Propensity score calipers and the overlap condition

Ben Hansen, UMich Statistics (bbh@umich.edu). Project site -
github.com/benthestatistician/PISE (code/manuscript), benthestatistician.github.io/PISE (slides)
Outline

1. Problem: determining the region of overlap
2. Propensity score caliper matching
3. PPSE calipers
4. Applications & discussion
Overlap & bias in comparisons w/ 1 covariate

- The quasi-experimental setup: $X, Y; Z \in \{0, 1\}, Y_c$.
- Overlap?
  - Observations well outside region of overlap should be excluded.
  - Wasteful to exclude members of $\{i : Z_i = 1\}$ because their $x$ is just outside $\text{range}(\{x_j : Z_j = 0\})$.
  - May distort research question, too.
  - Close/far distinctions presuppose a model.
- Balance - i.e., is $\bar{x}_t \approx \bar{x}_c$?
  - The question presupposes overlap. (OW mean comparisons can be misleading.)
  - A topic for another day...
Statistical assumptions for general Xes

- Now, multiple Xes: $X$, not $x$.
- Strong ignorability conditions:
  - No unmeasured confounding: $Y_c \perp Z|X$.
  - Overlap: $\Pr(Z = 1|X) < 1$.
- No unmeasured confounding says, there are natural experiments in each "cell" of $X$.
- Problem: if multiple Xes, many observations will be all alone in their cells.
- Rosenbaum & Rubin's (1983) solution is match on $\Pr(Z = 1|X = x)$, the PS.
- Remaining problem: In a cell $x_0$ s.t. $\Pr(Z = 1|X = x_0) = 1$, there's no data to estimate $\Pr(Y_c \in \cdot|X = x_0)$, even as $n \uparrow \infty$. (Whether or not you match or otherwise use PSes.)
- The second problem leads people to reduce analytic sample to region of common support on $\hat{PS}$ (the method of "strict overlap").
Example 1: violence & public infrastructure in Medellín

- Medellin, Colombia (population 2 million)
- As of early 2000s, 60\% poverty rate, 20\% unemployment, homicide 185 per 100K
- High residential segregation, w/ concentrated poverty in surrounding hills.
- 2004-2006: gondolas connect some but not all to city center.
Propensity scores in the Medellin study

- Small matching problem: \( n_t = 25, \ n_c = 23 \) (neighborhoods).
- \( \bar{x}_t \)'s and \( \bar{x}_c \)'s not too far apart, but PS matching brought them closer.
- Figure shows PS we matched on - the \( X\hat{\beta} \) from a logistic regression (Cerda et al 2012, Am J Epi).
- Region of strict overlap contains only 6/25 \( t \)'s and 4/23 \( c \)'s!
- Yet Cerda et al full-matched all 48 neighborhoods.
Outline

1. Problem: determining the region of overlap
2. Propensity score caliper matching
3. PPSE calipers
4. Applications & discussion
Matching within PS calipers (R.& R.)

- For matching, useful to specify a tolerance or caliper.
- Absent unmeasured confounding, matches with small PS differences mimic paired random assignment.
- In effect, caliper adjudicates "closeness" to region of common support.

- (Not \( \widehat{PS} \) the estimated probability, \( \widehat{\text{PS}} \) the estimated index function. I.e., logits of estimated probabilities.)
- The other widely used method (e.g. Austin & Lee 2009, Lunt 2013), besides requiring strict overlap.
Room for improvement (in caliper = .25s_p)

- In the Medellin study, .25s_p(\(\hat{PS}\)) caliper still excludes 14/25 treatment neighborhoods.
- At the same time, in that study it's tenable (\(p = .10\)) that all PSes are the same!

- Moral: in small studies, .25s_p(\(\hat{PS}\)) may be too strict.
- In large studies, PS may be estimable w/ much more precision than .25s_p(\(\hat{PS}\)). At same time, more potential controls.
- Moral: in large studies, .25s_p(\(\hat{PS}\)) may be too loose.
Outline

1. Problem: determining the region of overlap
2. Propensity score caliper matching
3. PPSE calipers
4. Applications & discussion
Calipers as insurance that PS differences tend to 0

- Pairings (denoted "i ∼ j") should satisfy \( \sup_{i \sim j} |(\tilde{x}_i - \tilde{x}_j)\beta| \downarrow 0 \) as \( n \uparrow \infty \).

- Can this be accomplished with a requirement of form \( |(\tilde{x}_i - \tilde{x}_j)\beta| \leq w_n \), since \( |(\tilde{x}_i - \tilde{x}_j)\beta| \leq |(\tilde{x}_i - \tilde{x}_j)(\hat{\beta} - \beta)| + |(\tilde{x}_i - \tilde{x}_j)\hat{\beta}| \)

- \( (w_n = s_p/4 \) won't do. We need \( w_n \xrightarrow{P} 0. \) In itself, even this won't be quite enough.)

- Scanning pairs \( i, j \leq n, |(\tilde{x}_i - \tilde{x}_j)(\hat{\beta} - \beta)| \leq |\tilde{x}_i - \tilde{x}_j|_2 |\hat{\beta} - \beta|_2 \).

- Expect \( \sup_{i,j} |(\tilde{x}_i - \tilde{x}_j)(\hat{\beta} - \beta)| \approx (\sup_{i,j} |\tilde{x}_i - \tilde{x}_j|_2) |\hat{\beta} - \beta|_2 \)

- Even \( w/ |x_{ij}| \) bounded, uniformly in \( i, j, \) and \( n \), this is \( O(p^{1/2})O_P([p/n]^{1/2}) = O_P(p/n^{1/2}) \). We'll need to assume \( p/n^{1/2} \downarrow 0 \).

- In that case, a \( w_n \) that's \( O_P(p/n^{1/2}) \) will do the trick.

- E.g., a \( w_n \) measuring average size of \( |(\tilde{x}_i - \tilde{x}_j)(\hat{\beta} - \beta)| \).
Sampling variability of paired PS differences

- Intuition: Even if we had matched perfectly on the PS, there would still be matched differences in \( \hat{\text{PS}} \). Let's estimate the size of these differences & use result to define caliper.
  - caliper \( \downarrow 0 \) as \( n \uparrow \infty \). So, \( \hat{\beta} \rightarrow \beta \Rightarrow \text{paired PS differences} \downarrow 0 \).
  - excludes treatment subjects whose PSes are distinguishably different from all controls'.

- To limit model sensitivity, work on index scale.

- If \( \mathbf{d} = \mathbf{x}_1 - \mathbf{x}_2 \), \( \mathbf{d}(\hat{\beta} - \beta) \) is the error of estimation of paired PS difference \( \mathbf{d}\beta \).

- Considering all \( \binom{n}{2} \) possible pairs \( r \), the expected MS estimation error of paired PS differences can be expressed as a Frobenius inner product:

\[
\mathbb{E}_{\beta} \sum_r (\mathbf{d}_r(\hat{\beta} - \beta))^2 \binom{n}{2} = 2 \langle \Sigma^{(\alpha)} , \mathbb{E}_{\beta} \{ (\hat{\beta} - \beta)(\hat{\beta} - \beta)' \} \rangle_F.
\]
Sampling variability of paired PS diffs (ii)

· But all possible pairs include many that are quite different on the PS. We wanted to characterize $E(d_r(\hat{\beta} - \beta))^2$ among pairs $r$ s.t. $d_r\beta \approx 0$.

· Since there may be few or no such pairs, instead "residualize" all pairs for the true PS, leaving remaining differences in place. For $i = 1, \ldots, p$, define $d_{(i)}^\perp = \text{residual of } d_{(i)}$ regression on $d\beta$; $d_1^\perp = (d_1^\perp \ d_2^\perp \ \ldots \ d_p^\perp)$. By construction, $d_r^\perp \beta = 0$ for all $r$.

· Rather than the mean-square of PS-difference estimation errors, $\binom{n}{2}^{-1} E \sum_r (d_r(\hat{\beta} - \beta))^2$, consider the mean-square PS-difference estimation error, net of differences in the true PS:

$$\binom{n}{2}^{-1} E \beta \sum_r (d_r^\perp(\hat{\beta} - \beta))^2 = 2\langle S^\perp, C_{\hat{\beta}} \rangle_F,$$

where $S^\perp = \frac{1}{2} \text{Cov}(d^\perp) = \text{Cov}(x^\perp x\beta), x^\perp x\beta = e(x|1, x\beta)$.

· Plugging in $\hat{S}^\perp = \frac{1}{n-1} e(x|1, x\hat{\beta})' e(x|1, x\hat{\beta})$ and a $\hat{C}_{\hat{\beta}}$ extracted from the regression fit gives the "propensity-paired standard error" (PPSE).
Outline

1. Problem: determining the region of overlap
2. Propensity score caliper matching
3. PPSE calipers
4. Applications & discussion
Example 2: Real or cosmetic reform in police practices?

Vagrancy arrests in the 60s and 70s

- Before 1960, arrests for vagrancy were relatively common in U.S.; by 1980, relatively rare.
- Increasingly forbidden by laws, court rulings, administrative decisions.
- Did pretextual arrests decline, a civil rights victory, or were they simply shifted to other categories (displacement)?
- Using administrative records assembled by R. Goluboff, D. Thacher and I identified 121 (local agency, year) instances, of \( n = 5600 \) at-risk agency-years, that experienced a sharp, anomalous decline in vagrancy arrests.
- and we matched them to otherwise similar agencies in terms of year, region of U.S., population, prior use of vagrancy arrests and a propensity score.
Example 3: Vascular closure devices vs manual closure following percutaneous coronary intervention

- Once this stent has been threaded up into your heart, the hole in your femoral artery needs to be closed.
- VCDs are more comfortable than manual closure -- are they as safe?
- Large matching problem: $n_t = 31K$; $n_c = 54K$. 

[Diagrams of stent insertion and related processes]
Applications

- For the Medellin example, 2.5 PPSEs works out to 4 logits ($3.9s_p$).
- No neighborhoods excluded. Recall that the alternatives excluded at least 12!
- In the vagrancy study, 2.5 PPSEs = 3.8 logits ($0.5s_p$).
- In the VCD example 2.5 PPSEs= 0.15 logits ($0.3s_p$).
- Larger sample $\Rightarrow$ smaller PPSE.
Multiple PS calipers in the vagrancy study (Ex 2)

$pse(\text{psmod3e}) = 1.5$, psmod3e's specs:
Stopped $\sim \text{ns}((\text{arrptOM.2.yrs.ago,3}) + \text{ns}((\text{arrptOM.2.yrs.ago,3}) + <..$

$pse(\text{psmod}) = 2.8e+14$, psmod's specs:
$\text{ppped} \sim \text{arrptOM.2.yrs.ago} + \text{arrptOM2.2.yrs.ago} + <..$, weights=arrTOT.
Discussion

- Getting consistency out of PPSE caliper requires \( p^2 / n \downarrow 0 \). If \( p/n \downarrow 0 \), still interpretable as an SE of sorts.
- No need to restrict yourself to a single index score.
- With a saturated PS model, observed information may be poorly conditioned. The method of PPSE estimation as described here requires some elaboration (see code on github).
- Likeness of observations ("like to like") is interpreted in terms of PS variables. If too strict or too loose, then adjust PS variables.

Recommendations:

- I like calipers of 2.5 PPSEs
- I'm using sandwich estimates of \( \text{Cov}(\hat{\beta}) \), w/ some special sauce to limit numerical instability.
- Our \texttt{optmatch} R package (Fredrickson et al, 2016) does optimal pair and full matching (Gu & Rosenbaum, 1993; Hansen & Klopfer, 2006; Stuart & Green, 2008), readily accommodating PS calipers.