Doubly Robust Estimation of Causal Effects

Francesca Vescia

Spring 2024

Welcome! This is a conversational introduction to doubly robust estimation. Largely following Funk et al. 2011, we'll talk about what doubly robust estimation is and why we do it; break down a simple doubly robust estimator to understand how it works; and run some simulations to see the estimator in action.

What and why

I like to think of "causal inference" as a catch-all term for the tools we have to figure out whether relationships we see in the world are cause and effect or just coincidence. Often, we learn about these relationships by modeling them: we collect data on something we're curious about and fit models to that data to help us see the patterns in it. Models capture patterns in the data in the form of numerical estimates.

But fitting models to data almost always involves making some assumptions about how the world works. For example, when we fit a linear regression model, we assume the relationship we are quantifying is linear. And when we misspecify our models – that is, when our assumptions about the world are wrong – the estimates our models produce are poor.

Doubly robust estimation is one way to help insure our estimates against bad assumptions. It combines two modeling strategies, outcome regression modeling and propensity score estimation, in such a way that as long as one of the two models is correct, our estimates will be unbiased, even if the other model is misspecified.

You can think of doubly robust estimation as statistics' version of the Swiss cheese model. Each modeling strategy – propensity score estimation and outcome regression modeling – is a layer of defense against bias. Neither is perfect, but by layering the strategies on top of each other, we give ourselves more opportunities to block bias from degrading our estimates.

Some housekeeping

First things first, notation. I follow Funk et al., except that I use D instead of X to denote treatment, to avoid confusion given X is often used elsewhere to denote covariates. Here is a full list of the terms we're going to use:

Z are characteristics (that is, covariates) of individuals in our data

D is a binary treatment indicator equal to 1 if an individual is exposed to treatment or 0 if they aren't

$Y_{D=1}$ and $Y_{D=0}$ are *observed* outcomes for individuals in the treatment and control groups, respectively

$\hat{Y_1}$ and $\hat{Y_0}$ are *predicted* outcomes under treatment and control, respectively [1]

$m1(Z_i, \hat{\alpha}_1)$ and $m0(Z_i, \hat{\alpha}_0)$ are the models we'll use to predict $\hat{Y}_{D=1}$ and $\hat{Y}_{D=0}$ for each individual, where Z_i are that individual's characteristics and $\hat{\alpha}_1$ and $\hat{\alpha}_0$ are coefficients that describe the estimated relationships between characteristics and outcomes in the treatment and control groups; we get $\hat{\alpha}_1$ and $\hat{\alpha}_0$ by fitting linear regression models that predict outcomes using characteristics

[1] Per usual, any term with a "hat" is an estimate of a true quantity we can't measure directly

$PS = P[D = 1|Z]$ is the propensity score, which measures how likely an individual is to be exposed to treatment based on their characteristics

$e(Z_i, \hat{\beta})$ is the model we'll use to estimate the propensity score, where Z_i are once again an individual's characteristics and $\hat{\beta}$ is a coefficient that describes the estimated relationship between characteristics and treatment propensity; we get $\hat{\beta}$ by fitting a logistic regression model that predicts treatment using characteristics

DR_1 and DR_0 are the doubly robust estimated outcomes under treatment and control, respectively

Also, a few assumptions. Briefly, **exchangeability** assumes we would see the same outcome distributions if the treatment assignments had been flipped. **Positivity** or overlap assumes every individual in the data had some chance of being treated: formally, $0 \leq P[D_i = 1] \leq 1$. It ensures we have individuals in the treatment and control groups to contrast. **Consistency** or the stable unit treatment assumption (SUTVA) assumes every individual has a single, stable potential outcome for each possible treatment: formally, $Y = Y_{D=1}$ if $D = 1$. SUTVA has several implications; see Anton Strezhnev's introduction to the potential outcomes framework for an excellent summary.

A simple doubly robust estimator

Let's break down the doubly robust estimator mathematically and programmatically. We'll use the notation defined above for our proofs, and we'll fit our models on simulated data. If you're curious, see Funk et al. Web Appendix 3 for the exact details of the data-generating process and GitHub for the code. The main things to know are that treatment (D) and outcomes (Y) are both a function of individual characteristics[2] (Z), but outcomes are *not* a function of treatment, so the true treatment effect is zero.

Recall that doubly robust estimation combines two modeling strategies, outcome regression modeling and propensity score estimation. We're going to correctly specify our outcome model and intentionally *misspecify* our propensity score model [3], so we can see double robustness at work.

Let's start with our outcomes: we model the relationships between characteristics (Z) and outcomes (Y) in each exposure group, then use those models to predict outcomes ($\hat{Y}$) for everyone. This means each individual has one *observed* outcome, for the condition they are assigned to, and two *predicted* outcomes, one each for treatment and control.

```r
# Fit correct outcome regression models within exposure groups
mod_t <- lm(y ~ z1 + z3, data = data %>% filter(d == 1))
mod_c <- lm(y ~ z1 + z3, data = data %>% filter(d == 0))

# Predict outcomes for everyone
data <- data %>% mutate(yhat1 = predict(mod_t, data))
data <- data %>% mutate(yhat0 = predict(mod_c, data))
```

Now let's fit our propensity score model, purposefully leaving out Z_3, and use the resulting, *misspecified* model to predict propensity scores:

```r
# Fit incorrect propensity score logit model (omit z3)
ps_mod <- glm(d ~ z1, data = data, family = 'binomial')

# Predict propensity scores
data <- data %>% mutate(ps_hat = predict(ps_mod, data))
```

[2] The Z_2 / `z2` characteristic is only used for data generation, so you won't see it anywhere in this document.

[3] For anyone following along in Funk et al., this is their Scenario 3.

And now, the fun part! We take our estimated outcomes and propensity scores and combine them to calculate doubly robust outcome estimates as shown in the table below. Just like when we calculated our regression-based $\hat{Y}s$, each person gets two estimates, one each for treatment and control.

	DR_1	DR_0
For $D = 1$	$\frac{Y_{D=1}}{PS} - \frac{\hat{Y}_1(1-PS)}{PS}$	$\hat{Y}_0$
For $D = 0$	$\hat{Y}_1$	$\frac{Y_{D=0}}{1-PS} - \frac{\hat{Y}_0(PS)}{1-PS}$

```
# Compute doubly robust estimates
data <- data %>% mutate(dr1 =
                          ifelse(d == 1,
                                 ((y / ps_hat) - ((yhat1 * (1 - ps_hat)) / (ps_hat))),
                                 (yhat1)),
                        dr0 =
                          ifelse(d == 1,
                                 (yhat0),
                                 ((y / (1 - ps_hat) - ((yhat0 * ps_hat) / (1 - ps_hat))))))
```

Finally, we take the averages of DR_1 and DR_0 across our entire sample and take the difference of those two averages as our estimated average treatment effect:

```
# Compute ATE
(mean(data$dr1) - mean(data$dr0))
```

```
## [1] -0.1667564
```

It's small, close to the true value of zero even though we misspecified our propensity score model. Let's think through what exactly is going on in this estimator. You'll notice that that our doubly robust estimates of DR_1 for individuals in the control group and DR_0 for individuals in the treatment group are just $\hat{Y}_1$ and $\hat{Y}_0$, the predictions from our regression models. But our "within-condition" doubly robust estimates - that is, our DR_1s for individuals in the treatment group and our DR_0s for individuals in the control group - are a little fancier. We observe individuals' true outcomes under their assigned conditions, and we can use that information together with our propensity score estimates to adjust our predictions.

Some rearranging helps demonstrate how these adjustments block bias [4]. The ordering and notation are a little different, but this equation ultimately does exactly the same math as the equations in the table above:

$$\hat{\Delta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}[\frac{D_iY_i}{e(Z_i,\hat{\beta})} - \frac{(D_i - e(Z,\hat{\beta}))}{e(Z,\hat{\beta})}m_1(Z_i,\hat{\alpha}_1)] - \frac{1}{n}\sum_{i=1}^{n}[\frac{(1-D_i)Y_i}{1-e(Z_i,\hat{\beta})} + \frac{(D_i - e(Z_i,\hat{\beta}))}{1-e(Z,\hat{\beta})}m_0(Z_i,\hat{\alpha}_0)]$$

The first term in each average is the inverse-propensity-weighted (IPW) estimator for the expected outcome under the a given condition (treatment on the left-hand side and control on the right). The second "augmentation" term serves two purposes: it increases efficiency, which is beyond the scope of this tutorial, and also gives us double robustness. To see how it does the latter, let's return to our scenario from before - a correctly specified outcome regression model and a misspecified propensity score model - and consider the estimated outcome under treatment (the left-hand side average in the equation above; similar logic holds for the estimated outcome under control, the right-hand side average in the equation above).

[4] The following proof is excerpted with minimal adaptation from Funk et al. Web Appendix 1, which is in turn adapted from Tsiatis 2006.

Here's another look at that left-hand side term. When n is large, this sample average (top) estimates the population average (bottom), thanks to the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}[\frac{D_iY_i}{e(Z_i,\hat{\beta})} - \frac{(D_i - e(Z_i,\hat{\beta}))}{e(Z_i,\hat{\beta})}m_1(Z_i,\hat{\alpha}_1)]$$

$$\approx E[Y_{D=1}] + E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(Y_{D=1} - m_1(Z,\alpha_1))]$$

The first term, $E[Y_{D=1}]$, is the quantity we're interested in, the expected outcome under treatment. If the second term reduces to zero, the equation reduces to our quantity of interest. Let's see how the second term behaves in our scenario. Since we have the correct outcome regression model, we can replace the model $m_1(Z,\alpha_1)$ with the quantity it estimates, $E[Y|D = 1, Z]$:

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(Y_{D=1} - m_1(Z,\alpha_1))]$$

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(Y_{D=1} - E[Y|D = 1, Z])]$$

Now let's rearrange:

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(Y_{D=1} - E[Y|D = 1, Z])|X, Z]$$

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(E[Y_{D=1} - E[Y|D = 1, Z]])|D, Z]$$

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(E[Y_{D=1}|D, Z] - E[Y|D = 1, Z])]$$

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(E[Y_{D=1}|Z] - E[Y_{D=1}|Z])]$$

$$E[\frac{(D - e(Z,\beta))}{e(Z,\beta)}(0)] = E[0] = 0$$

$E[Y_{D=1}|Z] - E[Y_{D=1}|Z]$, which comes from our correctly specified outcome regression model, reduces to zero, taking any bias our misspecified propensity score model, $e(Z,\beta)$, would have otherwise introduced out of the equation.

So, that's what's going on under the hood. To see it in action, I simulated 1,000 samples each of 100, 500, 1,000, 2,000, and 10,000 observations and estimated average treatment effects for each sample under three scenarios, the good outcome regression/bad propensity score model scenario we have explored together (Funk et al.'s Scenario 3), its inverse (bad outcome regression/good propensity score model, Scenario 2), and the best-case good outcome regression/good propensity score model scenario (Scenario 1). The simulation results are below, and the code is on GitHub. For standard error and confidence interval estimates under these same three scenarios, see Funk et al.

There is no true treatment effect by design, so the average ATEs from the doubly robust estimator should be close to zero in all scenarios. These averages are larger than I was expecting based on theory - and the results from Funk et al.'s simulations with the same, relatively small sample sizes - and I am surprised a relatively large portion of them are negative, since an unbiased estimator should not systematically bias estimates in a particular direction. Thoughts and feedback on these somewhat unexpected results are welcome to fvescia@uchicago.edu

	Sample Size	Average ATE Estimate
Scenario 1	100	-0.1314
	500	-1.2578
	1,000	-1.0912
	2,000	0.0189
	10,000	-0.0190
Scenario 2	100	-0.1161
	500	-1.1239
	1,000	-1.3463
	2,000	-0.0645
	10,000	-0.0812
Scenario 3	100	0.2050
	500	0.6942
	1,000	-0.0344
	2,000	-0.1411
	10,000	-0.1106