

Economics 696F, Causal Inference and Program Evaluation

Lecture Note 4: The Propensity Score

We continue to make the unconfoundedness assumption, and focus on estimating the ATE and the TT.

Note: in order to simplify the notation a bit, I will sometimes drop the i subscript on random variables. So X_i becomes X , etc.

The nonparametric methods discussed in LN3 run into problems if the covariate vector X is high-dimensional. Unless the sample size is huge, it is hard to find two observations that are really “close” to each other along every dimension of X . One alternative, as we saw, was to make some restrictions about the functional form of $E[Y|T, X]$.

Rosenbaum and Rubin (1983) suggest an alternative approach to reduce the dimensionality of the problem. They note that if the treatment is independent of the potential outcomes given X , it might also be independent given a (lower-dimensional) function of X .

The particular function they propose is called the propensity score, defined as:

$$p(x) := Pr(T = 1|X = x).$$

Notice that this is also equal to $E[T|X = x]$.

One way to estimate $p(x)$ is to use the logit series estimator of LN3. For now, let’s suppose that the function $p(x)$ is actually known.

Theorem 1 *Suppose that*

$$Y(0), Y(1) \perp T \mid X,$$

and that

$$0 < Pr(T = 1|X = x) < 1$$

for all x in the support of X . Then

$$Y(0), Y(1) \perp T \mid p(X).$$

Proof: We will show that $Pr(T = 1|Y(0), Y(1), p(X)) = Pr(T = 1|p(X)) = p(X)$, which implies that T is independent of $(Y(0), Y(1))$ conditional on $p(X)$. First,

$$\begin{aligned} Pr(T = 1|Y(0), Y(1), p(X)) &= E[T|Y(0), Y(1), p(X)] \\ &= E\left[E[T|Y(0), Y(1), p(X), X] \mid Y(0), Y(1), p(X)\right] \\ &= E\left[E[T|Y(0), Y(1), X] \mid Y(0), Y(1), p(X)\right] \\ &= E\left[E[T|X] \mid Y(0), Y(1), p(X)\right] \end{aligned}$$

$$\begin{aligned}
&= E\left[p(X)\middle|Y(0), Y(1), p(X)\right] \\
&= p(X).
\end{aligned}$$

The same argument shows that

$$\begin{aligned}
Pr(T = 1|p(X)) &= E[T|p(X)] \\
&= E\left[E[T|X, p(X)]\middle|p(X)\right] \\
&= E[p(X)|p(X)] = p(X).
\end{aligned}$$

□

This result implies that we can replace the covariate X with the scalar $p(X)$ in our previous regression-based approach. More specifically, let

$$\begin{aligned}
s_0(p) &:= E[Y(0)|p(X) = p], \\
s_1(p) &:= E[Y(1)|p(X) = p].
\end{aligned}$$

By Theorem 1,

$$\begin{aligned}
s_0(p) &= E[Y|T = 0, p(X) = p] \\
s_1(p) &= E[Y|T = 1, p(X) = p].
\end{aligned}$$

If we know the propensity score, then we can use any of the nonparametric regression techniques to estimate s_0 and s_1 . Then we can form the following estimator of the ATE:

$$\hat{\tau} := \frac{1}{n} \sum_{i=1}^n [\hat{s}_1(p(X_i)) - \hat{s}_0(p(X_i))].$$

For the effect of the treatment on the treated, we could use

$$\hat{\tau}_t = \frac{\sum_{i=1}^n 1(T_i = 1) [Y_i - \hat{s}_0(p(X_i))]}{\sum_{i=1}^n 1(T_i = 1)}.$$

As in the previous LN, think about the case where we use the single-nearest neighbor estimator of $s_0(p)$. Then this amounts to finding the untreated individual with propensity score closest to treated individual i . Now we are “matching” individuals based on the scalar propensity score, instead of the multidimensional variable X_i .

Alternatively, we can use a weighting-based estimator. Consider the following expectation:

$$\begin{aligned}
E\left[\frac{T \cdot Y}{p(X)}\right] &= E\left[\frac{T \cdot Y(1)}{p(X)}\right] \\
&= E\left[E\left[\frac{T \cdot Y(1)}{p(X)} \middle| X\right]\right] \\
&= E\left[\frac{1}{p(X)} E[T \cdot Y(1) | X]\right]
\end{aligned}$$

$$\begin{aligned}
&= E \left[\frac{1}{p(X)} E[T | X] \cdot E[Y(1) | X] \right] \\
&= E \left[\frac{p(X)}{p(X)} \cdot E[Y(1) | X] \right] \\
&= E [E[Y(1) | X]] \\
&= E[Y(1)].
\end{aligned}$$

Using a similar argument,

$$E \left[\frac{(1-T) \cdot Y}{1-p(X)} \right] = E[Y(0)].$$

Thus,

$$ATE = E \left[\frac{T \cdot Y}{p(X)} - \frac{(1-T) \cdot Y}{1-p(X)} \right].$$

This suggests the following estimator for ATE, assuming the propensity score function is known:

$$\tilde{\tau} := \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i Y_i}{p(X_i)} - \frac{(1-T_i) Y_i}{1-p(X_i)} \right).$$

We can think of the terms $T_i/(np(X_i))$ and $(1-T_i)/n(1-p(X_i))$ as weights, making the estimator the difference between two weighted averages. One problem in practice with this estimator is that the weights do not necessarily add up to 1. We could modify the estimator by normalizing the weights so they add up to one:

$$\tilde{\tau}_{renorm} := \left[\frac{\sum_{i=1}^n \frac{T_i Y_i}{p(X_i)}}{\sum_{i=1}^n \frac{T_i}{p(X_i)}} \right] - \left[\frac{\sum_{i=1}^n \frac{(1-T_i) Y_i}{1-p(X_i)}}{\sum_{i=1}^n \frac{(1-T_i)}{1-p(X_i)}} \right].$$

In practice, this seems to make the estimator a bit more stable.

Now consider TT . We can construct a weighting estimator for this quantity as well, building upon our approach for ATE . First, note that

$$\begin{aligned}
E[Y(1) - Y(0) | T = 1] &= E [E[Y(1) - Y(0) | T = 1, X] | T = 1] \\
&= E [E[Y(1) - Y(0) | X] | T = 1] \\
&= E [m_1(X) - m_0(X) | T = 1] \\
&= \int [m_1(x) - m_0(x)] f(x|T = 1) dx,
\end{aligned}$$

where $f(x|T = 1)$ denotes the conditional density of X given $T = 1$. By Bayes' Rule, we can write

$$\begin{aligned}
f(x|T = 1) &= \frac{P(T = 1|x)f(x)}{\int P(T = 1|x)f(x)dx} \\
&= \frac{p(x)f(x)}{\int p(x)f(x)dx}.
\end{aligned}$$

Here $f(x)$ is the marginal density of X , and $p(x)$ is the propensity score. So

$$\begin{aligned} E[Y(1) - Y(0) \mid T = 1] &= \frac{\int p(x) (m_1(x) - m_0(x)) f(x) dx}{\int p(x) f(x) dx} \\ &= \frac{E[p(X) (m_1(X) - m_0(X))]}{E[p(X)]}. \end{aligned}$$

By iterated expectations, this equals

$$\frac{E[p(X) (Y(1) - Y(0))]}{E[p(X)]},$$

and using our earlier argument, this equals

$$\frac{E\left[p(X) \left[\frac{YT}{p(X)} - \frac{Y(1-T)}{1-p(X)}\right]\right]}{E[p(X)]}.$$

This suggests the estimator:

$$\tilde{\tau}_t = \frac{\frac{1}{n} \sum_{i=1}^n p(X_i) \left(\frac{T_i Y_i}{p(X_i)} - \frac{(1-T_i) Y_i}{1-p(X_i)} \right)}{\frac{1}{n} \sum_{i=1}^n p(X_i)}.$$

Estimating the Propensity Score

For both the regression/matching approach and the weighting approach, we need to know the propensity score. Usually, however, this must be estimated from data.

There are different ways to estimate the propensity score. One method proposed in Hirano, Imbens, and Ridder (2002) is to use the logit series estimator we discussed in LN3. Alternatively, we could use kernel-type regression estimators (see, e.g. Heckman, Ichimura, and Todd (1998)).

Let $\hat{p}(x)$ denote the estimated version of the propensity score. Then we could adapt the regression estimator as:

$$\hat{\tau} := \frac{1}{n} \sum_{i=1}^n [\tilde{s}_1(\hat{p}(X_i)) - \tilde{s}_0(\hat{p}(X_i))].$$

Here, we first have to estimate $p(x)$ with $\hat{p}(x)$. Then we use the estimated propensity scores to obtain estimates $\tilde{s}_0(p)$ and $\tilde{s}_1(p)$. We could construct a similar estimator for TT , using $\hat{\tau}_t$ with the estimated propensity score in place of the true propensity score.

Heckman, Ichimura, and Todd (1998), and Hahn (1998) consider this type of estimator. One drawback is the need to do 3 nonparametric regressions: the propensity score estimation, and two regressions of the outcome on the (estimated) propensity score.

We could alternatively use a weighting estimator, which can be written as:

$$\tilde{\tau} := \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i Y_i}{\hat{p}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{p}(X_i)} \right).$$

(Or we could renormalize the weights as discussed before.)

In general, the estimation of the propensity score does affect the distribution of the estimator, even asymptotically. We'll return to this point in a bit.

For estimating TT , we can also use the previous weighting estimator $\tilde{\tau}_t$, replacing the unknown $p(x)$ with the estimated version.

Semiparametric Variance Bound

Since we have so many different possible estimators for ATE and TT , it's useful to compare them and try to find ones that are, in some sense, optimal. Many of the estimators given above will satisfy consistency and asymptotic normality:

$$\hat{\tau} \xrightarrow{p} \tau,$$

and

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V),$$

for some asymptotic variance V . Here, τ is either ATE or TT as appropriate.

(Note: the single-nearest neighbor approach does not, in general, have a zero-mean asymptotic normal distribution. So while it is intuitively appealing, its asymptotic properties are not as good as some other estimators, and there are some problems with working out valid standard errors, confidence intervals, etc., for this estimator.)

Hahn (1998) shows that the semiparametric variance bound for estimating ATE is

$$V^* = E \left[(m_1(X) - m_0(X) - ATE)^2 + \frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{(1 - p(X))} \right],$$

where $\sigma_1^2(X) = Var[Y(1)|X]$ and $\sigma_0^2(X) = Var[Y(0)|X]$.

What this means, loosely, is that for estimators satisfying consistency and asymptotic normality as above, the asymptotic variance $V \geq V^*$. So if we can find an estimator which has variance equal to the bound, it is "optimal." An interesting result is that the variance bound doesn't change if we happen to know the true propensity score function.

Not all matching-type estimators achieve the variance bound, but Hahn's estimator (which does not use propensity scores, and uses a series regression approach), does. HIR show that the estimator $\tilde{\tau}$, using the series-logit based propensity score estimates, also achieves the bound. In contrast, using the true propensity score, if that were available, would lead to an inefficient estimator.

The corresponding variance bound for TT is similar. It turns out to depend on whether or not the propensity score is known. In the case where $p(x)$ is not known, the variance bound is:

$$V^* = \frac{1}{E[p(X)]^2} E \left[p(X) (m_1(X) - m_0(X) - ATE)^2 + p(X) \sigma_1^2(X) + \frac{p(X)^2 \sigma_0^2(X)}{(1 - p(X))} \right].$$

A matching estimator proposed by Hahn achieves this bound, and so does the weighting estimator $\tilde{\tau}_t$ with estimated propensity scores.

Support Condition

To be able to estimate ATE or TT , we require that

$$0 < Pr(T = 1|X) < 1.$$

In other words, the propensity score must be strictly bounded away from 0 and 1. If that were not the case, then there would be some subgroups of the population for which we never observe one of the two possible treatment values.

Note that if we only want to estimate the TT , then we could make do with the weaker requirement that

$$Pr(T = 1|X) < 1.$$

since we do not care about the part of the population with X values such that $P(T = 1|X) = 0$.

Since we observe T and X , this support condition can be checked in the data. A simple, rough way to do this is to estimate the propensity score, and then plot the propensity score values for the two treatment groups. If there is insufficient overlap, then this suggests that there are some treated observations that are not really comparable to any control observations, or vice versa. However, if the propensity score specification is not correct, then it is possible that we could miss potential areas of non-overlap.