Causal Inference: Estimation via Z Estimators

Hyunseung Kang

April 24, 2024

Abstract

Once the causal estimand is identified (i.e. the causal estimand is a function of the observed data), the next natural step is to estimate it with data. Here, we discuss an estimation approach using Z estimators. We'll first review Z estimators. Next, we'll show how to frame estimation of causal estimands as an instance of an Z estimation problem. This document assumes that you have taken an undergraduate course in mathematical statistics and an undergraduate course in linear algebra.

1 Review: Causal Identification

Causal identification is the exercise of equating a causal estimand into another estimand that is defined with only the observed data only. For example, we identified the average treatment effect under strong ignorability (i.e. $Y(1), Y(0) \perp A \mid X$ and $0 < P(A = 1|X = x) < 1$ for all x) and SUTVA (i.e $Y = AY(1) + (1 - A)Y(0)$) as follows:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mu_1(X) - \mu_0(X)], \quad \mu_a(X) = \mathbb{E}[Y \mid A = a, X]$$

Once the causal estimands is identified, estimation focuses on estimating the estimand defined with the observed data, often referred to as *a functional of the observed data*. In the ATE example above, estimation involves estimating the functional $\mathbb{E}[\mu_1(X) - \mu_0(X)]$, which only consists of the observed data. We discuss how to do this using Z estimators. Throughout the document, we assume that we collect n i.i.d. samples from some common distribution.

2 Review: Z Estimators

2.1 Definition and Examples

Suppose we observe n i.i.d. samples of data $O_1, \ldots, O_n \overset{\text{iid}}{\sim} F$ from some distribution, denoted as F; note that O_i can be a scalar or a vector. Consider an estimator $\hat{\boldsymbol{\theta}}$ that satisfies the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} f(O_i, \hat{\theta}) = 0, \quad f(O_i, \boldsymbol{\theta}) \in \mathbb{R}^m, \ \theta \in \mathbb{R}^d \tag{1}$$

where R_n is a sequence of random variables. Some estimators that satisfy equation (1) include

1. The sample mean can be written as

   $$\frac{1}{n}\sum_{i=1}^{n} O_i - \hat{\theta} = 0, \quad f(O_i, \mu) = O_i - \theta, \ m = d = 1$$

   Notice that this form remain the same for any distribution F (e.g. Normal, Poisson, Exponential, Exponential family, etc.).

2. The sample variance where the mean of F is unknown

   $$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} O_i - \hat{\theta}_1 = 0 \\ \frac{1}{n}\sum_{i=1}^{n}(O_i - \hat{\theta}_1)^2 - \hat{\theta}_2 = 0 \end{pmatrix}, \quad f(O_i, \theta) = \begin{pmatrix} O_i - \theta_1 \\ (O_i - \theta_1)^2 - \theta_2 \end{pmatrix}, m = d = 2$$

   Notice again that this form remains the same for any distribution F.

3. The maximum likelihood estimator (MLE) for a parametric distribution of O_i with density $p(o,\theta)$, $\theta \in \mathbb{R}^d$:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\text{argmax}} \prod_{i=1}^{n} p(O_i.\theta) = \underset{\theta \in \mathbb{R}^d}{\text{argmax}} \sum_{i=1}^{n} \log(p(O_i,\theta))$$

   To solve the optimization problem, we usually have to take the partial derivative of the log likelihood with respect to θ and set it equal to zero:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{p(O_i,\hat{\theta})} \nabla_\theta p(O_i,\hat{\theta}) = 0, \quad f(O_i,\theta) = \frac{1}{p(O_i,\theta)} \nabla_\theta p(O_i,\theta), \; m = d$$

   The function f above (i.e. the derivative of the log likelihood) is called the *score function.* The score function has the unique property that at the true value of the density, denoted as $p(o,\theta^*)$, its expectation is

$$\mathbb{E}\left[\frac{1}{p(O_i,\theta^*)} \nabla_\theta p(O_i,\theta^*)\right] = \int \frac{1}{p(o,\theta^*)} \nabla_\theta p(o,\theta^*) p(o,\theta^*) do =_a \int \nabla_\theta p(o,\theta^*) do = \nabla_\theta \int p(o,\theta^*) do = \nabla_\theta 1 = 0$$

   The equality $=_a$ assumes that we can switch integration with differentiation.

4. Linear regression where given the outcome $Y_i \in \mathbb{R}$ and d predictors $X_i \in \mathbb{R}^d$, we solve the following optimization problem:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\text{argmax}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i^\intercal \theta)^2.$$

   We usually find the OLS estimator by taking the partial derivative of the objective with respect to θ and setting it equal to zero:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i^\intercal \hat{\theta}) = 0, \quad f(O_i,\theta) = X_i(Y_i - X_i^\intercal \theta), \; m = d$$

   Note that $O_i = (Y_i, X_i)$.

5. Logistic regression where given a binary outcome A and d predictors $X \in \mathbb{R}^d$, we solve the following likelihood problem:

$$\begin{aligned}
\hat{\theta} &= \underset{\theta \in \mathbb{R}^d}{\text{argmax}} \prod_{i=1}^{n} \pi(X_i,\theta)_i^A (1 - \pi(X_i,\theta)^{1-A_i} \\
&= \underset{\theta \in \mathbb{R}^d}{\text{argmax}} \sum_{i=1}^{n} A_i \log(\pi(X_i,\theta)) + (1-A_i)\log(1-\pi(X_i,\theta)), \quad \pi(X_i,\theta) = \frac{\exp(X_i^\intercal \theta)}{1+\exp(X_i^\intercal \theta)}
\end{aligned}$$

   We can solve the above optimization by taking the derivative with respect to θ, i.e.

$$\begin{aligned}
\nabla_\theta \sum_{i=1}^{n} A_i \log(\pi(X_i,\theta)) + (1-A_i)\log(1-\pi(X_i,\theta)) &= \sum_{i=1}^{n} \nabla_\theta \pi(X_i,\theta) \left(\frac{A_i}{\pi(X_i,\theta)} - \frac{1-A_i}{1-\pi(X_i,\theta)}\right) \\
&= \sum_{i=1}^{n} \pi(X_i,\theta)(1-\pi(X_i,\theta))X_i \left(\frac{A_i}{\pi(X_i,\theta)} - \frac{1-A_i}{1-\pi(X_i,\theta)}\right) \\
&= \sum_{i=1}^{n} X_i(A_i - \pi(X_i,\theta)),
\end{aligned}$$

   and setting the above to zero, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} X_i(A_i - \pi(X_i,\theta)) = 0, \quad f(O_i,\theta) = X_i(A_i - \pi(X_i,\theta)), \; m = d = p$$

2.2 Key Result

The following theorem characterize the asymptotic behavior of estimators that satisfy equation (1). Throughout all the theory, we'll let θ^* be a solution to the equation:

$$\mathbb{E}[f(O_i,\theta^*)] = 0 \tag{2}$$

You want θ^* to be equal to the true parameter you want to estimate (e.g. population mean, population variance, true β in regression). We'll show that the estimator $\hat{\theta}$ in equation (1) is asymptotically Normal around θ^*.

Theorem 1. *Consider the case when $m = d$. Suppose (A1) there exists $\theta^* \in \mathbb{R}^d$ where $\mathbb{E}[f(O_i, \theta^*)] = 0$, (A2) $\mathbb{E}[\|f(O_i, \theta^*)\|_2^2] < \infty$, (A3) $\mathbb{E}[\nabla_\theta f(O_i, \theta) \mid_{\theta=\theta^*}]$ exists and is non-singular, (A4) for each θ in an open subset of $\mathbb{R}^d$, $\frac{\delta^2}{\delta\theta_j\theta_k} f(o, \theta)$ exists for every j, k, o and is continuous in θ, and (A5) for every $h = 1, \dots, m$, there exists a fixed function $g(o)$ where $\mathbb{E}[|g(O_i)|] < \infty$ and $|\frac{\delta^2}{\delta\theta_j\theta_k} f_h(o, \theta)| \leq g(o)$ for every θ in a neighborhood of θ^*. Then, as long as the solution to $n^{-1}\sum_{i=1}^n f(O_i, \hat{\theta}) = 0$ is unique for every n, we have $\hat{\theta} \to \theta^*$ and*

$$\sqrt{n}(\hat{\theta} - \theta^*) \to N\left(0, \mathbb{E}[\nabla_\theta f(O_i, \theta) \mid_{\theta=\theta^*}]^{-1} \mathbb{E}[f(O_i, \theta^*) f^\intercal(O_i, \theta^*)] \mathbb{E}[\nabla_\theta f(O_i, \theta) \mid_{\theta=\theta^*}]^{-\intercal}\right)$$

Proof. See van der vaart, Theorem 5.41 and Theorem 5.42. In particular, the condition that there is a unique root to $n^{-1}\sum_{i=1}^n f(O_i, \hat{\theta}) = 0$ guarantees that the estimator $\hat{\theta}$ that the statistician actually obtains is consistent. ☐

This is not the most general theorem for Z estimators, but it's the easiest to understand[1]. For causal inference, the goal is to apply this theorem by checking the conditions (A1)-(A5) and if necessary, making additional assumptions to make sure (A1)-(A5) are satisfied.

[add next year: asymptotic efficiency via MLE; local asymptotic minimaxity from Chamberlain 1987]

2.3 Application of Theorem 1

2.3.1 Sample Mean

For the sample mean, we show that (A1)-(A5) in Theorem 1 are satisfied if O_i has a finite second moment.

- (A1) We can take the expectation of f, i.e. $\mathbb{E}[f(O_i, \theta)] = \mathbb{E}[O_i] - \theta = 0$. The values of θ that will make this equation equal to zero is $\theta = \mathbb{E}[O_i]$. In other words, $\theta^* = \mathbb{E}[O_i]$.
- (A2) We can evaluate $\mathbb{E}[(O_i - \theta^*)^2] = \text{Var}(O_i)$, which is finite because O_i has second moments.
- (A3) We have $\frac{\delta}{\delta\theta} f(O_i, \theta) \mid_{\theta=\theta^*} = -1$. An expectation of this derivative is finite and this value is non-singular.
- (A4) We have $\frac{\delta^2}{\delta^2\theta} f(O_i, \theta) \mid_{\theta=\theta^*} = 0$ and thus, the second derivative s continuous for every θ.
- (A5) From (A4), since the second derivative is zero, it is dominated by a function $g(o) = 1$.

Also, for every n, the solution to $\frac{1}{n}\sum_{i=1}^n O_i - \hat{\theta} = 0$ is unique, namely that $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n O_i$.

2.3.2 Sample Variance

For the sample variance, we show that (A1)-(A5) in Theorem 1 are satisfied if O_i has a finite fourth moment.

- (A1) We can take the expectation of f:

$$\mathbb{E}[f(O_i, \theta)] = \begin{pmatrix} \mathbb{E}[O_i] - \theta_1 \\ \mathbb{E}[(O_i - \theta_1)^2] - \theta_2 \end{pmatrix} = 0.$$

  The values of θ_1, θ_2 that will make the above equation equal to zero is $\theta_1 = \mathbb{E}[O_i]$ and $\theta_2 = \text{Var}[O_i]$. In other words, $\theta^* = (\mathbb{E}[O_i], \text{Var}[O_i])$.
- (A2) For the first component of f, we have $\mathbb{E}[(O_i - \theta_1^*)^2] = \text{Var}(O_i)$, which is finite because O_i has finite fourth moments. For the second component of f, $\mathbb{E}[\left((O_i - \theta_1^*)^2 - \theta_2^*\right)^2] = \text{Var}[(O_i - \theta_1^*)^2]$ where the equality is by the definition of variance. If the fourth moment of O_i exists, $\text{Var}[(O_i - \theta_1^*)^2] \leq \mathbb{E}[(O_i - \theta_1^*)^4]$ is finite.
- (A3) For each partial derivative, we have

$$\frac{\delta}{\delta\theta_1} f(O_i, \theta) \mid_{\theta=\theta^*} = \begin{pmatrix} -1 \\ -2(O_i - \theta_1^*) \end{pmatrix}, \quad \frac{\delta}{\delta\theta_2} f(O_i, \theta) \mid_{\theta=\theta^*} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

  The expectation of these quantities are finite. Also, the matrix $E[\nabla_\theta f(O_i, \theta) \mid_{\theta=\theta^*}] = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ is non-singular.
- (A4) From (A3), we see that any second partial derivatives of $f(o, \theta)$ with respect to θ will be constant and hence, is continuous.
- (A5) From (A4), since all second partial derivatives will be constant, they will be dominated by a function $g(o) = 1$.

Also, for every n, the solution to $\frac{1}{n}\sum_{i=1}^n f(O_i, \hat{\theta}) = 0$ is unique since $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n O_i$ and thus, $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n (O_i - \hat{\theta}_1)^2$

[1] In Theorem 1, for those with a weak background in real analysis, you can replace "for each θ in an open subset of $\mathbb{R}^d$" with "for every $\theta \in \mathbb{R}^d$" and "for every θ in a neighborhood of θ^*" with "for every θ." These changes are more stringent than Theorem 1.

2.3.3 Linear Regression

For linear regression, we show that if

(a) (Y_i, X_i) is generated from the following model: $Y_i = X_i^\intercal \theta^* + \epsilon_i$ where $\mathbb{E}[\epsilon_i \mid X_i] = 0$ and $\text{Var}[\epsilon_i \mid X_i] = (\sigma^*)^2 < \infty$

(b) the covariance matrix of X , denoted as $\Sigma_X = \mathbb{E}[X_i X_i^\intercal]$, is finite and positive definite

(c) for every n, the matrix $\frac{1}{n}\sum_{i=1}^n X_i X_i^\intercal$ is invertible

the conditions (A1)-(A5) are satisfied.

(A1) The values of θ that satisfies $\mathbb{E}[f(O_i,\theta)] = \mathbb{E}[X_i(Y_i - X_i^\intercal\theta)] = 0$ is $\theta^* = \beta^*$ because $\mathbb{E}[X_i(Y_i - X_i^\intercal\beta^*)] = \mathbb{E}[X_i\epsilon_i] = \mathbb{E}[X_i\mathbb{E}[\epsilon_i \mid X_i]] = 0$

(A2) We have
$$\mathbb{E}[\|X_i(Y_i - X_i^\intercal\beta^*)\|_2^2] = \mathbb{E}[\|X_i\|_2^2\epsilon_i^2] = \mathbb{E}[\|X_i\|_2^2\mathbb{E}[\epsilon_i^2 \mid X_i]\|] = \mathbb{E}[\|X_i\|_2^2](\sigma^*)^2$$
Since Σ_X is finite, $\mathbb{E}[\|X_i\|_2^2]$ is bounded and thus, the whole expression is bounded.

(A3) We have $\nabla_\theta f(O_i,\theta) = -X_iX_i^T$, whose expectation exists and is non-singular by assumption on Σ_X

(A4) From (A3), $\frac{\delta^2}{\delta\theta_j\theta_k} f(o,\theta) = 0$ for any j, k, o and is trivially continuous for all θ.

(A5) From (A4), the second partial derivatives are all bounded above by the constant function $g(o) = 1$.

Finally, the solution $\frac{1}{n}\sum_{i=1}^n X_i(Y_i - X_i^\intercal\hat{\theta}) = 0$ is unique because $\frac{1}{n}\sum_{i=1}^n X_iY_i = \frac{1}{n}\sum_{i=1}^n X_iX_i^\intercal\hat{\theta}$ and so long as $\frac{1}{n}\sum_{i=1}^n X_iX_i^\intercal$ is invertible for every n, we have $\hat{\theta} = (\sum_{i=1}^n X_iX_i^\intercal)^{-1}\sum_{i=1}^n X_iY_i$.

3 Z Estimators of Causal Estimands

3.1 Basic Idea

We can construct a Z estimator of the ATE as follows. Suppose we pretend for a moment that we actually know the true $\mu_a(X)$. Then, a natural estimator of the ATE, denoted as $\hat{\theta}$, is simply the sample equivalent of $\mathbb{E}[\mu_1(X) - \mu_0(X)]$ or

$$\frac{1}{n}\sum_{i=1}^n \mu_1(X_i) - \mu_0(X_i) - \hat{\theta} = 0, \quad f(O_i,\theta) = \mu_1(X_i) - \mu_0(X_i) - \theta,\ m = d = 1. \tag{3}$$

Note that $O_i = X_i \in \mathbb{R}^p$. In other words, the estimator the ATE is equivalent to the sample mean of $\mu_1(X_i) - \mu_0(X_i)$. Then, applying Theorem 1, we arrive at the following corollary

Corollary 1 (Asymptotic Normality Under Known μ_a)**.** *Suppose the function $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$ is known a priori. Let $\hat{\theta} = n^{-1}\sum_{i=1}^n \mu_1(X_i) - \mu_0(X_i)$ and $\theta^* = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)]$, which also equal the ATE under SUTVA and strong ignorability. If $\mu_1(X_i) - \mu_0(X_i)$ have finite second moments, we have*

$$\sqrt{n}(\hat{\theta} - \theta^*) \to N\left(0, \text{Var}[\mu_1(X_i) - \mu_0(X_i)]\right).$$

Proof. This is a direct consequence of the sample mean example in Section 2.3.1 where $\mu_1(X_i) - \mu_0(X_i)$ is the new O_i. □

Also, let $e(X_i) = P(A_i = 1 \mid X_i)$ be the propensity score and suppose this function is known; this would be the case in a randomized experiment. Consider the following estimator of the ATE, sometimes referred to as the inverses probability weighted (IPW) estimator:

$$\frac{1}{n}\sum_{i=1}^n \frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)} - \hat{\theta} = 0, \quad f(O_i,\theta) = \frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)} - \theta,\ m = d = 1 \tag{4}$$

Note that $O_i = (Y_i, A_i, X_i)$. We can apply Theorem 1 and arrive at the following:

Corollary 2 (Asymptotic Normality Under Known Propensity Score)**.** *Suppose the function $\pi(x) = P(A_i = 1 \mid X_i = x)$ is known a priori and $0 < \pi(x) < 1$. Let $\hat{\theta} = n^{-1}\sum_{i=1}^n \frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}$. If $\mathbb{E}[Y^2 \mid A = a, X]$ has finite second moments for $a = 0, 1$, we have*

$$\sqrt{n}(\hat{\theta}-\theta^*) \to N\left(0, \mathbb{E}\left[\frac{\mathbb{E}[Y_i^2 \mid A_i = 1, X]}{e(X_i)}\right] + \mathbb{E}\left[\frac{\mathbb{E}[Y_i^2 \mid A_i = 0, X]}{1-e(X_i)}\right] - (\mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]])^2\right).$$

where $\theta^* = \mathbb{E}[\mathbb{E}[Y_i \mid A_i = 1, X_i] - \mathbb{E}[Y_i \mid A_i = 1, X_i]]$

Proof. We go through each of the conditions in Theorem 1 below.

(A1) We see that

$$\mathbb{E}\left[\frac{Y_iA_i}{e(X_i)}\right] = \mathbb{E}\left[\frac{1}{e(X_i)}\mathbb{E}[Y_iA_i \mid X_i]\right] = \mathbb{E}\left[\frac{1}{e(X_i)}\mathbb{E}[Y_i \mid X_i, A_i = 1]P(A_i = 1 \mid X_i)\right] = \mathbb{E}[\mathbb{E}[Y_i \mid X_i, A_i = 1]]$$

A similar logic reveals $\mathbb{E}\left[\frac{Y_i(1-A_i)}{1-e(X_i)}\right] = \mathbb{E}[\mathbb{E}[Y_i \mid X_i, A_i = 0]]$. Then, the solution to the equation $\mathbb{E}[f(O_i, \theta^*)] = \mathbb{E}[\mathbb{E}[Y_i \mid X_i, A_i = 1]] - [\mathbb{E}[Y_i \mid X_i, A_i = 0]] - \theta^* = 0$ is equal to $\theta^* = \mathbb{E}[\mathbb{E}[Y_i \mid X_i, A_i = 1] - \mathbb{E}[Y_i \mid X_i, A_i = 0]]$

(A2) We have $\mathbb{E}[f(O_i, \theta^*)^2] = \text{Var}\left[\frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right]$, which is bounded by assumption.

(A3) We have $\frac{\delta}{\delta\theta}f(O_i, \theta) = -1$, which is non-singular.

(A4) The second partial derivative $\frac{\delta^2}{\delta^2\theta}f(O_i, \theta) = 0$, which is continuous for all θ

(A5) The second partial derivative is always bounded above by the constant function $g(o) = 1$

Finally, it's obvious that the solution to $\frac{1}{n}\sum_{i=1}^n f(O_i, \hat{\theta}) = 0$ is unique.

For the asymptotic variance, we have

$$\begin{aligned}
\text{Var}\left[\frac{Y_iA_i}{e(X_i)}\right] &= \mathbb{E}\left[\frac{Y_i^2A_i}{e^2(X_i)}\right] - \mathbb{E}\left[\frac{Y_iA_i}{e(X_i)}\right]^2 = \mathbb{E}\left[\frac{\mathbb{E}[Y_i^2 \mid A_i = 1, X]}{e(X_i)}\right] - \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]]^2 \\
\text{Var}\left[\frac{Y_i(1-A_i)}{1-e(X_i)}\right] &= \mathbb{E}\left[\frac{\mathbb{E}[Y_i^2 \mid A_i = 0, X]}{1-e(X_i)}\right] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]]^2 \\
\text{Cov}\left[\frac{Y_iA_i}{e(X_i)}, \frac{Y_i(1-A_i)}{1-e(X_i)}\right] &= -\mathbb{E}\left[\frac{Y_iA_i}{e(X_i)}\right]\mathbb{E}\left[\frac{Y_i(1-A_i)}{1-e(X_i)}\right] = -\mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] \cdot \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]]
\end{aligned}$$

Combining the above results, we get

$$\begin{aligned}
&\text{Var}\left[\frac{Y_iA_i}{e(X_i)} - \frac{Y_i(1-A_i)}{1-e(X_i)}\right] \\
=&\text{Var}\left[\frac{Y_iA_i}{e(X_i)}\right] + \text{Var}\left[\frac{Y_i(1-A_i)}{1-e(X_i)}\right] - 2\text{Cov}\left[\frac{Y_iA_i}{e(X_i)}, \frac{Y_i(1-A_i)}{1-e(X_i)}\right] \\
=&\mathbb{E}\left[\frac{\mathbb{E}[Y_i^2 \mid A_i = 1, X]}{e(X_i)}\right] + \mathbb{E}\left[\frac{\mathbb{E}[Y_i^2 \mid A_i = 0, X]}{1-e(X_i)}\right] - (\mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]])^2
\end{aligned}$$

□

3.2 Estimation of the ATE with Estimated Nuisance Functions

Now, consider a more realistic scenario where $\mu_a(X)$ is unknown and must be estimated. For each $A = a$, suppose we use OLS to estimate $\mu_a(X)$, which can be written as the following Z estimators

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i(Y_i - X_i^\intercal\hat{\beta} - \hat{\theta}) = 0 \\ \frac{1}{n}\sum_{i=1}^n A_iX_i(Y_i - X_i^\intercal\hat{\beta}_1) = 0 \end{pmatrix}, \quad f(O_i, (\beta_0, \beta_1)) = \begin{pmatrix} (1-A_i)X_i(Y_i - X_i^\intercal\beta_0) \\ A_iX_i(Y_i - X_i^\intercal\beta_1) \end{pmatrix}, \ m = d = 2p$$

We then plug in the predictions from the OLS estimator into equation (3). This plug-in estimator plus the OLS estimators of $\mu_a(\cdot)$ can be written as a Z estimator

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^\intercal\hat{\beta}_1 - X_i^\intercal\hat{\beta}_0 - \hat{\theta} = 0 \\ \frac{1}{n}\sum_{i=1}^n (1-A_i)X_i(Y_i - X_i^\intercal\hat{\beta}_0) = 0 \\ \frac{1}{n}\sum_{i=1}^n A_iX_i(Y_i - X_i^\intercal\hat{\beta}_1) = 0 \end{pmatrix}, \quad f(O_i, (\theta, \beta_0, \beta_1)) = \begin{pmatrix} X_i^\intercal\beta_1 - X_i^\intercal\beta_0 - \theta \\ (1-A_i)X_i(Y_i - X_i^\intercal\beta_0) \\ A_iX_i(Y_i - X_i^\intercal\beta_1) \end{pmatrix}, \ m = d = 2p \quad (5)$$

Here, $O_i = (Y_i, A_i, X_i)$ and the first element of the vector f is the plug-in estimator $\hat{\theta}$ based on the OLS estimates of $\mu_a(X_i)$. In other words, the only difference between equation (3) and the equation (5) is that we are taking an average of estimated $\mu_a(X_i)$.

The following corollary shows that $\hat{\theta}$ in equation 5 is asymptotically Normal.

Proposition 1. *Consider the estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i^\intercal\hat{\beta}_1 - X_i^\intercal\hat{\beta}_0$ where $\hat{\beta}_a$ is defined in equation 3. Suppose the following conditions hold:*

- *(a) (Y_i, X_i, A_i) follows the model*

$$Y_i = \begin{cases} X_i^\intercal \beta_0^* + \epsilon_{0,i}, \ \mathbb{E}[\epsilon_{0,i} \mid X_i, A_i = 0] = 0, \ \mathrm{Var}[\epsilon_{0,i} \mid X_i, A_i = 0] = (\sigma_0^*)^2 & \text{if } A_i = 0 \\ X_i^\intercal \beta_1^* + \epsilon_{1,i}, \ \mathbb{E}[\epsilon_{0,i} \mid X_i, A_i = 1] = 0, \ \mathrm{Var}[\epsilon_{0,i} \mid X_i, A_i = 1] = (\sigma_1^*)^2 & \text{if } A_i = 1 \end{cases}$$

  and the variances $(\sigma_a^*)^2$ are finite. Note that this assumption implicitly assumes positivity, i.e. $0 < P(A_i \mid X_i = x) < 1$ for every x
- *(b) The covariance matrix of X given $A = a$, denoted as $\Sigma_{X|a} = \mathrm{Cov}[X_i \mid A_i = a]$, is finite and is non-singular for each $a = 0, 1$*
- *(c) For every n, the matrices $\frac{1}{n}\sum i = 1^n A_i X_i X_i^\intercal$ and $\frac{1}{n}\sum_{i=1}^n (1 - A_i) X_i X_i^\intercal$ are invertible.*

Then, we have

$$\sqrt{n}(\hat{\theta} - \theta^*) \to N\left(0, \mathrm{Var}[X_i^\intercal(\beta_1^* - \beta_0^*)]\right).$$

where $\theta^* = \mathbb{E}[X_i^\intercal \beta_1^* - X_i^\intercal \beta_0^*]$

Proof. We show that the five conditions in Theorem 1 hold for the Z-estimator written in equation (5).

(A1) If we define $\theta^* = (\mathbb{E}[X_i^\intercal \beta_1^* - X_i^\intercal \beta_0^*], \beta_0^*, \beta_1^*)$, the first element of f is zero. The other parts of f become zero because

$$\begin{aligned}
\mathbb{E}[(1 - A_i) X_i (Y_i - X_i^\intercal \beta_0^*)] &= \mathbb{E}[X_i (Y_i - X_i^\intercal \beta_0) \mid A_i = 0] \mathbb{P}(A_i = 0) \\
&= \mathbb{E}[X_i (\mathbb{E}[Y_i \mid X_i, A_i = 0] - X_i^\intercal \beta_0^*) \mid A_i = 0] \mathbb{P}(A_i = 0) \\
&= \mathbb{E}[X_i (X_i^\intercal \beta_0^* - X_i^\intercal \beta_0^*) \mid A_i = 0] \mathbb{P}(A_i = 0) \\
&= 0,
\end{aligned}$$

and the same argument can show that $\mathbb{E}[A_i X_i (Y_i - X_i^\intercal \beta_1^*)] = 0$.

(A2) Let $q = \max((\sigma_0^*)^2, (\sigma_1^*)^2)$, which must be bounded. Then,

$$\begin{aligned}
&\mathbb{E}[\|f(O_i, \theta^*)\|_2^2] \\
=&\mathrm{Var}[X_i^\intercal(\beta_1^* - \beta_0^*)] + \mathbb{E}[\|X_i\|_2^2 (1 - A_i)(Y_i - X_i^\intercal \beta_0^*)^2] + \mathbb{E}[\|X_i\|_2^2 A_i (Y_i - X_i^\intercal \beta_1^*)^2] \\
=&\mathrm{Var}[X_i^\intercal(\beta_1^* - \beta_0^*)] + \mathbb{E}[\|X_i\|_2^2 (Y_i - X_i^\intercal \beta_0^*)^2 \mid A_i = 0] \mathbb{P}(A_i = 0) + \mathbb{E}[\|X_i\|_2^2 (Y_i - X_i^\intercal \beta_1^*)^2 \mid A_i = 1] \mathbb{P}(A_i = 1) \\
=&\mathrm{Var}[X_i^\intercal(\beta_1^* - \beta_0^*)] + \mathbb{E}[\|X_i\|_2^2 \mid A_i = 0](\sigma_0^*)^2 \mathbb{P}(A_i = 0) + \mathbb{E}[\|X_i\|_2^2 \mid A_i = 1](\sigma_1^*)^2 \mathbb{P}(A_i = 1)
\end{aligned}$$

Since $\Sigma_{X|a}$ is finite, the above term are bounded.

(A3) The partial derivatives with respect to θ, β_1, β_0 yield

$$\begin{aligned}
&\frac{\delta}{\delta\theta} f(O_i, \theta) \mid_{\theta=\theta^*} = (-1, \mathbf{0}) \in \mathbb{R}^{2p+1} \\
&\frac{\delta}{\delta\beta_{1,k}} f(O_i, \theta) \mid_{\theta=\theta^*} = (X_{ik}, -A_i X_i X_{ik}, \mathbf{0}) \in \mathbb{R}^{2p+1}, \quad k = 1, \dots, p \\
&\frac{\delta}{\delta\beta_{0,k}} f(O_i, \theta) \mid_{\theta=\theta^*} = (-X_{ik}, \mathbf{0}, -(1 - A_i) X_i X_{ik}) \in \mathbb{R}^{2p+1}, \quad k = 1, \dots, p.
\end{aligned}$$

This implies the gradient

$$\mathbb{E}[\nabla_\theta f(O_i, \theta) \mid_{\theta=\theta^*}] = \begin{pmatrix} -1 & \mathbb{E}[X_i] & -\mathbb{E}[X_i] \\ \mathbf{0} & -\mathbb{E}[A_i X_i X_i^\intercal] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbb{E}[(1 - A_i) X_i X_i^\intercal] \end{pmatrix} = \begin{pmatrix} -1 & \mathbb{E}[X_i] & -\mathbb{E}[X_i] \\ \mathbf{0} & -\Sigma_{X|1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\Sigma_{X|0} \end{pmatrix}$$

Using the property of determinants, the determinant of the above matrix is $-1 * \det(\Sigma_{X|1})\det(\Sigma_{X|0})$, which is non-zero by the non-singularity of the covariance matrices and thus, the expectation of the gradient is non-singular.

(A4) From (A3), all of the second partial derivatives must be zero and thus, is continuous in θ

(A5) From (A4), every second partial derivative is uniformly bounded above by the function $g(0) = 1$

Finally, for every n, the solutions to $\hat{\beta}_1$ and $\hat{\beta}_0$ are unique in the equation $\frac{1}{n}\sum_{i=1}^n f(O_i, (\hat{\theta}, \hat{\beta}_1, \hat{\beta}_0)) = 0$ based on the arguments in the linear regression example. If $\hat{\beta}_1$ and $\hat{\beta}_0$ are unique, then the solution $\hat{\theta}$ is also unique.

Finally, the asymptotic variance can be derived as follows. First, inverting the gradient gives us equal to

$$\mathbb{E}[\nabla_\theta f(O_i, \theta) \mid_{\theta=\theta^*}]^{-1} = \begin{pmatrix} -1 & \mathbb{E}[X_i] & -\mathbb{E}[X_i] \\ \mathbf{0} & -\Sigma_{X|1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\Sigma_{X|0} \end{pmatrix}^{-1} = \begin{pmatrix} -1 & -\mathbb{E}[X_i]\Sigma_{X|1}^{-1} & \mathbb{E}[X_i]\Sigma_{X|0}^{-1} \\ \mathbf{0} & -\Sigma_{X|1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\Sigma_{X|1}^{-1} \end{pmatrix}$$

Second, the inner matrix $\mathbb{E}[f(O_i, \theta^*) f(O_i, \theta^*)^\intercal]$ simplifies to

$$\mathbb{E}[f(O_i, \theta^*) f(O_i, \theta^*)^\intercal] = \begin{pmatrix} \mathrm{Var}[X_i^\intercal(\beta_1^* - \beta_0^*)] & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{X|1}(\sigma_1^*)^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{X|0}(\sigma_0^*)^2 \end{pmatrix}$$

The off-diagonal elements use the fact that

$$\begin{aligned} \mathbb{E}[(X_i^\intercal\beta_1^* - X_i^\intercal\beta_0^* - \theta^*)(1 - A_i)(Y_i - X_i^\intercal\beta_0^*)] &= \mathbb{E}[(X_i^\intercal\beta_1^* - X_i^\intercal\beta_0^* - \theta^*)\mathbb{E}[(Y_i - X_i^\intercal\beta_0^*) \mid X_i, A_i = 0]] = 0 \\ \mathbb{E}[(X_i^\intercal\beta_1^* - X_i^\intercal\beta_0^* - \theta^*)A_i(Y_i - X_i^\intercal\beta_1^*)] &= \mathbb{E}[(X_i^\intercal\beta_1^* - X_i^\intercal\beta_0^* - \theta^*)\mathbb{E}[(Y_i - X_i^\intercal\beta_1^*) \mid X_i, A_i = 1]] = 0 \\ \mathbb{E}[A_i(Y_i - X_i^\intercal\beta_1^*)(1 - A_i)(Y_i - X_i^\intercal\beta_0^*)] &= 0 \end{aligned}$$

Third, multiplying the matrices above and extracting the $(1, 1)$ element gives us the desired result. ☐

An interesting phenomena occurs for the IPW estimator if we use an estimated $\hat{e}(X_i)$ instead of the true $\hat{e}$.

Proposition 2. *Consider the IPW estimator* $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \frac{Y_i A_i}{\hat{e}(X_i)} - \frac{Y_i(1-A_i)}{1-\hat{e}(X_i)}$ *where* $\hat{e}(X_i) = \hat{p} = \frac{1}{n}\sum_{i=1}^n A_i$. *Suppose the following conditions hold: (a)* $\mathbb{E}[Y^2 \mid A = a]$ *is bounded for* $a = 0, 1$, *(b)* $\mathbb{E}[A_i]$ *is far from* 0 *and* 1, *and (c) for every* n, $\hat{p}$ *is far from* 0 *and* 1. *Let* $\theta^* = \mathbb{E}[\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]]$. *Let* σ^2 *be the asymptotic variance under Corollary 2 where we use the true* $e(X_i)$. *Then, we have*

$$\sqrt{n}(\hat{\theta} - \theta^*) \to N(0, \sigma^2 - q^2)$$

where $q^2 \geq 0$.

Proposition 2 shows that the asymptotic variance of the IPW estimator is less than or equal to the asymptotic variance of the IPW estimator with a known p^*. This does not occur with the estimator based on the outcome regression.

Proof. The corresponding Z estimator is

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \frac{Y_i A_i}{\hat{p}} - \frac{Y_i(1-A_i)}{1-\hat{p}} - \hat{\theta} = 0 \\ \frac{1}{n}\sum_{i=1}^n A_i - \hat{p} = 0 \end{pmatrix}, \quad f(O_i, \theta, p) = \begin{pmatrix} \frac{Y_i A_i}{p} - \frac{Y_i(1-A_i)}{1-p} - \theta \\ A_i - p \end{pmatrix}, \quad m = d = 2$$

We show that the five conditions in Theorem 1 hold.

(A1) The solution to $\mathbb{E}[f(O_i, \theta^*, p^*)] = 0$ exists and they are $p^* = \mathbb{E}[A_i]$ and $\theta^* = \mathbb{E}[\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]]$.

(A2) We have

$$\begin{aligned} \mathbb{E}[\|f(O_i, \theta^*, p^*)\|_2^2] &= \mathbb{E}\left[\left(\frac{Y_i A_i}{p^*} - \frac{Y_i(1 - A_i)}{1 - p^*} - \theta^*\right)^2\right] + \mathbb{E}[(A_i - p^*)^2] \\ &= \mathrm{Var}\left[\frac{Y_i A_i}{p^*} - \frac{Y_i(1 - A_i)}{1 - p^*}\right] + p^*(1 - p^*) \end{aligned}$$

The last expression is bounded above by assumption.

(A3) The first-order partial derivatives are

$$\begin{aligned} \frac{\delta}{\delta\theta} f(O_i, \theta, p) \mid_{\theta=\theta^*, p=p^*} &= \{-1, 0\} \\ \frac{\delta}{\delta p} f(O_i, \theta, p) \mid_{\theta=\theta^*, p=p^*} &= \left\{-1\left(\frac{Y_i A_i}{(p^*)^2} + \frac{Y_i(1 - A_i)}{(1 - p^*)^2}\right), -1\right\} \\ \mathbb{E}\left[\frac{\delta}{\delta p} f(O_i, \theta, p) \mid_{\theta=\theta^*, p=p^*}\right] &= \left\{-1\left(\frac{\mathbb{E}[\mathbb{E}[Y_i \mid A_i = 1]]}{p^*} + \frac{\mathbb{E}[\mathbb{E}[Y_i \mid A_i = 0]]}{(1 - p^*)}\right), -1\right\} \end{aligned}$$

Thus, $\mathbb{E}[\nabla_{\theta,\beta} f(O_i, \theta, p) \mid_{\theta=\theta^*, p=p^*}]$ exists and is non-singular.

(A4) All of the second partial derivatives are zero except

$$\frac{\delta}{\delta^2 p} f(O_i, \theta, p) = \left\{-2\left(\frac{Y_i(1 - A_i)}{(1 - p)^3} - \frac{Y_i A_i}{p^3}\right), 0\right\}$$

This exists and is continuous within a neighborhood of p^* that is far from 0 and 1.

(A5) All elements of the Hessian matrix is bounded above by $g(o) = 1$ except for the Hessian corresponding to the second partial derivatives of p. For that component, since $\mathbb{E}[A_i]$ is far from 0 and 1, there exists $\delta > 0$ such that $\delta < \mathbb{E}[A_i] < 1 - \delta$. Consider a function g such that $g(o) = 2 * Y_i((1 - A_i)(1 - \delta)^3 + A_i\delta^3)$. This g function satisfies

$$\left| -2\left(\frac{Y_i(1 - A_i)}{(1 - p)^3} - \frac{Y_iA_i}{p^3}\right)\right| \leq 2 * |Y_i| \cdot |(1 - A_i)(1 - \delta)^3 + A_i\delta^3|$$

Furthermore, we have $\mathbb{E}[|g(O_i)|] = 2\mathbb{E}[(1 - \delta)^3\mathbb{E}[|Y_i| \mid A_i = 0](1 - p^*) + \delta^*\mathbb{E}[|Y_i| \mid A_i = 1]p^*]$, which is bounded above by the finite moment assumption on Y_i given $A_i = a$.

We also guarantee that the solution is unique at every n by ensuring that $\hat{p}$ is far from 0 and 1.

For the asymptotic variance, we get

$$\mathbb{E}[\nabla_{\theta,p} f(O_i, \theta, p) \mid_{\theta=\theta^*,p=p^*}] = \begin{pmatrix} -1 & -1\left(\frac{\mathbb{E}[\mathbb{E}[Y_i|A_i=1]]}{p^*} + \frac{\mathbb{E}[\mathbb{E}[Y_i|A_i=0]]}{(1-p^*)}\right) \\ 0 & -1 \end{pmatrix}$$
$$\mathbb{E}[\nabla_{\theta,p} f(O_i, \theta, p) \mid_{\theta=\theta^*,p=p^*}]^{-1} = \begin{pmatrix} -1 & \left(\frac{\mathbb{E}[\mathbb{E}[Y_i|A_i=1]]}{p^*} + \frac{\mathbb{E}[\mathbb{E}[Y_i|A_i=0]]}{(1-p^*)}\right) \\ 0 & -1 \end{pmatrix}$$
$$\mathbb{E}[f(O_i, \theta^*, p^*)f(O_i, \theta^*, p^*)^\intercal] = \begin{pmatrix} \mathrm{Var}\left[\frac{Y_iA_i}{p^*} - \frac{Y_i(1-A_i)}{1-p^*}\right] & \mathbb{E}[Y_i \mid A_i = 1](1 - p^*) + \mathbb{E}[Y_i \mid A_i = 0]p^* \\ \mathbb{E}[Y_i \mid A_i = 1](1 - p^*) + \mathbb{E}[Y_i \mid A_i = 0]p^* & p^*(1 - p^*) \end{pmatrix}$$

Putting it all together and some painful algebra leads to

$$\begin{aligned} &\mathbb{E}[f(O_i, \theta^*, p^*)f(O_i, \theta^*, p^*)^\intercal]^{-1}\mathbb{E}[f(O_i, \theta^*, p^*)f(O_i, \theta^*, p^*)^\intercal]\mathbb{E}[f(O_i, \theta^*, p^*)f(O_i, \theta^*, p^*)^\intercal]^{-\intercal} \\ =&\mathrm{Var}\left[\frac{Y_iA_i}{p^*} - \frac{Y_i(1 - A_i)}{1 - p^*}\right] - \frac{(\mathbb{E}[Y \mid A = 1](1 - p^*) + \mathbb{E}[Y \mid A = 0]p^*)^2}{p^*(1 - p^*)} \end{aligned}$$

□