Lecture 3 - Propensity Scores

Paul Goldsmith-Pinkham

January 21, 2026

For today, propensity scores. The end goal:

- Have a framework for discussing subpopulations being treated
- A way to link to an underlying economic model

This will provide structure for us later

Propensity score weighting

Begin by recalling our definition of conditional strong ignorability:

Definition 1
We say that D_i is strongly ignorable conditional on a vector $\mathbf{X}_i$ if

1. $Y_i(0), Y_i(1) \perp D_i|\mathbf{X}_i$
2. $\exists \varepsilon > 0$ *such that* $\varepsilon < Pr(D_i = 1|\mathbf{X}_i) < 1 - \varepsilon$.

The important feature that we will engage with today is how conditional strong ignorability can depend on a potential complex and high-dimensional vector $\mathbf{X}$. When $\mathbf{X}$ is complex, it can be challenging to consider how to implement the ATE estimator as we constructed in our proof. Namely, estimating the CATE $\tau(x)$ for all x may be challenging.

What we will explore today is how the *propensity score* can be used to elide this problem. Let $\pi(\mathbf{X}_i) \equiv E(D_i|\mathbf{X}_i) = Pr(D_i|\mathbf{X}_i)$ be the probability of treatment conditional on $\mathbf{X}$. The propensity score is a scalar summary of the high-dimensional $\mathbf{X}$.[1]

A key result from **?**[2] is that if

$$Y_i(0), Y_i(1) \perp D_i|\mathbf{X}_i$$

holds, then so does $Y_i(0), Y_i(1) \perp D_i|\pi(\mathbf{X})$. The propensity score here acts as the coarsest possible "balancing score" such that the distribution of $\mathbf{X}$ is the same for both the treated and control groups. Why is this useful? It solves a high-dimensional problem – instead of exactly matching on many different values in $\mathbf{X}$, we only have to worry about a single scalar $\pi(\mathbf{X})$.

Conditioning on a single propensity score is a slightly weaker condition than conditioning on the full vector $\mathbf{X}$. However, it opens up new questions and estimation issues.

[1] A student of linear regression might notice that the propensity score is analagous to the auxiliary regression of a regression setup of Y_i on $\mathbf{X}_i$ and D_i. In essence, the propensity score captures the bias in the coefficient on D_i that would occur from omitting $\mathbf{X}$ from the main regression.

[2] Aptly named "The Central Role of the Propensity Score in Observational Studies for Causal Effects."

1. First, how do we estimate the propensity score? When $\mathbf{X}$ is discrete, we can estimate $\pi(\mathbf{X})$ non-parametrically by calculating $E(D_i|\mathbf{X}_i = x)$ for every x value, but that may ask quite a bit of the data. An alternative approach is to assume a model for $\pi(\mathbf{X})$, such as a logistic regression model. This is a parametric approach, but it can be more efficient if the model is correctly specified. A third approach is to use a parametric model but to include flexible terms for $\mathbf{X}$ to allow for non-linearities. This is a semi-parametric approach, and it can be more flexible than the fully parametric approach.

2. Second, once we have an estimated $\pi(\mathbf{X})$, how do we use it to construct the ATE or other estimands? If we directly treat it as a covariate, it becomes a bit challenging, as we discover in Example 1. Instead, we will use another beautiful result from **?** that shows how we can use the Horvitz-Thompson estimator to construct the ATE using the propensity score.

Horvitz-Thompson Estimator

Recall our estimator for the average treatment effect from last lecture, the Horvitz-Thompson estimator:[3]

[3] This is sometimes referred to as the inverse propensity score (or weighting) (IPW) estimator. But if you want to seem fancy you can call it the HOrvitz-Thompson estimator. Impress your statistics colleagues!

Definition 2
*We observe a sample of (Y_i, X_i, D_i) triples for n observations. Let $\pi(\mathbf{X}_i) = Pr(D_i = 1|\mathbf{X}_i)$ be the **propensity score** and define the **Horvitz-Thompson estimator** for the average treatment effect as:*

$$\hat{\tau}_{HT} = n^{-1} \sum_{i=1}^{n} \frac{D_i Y_i}{\pi_i(X_i)} - n^{-1} \sum_{i=1}^{n} \frac{(1 - D_i) Y_i}{1 - \pi_i(X_i)}$$

This estimator is the direct empirical analog to the following population result:

$$E(\tau_i) = E\left(\underbrace{\frac{Y_i D_i}{\pi(\mathbf{X})}}_{E(Y_i(1))} - \underbrace{\frac{Y_i(1 - D_i)}{1 - \pi(\mathbf{X})}}_{E(Y_i(0))} \right)$$

This problem takes advantage of the estimand of interest, rather than tryign to address the problem literaly by estimating every CATE and weighting up accordingly.[4]

[4] Convince yourself that under discrete and few X, this collapses to what we would logically do anyway. With many X, you typically may be willing to make modeling assumptions on π for efficiency reasons.

Example 1 (Propensity score matching)
Consider the following example from **?**, *with 6 observations and two variables X_{i1} and X_{i2}:*

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	-	2	1	1	7	0.33
2	5	-	0	0	7	0.14
3	-	3	1	10	3	0.73
4	-	10	1	3	1	0.35
5	-	2	1	5	2	0.78
6	0	-	0	7	0	0.70

Ideally, we would match observations based on the exact propensity score, but as we can already see in this example, no two observations have the same $\pi(\mathbf{X})$. Instead, we have to approximate matches. This will create bias unless we assume that the distance will shrink as our number of observations grows.

It becomes a question of exactly how to construct the matches, and how many matches to choose. Do you pick just the closest neighbor? All neighbors within a fixed distance? This is a complicated problem that can affect inference and discussed in **?**. *As an example, consider what happens if we impute based on the closest neighbor for each observation:*

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	5	2	1	1	7	0.33
2	5	2	0	0	7	0.14
3	0	3	1	10	3	0.73
4	5	10	1	3	1	0.35
5	0	2	1	5	2	0.78
6	0	3	0	7	0	0.70

In this case, several units are used more than once as matches. Moreover, there are very close "ties": we picked unit $i = 1$ for unit 2, but unit 4 was very close as well. Why not pick that one, especially with $\pi(\mathbf{X})$ is noisy? This is a difficult problem to solve, and it is not clear that there is a single best answer.

The Horvitz-Thompson estimator works well but in smalll samples can be high variance if $\pi(\mathbf{X})$ is close to zero or one. This can be improved through the use of *stabilized weights*:

$$\hat{\tau}_{SIPW} = \frac{\frac{1}{n}\sum_i \frac{Y_i D_i}{\hat{\pi}(\mathbf{X}_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}} - \frac{\frac{1}{n}\sum_i \frac{Y_i(1-D_i)}{1-\hat{\pi}(\mathbf{X})}}{\frac{1}{n}\sum_i \frac{(1-D_i)}{1-\hat{\pi}(\mathbf{X})}}$$

This estimator benefits by adjusting for unusually high or low

values of $\pi(\mathbf{X})$ by constructing weights $w_i = \frac{\frac{D_i}{\hat{\pi}(\mathbf{X}_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}}$ for the treated group.[5] Similar to the IPW, this is also an unbiased estimator of the ATE.

[5] In the limit, recall that $\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}$ should converge to 1 (since $E(\pi^{-1}(\mathbf{X}_i)D_i) = 1$), and hence the SIPW converges to the IPW.

Comment 1 (True versus estimated propensity scores) *When an intervention is truly randomly assigned, the propensity score itself is known. However, in most non-experimental settings, the p-score is unknown and must be estimated. As we discussed above, we need to estimate $\pi(X)$ and it can be done in parametric, semi-parametric, or non-parametric fashion, depending on your assumptions about* $\mathbf{X}$.

It is useful to note that the model used to estimate the propensity score matters. For example, the linear probability model, which is commonly used for many binary outcomes, may predict probabilities for the propensity score that are outside the range of $[0, 1]$, thereby generating improper IPW estimates. The LPM will work if the model is fully saturated and non-parametric (e.g. a set of fully interacted dummies), but this is not always the case.

Another important result in this literature is that even if you know the true function $\pi(\mathbf{X})$, *you are better off using the estimated function than the true $\pi(\mathbf{X})$ [**?**]. The intuition for this result is that the deviations from the "true" propensity score ($\hat{\pi}(\mathbf{X}) - \pi(\mathbf{X})$) are informative for the estimation of the treatment effects (a la extra moment restrictions in GMM)*

Contrasting linear regression with propensity scores

Say we have strong ignorability conditional on $\mathbf{X}$ and we run the following regression:

$$Y_i = \gamma_0 + D_i\beta + X_{i1}\gamma_1 + X_{i2}\gamma_2 + u_i. \tag{1}$$

How should we contrast this to the propensity score methods we used above?

We can revisit the **?** example from Example 1, but instead of imputing for the missing counterfactual using matching, we impute using regression.

This tells us, roughly speaking, what this regression approach will assume for the missing counterfactuals. Notably, this approach will do well when the conditional expectation function for the outcome (e.g. $E(Y_i(0))$) is approximately linear in $\mathbf{X}$ and D_i – e.g. you are roughly correctly specified. Importantly, the approximation should not extrapolate too much across the support of $\mathbf{X}_i$.[6]

This is just another way to infer the missing data. However, a

[6] To concretely give an example of how this could be an issue: note that unit 3 is treated and is far out on the support of $\mathbf{X}_i$ (10,3). Imputing values for that unit's control requires extrapolating quite far out on the support of $\mathbf{X}_i$, which may be problematic unless the conditional mean is exactly correctly specified.

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	2	1	1	7	0.33
2	5	$\hat{\gamma}_0 + \hat{\beta} + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	0	0	7	0.14
3	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	3	1	10	3	0.73
4	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	10	1	3	1	0.35
5	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	2	1	5	2	0.78
6	0	$\hat{\gamma}_0 + \hat{\beta} + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	0	7	0	0.70

Table 1: Regression imputation in **?** example

key issue is that if we just use OLS to estimate Equation (1), we will not necessarily recover the ATE with τ. Recall that when $\mathbf{X}$ is just a constant, then *tau* is the ATE. Once we condition on covaraites, however, the estimand recovered by τ changes. We will revisit this in our linear regression lectures, but the key fact is that OLS will recover an estimand which is a different weighted average of the CATEs – namely one that variance-weights the CATEs.

First, it's useful to recall what we are assuming with strong ignorability. Recall that we are assuming that there is a function $E(D_i|\mathbf{X}) \equiv \pi(\mathbf{X})$ that we can condition on such that there is no remaining correlation between D_i and the potential outcomes. However, we do not know the function $\pi(\mathbf{X})$. If we are additionally willing to assume that $\pi(\mathbf{X}) = X_{i1}\gamma_1 + X_{i2}\gamma_2$, then by Frisch-Waugh-Lovell, we can show that τ will recover a weighted combination of the CATEs.[7] Specifically, let $\tilde{D}_i = D_i - \pi(\mathbf{X}_i)$ and $\tau_i = Y_i(1) - Y_i(0)$. Then,

$$\beta_{OLS} = \frac{E(\tilde{D}_i D_i \tau_i)}{E(\tilde{D}_i^2)}. \tag{2}$$

As a result, the OLS coefficient will not necessarily estimate the ATE estimand. Instead, it will estimate a weighted average of the CATEs, where the weights are the variance of the treatment assignment conditional on $\mathbf{X}$. We will revisit this during our linear regression lectures.

As a result, there are two key caveats with this approach: first, we need to be careful about how we specify the regression function with our $\mathbf{X}$ variables. Especially if there is significant extrapolation across the support of $\mathbf{X}$.[8] Second, we need to be careful about how we interpret the coefficient on D_i in the regression. It will not necessarily be the ATE, but instead a variance-weighted average of the CATEs.

Consequentially, an important question for us as practitioners is which approach to use: IPW or regression? There are good reasons to like the IPW estimator for the ATE: **?** show that the IPW estimator is semiparametrically efficient when the propensity score is unknown.[9] But, linear regression is nice! It is straightforward to run, and easy to interpret. Moreover, it has been endorsed. From Angrist and Pischke

[7] This discussion here follows **?**, and is generally inspired by **?**.

[8] Note that in cases where we have two sets of fixed effects (e.g. age and location), we often just include the marginal fixed effects (e.g. age and location separately) and not the fully saturated specification(e.g. age $\times$ location). This is because the interaction fixed effects require too much of the data. However, there will consequentially be extrapolation that may be wrong.

[9] However, the IPW estimator can be high variance when the propensity score is close to zero or one (this is known as weak overlap). This can create issues generally with estimating the ATE, and linear regression avoids this issue. See **?**.

(2009):

> We believe regression should be the starting point for most empirical projects. This is not a theorem; undoubtedly, there are circumstances where propensity score matching provides more reliable estimates of average causal effects. The first reason we don't find ourselves on the propensity-score bandwagon is practical: there are many details to be filled in when implementing propensity-score matching - such as how to model the score and how to do inference - these details are not yet standardized. Different researchers might therefore reach very different conclusions, even when using the same data and covariates. Moreover, as we've seen with the Horvitz-Thompson estimands, there isn't very much theoretical daylight between regression and propensity-score weighting.

Comment 2 (Ex-Post Weights vs. Ex-Ante Weights) *Note that in Equation (2),*

$$\phi_i(\mathbf{X}_i) = \tilde{D}_i D_i$$

is a function of $\mathbf{X}_i$ and can be negative. Then that means for some CATE ($E(\tau_i|\mathbf{X})$), there may be negative weights. In some special cases, this will not be the case (e.g. $\pi(\mathbf{X})$ is correctly specified and/or $\mathbf{X}$ is discrete and fully satuated). These negative weights can be problematic because the weighted TE could then reflect an effect that does not exist in the underlying population at all.

However, as shown succinctly in **?**, *the* expected *ex ante weights in a design-based approach are guaranteed to be positive:*

$$E(\phi_i|\mathbf{X}_i, \beta_i) = Var(D_i|\mathbf{X}_i, \beta_i) > 0.$$

This implies, intuitively, that all units with the same $\mathbf{X}_i$ will have the same weight prior to treatment. This is a key difference between the design-based approach and the model-based approach. This same statement cannot be done in a context where we model $E(Y_i(0))$ because we are not allowing the treatment to be randomly allocated across units (by)

When we start worrying about the propensity score, life gets more complicated. It forces us to think about overlap of covariates and balance, namely how comparable the treated and untreated groups are. This became a key issue around the seminal NSW paper.

NSW and Propensity score matching

There was a randomized intervention called the National Supported Work Demonstration (NSW), which was a temporary employment

program to give work experience. A seminal paper by Lalonde [**?**] showed that a non-experimental analysis of this program would have given biased estimates compared to experimental approach.[10]

?? reanalyze this data using propensity score methods, and argue that these results are more similar to the experimental results, relative to the non-experimental results proposed by **?**. Moreover, using propensity scores provides a form of diagnostics on how comparable the treated and control groups are.

It is worth digging in to the **?** paper to understand how they use propensity scores. Crucial to their approach is they included the lagged outcomes as covariates in their approach. As a consequence, they subsample the data to have two years of pre-treatment data. This is a key difference from the **?** approach, which used the full sample. **?** assess this approach, and argue two points: first, the specification itself for the propensity score is quite sensitive to the choice of included variables.[11] Second, they argue that the *subsampling* in **?** predisposes to a group where the analysis is "easy." Moreover, in this case difference-in-differences works best because it removes the time-invariant heterogeneity across units.

Dehijia's response in **?** is "of course!" – the point of propensity score matching is to transparently highlight the assumptions in the data. More verbosely, he says:

> A judgment-free method for dealing with problems of sample selection bias is the Holy Grail of the evaluation literature, but this search reflects more the aspirations of researchers than any plausible reality. In practice, the best one can hope for is a method that works in an identifiable set of circumstances, and that is self-diagnostic in the sense that it raises a red flag if it is not functioning well. Propensity score methods are applicable when selection is based on variables that are observed. In the context of training programs, Dehejia and Wahba (1999, 2002), following on a suggestion from the training program literature (Ashenfelter, 1978; Ashenfelter and Card, 1985), suggest that two or more years of pre-treatment earnings are necessary. In terms of the self-diagnosis, the method and its associated sensitivity checks successfully identify the contexts in which it succeeds and those in which it does not succeed, at least for the NSW data.
>
> **Propensity score matching does not provide a silver-bullet, black-box technique that can estimate the treatment effect under all circumstances**; neither the developers of the technique nor Dehejia and Wahba have claimed otherwise. However, with input and judgment from the researcher, it can be a useful and powerful tool. [Emphasis added]

What causes residual variation in treatment?

We initially motivated strong ignorablity under settings with random assignment or something approximating it. However, in many

[10] "This comparison shows that many of the econometric proce- dures do not replicate the experimentally determined results."

[11] In other words, deciding which variables to include in $\mathbf{X}$ is important and can affect bias significantly. This echoes Leamer's critique on model specification in **?**.

settings, the data researchers will use is exclusively observational and does not have random assignment. Despite that, they want to estimate a causal effect.

To quote **?**:

> Ironically, missing data give rise to the problem of causal inference, but missing data, i.e. the unobservables producing variation in D conditional on X, are also required to solve the problem of causal inference.

In other words, when we control for $\mathbf{X}_i$, there must be additional source of variation in D_i that we do not capture that drives differences in the choice of treatment, but is also unrelated to the potential outcomes.

Example 2 (Why do we need residual variation?)
Consider D to be a medical treatment selected by a doctor, with Y their subsequent health outcome. What would happen if D was perfectly predictable by $\mathbf{X}$: e.g., age of patient, the doctor's background, etc. In other words, if we know $\mathbf{X}$, then we know D.
In this setting where the $\mathbf{X}$ perfectly predict the treatment, is the effect of D on Y identifie after conditioning on $\mathbf{X}$? No. See this in two ways:

- *$Pr(D_i|\mathbf{X}_i) = 1$ or 0 and we fail overlap (and thus strong ignorability)*
- *$Y_i = D_i\tau + \mathbf{X}_i\gamma + \epsilon$ is our estimating equation, but $\mathbf{X}$ and D are perfectly collinear.*

To estimate the effect of D and Y, we need additional "exogeneous" variation.

A structural econometrician would describe the variation in D as driven by two pieces, V and X. Ideally, V is exogeneous. But what actually is V? Much of the time we don't know. This comes back to our research design question. Is there something "near-random" that caused a difference in treatment? Or if we choose to be pessimistic, if units are observably identical, but choose different outcomes, a purely rational model would suggest there are intrinsically different characteristics driving this decision. How will this bias our estimates (if at all)?

Consider Figure 1. There are many parts of $\pi(\mathbf{X})$ where there is lots of overlap between the treatment and control group. However, in some parts of the distribution there is significantly less, especially where $\pi(\mathbf{X}) < 0.5$. What does it mean to have so few treated units for the pscore less than 0.5? This suggests that ther are a small share of

units who are both treated and look observably similar to a large set of the control group. It seems plausible that these units might not be comparable. If we choose not to use the units at the "extremes", e.g. by selecting on only propensity scores between 0.5 and 0.75, what would that imply about our model estimates? This would be targeting a very particular estimand that may or may not be of interest.

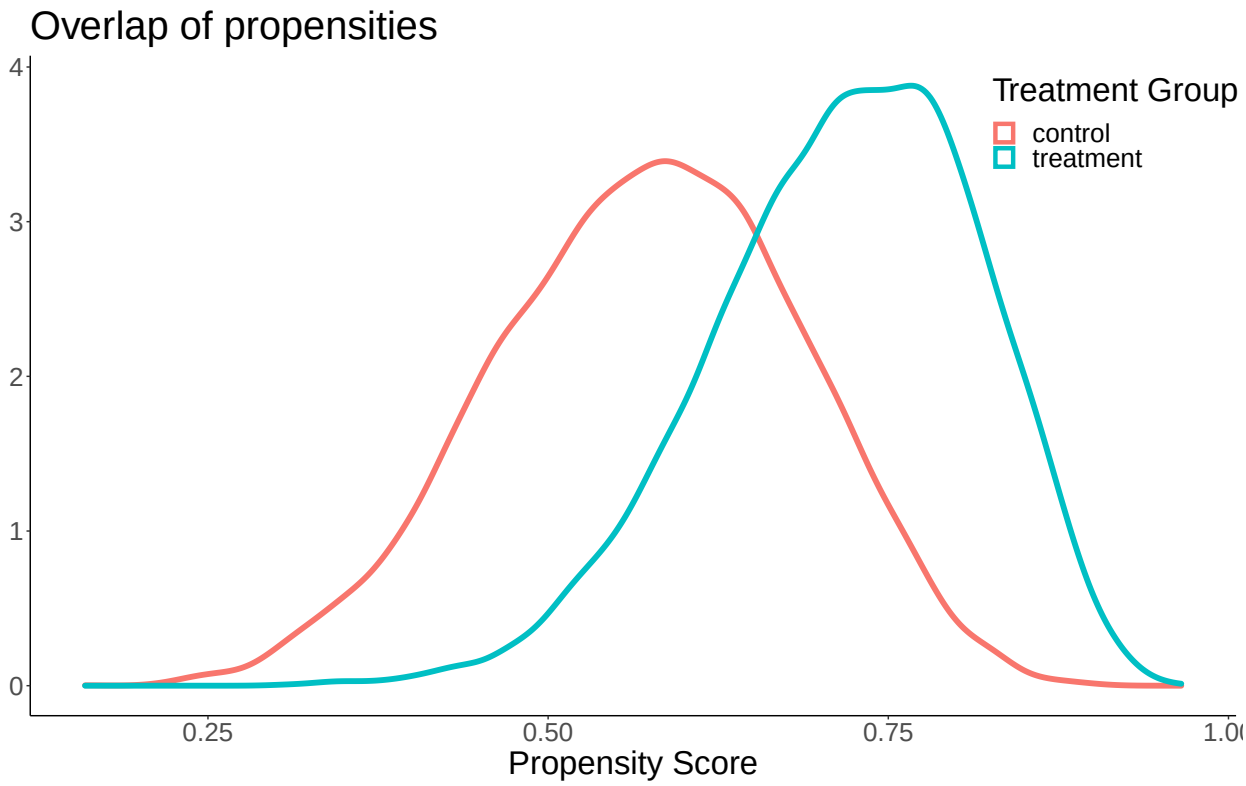


Figure 1: Overlap in the propensity scores for example treated and control populations

As we go forward in thinking about random variation in D_i, it is convenient to consider the following economic model (from **?**):

$$Y_i(0) = g(X_i, D_i = 0) + U_{i0}$$
$$Y_i(1) = g(X_i, D_i = 1) + U_{i1}$$
$$Y_i = g(X_i, 0) + D_i \left(\underbrace{g(X_i, 1) - g(X_i, 0)}_{\text{Average Population Gain}} + \underbrace{U_{i1} - U_{i0}}_{\text{idiosyncratic gain}} \right) + U_{i0}$$

Now, we consider what drives the decision making for D_i:

$$D_i = 1((Y_i(1) - Y_i(0)) + \kappa + V_i > 0)$$

In other words, when the value is sufficiently high (above some overall + idiosyncratic cost $\kappa + V_i$), I choose to take the program. This creates obvious correlation between D_i and $(Y_i(0), Y_i(1))$.

WIth this setup, we can consider under what settings controlling for **X** will be sufficient to recover a causal estimand. The easiest is when there are constant effects, e.g. $U_{i1} - U_{i0} = 0$ for everyone. In this case, random variation in V_i is what drives takeup, and is unrelated to the outcome. This makes life easy for us, but is not very interesting and also means there is no underlying heterogeneity in the treatment effect beyond what we observe in the characteristics.

The other case is when the *expected* gains to the program is the same for everyone ($E(U_{i1} - U_{i0}|X_i) = 0$), perhaps because of lack of information on the part of the individuals'. Then, while there may be ex post differences in treatment effects, they are not ex ante anticipated by the individuals and consequentially selected on.

The propensity score in this model is:

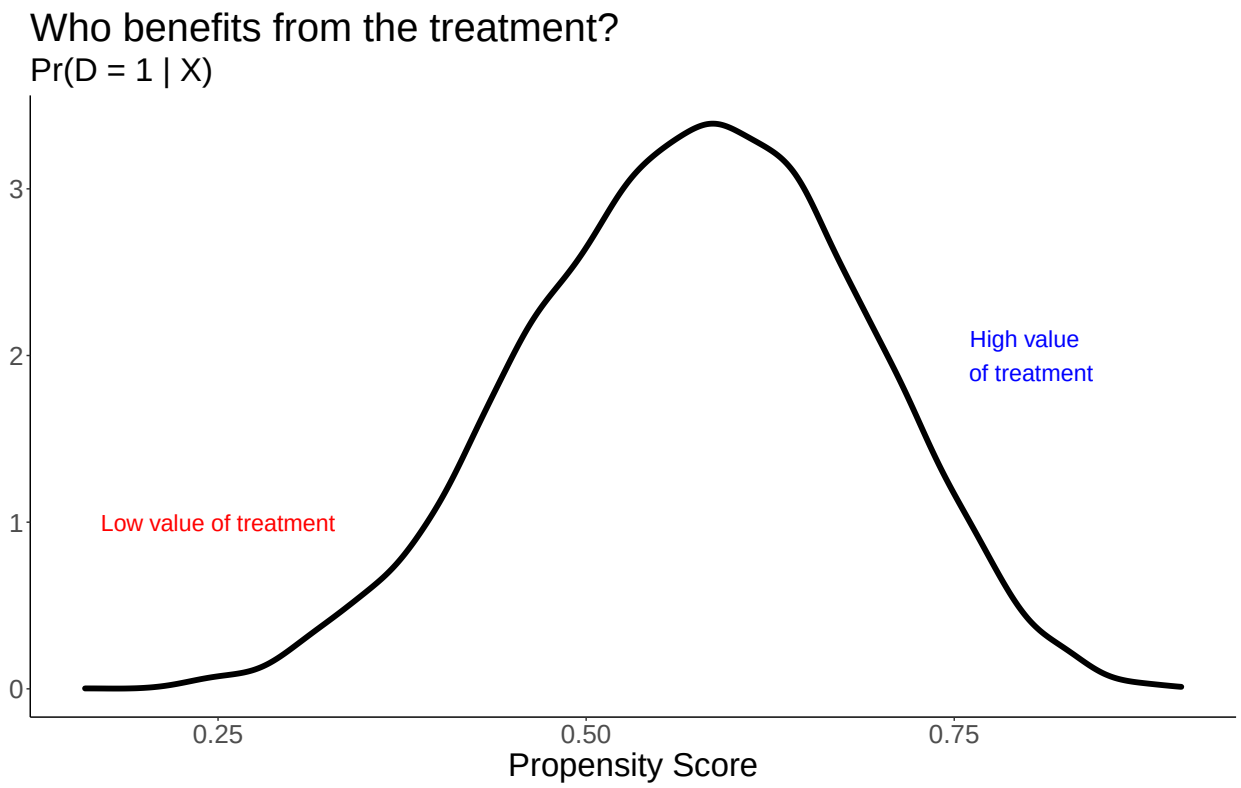
$$Pr(D_i = 1|X_i) = Pr\left(g(X_i, 1) - g(X_i, 0) + \kappa + (U_{i1} - U_{i0}) > V_i\right)$$

This gives us a framework to consider the economic returns to individuals to take a program, as in **??** . What would it take to switch them into the program?

1. Lack of choice: not always available
2. Large incentive: expensive
3. High personal returns: selects into a particular type of person.

It is useful to remember this graph when considering how to induce participation. Some folks may just not want to participate. This could be perceptions on the returns (e.g. $Y(1) - Y(0)$), rightly or wrongly, but as a consequence, they will be expensive to move. The estimand of interest will be considering parts of this distribution, and hence it si important to consider this when thinking about external validity.

Going forward, it is helpful to consider the propensity score as an index of valuation. We are hence looking for things that vary individuals' valuation and do not correlate with the potential outcome. In design-based work in economics, this is often referred to as an *instrument*, and is a crucial part of the design-based research.


Who benefits from the treatment?
Pr(D = 1 | X)
3
2
1
0
High value
of treatment
Low value of treatment
0.25
0.50
0.75
Propensity Score