A note on reducing variance by estimating the nuisance

Yen-Chi Chen
University of Washington
November 7, 2023

Some people may have heard of a strange phenomenon in causal inference or missing data problem that an estimator based on estimated nuisance could have a smaller variance than an estimator based on oracle nuisance (knowing the true nuisance parameter such as the propensity score).

In this note, I will try to explain how this phenomenon occurs in the setting of estimating equations.

The main reference is

> Lok, Judith J. "How estimating nuisance parameters can reduce the variance (with consistent variance estimation)." arXiv preprint arXiv:2109.02690 (2021).

1 Setup: estimating equations

We consider a general setup for the estimating equations with nuisance parameter.

Suppose our data consists of IID triplets

$$(X_1, Y_1, Z_1), \cdots, (X_n, Y_n, Z_n) \sim \mathbb{P}.$$

Let $\theta \in \mathbb{R}^d$ be the parameter of interest in our model. Let $\eta \in \mathbb{R}^q$ be the nuisance parameter that only involves X, Z.

Let $\mathbb{P}_n$ be the empirical measure from the data and $\mathbb{P}$ be the probability measure.

We consider the following two functions for identifying θ, η:

$$\begin{aligned} S_1(x, y, z; \theta, \eta) = S_1(\theta, \eta) \in \mathbb{R}^d \\ S_2(x, z; \eta) = S_2(\eta) \in \mathbb{R}^q. \end{aligned}$$

The estimators of θ and η are parameters that solves the estimating equations:

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^n S_1(X_i, Y_i, Z_i; \widehat{\theta}, \widehat{\eta}) = \mathbb{P}_n S_1(\widehat{\theta}, \widehat{\eta}) = S_{n,1}(\widehat{\theta}, \widehat{\eta}) \\ 0 &= \frac{1}{n}\sum_{i=1}^n S_2(X_i, Z_i; \widehat{\eta}) = \mathbb{P}_n S_2(\widehat{\eta}) = S_{n,2}(\widehat{\eta}). \end{aligned} \tag{1}$$

For abbreviation, we sometimes write

$$S_1(\cdot; \theta, \eta) = S_1(X, Y, Z; \theta, \eta), \qquad S_2(\cdot; \eta) = S_2(X, Z; \eta).$$

The population version of these quantities are

$$\begin{aligned} 0 &= \mathbb{E}[S_1(X_i, Y_i, Z_i; \theta_0, \eta_0)] = \mathbb{P}S_1(\theta_0, \eta_0) \\ 0 &= \mathbb{E}[S_2(X_i, Z_i; \eta_0)] = \mathbb{P}S_2(\eta_0). \end{aligned} \tag{2}$$

Key assumption on the estimating equations: nuisance generative model. We consider the following additional requirement on the second estimating equation:

$$S_2(x, z; \eta) = \nabla_\eta \log p(z|x; \eta), \tag{3}$$

where $p(z|x;\eta)$ is the conditional PDF/PMF of Z given X. Namely, the nuisance parameter is identified from the log-likelihood model.

Equation (3) together with equation (2) implies that the joint distribution $p(x,y,z)$ that generates our data can be factored as

$$p(x, y, z) = p(y|x, z; \theta_0, \gamma_0) p(z|x; \eta_0) p(x),$$

where γ_0 is another set of nuisance parameters such that (θ_0, γ_0) together determines the conditional distribution $p(y|x,z;\theta_0,\gamma_0)$. This is because the parameter of interest θ may not uniquely determine a generative model. Thus, we allow another set of parameters γ in the conditional distribution. Note that $p(x)$ itself is another nuisance function that is not specified in the estimating equations.

Example: Inverse probability weighting in binary treatment problem. Consider a binary treatment effect problem where our Y is outcome of interest, X is confounders, $Z \in \{0,1\}$ is the treatment indicator. A common estimator of the average treatment effect (ATE) θ is

$$\theta = \mathbb{E}\left(\frac{YZ}{\pi(X;\eta)} - \frac{Y(1-Z)}{1-\pi(X;\eta)} \right),$$

where $\pi(x;\eta) = P(Z = 1|X = x;\eta)$. A common model of η is the logistic regression, i.e.,

$$\pi(x;\eta) = \frac{e^{\eta^T x}}{1 + e^{\eta^T x}}.$$

A common estimator of θ is

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i Z_i}{\pi(X_i; \widehat{\eta})} - \frac{Y_i(1-Z_i)}{1-\pi(X_i;\widehat{\eta})},$$

which can be written as $\widehat{\theta}$ from solving

$$0 = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\theta} - \frac{Y_i Z_i}{\pi(X_i; \widehat{\eta})} - \frac{Y_i(1-Z_i)}{1-\pi(X_i;\widehat{\eta})} \right),$$

so this implies

$$S_1(X, Y, Z; \theta, \eta) = \theta - \frac{Y_i Z_i}{\pi(X_i;\eta)} - \frac{Y_i(1-Z_i)}{1-\pi(X_i;\eta)} = \theta - Y_i Z_i (1 + e^{-\eta^T X_i}) - Y_i(1-Z_i)(1+e^{\eta^T X_i}).$$

The estimator $\widehat{\eta}$ is often from the MLE, which in the case of logistic regression, is from the following score equation:

$$0 = \frac{1}{n} \sum_{i=1}^{n} Z_i X_i - \frac{X_i e^{\widehat{\eta}^T X_i}}{1 + e^{\widehat{\eta}^T X_i}}.$$

Thus,

$$S_2(X,Z;\eta) = ZX - \frac{Xe^{\eta^T X}}{1+e^{\eta^T X}}.$$

2 Asymptotic variance

To investigate the asymptotic variance of $\widehat{\theta},\widehat{\eta}$, we use a standard procedure. First, let

$$S_n(\theta,\eta) = \begin{pmatrix} \mathbb{P}_n S_1(\theta,\eta) \\ \mathbb{P}_n S_2(\eta) \end{pmatrix} \in \mathbb{R}^{d+q}, \qquad \bar{S}(\theta,\eta) = \begin{pmatrix} \mathbb{P} S_1(\theta,\eta) \\ \mathbb{P} S_2(\eta) \end{pmatrix} \in \mathbb{R}^{d+q}.$$

Equations (1) and (2) imply that

$$0 = S_n(\widehat{\theta},\widehat{\eta}) = \bar{S}(\theta_0,\eta_0).$$

Thus, Taylor expansion shows that

$$\begin{aligned} S_n(\theta_0,\eta_0) - \bar{S}(\theta_0,\eta_0) &= S_n(\theta_0,\eta_0) - S_n(\widehat{\theta},\widehat{\eta}) \\ &\approx -\nabla S_n(\theta_0,\eta_0) \begin{pmatrix} \widehat{\theta}-\theta_0 \\ \widehat{\eta}-\theta_0 \end{pmatrix}, \end{aligned}$$

which further implies that

$$\begin{pmatrix} \widehat{\theta}-\theta_0 \\ \widehat{\eta}-\theta_0 \end{pmatrix} \approx -[\bar{H}(\theta_0,\eta_0)]^{-1}(S_n(\theta_0,\eta_0) - \bar{S}(\theta_0,\eta_0)), \tag{4}$$

where

$$\bar{H}(\theta,\eta) = \nabla \bar{S}(\theta,\eta)$$

is the asymptotic limit of $\nabla S_n(\theta_0,\eta_0)$.

The Hessian matrix $\bar{H}(\theta,\eta)$ has a block diagonal feature that

$$\begin{aligned} \bar{H}(\theta_0,\eta_0) &= \begin{pmatrix} \mathbb{E}[\nabla_\theta S_1(X,Y,Z;\theta_0,\eta_0)] & \mathbb{E}[\nabla_\eta S_1(X,Y,Z;\theta_0,\eta_0)] \\ \mathbb{E}[\nabla_\theta S_2(X,Z;\eta_0)] & \mathbb{E}[\nabla_\eta S_2(X,Z;\eta_0)] \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{E}[\nabla_\theta S_1(X,Y,Z;\theta_0,\eta_0)] & \mathbb{E}[\nabla_\eta S_1(X,Y,Z;\theta_0,\eta_0)] \\ 0 & \mathbb{E}[\nabla_\eta S_2(X,Z;\eta_0)] \end{pmatrix}. \end{aligned}$$

This implies that the inverse matrix will have a similar block-diagonal structure:

$$\bar{H}(\theta_0,\eta_0)^{-1} = \begin{pmatrix} \bar{\Omega}_{11}(\theta_0,\eta_0) & \bar{\Omega}_{12}(\theta_0,\eta_0) \\ 0 & \bar{\Omega}_2 2(\eta_0) \end{pmatrix}.$$

Using the matrix algebra, one can show that the off-diagonal block is

$$\begin{aligned} \bar{\Omega}_{12}(\theta_0,\eta_0) &= -\mathbb{E}[\nabla_\theta S_1(\cdot;\theta_0,\eta_0)]^{-1}\mathbb{E}[\nabla_\eta S_1(\cdot;\theta_0,\eta_0)]\mathbb{E}[\nabla_\eta S_2(\cdot;\eta_0)]^{-1} \\ &= -H_{11}^{-1}(\theta_0,\eta_0)\mathbb{E}[\nabla_\eta S_1(\cdot;\theta_0,\eta_0)]H_{22}^{-1}(\eta_0). \end{aligned} \tag{5}$$

An interesting note is that

$$H_{11}^{-1}(\theta_0,\eta_0)=\Omega_{11}(\theta_0,\eta_0),\qquad H_{22}^{-1}(\eta_0)=\Omega_{22}(\eta_0).$$

The quantity $\bar{\Omega}_{12}(\theta_0,\eta_0)$ will play a key role in the asymptotic variance of $\widehat{\theta}$. To see this, putting equation (5) into equation (4) leads to

$$\begin{aligned}\widehat{\theta}-\theta_0 &= -\bar{\Omega}_{11}(\theta_0,\eta_0)(S_{n,1}(\theta_0,\eta_0)-\underbrace{\bar{S}_1(\theta_0,\eta_0)}_{=0})-\bar{\Omega}_{12}(\theta_0,\eta_0)(S_{n,2}(\eta_0)-\underbrace{\bar{S}_2(\eta_0)}_{=0})\\ &= -\bar{\Omega}_{11}(\theta_0,\eta_0)[S_{n,1}(\theta_0,\eta_0)-\underbrace{\mathbb{E}[\nabla_\eta S_1(\cdot;\theta_0,\eta_0)]\Omega_{22}(\eta_0)S_{n,2}(\eta_0)}_{=(A)}].\end{aligned}\tag{6}$$

The quantity (A) is how estimating the nuisance (from $S_{n,2}$) influences the variance of the estimator $\widehat{\theta}$.

As a reference, if we are using the oracle estimator, i.e., knowing η_0, the asymptotic expansion will be

$$\widetilde{\theta}-\theta_0\approx -\bar{\Omega}_{11}(\theta_0,\eta_0)S_{n,1}(\theta_0,\eta_0).\tag{7}$$

3 Invariance under linear transformation

To investigate how the asymptotic variance from equation (6) can be smaller than the oracle in equation (7), we use the following insight:

> the estimating equations remain unchanged under some linear transformation.

Let $A\in\mathbb{R}^{d\times p}$ be a none-random matrix. The estimators from equation (1) will be the same if we replace $S_1(X,Y,Z;\theta,\eta)$ by

$$S_{1,A}(X,Y,Z;\theta,\eta)=S_1(X,Y,Z;\theta,\eta)+AS_2(X,Z;\eta).$$

After this change, we can redo all the derivation, which leads to another asymptotic linear representation of $\widehat{\theta}$:

$$\widehat{\theta}-\theta_0\approx -\bar{\Omega}_{11,A}(\theta_0,\eta_0)[S_{n,1,A}(\theta_0,\eta_0)-\mathbb{E}[\nabla_\eta S_{1,A}(\cdot;\theta_0,\eta_0)]\Omega_{22,A}(\eta_0)S_{n,2}(\eta_0)],$$

where $\bar{\Omega}_{11,A}$ and $\bar{\Omega}_{22,A}$ are the corresponding quantities of $\bar{\Omega}_{11}$ and $\bar{\Omega}_{11}$ under the linear transformation. An important note is that $\bar{\Omega}_{11,A}=\bar{\Omega}_{11}$ and $\bar{\Omega}_{22,A}=\bar{\Omega}_{22}$!

Thus, we conclude that

$$\widehat{\theta}-\theta_0\approx -\bar{\Omega}_{11}(\theta_0,\eta_0)[\underbrace{S_{n,1,A}(\theta_0,\eta_0)}_{=(B)}-\underbrace{\mathbb{E}[\nabla_\eta S_{1,A}(\cdot;\theta_0,\eta_0)]\Omega_{22}(\eta_0)S_{n,2}(\eta_0)}_{=(C)}],\tag{8}$$

where

$$\begin{aligned}S_{n,1,A}(\theta_0,\eta_0)&=\mathbb{P}_n[S_1(\theta_0,\eta_0)+A\cdot S_2(\eta_0)]\\ S_{1,A}(\cdot;\theta_0,\eta_0)&=S_1(\cdot;\theta_0,\eta_0)+A\cdot S_2(\cdot;\eta_0).\end{aligned}$$

This holds for any non-random A. So the question is: how can we choose A to minimize this variance.

We will consider the following choice:

$$A^* = \mathbb{E}[S_1(X,Y,Z;\theta_0,\eta_0)S_2(X,Z;\eta_0)^T]\mathbb{E}[S_2(X,Z;\eta_0)S_2(X,Z;\eta_0)^T]^{-1}$$

Note that equation (3) implies the following two powerful results:

$$\begin{aligned}\mathbb{E}(\nabla_\eta S_2(X,Z;\eta)) &= \mathbb{E}(S_2(X,Z;\eta)S_2(X,Z;\eta)^T),\\ \mathbb{E}(\nabla_\eta S_1(X,Y,Z;\theta_0,\eta_0)) &= -\mathbb{E}[S_1(X,Y,Z;\theta_0,\eta_0)S_2(X,Z;\eta_0)^T].\end{aligned} \tag{9}$$

This will imply a powerful fact about $\mathbb{E}[\nabla_\eta S_{1,A}(\cdot;\theta_0,\eta_0)]$ that is a key quantity in term (C). A simple derivation shows that

$$\begin{aligned}\mathbb{E}[\nabla_\eta S_{1,A^*}(\cdot;\theta_0,\eta_0)] &= \mathbb{E}[\nabla_\eta S_1(\cdot;\theta_0,\eta_0) + A^* \cdot \nabla_\eta S_2(\cdot;\eta_0)]\\ &= \mathbb{E}[\nabla_\eta S_1(\cdot;\theta_0,\eta_0)] + \mathbb{E}[S_1(\cdot;\theta_0,\eta_0)S_2(\cdot;\eta_0)^T]\underbrace{\mathbb{E}[S_2(\cdot;\eta_0)S_2(\cdot;\eta_0)^T]^{-1}\mathbb{E}[\nabla_\eta S_2(\cdot;\eta_0)]}_{=I_d}\\ &= \mathbb{E}[\nabla_\eta S_1(\cdot;\theta_0,\eta_0)] + \mathbb{E}[S_1(\cdot;\theta_0,\eta_0)S_2(\cdot;\eta_0)^T]\\ &= 0.\end{aligned}$$

Therefore, under $A = A^*$, (C) = 0.

Therefore, the only quantity left is term (B). So we can rewrite equation (8) as

$$\widehat{\theta} - \theta_0 \approx -\bar{\Omega}_{11}(\theta_0,\eta_0)S_{n,1,A^*}(\theta_0,\eta_0),$$

where

$$\begin{aligned}S_{n,1,A^*}(\theta_0,\eta_0) &= \mathbb{P}_n S_{1,A^*}(\theta_0,\eta_0)\\ S_{1,A^*}(\cdot;\theta_0,\eta_0) &= S_1(\cdot;\theta_0,\eta_0) - \mathbb{E}(S_1(\cdot;\theta_0,\eta_0)S_2(\cdot;\eta_0)^T)\mathbb{E}(S_2(\cdot;\eta_0)S_2(\cdot;\eta_0)^T)^{-1}S_2(\cdot;\eta_0).\end{aligned}$$

The quantity $S_{1,A^*}(\theta_0,\eta_0)$ is the orthogonal component of $S_1(\theta_0,\theta_0)$ of $S_2(\eta_0)$. Thus, the covariance matrix of $S_{1,A^*}(\theta_0,\eta_0)$ will be smaller than the covariance matrix of $S_1(\theta_0,\eta_0)$! Namely, one can show that

$$\mathbb{E}[S_{1,A^*}(\cdot;\theta_0,\eta_0)S_{1,A^*}(\cdot;\theta_0,\eta_0)^T] - \mathbb{E}[S_1(\cdot;\theta_0,\eta_0)S_1(\cdot;\theta_0,\eta_0)^T]$$

is negative definite. This implies that the estimator $\widehat{\theta}$ has a smaller asymptotic variance than the oracle estimator $\widetilde{\theta}$!

4 Derivation of equation (9)

This derivation is similar to the derivation in the maximum likelihood estimator theory. For the completeness, we provide a derivation on the second equation. The first equation can be obtained in a similar manner.

Equation (3) requires

$$S_2(x,z;\eta) = \nabla_\eta \log p(z|x;\eta).$$

Using the population estimating equations,

$$0 = \mathbb{E}(S_1(X,Y,Z;\theta_0,\eta_0)) = \int S_1(x,y,z;\theta_0,\eta_0)p(y|x,z;\theta_0,\gamma_0)p(z|x;\eta_0)p(x)dydzdx$$

assuming that the data is from $p(x,y,z;\theta_0,\gamma_0,\eta_0)$, where γ_0 is another set of nuisance parameter such that (θ_0,γ_0) together determines the distribution $p(y|x,z;\theta_0,\gamma_0)$. See the discussion after Equation (3).

The above result implies that if the data is from $p(x,y,z;\theta,\gamma,\eta)$, we have the integral relation

$$0 = \int S_1(x,y,z;\theta,\eta)p(y|x,z;\theta,\gamma)p(z|x;\eta)p(x)dydzdx.$$

for any η,θ. Taking derivative with respect to η leads to

$$\begin{aligned}
0 &= \nabla_\eta 0 \\
&= \nabla_\eta \int S_1(x,y,z;\theta,\eta)p(y|x,z;\theta,\gamma)p(z|x;\eta)p(x)dydzdx \\
&= \int [\nabla_\eta S_1(x,y,z;\theta,\eta)]p(y|x,z;\theta,\gamma)p(z|x;\eta)p(x)dydzdx \\
&\quad + \int S_1(x,y,z;\theta,\eta)p(y|x,z;\theta,\gamma)[\nabla_\eta \log p(z|x;\eta)]\log p(z|x;\eta)p(x)dydzdx \\
&= \mathbb{E}[\nabla_\eta(S_1(X,Y,Z;\theta,\eta))] + \mathbb{E}[S_1(X,Y,Z;\theta,\eta)S_2(X,Z;\eta)^T],
\end{aligned}$$

which is the desired result.