Propensity Scores

Paul Goldsmith-Pinkham

January 20, 2026

This week

- Discuss two points: propensity scores and interference
  1. Propensity scores
  2. Interference and violations of SUTVA
- 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

Estimation of treatment effects

- Recall two results:
  - Horvitz-Thompson Estimator

$$\hat{\tau}_{HT} = n^{-1} \sum_i \pi_i^{-1} Y_i D_i - (1 - \pi_i)^{-1} Y_i (1 - D_i)$$

    Unbiased estimator of τ_{ATE} , here $\pi_i = Pr(D_i = 1|X_i)$
  - Conditional strong ignorability: D_i is *strongly ignorable* conditional on a vector $\mathbf{X}_i$ if
    1. $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|\mathbf{X}_i$
    2. $\exists \epsilon > 0$ s.t. $\epsilon < \Pr(D_i = 1|X_i) < 1 - \epsilon_i$

- Key: $\pi(X_i) = Pr(D_i = 1|X_i)$ is important
  - This is the *propensity score*.
  - We will dive into this today

Why does the propensity score matter? Rosenbaum-Rubin (1983)

- Note our strong ignorability condition, $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | \mathbf{X}_i$ conditions on $\mathbf{X}_i$, which can be quite high dimensional
- Key result from Rosenbaum-Rubin: if the above holds, then so does $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i | \pi(\mathbf{X}_i)$.
  - The intuition comes from the fact that conditional on $\pi(\mathbf{X}_i)$, the distribution of X is the same for the treated and untreated, and thus X_i and D_i are independent.
- Why does this matter? Crucially, solves a high-dimensional problem – now we just need to condition on a single scalar value (π_i)

Layering complications – how to match?

- You ideally match on exactly the propensity score
  - However, "Unfortunately, exact matches even on a scalar balancing score are often impossible to obtain, so methods which seek approximate matches must be used."
  - This creates bias. How? Consider this example from Aronow and Miller

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

Layering complications – how to match?

- You ideally match on exactly the propensity score
  - However, "Unfortunately, exact matches even on a scalar balancing score are often impossible to obtain, so methods which seek approximate matches must be used."
  - This creates bias. How? Consider this example from Aronow and Miller

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

- We need to construct $E(Y_i(1)|\pi(X))$ and $E(Y_i(0)1)|\pi(X))$ for each observation. How do we pick now? Closest p-value? What are the issues with this?
  - We picked unit $i = 2$ for unit 1, but unit 4 was very close. Why not pick that one?
  - More generally, this approach has challenges for inference, especially with $\pi(\mathbf{X})$ unknown (see Abadie and Imbens (2008))

What to do instead of matching?

- Note that matching addresses the problem literally
  - Since the ignorability statement is with respect to X, natural to match on it
  - But this ignores the estimand – always focus on the estimand!

- Key result: consider the follow result about the population version of the Horvitz-Thompson estimator, sometimes referred to as the inverse probability weighting estimator:

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

- This is an amazing result!
  - Under discrete $\mathbf{X}$, this collapses to what we would logically do anyway

A more stable IPW estimator

- The IPW approach works well, but in small samples can be high variance if you get big $\pi(\mathbf{X})$ values.
  - We can slightly improve on it using the stabilized IPW estimator:

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

- This estimator benefits by adjusting for unusually high or low values of $\pi(\mathbf{X})$
  - Note that this estimator effectively constructs $w_i = \frac{\frac{D_i}{\hat{\pi}(\mathbf{X}_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}}$ for the treated group – reweighting by the average density within the sample
  - In the limit, this just goes to one but works well in finite samples

True vs. estimated propensity scores?

- True propensity scores are only known sometimes (e.g. randomized experiments)

- In most non-experimental settings, the p-score is unknown and must be estimated

- When estimating, we have two cases:
  - If X is discrete, we know that $\hat{\pi}(X)$ can be an exact approximation (why?)
  - If X is *not* discrete (or high-dimensional), how should we approximate it?

- We need to estimate $\pi(X)$ in a way that is flexible and will converge to the truth in the limit – e.g. semi-parametric estimation of π
  - Note a linear model of π will inherently be wrong b/c probabilities are bounded between 0 and 1
  - Practical implication: logit estimation of $\pi(X)$ is reasonable, allowing for flexible specification of X
    - As dimension of X grows, ML / lasso style models grow in value

True vs. estimated propensity scores?

- Important result: *even if you know the true function* $\pi(\mathbf{X})$, better to use the estimated function than the truth (Imbens, Hirano and Ridder (2002))
  - Intuition: 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)

- Clear tension – as dimension of controls increases, the noisiness in π grows as well
  - Is it reasonable to consider this a good research design?

- We will consider this in a basic way on the homework

Contrasting propensity scores with regression

- Say we have strong ignorability and we run the following regression

$$Y_i = \gamma_0 + D_i\tau + X_{i1}\gamma_1 + X_{i2}\gamma_2 + u_i$$

How should we contrast this to some pscore approach?

Contrasting propensity scores with regression

- Say we have strong ignorability and we run the following regression

$$Y_i = \gamma_0 + D_i\tau + X_{i1}\gamma_1 + X_{i2}\gamma_2 + u_i$$

How should we contrast this to some pscore approach?

- Revisit the Aronow and Miller example

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

Contrasting propensity scores with regression

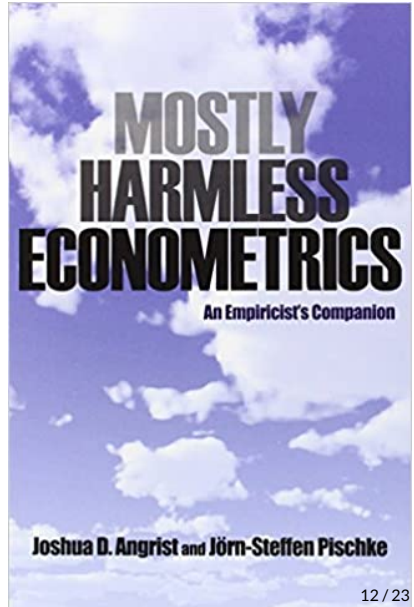
- When will this approach do well?
  - If the outcome conditional expectation function is approximately linear
  - We don't have to extrapolate too much across the support of $\mathbf{X}_i$
  - Put differently – OLS is the best linear predictor. We might as well take advantage of that fact!

- Key point: this is just another way to infer the missing data. We will rely on regression heavily, but important to not forget that this is just another way to infer values of missing data.

- Substantial *empirical* debate about which approach is best.
  - Debatable in finite samples but Imbens Hirano & Ridder (2003) shows that if pscore is unknown, using the estimated non-parametric pscore is semiparametric efficient
  - We will revisit in linear regression

Angrist and Pischke's advocation for regression

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.*



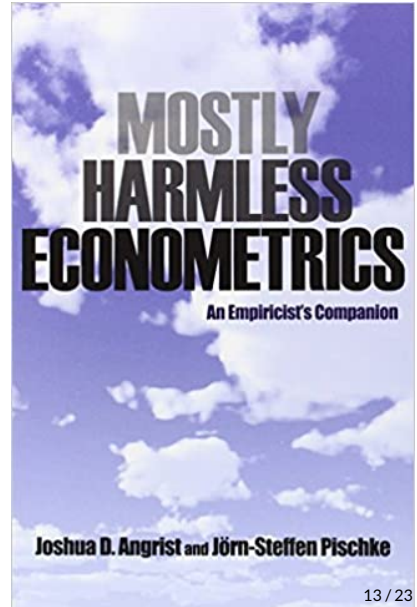
Angrist and Pischke's advocation for regression

- This point about the IPW estimator comes from the fact that when the covariates are discrete, a fully saturated regression model can be written as an IPW estimator.

- See MHE for this discussion, but the intuition comes from a correctly specified propensity score (due to the full saturation) and residual regression.

- The punchline is that when we start having to worry about the pscore (i.e. in worlds with many covariates and/or continuous covariates), life gets complicated.

- As we will discuss below, forces us to think about overlap of covariates and balance
  - E.g. comparability between treated and untreated groups



Crucial empirical context: Lalonde (1986), Dehijia and Wahba (1999,2002), Smith and Todd (2005)

- There was a randomized intervention called the NSW (National Supported Work Demonstration) – temporary employment program to give work experience
- Key takeaway from Lalonde (1986) – non-experimental analaysis of this program (e.g. defining control gorup using non-experimental data) would have given biased estimates compared to experimental approach
  - That's bad! "This comparison shows that many of the econometric proce- dures do not replicate the experimentally determined results."
- Dehijia and Wahba reanalyze this data using pscore methods
  - Key point – using pscores gets you closer *and* provides a form of diagnostics on how comparable the groups
  - Necessary consequence of these methods – need to subsample the data to have 2 years of pre-treatment data to match well
- Smith and Todd reanalyze this approach, and argue that the *subsampling* predisposes to a group where the analysis is "easy."
  - Dehijia's response – "of course!"

Dehijia's conclusion

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.

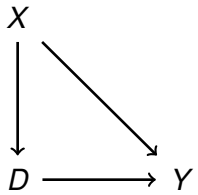
The problem with propensity scores/matching/observational data

- We initially motivated strong ignorablity under settings with random assignment or something approximating it
  - However, in many settings, researchers will use exclusively observational data and use this to estimate a causal effect

- To quote Heckman, Todd and Ichimura: “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 – if we’re controlling for X, there must be additional source of variation in D that we’re not capturing.

  - Why? Think on this, then example next slide



Example of variation necessary in D

- Consider D to be a medical treatment selected by a doctor, with Y their subsequent health outcome
  - What if D was perfectly predictable by $\mathbf{X}$: e.g., age of patient, the doctor's background, etc.
  - In other words, if we know $\mathbf{X}$, we know D.

- Is the effect of D on Y identified, conditional on $\mathbf{X}$?

Example of variation necessary in D

- Consider D to be a medical treatment selected by a doctor, with Y their subsequent health outcome
  - What if D was perfectly predictable by $\mathbf{X}$: e.g., age of patient, the doctor's background, etc.
  - In other words, if we know $\mathbf{X}$, we know D.

- Is the effect of D on Y identified, conditional on $\mathbf{X}$?

- No. See this in two ways:
  - $Pr(D_i|\mathbf{X}_i)$ = 1 or 0 (fails strong ignorability)
  - $Y_i = D_i\tau + \mathbf{X}_i\gamma + \epsilon$ – perfectly collinear!

- Need additional "exogeneous" variation

Wanted: exogeneous variation

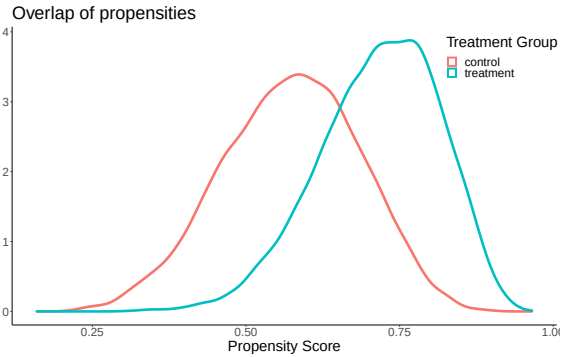
- A structural econometrician would describe the variation in D as driven by two pieces, V and X. Ideally, V is exogeneous.

- But what 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?

- More worryingly – if units are observably identical, but choose different outcomes, a purely rational model would suggest there are intrinsically different characteristics driving this decision. Will this bias our estimates?

Consider a p-score overlap example

- There are many parts of $\pi(\mathbf{X})$ where there is lots of overlap
- In some parts it becomes less common
- What does it mean to have so few treated units for the pscore less than 0.5?
  - I would worry that these units are somehow not comparable.
  - If we select away from them, what does that imply about our model estimates?



Where we will go with this

- A convenient economic model to consider (from Heckman (1997)):

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

- 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 $(Y_i(0), Y_i(1)$

Where we will go with this

- Useful to identify when conditioning works:
  - Constant effects (e.g. $U_{i1} - U_{i0}$) for everyone
  - Expectation is the same for everyone ($E(U_{i1} - U_{i0}|X_i) = 0$, because of lack of info

- The pscore is:

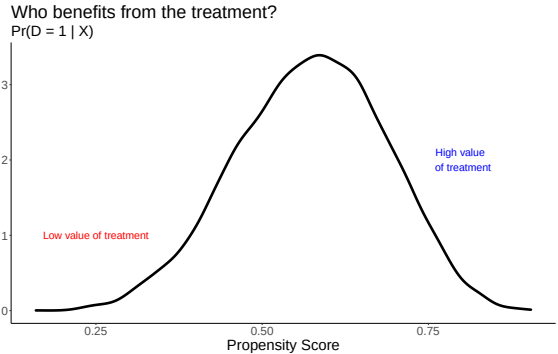
$$Pr(D_i = 1|X_i) = Pr\left(g(X_i, 1) - g(X_i, 0) + \kappa + (U_{i1} - U_{i0}) > V_i\right)$$

- Why is this useful? Gives us a framework to consider the economic returns to individuals take a program

- 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 – that's good, but selects into a particular type of person

How do we randomly vary people's incentives to move in their pscores?

- Consider the pscore as an index of valuation
- We now want something that varies individuals' valuation
  - Will give us "real" variation (e.g., avoiding Heckman et al. critique)
  - Will identify a particular subspace of treated individuals
- This is instrumental variables – a shifter that moves our propensity score values in an exogeneous fashion



How do we randomly vary people's incentives to move in their pscores?

- Useful to remember this graph when considering how to induce participation
- Some folks do not want to participate!
  - Could be perceptions on the returns (e.g. $Y(1) - Y(0)$), rightly or wrongly
  - *They will be expensive to move*
- Your estimand of interest will be considering parts of this distribution
  - Useful when considering external validity