.5)}
\end{aligned}$$

By Hoeffding's inequality (1963) for U-statistics, the first term in the right hand side of (S1.5) can be bounded by $O(n^{2d-(M-1)^2/4})$. Using Chebyshev's inequality, the second term in the right hand side of (S1.5) is bounded by $O(\frac{1}{n^2})$.

Now we have shown that

$$\begin{aligned}
& P(\sup_{\beta}[g_n(\beta) - g(\beta)] > M\sqrt{\frac{\log n}{n}}) \\
=\ & O(n^{2d-(M-1)^2/4}) + O(\frac{1}{n\log n}). \qquad \text{(S1.6)}
\end{aligned}$$

Since the last equality still holds if we replace g_n and g by $-g_n$ and $-g$, it can be written as

$$\begin{aligned}
& P(\sup_{\beta}|g_n(\beta) - g(\beta)| > M\sqrt{\frac{\log n}{n}}) \\
=\ & O(n^{2d-(M-1)^2/4}) + O(\frac{1}{n\log n}). \qquad \text{(S1.7)}
\end{aligned}$$

Then it follows equality (S1.4).

Step 3. We show that $\hat{\beta}_n$ converges to β_0 in probability.

Since β_0 is the unique maximizer of g, and $\hat{\beta}_n$ is the maximizer of g_n, we have

$$\begin{aligned}
0 &\leq g(\beta_0) - g(\hat{\beta}_n) \\
&= [g(\beta_0) - g_n(\beta_0)] - [g(\hat{\beta}_n) - g_n(\hat{\beta}_n)] - [g_n(\hat{\beta}_n) - g_n(\beta_0)] \\
&\leq [g(\beta_0) - g_n(\beta_0)] - [g(\hat{\beta}_n) - g_n(\hat{\beta}_n)] \\
&= O_p(\sqrt{\frac{\log n}{n}}) + O_p(\sqrt{\frac{\log n}{n}}) \\
&= O_p(\sqrt{\frac{\log n}{n}}) \qquad \text{(S1.8)}
\end{aligned}$$

On the other hand, by the differentiability of density functions of $\tilde{Z}$ and X, note that β_0 is the unique maximizer of g and $\dot{g}(\beta_0) = 0$, the Taylor expansion can then be written as

$$g(\hat{\beta}_n) - g(\beta_0) = -(\hat{\beta}_n - \beta_0)' A(\hat{\beta}_n - \beta_0) + o_p(\hat{\beta}_n - \beta_0)^2, \qquad \text{(S1.9)}$$

where A is the negative hessian matrix of g at β_0, which is a positive definite matrix.

Compare the last two equations, it follows that

$$\hat{\beta}_n - \beta_0 = O_p(\sqrt[4]{\frac{\log n}{n}}) = o_p(n^{-1/5}). \qquad \text{(S1.10)}$$

The consistency is proved.

Asymptotic normality:

We still use the notation of g and g_n as above. Furthermore, denote

$$\epsilon_n(\beta) = g_n(\beta) - g(\beta). \tag{S1.11}$$

Standard decomposition of U-statistics gives

$$\epsilon_n(\beta) - \epsilon_n(\beta_0) = \frac{1}{n}\sum_{i=1}^{n} b_i(\beta) + \frac{1}{n^2-n}\sum_{i<j} d_{ij}(\beta), \tag{S1.12}$$

where

$$b_i(\beta) = E[a_{ij}(\beta) + a_{ji}(\beta) - 2Ea_{ij}(\beta)|Z_i, X_i, Y_i], \tag{S1.13}$$

$$d_{ij}(\beta) = a_{ij}(\beta) + a_{ji}(\beta) - 2Ea_{ij}(\beta) - b_i(\beta) - b_j(\beta). \tag{S1.14}$$

and

$$a_{ij}(\beta) = [I\{Z_i + \beta' X_i > Z_j + \beta' X_j\} - I\{Z_i + \beta_0' X_i > Z_j + \beta_0' X_j\}] \\ I\{Y_i > Y_j\}. \tag{S1.15}$$

Note that $Eb_i(\beta) \equiv 0$, Taylor expansion gives

$$\frac{1}{n}\sum_{i=1}^{n} b_i(\beta) = (\beta - \beta_0)'\frac{1}{n}\sum_{i=1}^{n} \dot{b}_i(\beta_0) + o_p(|\beta - \beta_0|)^2. \tag{S1.16}$$

Using exponential inequality again, similar to the step 2 in the proof of consistency, we have

$$\sup_{|\beta-\beta_0|=o_p(n^{-1/5})} |\frac{1}{n^2-n}\sum_{i<j} d_{ij}(\beta)| = o_p(n^{-1}). \tag{S1.17}$$

So far we have shown that

$$
\begin{aligned}
& g_n(\beta) \\
= \;& g(\beta) + \epsilon_n(\beta) \\
= \;& g(\beta_0) - \frac{1}{2}(\beta-\beta_0)'A(\beta-\beta_0) + (\beta-\beta_0)'\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + \epsilon_n(\beta_0) + o_p(|\beta-\beta_0|)^2 \\
& \qquad + o_p(n^{-1}) \\
= \;& f_n(\beta) + \epsilon_n(\beta_0) + o_p(n^{-1}), \qquad \text{(S1.18)}
\end{aligned}
$$

where

$$
\begin{aligned}
& f_n(\beta) \\
= \;& g(\beta_0) - \frac{1}{2}(\beta-\beta_0)'A(\beta-\beta_0) + (\beta-\beta_0)'\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + o_p(|\beta-\beta_0|)^2 \\
= \;& g(\beta_0) - \frac{1}{2}(\beta-\beta_0)'A_n(\beta-\beta_0) + (\beta-\beta_0)'\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) \\
= \;& g(\beta_0) - \frac{1}{2}\{A_n^{1/2}[\beta-\beta_0-A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0)]\}'\{A_n^{1/2}[\beta-\beta_0-A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0)]\} \\
& + \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0))'A_n^{-1}(\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0)), \qquad \text{(S1.19)}
\end{aligned}
$$

where we let $o_p(|\beta-\beta_0|)^2 = c_n|\beta-\beta_0|^2$ with $c_n = o_p(1)$ and $A_n = A - 2c_nI$.

So the maximizer of f_n is

$$\hat{\gamma}_n = \beta_0 + A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) \qquad \text{(S1.20)}$$

Suppose that $\hat{\beta}_n$ is the maximizer of g_n, then

$$\begin{aligned}
0 &\leq f_n(\hat{\gamma}_n) - f_n(\hat{\beta}_n) \\
&= [f_n(\hat{\gamma}_n) + \epsilon_n(\beta_0) - g_n(\hat{\gamma}_n)] - [f_n(\hat{\beta}_n) + \epsilon_n(\beta_0) - g_n(\hat{\beta}_n)] - [g_n(\hat{\beta}_n) - g_n(\hat{\gamma}_n)] \\
&\leq [f_n(\hat{\gamma}_n) + \epsilon_n(\beta_0) - g_n(\hat{\gamma}_n)] - [f_n(\hat{\beta}_n) + \epsilon_n(\beta_0) - g_n(\hat{\beta}_n)] \\
&= o_p(n^{-1}) + o_p(n^{-1}) \\
&= o_p(n^{-1}). \qquad \text{(S1.21)}
\end{aligned}$$

On the other hand, from the expression of f_n,

$$\begin{aligned}
& f_n(\hat{\gamma}_n) - f_n(\hat{\beta}_n) \\
= \; & \frac{1}{2}\{A_n^{1/2}[\hat{\beta}_n - \beta_0 - A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0)]\}'\{A_n^{1/2}[\hat{\beta}_n - \beta_0 - A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0)]\}. \\
& \qquad \text{(S1.22)}
\end{aligned}$$

Compare (S1.21) and (S1.22), finally we have

$$\begin{aligned}
\hat{\beta}_n &= \beta_0 + A_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + o_p(n^{-1/2}) \\
&= \beta_0 + A^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + (A_n^{-1} - A^{-1})\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + o_p(n^{-1/2}) \\
&= \beta_0 + A^{-1}\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + o_p(n^{-1/2}), \qquad \text{(S1.23)}
\end{aligned}$$

where the last equation comes from that

$$A_n^{-1} - A^{-1} = o_p(1)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\dot{b}_i(\beta_0) = O_p(n^{-1/2})$$

by the definition of A_n and the central limit theorem.

Therefore,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = A^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{b}_i(\beta_0) + o_p(1) \rightarrow N(0, \Sigma)$$

in distribution, where

$$\Sigma = A^{-1}Var\{\dot{b}_1(\beta_0)\}(A^{-1})'.$$

We further define $B = Var\{\dot{b}_1(\beta_0)\}$ and the proof is done.