Doubly Robust Estimator for ATE

João Dimas
September 2025

Derivation of the Doubly Robust Estimator for the Average Treatment Effect (ATE)

1. Introduction and Objective

This document provides a step-by-step derivation of the doubly robust (DR) estimator for the Average Treatment Effect (ATE). The primary goal is to show how the final, celebrated formula is constructed by combining two simpler identification strategies: Outcome Regression (OR) and Inverse Probability Weighting (IPW).

The practical advantage of the DR approach is not that it solves omitted variable bias from unobserved confounders; if the unconfoundedness assumption fails, all three identification strategies are inconsistent. Instead, its strength lies in providing insurance against functional form misspecification of the nuisance models. For instance, if the true relationship between covariates and the outcome is non-linear, a simple linear regression for the outcome model would be misspecified. Similarly, logistic regression for the propensity score might be misspecified if the true relationship is more complex. A DR estimator remains consistent if either the propensity score model is correctly specified or the outcome regression is correctly specified (meaning both $\mu_0(X)$ and $\mu_1(X)$ are correctly specified), a property neither the OR nor IPW estimator possesses alone.

2. Setup: Parameter and Assumptions

Our target parameter is the **Average Treatment Effect (ATE)**, defined within the potential outcomes framework as:

$$\tau_{ATE} \equiv \mathbb{E}[Y(1) - Y(0)]$$

where $Y(d)$ is the potential outcome for an individual under treatment status $d \in \{0, 1\}$ and $D \in \{0, 1\}$ is the binary treatment indicator.

Core Assumptions:

1. **SUTVA (Stable Unit Treatment Value Assumption):** (i) The observed outcome is a function of the individual's own treatment status: $Y = D\,Y(1) + (1 - D)\,Y(0)$ (consistency), and (ii) there is no interference across units (one unit's potential outcomes are unaffected by other units' treatment assignments).
2. **Ignorability (Unconfoundedness):** The potential outcomes are jointly independent of treatment assignment, conditional on a set of observable covariates X: $(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$.

3. **Positivity (Strong Overlap):** For all relevant values of X, the probability of treatment is bounded away from 0 and 1: $0 < \epsilon \leq \pi(X) \leq 1 - \epsilon < 1$ almost surely for some $\epsilon > 0$.
4. **Random Sampling and Finite Moments:** The sample $\{(Y_i, D_i, X_i)\}_{i=1}^n$ is i.i.d., and $\mathbb{E}[Y^2] < \infty$. Combined with positivity, this ensures finite variance of the IPW and DR estimators.

Nuisance Functions:

Our identification relies on the following conditional expectation functions:

- **Propensity Score:** $\pi(X) \equiv \Pr(D = 1 \mid X)$
- **Outcome Model (Control):** $\mu_0(X) \equiv \mathbb{E}[Y \mid D = 0, X]$
- **Outcome Model (Treated):** $\mu_1(X) \equiv \mathbb{E}[Y \mid D = 1, X]$

Note that the ignorability assumption provides the bridge between potential outcome expectations $\mathbb{E}[Y(a) \mid X]$ and these observed conditional expectations $\mu_a(X)$.

3. Foundational Identification Strategies

3.1 Identification via Outcome Regression (OR)

The OR strategy uses the unconfoundedness assumption to identify the ATE by integrating the difference in conditional outcome models over the full distribution of covariates.

$$\begin{aligned}
\tau_{ATE} &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X]] \quad (\text{by law of iterated expectations}) \\
&= \mathbb{E}[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]] \quad (\text{by linearity of expectation}) \\
&= \mathbb{E}[\mathbb{E}[Y(1) \mid D = 1, X] - \mathbb{E}[Y(0) \mid D = 0, X]] \quad (\text{by ignorability}) \\
&= \mathbb{E}[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]] \quad (\text{by SUTVA}) \\
&= \mathbb{E}[\mu_1(X) - \mu_0(X)] \quad (\text{ by definition of } \mu_1(X) \text{ and } \mu_0(X))
\end{aligned}$$

This gives the OR estimand. Note that this identification requires positivity to ensure that the conditional expectations $\mu_1(X) = \mathbb{E}[Y \mid D = 1, X]$ and $\mu_0(X) = \mathbb{E}[Y \mid D = 0, X]$ can be identified from observed data. The consistency of an estimator based on it depends on the correct specification of the outcome model for both $\mu_1(X)$ and $\mu_0(X)$.

$$\tau_{ATE}^{OR} = \mathbb{E}[\mu_1(X) - \mu_0(X)]$$

3.2 Identification via Inverse Probability Weighting (IPW)

The IPW strategy uses the propensity score to reweight the observed outcomes, creating a reweighted population where covariate distributions are balanced between treatment groups in expectation. The independence assumption (ignorability) is used to link potential outcome expectations to observed conditional expectations, while the reweighting identities themselves require only the definition of the propensity score. We derive this step-by-step by showing how reweighting creates unbiased estimates of the potential outcome expectations.

Starting from the ATE definition, we need to identify both $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$:

$$\tau_{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Identifying $\mathbb{E}[Y(1)]$:

$$\begin{aligned}\mathbb{E}[Y(1)] &= \mathbb{E}[\mathbb{E}[Y(1) \mid X]] \quad \text{(by law of iterated expectations)} \\ &= \mathbb{E}[\mathbb{E}[Y(1) \mid D=1, X]] \quad \text{(by ignorability)} \\ &= \mathbb{E}[\mu_1(X)] \quad \text{(by definition of } \mu_1(X) \text{ and SUTVA)}\end{aligned}$$

To express this using unconditional expectations, we can show that:

$$\mathbb{E}[\mu_1(X)] = \mathbb{E}[\mathbb{E}[Y \mid D=1, X]] = \mathbb{E}\left[\frac{DY}{\pi(X)}\right]$$

The idea is that reweighting the treated observations by $\frac{1}{\pi(X)}$ gives the correct expectation. Importantly, the reweighting identity $\mathbb{E}[DY/\pi(X)] = \mathbb{E}[\mu_1(X)]$ does not require ignorability, as it follows purely from the definition of the propensity score. Ignorability was used earlier to establish $\mathbb{E}[Y(1)] = \mathbb{E}[\mu_1(X)]$. The identity $\mathbb{E}[DY \mid X] = \pi(X)\,\mathbb{E}[Y \mid D=1, X]$ is proven below:

$$\begin{aligned}\mathbb{E}\left[\frac{DY}{\pi(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{DY}{\pi(X)}\middle| X\right]\right] \quad \text{(law of iterated expectations)} \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}[DY \mid X]\right] \quad \text{(functions of X are constant conditional on X)} \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}\big[\mathbb{E}[DY \mid D, X] \mid X\big]\right] \quad \text{(law of iterated expectations)} \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}\big[D\,\mathbb{E}[Y \mid D, X] \mid X\big]\right] \quad \text{(since D is measurable given } (D, X)) \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\Pr(D=1 \mid X)\,\mathbb{E}[Y \mid D=1, X]\right] \quad \text{(definition of conditional expectation)} \\ &= \mathbb{E}\left[\frac{1}{\pi(X)}\pi(X)\mu_1(X)\right] \quad \text{(by definitions)} \\ &= \mathbb{E}[\mu_1(X)]\end{aligned}$$

Identifying $\mathbb{E}[Y(0)]$:

By a parallel argument:

$$\begin{aligned}\mathbb{E}[Y(0)] &= \mathbb{E}[\mathbb{E}[Y(0) \mid X]] \quad \text{(by law of iterated expectations)} \\ &= \mathbb{E}[\mathbb{E}[Y(0) \mid D=0, X]] \quad \text{(by ignorability)} \\ &= \mathbb{E}[\mu_0(X)] \quad \text{(by definition of } \mu_0(X) \text{ and SUTVA)}\end{aligned}$$

And we can show that:

$$\mathbb{E}[\mu_0(X)] = \mathbb{E}\left[\frac{(1-D)Y}{1-\pi(X)}\right]$$

This follows from a similar proof. The reweighted control observations $\frac{(1-D)Y}{1-\pi(X)}$ give the correct expectation:

$$\begin{aligned}\mathbb{E}\left[\frac{(1-D)Y}{1-\pi(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{(1-D)Y}{1-\pi(X)}\middle| X\right]\right] \quad \text{(law of iterated expectations)} \\ &= \mathbb{E}\left[\frac{1}{1-\pi(X)}\mathbb{E}[(1-D)Y \mid X]\right] \quad \text{(functions of X are constant conditional on X)} \\ &= \mathbb{E}\left[\frac{1}{1-\pi(X)}\mathbb{E}\big[(1-D)\,\mathbb{E}[Y \mid D, X] \mid X\big]\right] \quad \text{(law of iterated expectations)} \\ &= \mathbb{E}\left[\frac{1}{1-\pi(X)}\Pr(D=0 \mid X)\,\mathbb{E}[Y \mid D=0, X]\right] \quad \text{(definition of conditional expectation)} \\ &= \mathbb{E}\left[\frac{1}{1-\pi(X)}(1-\pi(X))\mu_0(X)\right] \quad \text{(by definitions)} \\ &= \mathbb{E}[\mu_0(X)]\end{aligned}$$

Combining the Results:

Substituting the identified expressions for both potential outcome expectations back into the original ATE definition gives the IPW estimand:

$$\tau_{ATE}^{IPW} = \mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\right]$$

This expression identifies τ_{ATE} under ignorability and positivity. The consistency of an estimator based on it depends on the correct specification of the propensity score model for $\pi(X)$.

4. Derivation of the Doubly Robust Estimator

The DR estimator, also known as the Augmented IPW (AIPW) estimator, improves upon the IPW formula by adding a correction term. This term is designed to have an expectation of zero if the propensity score is correct, but it simultaneously corrects for errors if the outcome models are correct instead.

Step 1: Define a Zero-Expectation Term

Consider the following expression, which involves the outcome models:

$$\mathbb{E}\left[\left(1 - \frac{D}{\pi(X)}\right)\mu_1(X) - \left(1 - \frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]$$

Step 2: Prove the term has mean zero when $\pi(X) = \mathbb{E}[D \mid X]$ and $0 < \pi(X) < 1$ a.s.

We need to show that this expression has an expectation of zero if the propensity score $\pi(X)$ is correctly specified. We'll prove this by applying the law of iterated expectations and carefully working through each component.

Starting with the original expression:

$$\mathbb{E}\left[\left(1 - \frac{D}{\pi(X)}\right)\mu_1(X) - \left(1 - \frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]$$

We can split this into two separate terms and analyze each:

Term 1: $\mathbb{E}\left[\left(1 - \frac{D}{\pi(X)}\right)\mu_1(X)\right]$

Applying the law of iterated expectations:

$$\begin{aligned}
\mathbb{E}\left[\left(1 - \frac{D}{\pi(X)}\right)\mu_1(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(1 - \frac{D}{\pi(X)}\right)\mu_1(X)\middle|X\right]\right] \quad \text{(law of iterated expectations)} \\
&= \mathbb{E}\left[\mu_1(X)\mathbb{E}\left[1 - \frac{D}{\pi(X)}\middle|X\right]\right] \quad \text{(functions of X are constant conditional on X)} \\
&= \mathbb{E}\left[\mu_1(X)\left(1 - \frac{\mathbb{E}[D|X]}{\pi(X)}\right)\right] \quad \text{(Linearity of conditional expectation)} \\
&= \mathbb{E}\left[\mu_1(X)\left(1 - \frac{\pi(X)}{\pi(X)}\right)\right] \quad \text{(by definition of } \pi(X) = \mathbb{E}[D|X]\text{)} \\
&= \mathbb{E}\left[\mu_1(X)\left(1-1\right)\right] = \mathbb{E}[0] = 0
\end{aligned}$$

Term 2: $\mathbb{E}\left[\left(1 - \frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]$

Similarly, applying the law of iterated expectations:

$$\begin{aligned}
\mathbb{E}\left[\left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\middle| X\right]\right] \quad \text{(law of iterated expectations)} \\
&= \mathbb{E}\left[\mu_0(X)\mathbb{E}\left[1-\frac{1-D}{1-\pi(X)}\middle| X\right]\right] \quad \text{(functions of X are constant conditional on X)} \\
&= \mathbb{E}\left[\mu_0(X)\left(1-\frac{\mathbb{E}[1-D|X]}{1-\pi(X)}\right)\right] \quad \text{(Linearity of conditional expectation)} \\
&= \mathbb{E}\left[\mu_0(X)\left(1-\frac{1-\pi(X)}{1-\pi(X)}\right)\right] \quad (\text{since } \mathbb{E}[1-D|X] = 1-\mathbb{E}[D|X] = 1-\pi(X)) \\
&= \mathbb{E}\left[\mu_0(X)\left(1-1\right)\right] = \mathbb{E}[0] = 0
\end{aligned}$$

Combining the Results:

Since both terms equal zero:

$$\mathbb{E}\left[\left(1-\frac{D}{\pi(X)}\right)\mu_1(X) - \left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right] = 0 - 0 = 0$$

This proves that the correction term has zero expectation. Importantly, this proof requires only that $\pi(X) = \Pr(D=1 \mid X)$ and positivity (to ensure denominators are well-defined), with no assumptions about ignorability or the correctness of the outcome models $\mu_0(X)$ and $\mu_1(X)$.

Step 3: Augment the IPW Estimator

Since the term is zero, we can add it to the IPW functional without changing its expected value:

$$\begin{aligned}
\tau_{ATE} &= \mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\right] + 0 \\
&= \mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\right] + \mathbb{E}\left[\left(1-\frac{D}{\pi(X)}\right)\mu_1(X) - \left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]
\end{aligned}$$

Step 4: Rearrange and Simplify

Now we carefully combine the IPW formula with the zero-expectation correction term and rearrange to obtain the final DR formula. Starting from:

$$\tau_{ATE} = \mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\right] + \mathbb{E}\left[\left(1-\frac{D}{\pi(X)}\right)\mu_1(X) - \left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]$$

Since both terms involve the same random variables, we can combine them into a single expectation:

$$\tau_{ATE} = \mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)} + \left(1-\frac{D}{\pi(X)}\right)\mu_1(X) - \left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]$$

Next, we distribute the terms involving the outcome models:

$$\tau_{ATE} = \mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)} + \mu_1(X) - \frac{D\mu_1(X)}{\pi(X)} - \mu_0(X) + \frac{(1-D)\mu_0(X)}{1-\pi(X)}\right]$$

Now we collect terms by grouping those involving the treated observations and those involving the control observations:

Treated terms (involving D):

$$\frac{DY}{\pi(X)} - \frac{D\mu_1(X)}{\pi(X)} = \frac{D(Y - \mu_1(X))}{\pi(X)}$$

Control terms (involving $1 - D$):

$$-\frac{(1-D)Y}{1-\pi(X)} + \frac{(1-D)\mu_0(X)}{1-\pi(X)} = -\frac{(1-D)(Y-\mu_0(X))}{1-\pi(X)}$$

Outcome model terms (the remaining terms):

$$\mu_1(X) - \mu_0(X)$$

Step 5: DR Check When Outcome Models Are Correct

We have shown that the AIPW expression equals τ_{ATE} when the propensity score is correctly specified. Now we establish the "other side" of double robustness: the expression also equals τ_{ATE} when the outcome models are correct, even if the propensity score is misspecified.

If $\mu_1(X) = \mathbb{E}[Y \mid D = 1, X]$ and $\mu_0(X) = \mathbb{E}[Y \mid D = 0, X]$ are correctly specified, then for any measurable $\pi^\star(X) \in (0, 1)$ (possibly misspecified):

$$\begin{aligned}\mathbb{E}\left[\frac{D\big(Y - \mu_1(X)\big)}{\pi^\star(X)}\middle| X\right] &= \frac{1}{\pi^\star(X)}\Big(\mathbb{E}[DY \mid X] - \mu_1(X)\,\mathbb{E}[D \mid X]\Big) \\ &= \frac{1}{\pi^\star(X)}\big(\pi(X)\mu_1(X) - \pi(X)\mu_1(X)\big) = 0\end{aligned}$$

Similarly:

$$\mathbb{E}\left[\frac{(1-D)\big(Y - \mu_0(X)\big)}{1-\pi^\star(X)}\middle| X\right] = 0$$

Therefore:

$$\begin{aligned}&\mathbb{E}\left[\frac{D(Y-\mu_1(X))}{\pi^\star(X)} - \frac{(1-D)(Y-\mu_0(X))}{1-\pi^\star(X)} + \mu_1(X) - \mu_0(X)\right] \\ &= \mathbb{E}[\mu_1(X) - \mu_0(X)] = \tau_{ATE}\end{aligned}$$

This establishes double robustness: correct outcome models $\mu_0(X), \mu_1(X)$ suffice for consistency even if $\pi^\star(X)$ is misspecified.

Combining all terms gives us the final DR identifying expression for the ATE:

$$\boxed{\tau_{ATE}^{DR} = \mathbb{E}\left[\frac{D(Y-\mu_1(X))}{\pi(X)} - \frac{(1-D)(Y-\mu_0(X))}{1-\pi(X)} + \mu_1(X) - \mu_0(X)\right]}$$

Under ignorability and positivity, the AIPW estimator is consistent if either the propensity model $\pi(X)$ is correct (as shown in Step 2) or **both** outcome models $\mu_0(X), \mu_1(X)$ are correct (as shown in Step 5).

6. Estimation

To construct an estimator, we use the **analogy principle**, replacing population expectations ($\mathbb{E}[\cdot]$) with sample averages ($\frac{1}{n}\sum_{i=1}^{n}$). This is a two-stage process:

1. **First Stage (Nuisance Estimation):** Obtain estimates for $\hat{\pi}(X_i)$, $\hat{\mu}_0(X_i)$, and $\hat{\mu}_1(X_i)$ using appropriate models (e.g., logistic regression, OLS, or flexible machine learning).
2. **Second Stage (Plug-in Estimator):** Plug the estimated nuisance functions into the sample analog of the DR estimand.

Theoretical Requirements: When using flexible machine learning methods for nuisance estimation, sample-splitting (cross-fitting) is required to maintain $\sqrt{n}$ consistency and asymptotic normality. The standard rate condition requires $\|\hat{\mu}_a - \mu_a\| \cdot \|\hat{\pi} - \pi\| = o_p(n^{-1/2})$ (e.g., each $o_p(n^{-1/4})$). If both nuisance models are correctly specified parametric models, cross-fitting is not required. Without sample-splitting, the same-sample plug-in can introduce bias that breaks the theoretical guarantees.

The final **ATE estimator** is:

$$\hat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right]$$

Asymptotic Inference: The asymptotic variance of the DR estimator is the variance of the influence function:

$$\varphi_i = \frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \hat{\tau}_{ATE}$$

where $\hat{\mu}_1(X_i)$, $\hat{\mu}_0(X_i)$, and $\hat{\pi}(X_i)$ are out-of-fold (cross-fitted) predictions when using flexible machine learning methods.
The variance of $\hat{\tau}_{ATE}$ is estimated as:

$$\widehat{\mathrm{Var}}(\hat{\tau}_{ATE}) = \frac{1}{n^2}\sum_{i=1}^{n}\hat{\varphi}_i^2 \quad \text{and} \quad \widehat{\mathrm{SE}}(\hat{\tau}_{ATE}) = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\hat{\varphi}_i^2}$$

Alternatively, use the sample variance form $\widehat{\mathrm{SE}} = \sqrt{\frac{1}{n(n-1)}\sum_i(\hat{\varphi}_i - \bar{\hat{\varphi}})^2}$. Standard normal (Wald) confidence intervals can then be constructed.

Practical Note: Near-violations of the overlap assumption can inflate the residual terms in the DR expression. Trimming observations with extreme propensity scores can improve the estimation.