OIT 661: Lecture 5
Heterogeneous Treatment Effect Estimation

(for class use only, please do not distribute)

Recap: Cross-Fitting for the ATE We have been discussing the following estimation problem. We observe data $(X_i, Y_i, W_i,) \in \mathcal{X} \times \mathbb{R} \times \{0, 1\}$ according to the potential outcomes model, such that there are potential outcomes $\{Y_i(0), Y_i(1)\}$ for which $Y_i = Y_i(W_i)$. We are not necessarily in a randomized controlled trial; however, we assume unconfoundedness, i.e., that treatment assignment is as good as random conditionally on the features X_i:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i. \tag{1}$$

Throughout, we write $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$ for the propensity score, $\mu_{(w)}(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right]$ for the conditional response surfaces of the potential outcomes, $\tau(x) = \mu_{(1)}(x) - \mu_{(0)}(x)$ for the conditional average treatment effect given $X_i = x$, and assume that $\sigma^2(x) = \text{Var}\left[Y_i(w) \,\middle|\, X_i = x\right]$ for not depend on w.

Our goal so far has been to estimate the average treatment effect $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. In the last lecture, we discussed cross-fitting, which uses

$$\begin{aligned}\hat{\tau}_{CF} = \frac{1}{n}\sum_{i=1}^{n} \Bigg(& \hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) \\ & + W_i \frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1 - \hat{e}^{(-k(i))}(X_i)}\Bigg),\end{aligned} \tag{2}$$

based on non-parametric pilot regressions $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$. We found that if the pilot regressions are uniformly consistent and

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}^{\mathcal{I}_2}(X) - \mu_{(w)}(X)\right)^2\right] \mathbb{E}\left[\left(\hat{e}^{\mathcal{I}_2}(X) - e(X)\right)^2\right] = o_P\left(\frac{1}{n}\right), \tag{3}$$

then $\hat{\tau}_{CF}$ is a $\sqrt{n}$-consistent estimator for τ and attains the semiparametric efficient variance.

As a concrete example of this result, suppose that we are willing to posit a high-dimensional linear model,

$$Y_i(w) = X_i \cdot \beta_{(w)} + \varepsilon_i(w), \quad \text{logit}\left(\mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]\right) = X_i \cdot \alpha \tag{4}$$

where the number of features p may be much larger than the number of observations, and that we estimate $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ via the lasso. Then, under assumptions, it is well known that the excess squared error of the regression adjustments scales as

$$\mathbb{E}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right] = \mathcal{O}_P\left(\frac{\left\|\beta_{(w)}\right\|_0 \log(p)}{n_w}\right),$$

etc., and so (3) holds whenever

$$\left\|\beta_{(w)}\right\|_0 \left\|\alpha\right\|_0 \ll \frac{n}{\log(p)^2},$$

i.e., (2) is $\sqrt{n}$-consistent whenever both the outcome and propensity models are sparse enough.[1]

Estimating treatment heterogeneity. In this class, we seek to go one step further, and estimate the conditional average treatment effect function $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$. Estimating $\tau(x)$ is harder than estimating the ATE τ. In particular, $\tau = \mathbb{E}\left[\tau(X)\right]$, so if we know the former we usually also know the latter.

As a first exposure to this problem, consider the high-dimensional linear model (4), where the conditional average treatment effect is $\tau(x) = x \cdot (\beta_{(1)} - \beta_{(0)})$. Earlier, we saw that, given enough sparsity, we can get an efficient ATE estimator by simply running separate cross-fit lassos for $\hat{\beta}_{(1)}$, $\hat{\beta}_{(0)}$ and $\hat{\alpha}$, and then combining them via (2). However, the resulting heterogeneous treatment effect (HTE) estimator,

$$\hat{\tau}(x) = x \cdot \left(\hat{\beta}^{lasso}_{(1)} - \hat{\beta}^{lasso}_{(0)}\right)$$

is terrible, because the two $\hat{\beta}_{(w)}$ estimators may shrink by slightly different amounts thus resulting in spurious heterogeneity. This phenomenon is often called regularization bias.

An algorithmic fix. To improve on this problem, a first idea is simply to re-parametrize the lasso, and use

$$\hat{\beta},\, \hat{\zeta} = \text{argmin}_{\beta,\,\zeta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i \cdot \beta - (2W_i - 1)X_i\zeta\right)^2 + \lambda_\beta \left\|\beta\right\|_1 + \lambda_\zeta \left\|\zeta\right\|_1\right\},$$

[1] Note that, because we now have a model for both the outcomes and the propensities, the semiparametric efficient variance seen last lecture may no longer be the lowest possible variance—since the argument used to justify that result assumed no modeling on the outcomes or propensities. See Athey et al. (2016), "approximate residual balancing" for more details.

resulting in $\hat{\beta}_{(0)} = \hat{\beta} - \hat{\zeta}$, $\hat{\beta}_{(1)} = \hat{\beta} + \hat{\zeta}$, and finally $\hat{\tau}(x) = 2x \cdot \hat{\zeta}$. The advantage of this estimator is that it can directly adapt to the sparsity of $\tau(x)$, and "share strength" between the two treatment arms to learn main effects. This estimator has been found to work much better than the "naive" difference-of-lassos estimator, especially in randomized trials.

Semiparametric modeling. In order to develop a more formal understanding of heterogeneous treatment effect estimation, it is helpful to consider the case where we have a model for $\tau(x)$,

$$Y_i(w) = f(X_i) + w\,\psi(X_i) \cdot \zeta + \varepsilon_i(w), \quad \mathbb{P}\left[W_i = 1 \,\middle|\, X_i\right] = e(x), \tag{5}$$

where $\psi(\cdot)$ is some pre-determined feature mapping. In this setup, we allow for non-parametric relationships between X_i, Y_i, and W_i; however, the treatment effect function itself takes on a simple form $\tau(x) = \psi(x) \cdot \zeta$.

This class of problems was studied by Robinson (1988), who showed that it is useful to write down the transformed sampling model:

$$Y_i - m(X_i) = (W_i - e(X_i))\,\psi(X_i) \cdot \zeta + \varepsilon_i, \text{ where } m(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right] \tag{6}$$

denotes the conditional expectation of the observed Y_i, marginalizing over W_i. This suggests the following "oracle" algorithm for estimating ζ:

1. Define transformed features $\widetilde{Y}_i^* = Y_i - m(X_i)$ and $\widetilde{Z}_i^* = (W_i - e(X_i))\psi(X_i)$.

2. Estimate $\hat{\zeta}^*$ by running the OLS regression $\widetilde{Y}_i^* \sim \widetilde{Z}_i^*$.

Robinson showed that this oracle procedure is $\sqrt{n}$-consistent and asymptotically normal,

$$\sqrt{n}\left(\hat{\zeta}^* - \zeta\right) \Rightarrow \mathcal{N}\left(0,\, V_\zeta\right), \tag{7}$$

where V_ζ is the semiparametric efficient variance for estimating ζ.

We of course can't use this oracle estimator in practice since we don't know $m(x)$ and $e(x)$. However, we can again use cross fitting to emulate the oracle:

1. Run non-parametric regressions $Y \sim X$ and $W \sim X$ using a method of our choice to get $\hat{m}(x)$ and $\hat{e}(x)$ respectively.

2. Define transformed features $\widetilde{Y}_i = Y_i - \hat{m}^{(-k(i)}(X_i)$ and $\widetilde{Z}_i = (W_i - \hat{e}^{(-k(i)}(X_i))\psi(X_i)$, using cross-fitting for $\hat{m}(x)$ and $\hat{e}(x)$ as usual.

3. Estimate $\hat{\zeta}_{CF}$ by running the OLS regression $\widetilde{Y}_i \sim \widetilde{Z}_i$.

Using a similar argument as discussed in class last time, we can verify that if all non-parametric regressions satisfy

$$\mathbb{E}\left[(\hat{m}(X) - m(X))^2\right]^{\frac{1}{2}},\ \mathbb{E}\left[(\hat{e}(X) - e(X))^2\right]^{\frac{1}{2}} = o_P\left(\frac{1}{n^{1/4}}\right), \tag{8}$$

then cross-fitting emulates the oracle: $\sqrt{n}(\hat{\zeta}^* - \hat{\zeta}_{CF}) \to_p 0$, and so $\hat{\zeta}_{CF}$ has the same distribution as in (7).

A better way to estimate $\tau(x)$ via the lasso. The above discussion only applies formally to the case where $\tau(x) = \psi(x) \cdot \zeta$ is a finite-dimensional object and $n \to \infty$. However, we can use the resulting construction more generally.

For example, in the case of the lasso, suppose first that we move to a slightly different model:

$$Y_i = X_i \cdot b + (W_i - e(X_i))X_i \cdot \zeta + \varepsilon_i. \tag{9}$$

Then, we could build a lasso-based estimator for ζ as follows:

1. Estimate $m(x) = x \cdot b$ and $e(x)$ via pilot lasso regressions.
2. Define transformed features $\widetilde{Y}_i = Y_i - \hat{m}^{(-k(i))}(X_i)$ and $\widetilde{Z}_i = (W_i - \hat{e}^{(-k(i))}(X_i))X_i$.
3. Estimate $\hat{\zeta}$ by running a lasso of $\widetilde{Y}_i$ on $\widetilde{Z}_i$.

In light of the results of Robinson, we should expect this estimator to behave better than the "algorithmic fix" estimator discussed above. However, a full theoretical description of these methods is still under development.

Finally, in order to use this "better" lasso with a direct connection to the theory of Robinson, we need to be willing to use the modified linear model (9). If this is not acceptable, we could try to estimate $m(x) = (1 - e(x))x\beta_{(0)} + e(x)x\beta_{(1)}$ directly in step one (not using a lasso), and then proceed analogously.

Closing thoughts. The problem of heterogeneous treatment estimation is richer, more difficult, and less well explored than that of average treatment effect estimation. In this lecture, we saw two ideas for building HTE estimators that should be kept in mind for new applications:

- Regularization bias (meaning biases that arise from divergent amounts of regularization for the treatment and control models) can be a real problem if not addressed up front. As a practical measure, we should make sure to mitigate the risk of regularization bias (e.g., use a joint lasso over two separate lassos, a causal tree over two separate trees, etc.).

- The construction of Robinson—originally motivated by asymptotics for semiparametric models—can also be used to motivate more general machine learning procedures for HTE estimation. There is considerable room for further work in several directions.

For further reading, see the following:

Imai, Kosuke, and Marc Ratkovic. **Estimating treatment effect heterogeneity in randomized program evaluation.** *The Annals of Applied Statistics*, 7(1), 2013.

Robinson, Peter M. **Root-N-consistent semiparametric regression.** *Econometrica*, 1988.