Program Evaluation (c)- Propensity Scores

Chris Conlon

March 29, 2021

Applied Econometrics

Matching

- Compare treated individuals to un-treated individuals with identical observable characteristics X_i.
- Key assumption: everything about $Y_i(1) - Y_i(0)$ is captured in X_i; or u_i is randomly assigned conditional on X_i.
- Basic idea: The treatment group and the control group don't have the same distribution of observed characteristics as one another.
- Re-weight the un-treated population so that it resembles the treated population.
- Once distribution of X_i is the same for both groups $X_i|T_i \sim X_i$ then we assume all other differences are irrelevant and can just compare means.
- Matching assumes all selection is on observables.

Why does any of this work?

Let $F^1(x)$ be the distribution of characteristics in the treatment group, we can define the ATT as

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0)|T = 1] = \mathbb{E}_{F^1(x)}[\mathbb{E}(Y(1) - Y(0)|T = 1, X)] \\ = \quad \mathbb{E}_{F^1(x)}[\mathbb{E}(Y(1)|T = 1, X)] - \mathbb{E}_{F^1(x)}[\mathbb{E}(Y(0)|T = 1, X)] \text{ linearity} \end{aligned}$$

The first part we observe directly:

$$= \quad \mathbb{E}_{F^1(x)}[\mathbb{E}(Y(1)|T = 1, X)]$$

But the counterfactual mean is not observed!

$$= \quad \mathbb{E}_{F^1(x)}[\mathbb{E}(Y(0)|T = 1, X)]$$

But conditional independence does this for us:

$$\mathbb{E}_{F^1(x)}[\mathbb{E}(Y(0)|T = 1, X)] = \mathbb{E}_{F^1(x)}[\mathbb{E}(Y(0)|T = 0, X)]$$

Matching in Practice: Inverse Probability Weighting

How do we actually do this?

- Calculate a smoothed estimate of the treatment probability $\pi(x) = Pr(T_i = 1|x)$.

$$\frac{1}{n}\sum_{t\in\text{Treatment}}\frac{y_t}{\pi(x_t)} - \frac{1}{n}\sum_{s\in\text{Control}}\frac{y_s}{1-\pi(x_s)}$$

- How to get $\pi(x)$? Run a logit or probit.
- We can stabilize the weights replace $w(x) = \frac{1}{\pi(x)}$ with:

$$w(x) = \frac{Pr(T=1)}{\pi(x)} \text{ for } T_i = 1 \quad w(x) = \frac{Pr(T=0)}{1-\pi(x)} \text{ for } T_i = 0$$

- This sometimes helps crazy big weights when treated group is small.

Higher Dimensions

So matching works great in dimension 1. But what if $dim(X) > 1$?

- True high-dimensional matching may be infeasible. There may be no set of weights such that: $f(X_i|T_i = 1) \equiv \int w_i f(X_i|T_i = 0)\partial w_i$.
- One solution is the nearest-neighbor approach in Abadie Imbens (2006).
- This is still cursed in that our nearest neighbors get further away as the dimension grows.
- Suppose instead we had a sufficient statistic

Propensity Score

- Rosenbaum and Rubin propose the propensity score

$$P(T_i = 1|X_i) \equiv P(X_i)$$

- They prove that the propensity score and any function of X, $b(X)$ which is finer serves as a balancing score.
- Finer implies that:

$$\begin{aligned} b(X^1) = b(X^2) \quad &\implies \quad P(X^1) = P(X^2) \\ P(X^1) = P(X^2) \quad &\not\Longrightarrow \quad b(X^1) = b(X^2) \end{aligned}$$

Propensity Score

- Main result: If treatment assignment is strongly ignorable conditional on X (CIA) then it is strongly ignorable $Y(1), Y(0) \perp T|X$ given any balancing score $b(X)$ including the propensity score:

$$\begin{aligned} Pr(T = 1|Y(1), Y(0), P(X)) &= E[Pr(T = 1|Y(1), Y(0), X)|P(X)] \\ &= E[Pr(T = 1|x)|P(X)] = P(X) \end{aligned}$$

- Also we require that $0 < P(X) < 1$ at each X which is known as the support condition.
- The theorem implies that given $P(X)$ we have as if random assignment.

Propensity Score

- Instead of matching on K dimensional X we can now match on a one-dimensional propensity score
- Thus the propensity score provides dimension reduction
- We still have to estimate the propensity score which is a high dimensional problem without *ad-hoc* parametric restrictions.
- Let us begin by assuming a can-opener.
- An easy way would be to use $\pi(x)$ from logit or probit.

Propensity Score

Just like in the matching case the problem arises because we do not observe the counterfactual mean:

$$E_{F^1(x)}[E(Y(0)|T=1,X)]$$

With conditional independence and the propensity score:

$$\begin{aligned} E_{F^1(x)}[E(Y(0)|T=1,X)] &= E_{F^1(x)}[E(Y(0)|T=0,X)] \\ &= E_{F^1(x)}[E(Y(0)|T=0,P(X))] \end{aligned}$$

Kernel Matching

How do we implement?

- Kernels are an obvious choice

$$\widehat{ATT} = \frac{1}{N_1} \sum_{i \in T=1} \left[Y_i - \frac{\sum_{j \in T=0} Y_j K\left(P(X_i) - P(X_j)\right)}{\sum_{s \in T=0} K\left(P(X_i) - P(X_s)\right)} \right]$$

  where N_1 is the sample size of the treatment group
  and $K(u)$ is a valid Kernel weight (people tend to use Gaussian Kernels here)

- As your propensity score gets further away from observation i you get less weight
- As $h \to 0$ (or σ_h) the window gets smaller and we use fewer neighbors.

Kernel Matching

- The usual caveats apply: h determines the bias-variance tradeoff
- Choice of Kernel effects finite-sample properties
- Here the common support is important. We can only learn about cases where $P(X) \neq 1$ and $P(X) \neq 0$. If you always get treated (or not-treated) we cannot learn from this observation.
- We also have to be careful in choosing X so as not to violate CIA (too many X's , too few X's) → have to think carefully!
- If you use propensity scores you will need a slide convincing us you have thought about why CIA holds for you!

Gotcha!

Under CIA we know

$$G(Y(1), Y(0)|X, T) = G(Y(1), Y(0)|X)$$

Suppose we add in Z, then we require that:

$$G(Y(1), Y(0)|X, Z, T) = G(Y(1), Y(0)|X, Z)$$

$$\begin{aligned} G(Y(1), Y(0)|X, T) = \int G(Y(1), Y(0)|X, Z, T)dF(Z|X, T) \\ = G(Y(1), Y(0)|X) \end{aligned}$$

where the last part follows by CIA.

- Thus each element can depend on T conditional on Z, X but the average may not.
- Mindless applications of matching can give you biased results!

Matching and OLS

- Recall that OLS is a special case of Kernel regression (and hence matching!)
- Think about

$$Y = \alpha + \beta T_i + u$$

- Assume that $E(u|T, X) = E(u|X)$ which is a conditional mean independence assumption
- The we can get β consistently (but not other variables) by running the following:

$$Y = \alpha + \beta T_i + \gamma X + v$$

- Again we are in the homogenous treatment world

A Matching Example

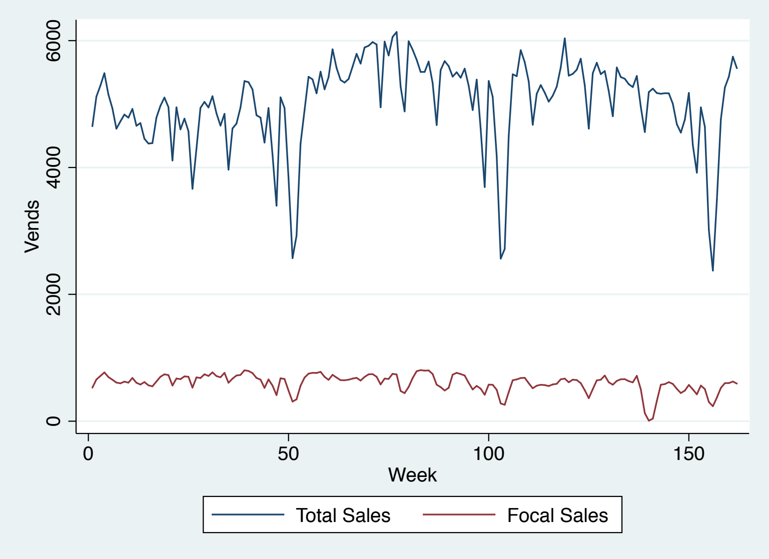
Here is an example where I found that matching was helpful in my own work with Julie Mortimer:

- We ran a randomized experiment where we removed Snickers bars from around 60 vending machines in office buildings in downtown Chicago.
- We have a few possible control groups:
  1. Same vending machine in other weeks (captures heterogeneous tastes in the cross section)
  2. Other vending machines in the same week (might capture aggregate shocks, ad campaigns, etc.)
- We went with #1 as #2 was not particularly helpful.

A Matching Example

Major problem was that there was a ton of heterogeneity in the overall level of (potential) weekly sales which we call M_t.

- Main source of heterogeneity is how many people are in the office that week, or how late they work.
- Based on total sales our average over treatment weeks was in the 74th percentile of all weeks.
- This was after removing a product, so we know sales should have gone down!
- How do we fix this without running the experiment for an entire year!
- Can't use shares instead of quantities. Why?


Vends
0
2000
4000
6000
0
50
100
150
Week
Total Sales
Focal Sales

A Matching Example

Ideally we could just observe M_t directly and use that as our matching variable X

- We didn't observe it directly and tried a few different measures:
  - Sales at the soda machine next to the snack machine
  - Sales of salty snacks at the same machine (not substitutes for candy bars).
  - We used k-NN with $k = 4$ to select control weeks – notice we re-weight so that overall sales are approximately same (minus the removed product).
- We also tried a more structured approach:
  - Define controls weeks as valid IFF
  - Overall sales were weakly lower
  - Overall sales were not less than Overall Sales less expected sales less Snickers Sales.

	Control	Control	Treatment	Treatment	Mean	
Product	Mean	%ile	Mean	%ile	Difference	% Δ
Vends						
Peanut M&Ms	359.9	73.6	478.3*	99.4	118.4*	32.9
Twix Caramel	187.6	55.3	297.1*	100.0	109.5*	58.4
Assorted Chocolate	334.8	66.7	398.0*	95.0	63.2*	18.9
Assorted Energy	571.9	63.5	616.2	76.7	44.3	7.8
Zoo Animal Cracker	209.1	78.6	243.7*	98.1	34.6*	16.5
Salted Peanuts	187.9	70.4	216.3*	93.7	28.4	15.1
Choc Chip Famous Amos	171.6	71.7	193.1*	95.0	21.5*	12.5
Ruger Vanilla Wafer	107.3	59.7	127.9	78.6	20.6*	19.1
Assorted Candy	215.8	43.4	229.6	60.4	13.7	6.4
Assorted Potato Chips	279.6	64.2	292.4*	66.7	12.8	4.6
Assorted Pretzels	548.3	87.4	557.7*	88.7	9.4	1.7
Raisinets	133.3	66.0	139.4	74.2	6.1	4.6
Cheetos	262.2	60.1	260.5	58.2	-1.8	-0.7
Grandmas Choc Chip	77.9	51.3	72.5	37.8	-5.4	-7.0
Doritos	215.4	54.1	203.1	39.6	-12.3*	-5.7
Assorted Cookie	180.3	61.0	162.4	48.4	-17.9	-10.0
Skittles	100.1	62.9	75.1*	30.2	-25.1*	-25.0
Assorted Salty Snack	1382.8	56.0	1276.2*	23.3	-106.7*	-7.7
Snickers	323.4	50.3	2.0*	1.3	-321.4*	-99.4
Total	5849.6	74.2	5841.3	73.0	-8.3	-0.1

Notes: Control weeks are selected through the-neighbor matching using four control observations for each treatment week. Percentiles are relative to the full distribution of control weeks.

Exercises

- This would be a good time to work through the vignette for `cobalt` `https://cran.r-project.org/web/packages/cobalt/vignettes/cobalt_A0_basic_use.html`
- Compare the ATE for the Lalonde data with the IPW, Nearest Neighbor, and Propensity Score estimates.
- Then start the homework