Additional Explanations for Athey and Imbens (2016)

Wooyong Park and Jeongwoo Yang

May 28, 2025

Some equalities in Athey and Imbens (2016) are challenging to understand at a single glance. In this guide, I add some comments for certain equations and properties to ease the comprehension.

Table of Equations

	No.	Equation
	HIPA-1	$-\text{EMSE}_\mu(\Pi) = -\mathbb{E}_{(Y_i,X_i),S^{est}}\left[(Y_i - \mu(X_i;\Pi))^2 - Y_i^2\right] - \mathbb{E}_{X_i,S^{est}}\left[(\hat{\mu}(X_i;S^{est},\Pi) - \mu(X_i;\Pi))^2\right]$
	HIPA-2	$-\mathbb{E}_{(Y_i,X_i),S^{est}}\left[(Y_i - \mu(X_i;\Pi))^2 - Y_i^2\right] - \mathbb{E}_{X_i,S^{est}}\left[(\hat{\mu}(X_i;S^{est},\Pi) - \mu(X_i;\Pi))^2\right] = \mathbb{E}_{X_i}[\mu^2(X_i;\Pi)] - \mathbb{E}_{S^{est},X_i}[\mathbb{V}(\hat{\mu}^2(X_i;S^{est},\Pi))]$
	HIPA-3	$\hat{\mathbb{E}}(\mu^2(x;\Pi)) = \hat{\mu}^2(x;S^{tr},\Pi) - \frac{S^2_{S^{tr}}(l(x;\Pi))}{N^{tr}_{(l(x;\Pi))}}$
	HIPA-4	$-\text{MSE}_\mu(S^{tr},S^{tr},\Pi) = \frac{1}{N^{tr}}\sum_{i\in S^{tr}}\hat{\mu}^2(X_i;S^{tr},\Pi)$
	HITE-1	$MSE_\mu(S^{te},S^{tr},\Pi) = -\frac{1}{N^{tr}}\sum_{i\in S^{te}}\hat{\mu}(X_i;S^{te},\Pi)\cdot\hat{\mu}(X_i;S^{tr},\Pi) + \frac{1}{N^{tr}}\sum_{i\in S}\hat{\mu}^2(X_i;S^{tr},\Pi)$
	HITE-2	$\hat{MSE}_\tau(S^{te},S^{tr},\Pi) = -\frac{1}{N^{tr}}\sum_{i\in S^{te}}\hat{\tau}(X_i;S^{te},\Pi)\cdot\hat{\tau}(X_i;S^{tr},\Pi) + \frac{1}{N^{tr}}\sum_{i\in S^{te}}\hat{\tau}^2(X_i;S^{tr},\Pi)$

Contents

1 Honest Splitting

1.1 Equation HIPA-1

By definition,

$$\text{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{(Y_i-\hat{\mu}(X_i;S^{est},\Pi))^2-Y_i^2\right\}\right] \tag{1}$$

Then,

$$\begin{aligned}
\text{EMSE}_\mu(\Pi) &= \mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{Y_i^2-2\hat{\mu}(X_i;S^{est},\Pi)Y_i+\hat{\mu}^2(X_i;S^{est},\Pi)-Y_i^2\right\}\right] \\
&= \mathbb{E}_{S^{te}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{Y_i^2-2\mu(X_i;\Pi)Y_i+\mu^2(X_i;\Pi)-Y_i^2\right\}\right] \\
&+ \mathbb{E}_{S^{te}.S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{2Y_i\big(\mu(X_i;\Pi)-\hat{\mu}(X_i;S^{est},\Pi)\big)\right\}\right] \\
&+ \mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{\hat{\mu}^2(X_i;S^{est},\Pi)-\mu^2(X_i;\Pi)\right\}\right] \\
&= \mathbb{E}_{S^{te}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{\big(Y_i-\mu(X_i;\Pi)\big)^2-Y_i^2\right\}\right] \\
&+ \mathbb{E}_{S^{te},S^{est}}\left[2Y_i\cdot\mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\big(\mu(X_i;\Pi)-\hat{\mu}(X_i;S^{est},\Pi)\big)|Y_i;i\in S^{te}\right]\right] \qquad (2) \\
&+ \mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\left\{\big(\hat{\mu}(X_i;S^{est},\Pi)-\mu(X_i;\Pi)\big)^2\right\}\right]+2\mu(X_i;\Pi)\mathbb{E}_{S^{te},S^{est}}\left[\hat{\mu}(X_i;S^{est},\Pi)-\mu(X_i;\Pi)\right] \qquad (3)
\end{aligned}$$

In 2, however,

$$\mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\big(\mu(X_i;\Pi)-\hat{\mu}(X_i;S^{est},\Pi)\big)|Y_i;i\in S^{te}\right]$$

$$=\mathbb{E}_{S^{te},S^{est}}\left[\frac{1}{N^{te}}\sum_{i\in S^{te}}\big(\mu(X_i;\Pi)-\hat{\mu}(X_i;S^{est},\Pi)\big)\right]=0$$

since the prediction $\hat{\mu}$ is made independently of the test set and $\hat{\mu}$ is an unbiased estimator of μ. For the same reason($\mathbb{E}_S[\hat{\mu}(x;S,\Pi)]=\mu(x;\Pi)$), the latter term of 3 is equal to zero, which gives us

$$\text{EMSE}_\mu(\Pi)=\mathbb{E}_{(Y_i,X_i),S^{est}}\left[(Y_i-\mu(X_i;\Pi))^2-Y_i^2\right]+\mathbb{E}_{X_i,S^{est}}\left[(\hat{\mu}(X_i;S^{est},\Pi)-\mu(X_i;\Pi))^2\right] \tag{4}$$

1.2 Equation HIPA-2

In this equation, the term $\mathbb{E}_{S^{est},X_i}[\mathbb{V}(\hat{\mu}^2(X_i;S^{est},\Pi))]$ seems like a typo. It should be $\mathbb{E}_{S^{est},X_i}[\mathbb{V}(\hat{\mu}(X_i;S^{est},\Pi))]$.

The first question you might have with this equation is "Why do we need an expectation over a parameter, $\mathbb{V}$?" One must know that $\mathbb{V}(\hat{\mu}(X_i;S^{est},\Pi))$ varies across different values of X_i. This is because $\mathbb{V}(\hat{\mu}(X_i;S^{est},\Pi))$ can be interpreted as $\mathbb{V}(\hat{\mu}(X_i|l(X_i;\Pi)))$, and thus the variance depends on the leaf $l(X_i;\Pi)$ and the estimation set, S^{te}. We take X_i and Π as given in $\mathbb{V}(\hat{\mu}(X_i|l(X_i;\Pi)))$, and then take the expectation of it across different X_i values given Π. We will have more discussions about $\mathbb{E}_{S^{est},X_i}[\mathbb{V}(\hat{\mu}(X_i;S^{est},\Pi))]$ in the next section. For now, we must show that

- $-\mathbb{E}_{(Y_i,X_i),S^{est}}\left[(Y_i-\mu(X_i;\Pi))^2-Y_i^2\right]=\mathbb{E}_{X_i}[\mu^2(X_i;\Pi)]$
- $-\mathbb{E}_{X_i,S^{est}}\left[(\hat{\mu}(X_i;S^{est},\Pi)-\mu(X_i;\Pi))^2\right]=-\mathbb{E}_{S^{est},X_i}[\mathbb{V}(\hat{\mu}(X_i;S^{est},\Pi))]$
  - To add a little on the above equation, the honest approach plays a key role in applying the definition. The unbiasedness of $\hat{\mu}$ is not guaranteed when the same sample is used for both training Π and estimating for each leaf. That is, $\mathbb{E}[\hat{\mu}(X_i;\mathcal{S}^{\text{est}},\Pi)]=\mu(X_i;\Pi)$ due to the fact that $\mathcal{S}^{\text{est}}$ and Π are independent

The second equality is direct from the definition of variance. For the first equality,

$$\begin{aligned}
-\mathbb{E}_{(Y_i,X_i),S^{est}}\left[(Y_i-\mu(X_i;\Pi))^2-Y_i^2\right] &= -\mathbb{E}_{(Y_i,X_i)}\left[Y_i^2-2Y_i\cdot\mu(X_i;\Pi)+\mu^2(X_i;\Pi)-Y_i^2\right] \\
&= -2\mathbb{E}_{(Y_i,X_i)}\left[\mu(X_i;\Pi)\mathbb{E}_{S^{est}}[Y_i|l(X_i;\Pi)]\right]+\mathbb{E}_{(Y_i,X_i)}[\mu^2(X_i;\Pi)] \\
&= -2\mathbb{E}_{X_i}\left[\mu^2(X_i;\Pi)\right]+\mathbb{E}_{X_i}\left[\mu^2(X_i;\Pi)\right] \\
&= -\mathbb{E}_{X_i}\left[\mu^2(X_i;\Pi)\right]
\end{aligned}$$

1.3 Equation HIPA-3

$\hat{\mathbb{E}}(\mu^2(x;\Pi))=\hat{\mu}^2(x;S^{tr},\Pi)-\frac{S^2_{S^{tr}}(l(x;\Pi))}{N^{tr}(l(x;\Pi))}$ is a direct result of $\mathbb{V}(X)=\mathbb{E}(X^2)-[\mathbb{E}(X)]^2$ for any random variable with finite variance - applying the rule to $\hat{\mu}$ will give the desired result.

1.4 Equation HIPA-4

We want to prove

$$-\text{MSE}_\mu(S^{tr},S^{tr},\Pi)=\frac{1}{N^{tr}}\sum_{i\in S^{tr}}\hat{\mu}^2(X_i;S^{tr},\Pi) \tag{5}$$

Equation 5 requires some extra illustration. To begin with, the MSE is defined as equation 6 since Athey and Imbens (2016) uses an adjusted MSE, which is equal to the conventional MSE subtracted by $\frac{1}{\#(S)}\sum_{i\in S} Y_i^2$.

$$-\text{MSE}_\mu(S^{tr}, S^{tr}, \Pi) \equiv -\frac{1}{N^{tr}} \sum_{i\in S^{tr}} \left\{ (Y_i - \hat{\mu}(X_i; S^{tr}, \Pi))^2 - Y_i^2 \right\} \tag{6}$$

$$= -\frac{1}{N^{tr}} \sum_{i\in S^{tr}} \left\{ Y_i^2 - Y_i^2 - 2\hat{\mu}(X_i; S^{tr}, \Pi) Y_i + \hat{\mu}^2 \right\}$$

$$= \frac{1}{N^{tr}} \sum_{i\in S^{tr}} \left\{ 2\hat{\mu}(X_i; S^{tr}, \Pi) Y_i - \hat{\mu}^2(X_i; S^{tr}, \Pi) \right\} \tag{7}$$

Recall that $\hat{\mu}(X_i; S^{tr}, \Pi))^2$'s are equal within the same leaf, and the estimation $\hat{\mu}$ is made by the sample mean of outcomes in each leaf from the training sample. Let $\Pi = \{l_1, l_2, \ldots, l_{\#(\Pi)}\}$ with $\cup_{j=1}^{\#(\Pi)} l_j = \mathbb{X}$. Also, let $\hat{\mu}_{l_j}(S^{tr}) = \frac{1}{N^{l_j}} \sum_{i\in l_j} Y_i$

Thus,

$$\frac{1}{N^{tr}} \sum_{i\in S^{tr}} \hat{\mu}(X_i; S^{tr}, \Pi) Y_i = \frac{1}{N^{tr}} \sum_{j=1}^{\#(\Pi)} N^{l_j} \cdot \hat{\mu}_{l_j}^2(S^{tr}) \tag{8}$$

$$= \frac{1}{N^{tr}} \sum_{i\in S^{tr}} \hat{\mu}^2(X_i; S^{tr}, \Pi) \tag{9}$$

Applying the result above to equation 7 results in equation 5.

2 Honest Inference for Treatment Effects

2.1 Equation HITE-1

In this equation, we suspect there are multiple typos; our revised version is as below:

$$MSE_\mu(S^{te}, S^{tr}, \Pi) = -\frac{1}{N^{te}} \sum_{i\in S^{te}} \hat{\mu}(X_i; S^{te}, \Pi) \cdot \hat{\mu}(X_i; S^{tr}, \Pi) + \frac{1}{N^{te}} \sum_{i\in S^{te}} \hat{\mu}^2(X_i; S^{tr}, \Pi) \tag{10}$$

Letting $\Pi = \{l_1, l_2, \ldots, l_{\#(\Pi)}\}$,

$$\hat{\mu}(X_i; S^{te}, \Pi) = \sum_{k=1}^{\#(\Pi)} \mathbb{I}(i \in l_k) \frac{1}{\#(l_k)^{te}} \sum_{j\in l_k; S^{te}} Y_j \tag{11}$$

where $\mathbb{I}(i \in l_k)$ is equal to 1 if $i \in l_k$ and 0 otherwise; and $\#(l_k)^{te}$ is the number of test samples that fall into the leaf (l_k). In plain words, this means that $\hat{\mu}(X_i; S^{te}, \Pi)$ is the sample mean of the test samples in the leaf where i belongs.

Since

$$
\begin{aligned}
MSE_{\mu}(S^{te}, S^{tr}, \Pi) &\equiv \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ (Y_i - \hat{\mu}(X_i; S^{tr}, \Pi))^2 - Y_i^2 \right\} \\
&= \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ -2\hat{\mu}(X_i; S^{tr}, \Pi) Y_i + \hat{\mu}^2(X_i; S^{tr}, \Pi) \right\}
\end{aligned}
$$

it suffices to show

$$
\frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left\{ -2\hat{\mu}(X_i; S^{tr}, \Pi) Y_i \right\} = -\frac{2}{\#(S^{te})} \sum_{i \in S^{te}} \hat{\mu}(X_i; S^{te}, \Pi) \cdot \hat{\mu}(X_i; S^{tr}, \Pi) \tag{12}
$$

Slightly abusing the notation, let $\hat{\mu}(l_k; S^{tr}, \Pi)$ denote $\hat{\mu}(l_k; S^{tr}, \Pi)$ for an individual that falls into leaf l_k. $\hat{\mu}(l_k; S^{te}, \Pi)$ is also defined similarly.

$$
\begin{aligned}
\sum_{i \in S^{te}} \left\{ -2\hat{\mu}(X_i; S^{tr}, \Pi) Y_i \right\} &= \sum_{k=1}^{\#(\Pi)} \left\{ \sum_{i \in l_k} -2\hat{\mu}(X_i; S^{tr}, \Pi) Y_i \right\} && (13) \\
&= \sum_{k=1}^{\#(\Pi)} \left\{ -2\hat{\mu}(l_k, S^{tr}, \Pi) \cdot \#(l_k)^{te} \sum_{i \in l_k} \frac{1}{\#(l_k)^{te}} Y_i \right\} && (14) \\
&= \sum_{k=1}^{\#(\Pi)} \left\{ -2\hat{\mu}(l_k, S^{tr}, \Pi) \cdot \#(l_k)^{te} \cdot \hat{\mu}(l_k; S^{te}, \Pi) \right\} && (15) \\
&= -2 \sum_{i \in S^{te}} \left\{ \hat{\mu}(X_i; S^{tr}, \Pi) \cdot \hat{\mu}(X_i; S^{te}, \Pi) \right\} && (16)
\end{aligned}
$$

2.2 Equation HITE-2

The derivation in section 2.1 to $\hat{\tau}$ under the fact that

$$
E_{S^{te}}[\tau_i | i \in S^{te} : i \in l(x, \Pi)] = E_{S^{te}}[\hat{\tau}(x; S^{te}, \Pi)]
$$

gives us the result directly.

Why $\hat{\text{MSE}}_{\tau}$, not MSE_{τ}?

This is puzzling. MSE, in the orthodox definition, is a function of random variables.

That is,

$$
MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
$$

Here, both Y_i and $\hat{Y}_i$ are random variables.

The difference between τ_i and Y_i is that the former is unobservable. In that aspect, τ_i is similar to a parameter, although it is a random variable defined as $Y_i(1) - Y_i(0)$. Our guess about the authors' intention is that they used the term $\hat{MSE}_{\tau}$ instead of MSE_{τ} to acknowledge the unobservability of τ_i.

- Technically speaking, we can ask if below equation holds without taking the expectation

$$\sum_{i \in l_k} \frac{1}{\#(l_k)^{\text{te}}} \tau_i = \hat{\tau}(l_k; \mathcal{S}^{\text{te}}, \Pi)$$

  Recall the definition of $\hat{\tau}(x; \mathcal{S}, \Pi) = \hat{\mu}(1, x; \mathcal{S}, \Pi) - \hat{\mu}(0, x; \mathcal{S}, \Pi)$. Now we can explicitly see that the analogous equality of (14) to τ does not hold without the expectation. While the estimator $\hat{\mu}$ is defined directly with Y_i, the estimator $\hat{\tau}$ is not defined with τ_i. Hence the fact $E_{S^{te}}[\tau_i | i \in S^{te} : i \in l(x, \Pi)] = E_{S^{te}}[\hat{\tau}(x; S^{te}, \Pi)]$ plays a critical role in keeping $\hat{MSE}_\tau$ an unbiased estimator of MSE_τ.

A Appendix: Typos

- (p.10) $-\text{EMSE}_\tau(\Pi) = \mathbb{E}[\tau^2(X_i; \Pi)] - \mathbb{E}_{\mathcal{S}^{\text{est}}, X_i}[\mathbb{V}(\hat{\tau}^2(X_i; \mathcal{S}^{\text{est}}, \Pi))]$

  $\rightarrow$ $-\text{EMSE}_\tau(\Pi) = \mathbb{E}[\tau^2(X_i; \Pi)] - \mathbb{V}_{\mathcal{S}^{\text{est}}, X_i}[\hat{\tau}(X_i; \Pi, \mathcal{S}^{\text{est}})]$

References

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.