Doubly robust estimation

1. Can the controls tell us how much our regression is wrong?

Prerequisites: Outcome regression and Inverse probability weighting. This note combines their two representations of the same missing mean.

We observe an outcome Y, a treatment indicator D, and pre-treatment covariates X. The target is the average treatment effect on the treated,

$$\tau = \mathbb{E}[Y(1) - Y(0) \mid D = 1].$$

Write $p_0(x) = \mathbb{P}(D = 1 \mid X = x)$, $\mu_0(x) = \mathbb{E}[Y \mid D = 0, X = x]$, and $\pi = \mathbb{E}[D] > 0$. Subscript zero means the true function. A function without that subscript, such as μ, is a candidate and may be wrong. These auxiliary functions are called **nuisance functions**: we need them to estimate the effect, but they are not themselves the target. A working model is a chosen specification for one of these functions.

The identifying assumptions are the observation rule $Y = DY(1) + (1 - D)Y(0)$, equality of untreated conditional means across groups,

$$\mathbb{E}[Y(0) \mid X, D = 1] = \mathbb{E}[Y(0) \mid X, D = 0],$$

and overlap $p_0(X) \leq 1-\epsilon$ for some $\epsilon > 0$. The mean equality is implied by the unconfoundedness assumption in the prerequisites and is all we need. It is required on treated support. Covariate values with $p_0(x) = 0$ contribute nothing to the treated target.

The prerequisite notes establish

$$\tau = \frac{\mathbb{E}[D(Y - \mu_0(X))]}{\pi} = \frac{\mathbb{E}[DY] - \mathbb{E}[(1 - D)o_0(X)Y]}{\pi}, \qquad o_0(X) = \frac{p_0(X)}{1 - p_0(X)}.$$

Here $o_0(X)$ is the treatment **odds** conditional on X, the number of treated people per control within a covariate group. Multiplying control outcomes by these odds makes their covariate composition match that of the treated.

Outcome regression needs the right μ_0; weighting needs the right p_0. Simply averaging their answers does not protect against either being wrong. Instead, we will use weighted control residuals to estimate the regression's bias.

Throughout, expectations are population averages. Assume finite absolute outcome means and integrable candidate regressions. Candidate probabilities satisfy $0 \leq p(X) \leq 1 - \epsilon$. These restrictions make the following averages finite.

2. The example and the two single-model biases

Retain the example from the prerequisites. X is uniform on $\{0, 1, 2\}$, $Y(0) = X^2 + U$, $\mathbb{E}[U \mid X, D] = 0$, and $Y(1) = Y(0) + 2X$.

x	$p_0(x)$	$\mathbb{P}(X = x \mid D = 1)$	$\mathbb{P}(X = x \mid D = 0)$	$\mu_0(x)$	$\mathbb{E}[Y \mid D = 1, x]$
0	1/4	1/6	1/2	0	0
1	1/2	1/3	1/3	1	3
2	3/4	1/2	1/6	4	8

Thus $\pi = 1/2$, the treated mean is 5, the control mean is 1, and $\tau = (0 + 2 \cdot 2 + 3 \cdot 4)/6 = 8/3$. These are population proportions, not a claim about what a random sample will contain.

Regression error passes through directly

We first derive the population bias of the outcome-regression estimate based on a possibly incorrect μ: its difference from the true ATT, τ. Section 3 then constructs a weighted control-residual correction to subtract from this estimate; correct treatment probabilities make that correction equal to the bias.

For $\tau_{\mathrm{OR}}(\mu) = \mathbb{E}[D(Y - \mu)]/\pi$, subtract the correct identity:

$$\begin{aligned}\tau_{\mathrm{OR}}(\mu) - \tau &= \frac{\mathbb{E}[D\{(Y - \mu) - (Y - \mu_0)\}]}{\pi} \\ &= \frac{\mathbb{E}[D(\mu_0 - \mu)]}{\pi} = \frac{\mathbb{E}[p_0(\mu_0 - \mu)]}{\pi}.\end{aligned}$$

The last equality conditions on X: the function error is fixed and $\mathbb{E}[D \mid X] = p_0(X)$.

Weighting error also passes through directly

We now derive the population bias of the inverse-probability-weighted estimate based on possibly incorrect treatment probabilities p. Combining weighting with an outcome regression will correct this bias when the regression is correct, as Section 4 shows.

Define $\tau_{\mathrm{IPW}}(p) = \{\mathbb{E}[DY] - \mathbb{E}[(1 - D)pY/(1 - p)]\}/\pi$. Here p, p_0, and μ_0 abbreviate functions evaluated at X. Subtract $\tau_{\mathrm{IPW}}(p_0) = \tau$. Both expressions have denominator π, so multiplying their difference by π clears that denominator; the treated outcome terms $\mathbb{E}[DY]$ cancel. Then condition on X, using $\mathbb{E}[(1 - D)Y \mid X] = (1 - p_0(X))\mu_0(X)$:

$$\begin{aligned}\pi\{\tau_{\mathrm{IPW}}(p) - \tau\} &= \mathbb{E}\left[(1 - D)\left\{\frac{p_0}{1 - p_0} - \frac{p}{1 - p}\right\} Y\right] \\ &= \mathbb{E}\left[\left\{p_0 - \frac{p(1 - p_0)}{1 - p}\right\}\mu_0\right] \\ &= \mathbb{E}\left[\frac{p_0(1 - p) - p(1 - p_0)}{1 - p}\mu_0\right] \\ &= \mathbb{E}\left[\frac{p_0 - p}{1 - p}\mu_0\right].\end{aligned}$$

Dividing by π gives the bias itself:

$$\tau_{\mathrm{IPW}}(p) - \tau = \frac{1}{\pi}\mathbb{E}\left[\frac{p_0 - p}{1 - p}\mu_0\right].$$

Denominator convention: both terms here use π. The prerequisite IPW note normalizes controls by their own total weight. The two population formulas agree at p_0, but need not agree at a wrong p. We keep the common denominator throughout this note; the panel DiD note derives normalization.

3. Build a correction from observable residuals

Fix a candidate regression μ. A control's residual is its observed outcome minus its predicted outcome. Within a covariate group its mean is

$$\mathbb{E}[Y - \mu(X) \mid X, D = 0] = \mu_0(X) - \mu(X).$$

We want this function averaged over treated people, because that is the regression bias on the previous page. Weighting gives exactly that change of population:

$$\begin{aligned}\mathbb{E}[(1-D)o_0(X)(Y-\mu(X)) \mid X] &= o_0(X)(1-p_0(X))(\mu_0(X)-\mu(X)) \\ &= p_0(X)(\mu_0(X)-\mu(X)).\end{aligned}$$

In the first line, $1 - p_0(X)$ is the chance of being a control and the last factor is the control's conditional mean residual. In the second, the denominator of $o_0 = p_0/(1-p_0)$ cancels. Averaging over X and dividing by π gives

$$\frac{\mathbb{E}[(1-D)o_0(Y-\mu)]}{\pi} = \tau_{\mathrm{OR}}(\mu) - \tau.$$

The right side contains an unknown bias. The left side tells us how to estimate it: use the control residuals and estimated treatment odds.

The two bias derivations in Section 2 show how regression and weighting can each fail on their own. We subtract an estimate of the regression's bias from the regression effect estimate. The correction weights control **residuals**, whereas the IPW effect estimate uses control outcomes.

This gives two ways for the corrected estimate to recover τ:

- **Correct treatment probabilities:** the weighted residual correction equals $\tau_{\mathrm{OR}}(\mu) - \tau$, so subtracting it leaves τ, even if the regression is wrong.
- **Correct outcome regression:** $\tau_{\mathrm{OR}}(\mu_0) = \tau$ already, and control residuals have mean zero at every X. The correction is therefore zero even if the treatment probabilities are wrong.

These are population statements. If both functions are wrong, the remaining bias is

$$\frac{1}{\pi}\mathbb{E}\left[\frac{p_0 - p}{1-p}(\mu_0 - \mu)\right],$$

as Section 5 derives. The product of the two function errors explains the double robustness: either correct function makes the bias zero.

Subtract the weighted control residual correction from the outcome-regression effect $\tau_{\mathrm{OR}}(\mu)$. Using candidate predictions $\mu(X)$ for untreated outcomes and $p(X)$ for treatment probabilities gives

$$T(\mu, p) = \underbrace{\frac{\mathbb{E}[D(Y-\mu)]}{\pi}}_{\text{regression effect}} - \underbrace{\frac{\mathbb{E}[(1-D)\frac{p}{1-p}(Y-\mu)]}{\pi}}_{\text{control residual correction}}.$$

This is the **augmented inverse probability weighting** (AIPW) formula for the ATT. $T(\mu, p)$ is the population effect obtained with those two candidate functions; it equals τ if either is correct (Section 4). Combining the terms,

$$\boxed{T(\mu, p) = \frac{1}{\pi}\mathbb{E}\left[\left\{D - (1-D)\frac{p}{1-p}\right\}(Y-\mu)\right].}$$

To estimate it from data, fit both functions, replace expectations by sample averages and π by the treated share (Section 7).

The braces equal 1 for a treated person and minus the odds for a control. They also have a compact expression:

$$D-\frac{(1-D)p}{1-p}=\frac{D-Dp-p+Dp}{1-p}=\frac{D-p}{1-p}.$$

There are two useful ways to read the construction. It is the treated mean residual minus a weighted control mean residual. Equivalently, it imputes the missing untreated mean as

$$\underbrace{\mathbb{E}[\mu(X)\mid D=1]}_{\text{initial imputation}}+\underbrace{\frac{\mathbb{E}[(1-D)o(X)(Y-\mu(X))]}{\pi}}_{\text{correction}}.$$

A positive weighted control residual says that the regression predicts too low. The correction raises the imputed counterfactual and lowers the estimated effect.

4. Why either correct function is enough

First branch: correct regression, possibly wrong odds

Set $\mu = \mu_0$. For any admissible candidate p,

$$
\begin{aligned}
\mathbb{E}[(1-D)o(X)(Y-\mu_0(X))] &= \mathbb{E}\left[o(X)\mathbb{E}[(1-D)(Y-\mu_0(X)) \mid X]\right] \\
&= \mathbb{E}\left[o(X)(1-p_0(X))\underbrace{(\mu_0(X)-\mu_0(X))}_{0}\right] \\
&= 0.
\end{aligned}
$$

Therefore the correction vanishes and

$$
T(\mu_0, p) = \frac{\mathbb{E}[D(Y-\mu_0)]}{\pi} = \tau.
$$

Even wrong odds cannot change a conditional residual mean of zero.

Second branch: correct odds, possibly wrong regression

Set $p = p_0$ in the definition of $T(\mu, p)$. Combining the two expectations and factoring out the common residual gives

$$
\begin{aligned}
T(\mu, p_0) &= \frac{\mathbb{E}[D(Y-\mu(X))]}{\pi} - \frac{\mathbb{E}[(1-D)\frac{p_0(X)}{1-p_0(X)}(Y-\mu(X))]}{\pi} \\
&= \frac{1}{\pi}\mathbb{E}\left[\left\{D-(1-D)\frac{p_0(X)}{1-p_0(X)}\right\}(Y-\mu(X))\right] \\
&= \frac{\mathbb{E}[A_0(Y-\mu(X))]}{\pi},
\end{aligned}
$$

where $A_0 = D-(1-D)o_0(X)$ abbreviates the signed weight in braces, using $o_0(X) = p_0(X)/(1-p_0(X))$. Its conditional mean is

$$
\mathbb{E}[A_0 \mid X] = p_0 - (1-p_0)\frac{p_0}{1-p_0} = 0.
$$

Expand the residual before taking the expectation:

$$
\begin{aligned}
T(\mu, p_0) &= \frac{\mathbb{E}[A_0 Y]}{\pi} - \frac{\mathbb{E}[A_0\mu(X)]}{\pi} \\
&= \tau - \frac{\mathbb{E}[\mu(X)\mathbb{E}[A_0 \mid X]]}{\pi} = \tau.
\end{aligned}
$$

The first term equals τ because expanding A_0 recovers the IPW identity $\{\mathbb{E}[DY] - \mathbb{E}[(1-D)o_0(X)Y]\}/\pi = \tau$ from Section 1. The second is zero because $\mu(X)$ is fixed when we condition on X.

The two protections differ. Correct regression makes the control residual mean zero. Correct weighting makes the signed weight mean zero. In each case, the other candidate function multiplies a quantity that has conditional mean zero.

These are exact population identities. They do not assert exact equality in a finite sample. An estimated regression generally leaves nonzero residual means on new data, and estimated odds generally fail to balance every function of X. Consistency requires an additional argument connecting sample averages to these population identities; we give it in Section 8.

5. When both functions are wrong: derive the bias

Let $m_1(X) = \mathbb{E}[Y \mid X, D = 1]$ on treated support. Conditioning on X gives

$$\mathbb{E}[D(Y-\mu) \mid X] = p_0(m_1 - \mu),$$

$$\mathbb{E}[(1-D)o(Y-\mu) \mid X] = \frac{p(1-p_0)}{1-p}(\mu_0 - \mu).$$

Thus

$$\pi T(\mu, p) = \mathbb{E}\left[p_0(m_1 - \mu) - \frac{p(1-p_0)}{1-p}(\mu_0 - \mu)\right].$$

Identification gives $\pi\tau = \mathbb{E}[p_0(m_1 - \mu_0)]$. Subtracting, the m_1 terms cancel:

$$\begin{aligned}
\pi\{T(\mu, p) - \tau\} &= \mathbb{E}\left[p_0(\mu_0 - \mu) - \frac{p(1-p_0)}{1-p}(\mu_0 - \mu)\right] \\
&= \mathbb{E}\left[\left\{p_0 - \frac{p(1-p_0)}{1-p}\right\}(\mu_0 - \mu)\right] \\
&= \mathbb{E}\left[\frac{p_0 - p}{1-p}(\mu_0 - \mu)\right].
\end{aligned}$$

The last step uses the common-denominator calculation from Section 2. Dividing by the positive treated share yields the central identity:

$$\boxed{T(\mu, p) - \tau = \frac{1}{\pi}\mathbb{E}\left[\frac{p_0 - p}{1-p}(\mu_0 - \mu)\right].}$$

The bias involves a **product of function errors**. Setting either error to zero recovers the corresponding proof from the previous page.

There are also two checks against the single-model formulas. If $p = 0$, the control correction disappears and the bias is $\mathbb{E}[p_0(\mu_0 - \mu)]/\pi$. If $\mu = 0$, the residual is simply Y and the bias is $\mathbb{E}[(p_0 - p)\mu_0/(1-p)]/\pi$. These are precisely the OR and IPW biases.

A product need not be small. Large errors in the same covariate region can reinforce each other, especially when $1-p$ is small. Errors may also cancel inside the expectation. Double robustness supplies sufficient conditions for zero bias; it does not say these are the only conditions, or that the combined estimator always improves on a single-model estimator.

6. Put specific wrong models into the example

Fit a line to the control regression x^2. In the control distribution $q = (1/2, 1/3, 1/6)$,

$$\mathbb{E}_q[X] = \frac{2}{3}, \quad \mathbb{E}_q[X^2] = 1, \quad \mathbb{E}_q[X^3] = \frac{5}{3}.$$

The population least-squares slope and intercept are therefore

$$b = \frac{\mathbb{E}_q[X^3] - \mathbb{E}_q[X]\mathbb{E}_q[X^2]}{\mathbb{E}_q[X^2] - \mathbb{E}_q[X]^2} = \frac{5/3 - 2/3}{1 - 4/9} = \frac{9}{5}, \qquad a = 1 - \frac{9}{5}\frac{2}{3} = -\frac{1}{5}.$$

The wrong regression is $\mu^\dagger(x) = -1/5 + 9x/5$. Its errors and its resulting ATT bias are

$$\mu_0 - \mu^\dagger = (1/5, -3/5, 3/5),$$

$$\tau_{\mathrm{OR}}(\mu^\dagger) - \tau = \frac{1}{6}\frac{1}{5} - \frac{1}{3}\frac{3}{5} + \frac{1}{2}\frac{3}{5} = \frac{1 - 6 + 9}{30} = \frac{2}{15}.$$

Consequently $\tau_{\mathrm{OR}}(\mu^\dagger) = 8/3 + 2/15 = 14/5$.

Choose a second, deliberately wrong candidate,

$$p^\dagger(x) = \frac{1}{3} + \frac{x}{6} = (1/3, 1/2, 2/3).$$

This is simply a specified misspecification; no measurement-error model is needed to justify choosing it. It has the correct marginal mean but the wrong slope. Its odds are $(1/2, 1, 2)$ and

$$\frac{p_0 - p^\dagger}{1 - p^\dagger} = (-1/8, 0, 1/4).$$

The common-denominator IPW bias is

$$\frac{1}{\pi}\mathbb{E}\left[\frac{p_0 - p^\dagger}{1 - p^\dagger}\mu_0\right] = \frac{2}{3}\left[-\frac{1}{8} \cdot 0 + 0 \cdot 1 + \frac{1}{4} \cdot 4\right] = \frac{2}{3}.$$

Thus $\tau_{\mathrm{IPW}}(p^\dagger) = 10/3$. For AIPW with both candidates wrong,

$$T(\mu^\dagger, p^\dagger) - \tau = \frac{2}{3}\left[-\frac{1}{8}\frac{1}{5} + 0 + \frac{1}{4}\frac{3}{5}\right] = \frac{2}{3}\frac{5}{40} = \frac{1}{12}.$$

The result is $8/3 + 1/12 = 11/4$. This example improves on both single-model biases. That is a property of these numbers, not a general guarantee.

7. From the correction to a sample estimator

For observations $i = 1, \dots, n$, let $n_1 = \sum_i D_i$ and $\hat{o}_i = \hat{p}(X_i)/(1 - \hat{p}(X_i))$. The estimator is

$$\boxed{\hat{\tau} = \frac{\sum_i [D_i - (1 - D_i)\hat{o}_i](Y_i - \hat{\mu}(X_i))}{n_1}.}$$

Fit the propensity on both groups and the untreated-outcome regression on controls. Predict both functions at the relevant observations, form residuals, then compute the treated residual total minus the odds-weighted control residual total. Divide the difference by the treated count. The next note explains how to fit on separate folds when using flexible models.

This formula requires $n_1 > 0$ and fitted propensities below one. A covariate cell with treated people but no controls cannot supply an empirical control mean; a model prediction in that cell is an extrapolation. Population overlap does not guarantee good support in a particular small sample.

Check the correction itself

At correct odds and wrong regression, the odds-weighted control residual mean is

$$\frac{1}{3}\frac{1}{5}\frac{1}{2} + 1\left(-\frac{3}{5}\right)\frac{1}{3} + 3\frac{3}{5}\frac{1}{6} = \frac{2}{15}.$$

Because the treated and control shares are both 1/2, this conditional control average also equals the correction divided by π. Subtracting it from 14/5 gives 8/3 exactly. At correct regression, every conditional control residual mean is zero, including under the wrong odds.

Population construction	Outcome candidate	Propensity candidate	Value	Bias
Unadjusted difference	—	—	4	4/3
OR	$\mu^\dagger$	—	14/5	2/15
IPW, common denominator	—	$p^\dagger$	10/3	2/3
AIPW	$\mu^\dagger$	p_0	8/3	0
AIPW	μ_0	$p^\dagger$	8/3	0
AIPW	$\mu^\dagger$	$p^\dagger$	11/4	1/12

These are population limits evaluated at specified functions. They are not finite-sample expectations of fitted estimators. Sampling variation remains even when the population bias is zero.

8. Why estimating the functions need not change the limit

Here is a complete sufficient argument for consistency. Assume i.i.d. observations and fixed limiting functions (μ^*, p^*), with $\mathbb{E}|\mu^*(X)| < \infty$. Suppose

$$a_n = \sup_x |\hat{\mu}(x) - \mu^*(x)| \xrightarrow{p} 0, \qquad b_n = \sup_x |\hat{p}(x) - p^*(x)| \xrightarrow{p} 0,$$

and $0 \le \hat{p}, p^* \le 1 - \epsilon$. Uniform convergence is stronger than necessary, but makes the proof work even when fitting and averaging use the same data.

Let $A_p = D - (1-D)p/(1-p)$ and $B(\mu, p) = A_p(Y - \mu)$. Since $|A_{\hat{p}}| \le M := \max\{1, (1-\epsilon)/\epsilon\}$ and

$$|A_{\hat{p}} - A_{p^*}| \le \frac{|\hat{p} - p^*|}{(1-\hat{p})(1-p^*)} \le \frac{b_n}{\epsilon^2},$$

adding and subtracting $A_{\hat{p}}(Y - \mu^*)$ gives the exact bound

$$|B(\hat{\mu}, \hat{p}) - B(\mu^*, p^*)| \le M a_n + \frac{b_n}{\epsilon^2}|Y - \mu^*(X)|.$$

Write $\mathbb{P}_n f = n^{-1}\sum_i f(W_i)$ for a sample average. Averaging the bound,

$$|\mathbb{P}_n B(\hat{\mu}, \hat{p}) - \mathbb{P}_n B(\mu^*, p^*)| \le M a_n + \frac{b_n}{\epsilon^2}\mathbb{P}_n|Y - \mu^*(X)| \xrightarrow{p} 0.$$

The last sample average converges to a finite mean by the law of large numbers; the other factors converge to zero. The same law gives $\mathbb{P}_n B(\mu^*, p^*) \xrightarrow{p} \mathbb{E}B(\mu^*, p^*)$ and $\mathbb{P}_n D \xrightarrow{p} \pi > 0$. Taking the ratio,

$$\hat{\tau} \xrightarrow{p} \frac{\mathbb{E}B(\mu^*, p^*)}{\pi} = T(\mu^*, p^*).$$

By Section 4 this limit is τ if either $\mu^* = \mu_0$ or $p^* = p_0$. The other fitted function must still stabilize and satisfy the stated bounds.

What we have established. The causal assumptions identify the target; the residual correction gives two routes to the right population answer; and convergence of the fitted averages gives consistency. None of these alone supplies a standard error, repairs a false causal assumption, or guarantees small-sample accuracy.

9. Computing the ATT in R and Python

DoubleML provides the doubly robust ATT estimator in both R and Python. Select `score = "ATTE"`: the package calls the ATT the *average treatment effect on the treated* (ATTE), while its default target is the ATE. Its ATT score uses the regression-minus-weighted-residual construction derived above.

R

The DoubleML package provides the `DoubleMLIRM` estimator. Supply the outcome, treatment indicator, and pre-treatment covariates, and choose `mlr3` learners for the outcome regression and propensity score. Set `score = "ATTE"` to estimate the ATT.

Python

The doubleml package provides the corresponding `DoubleMLIRM` estimator with scikit-learn-compatible learners: a regressor for the outcome and a classifier that predicts treatment probabilities. Again, select `score="ATTE"`.

Both packages handle nuisance-model fitting, cross-fitting, and estimation, and provide standard errors and confidence intervals. Cross-fitting means fitting the nuisance models outside each fold and predicting for observations inside it before combining the residual contributions.