DECART Summer School 2018:

Causal Inference Module

Matching

Why match?

- We need to adjust for covariate X_i in observational study, even when the assumption of no unmeasured confounding holds.

- Easy solution: Use a parametric model for $E(Y_i(A)|X_i)$.

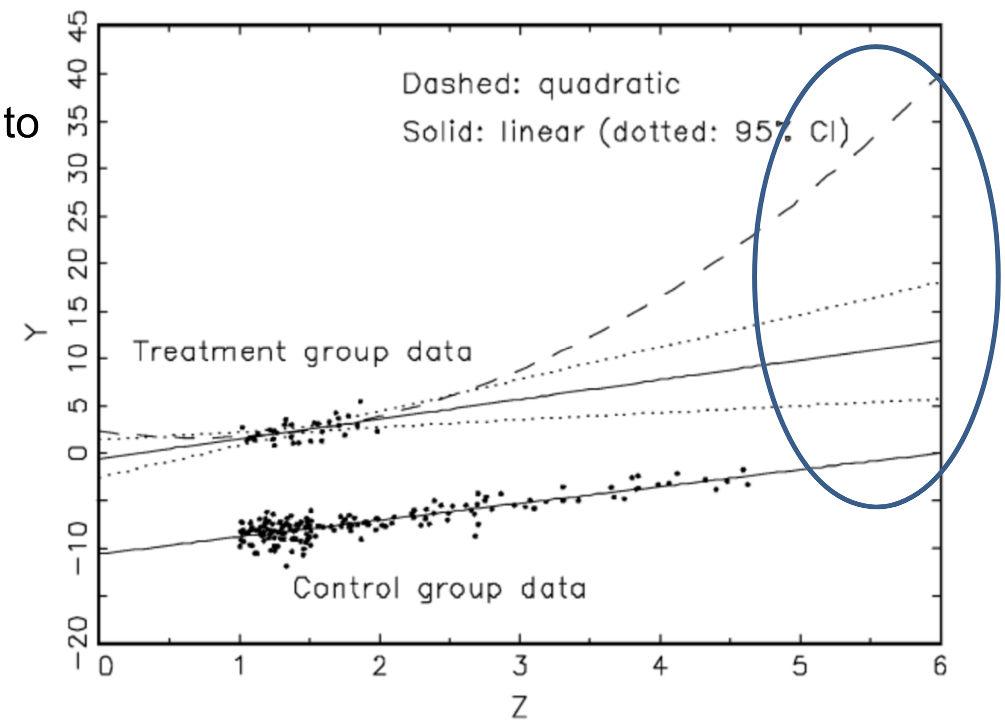
- But misspecified parametric model will lead to wrong causal estimates.

- Matching is a natural alternative solution with two benefits:
  - Reduces the dependence of estimates on parametric models.
  - Can simplify the analysis of causal effects.

Why parametric model failed?

- Often lead to large variation in the estimates of interest.
- Why does this occur?

  - Parametric model will extrapolate to regions with only treated or only control.
  - Modeling assumption will affect these extrapolation.



What does matching do?

- Allows for relatively nonparametric ways of estimating the casual effect.

- **Caution**: Matching doesn't justify a causal effect automatically.

- **Important Note:** Without strong design (such as clinical trial), no statistical modeling could completely make the move from correlation to causation persuasive.

Assumptions

1. Consistency:

$$Y_i = Y_i(A_i)$$

2. Unmeasured Confounders Assumption:

$$A_i \perp (Y_i(0), Y_i(1)) | X_i$$

3. Positivity

$$0 < P(A_i = 1 | X_i) < 1$$

(Simple) Exact Matching Without Replacement

- Let X_i take on a finite number of values, x.
- Let $I_t = \{1,2,\dots,N_t\}$ be the set of treated units.
- Exact matching without replacement:

For each treated unit, $i \in I_t$

1. Find the set M_i^c of unmatched control units j such that $X_i = X_j$, for $j \in M_i^c$
2. Randomly select one of these control units to be the match, indicated $j(i)$.

- Let $I_c = \{j(1), j(2), \dots, j(N_t)\}$ be the set of control units.
- The distribution of X_i will be exactly the same for the treated and matched control:

$$P(X_i = x | A_i = 1) = P(X_i = x | A_i = 0, I_c)$$

If the data is exactly matched, then an unbiased estimator for the average treatment effect for the treated (ATT) is:

$$\hat{\tau}_t^{match} = \frac{1}{N_t}\sum_{i:A_i=1} \hat{\tau}_i^{match} = \frac{1}{N_t}\sum_{i:A_i=1} \left(Y_i^{obs} - Y_{m_i^c}^{obs}\right)$$

i.e.

$$E\left[\hat{\tau}_t^{match}\right] = \tau_{ATT} = E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1]$$

Proof:

$$E[\hat{\tau}_t^{match}] = E[\frac{1}{N_t}\sum_{i:A_i=1}\left(Y_i^{obs} - Y_{m_i^c}^{obs}\right)]$$

$$= E[\frac{1}{N_t}\sum_{i:A_i=1}\left(Y_i^{obs}\right)] - E[\frac{1}{N_t}\sum_{i:A_i=1}E\left(Y_{m_i^c}^{obs}\right)]$$

$$= \int E[Y_i|A_i = 1, X_i = x]dP(X_i = x|A_i = 1)$$

$$- \int E[Y_i|A_i = 0, X_i = x]dP(X_i = x|A_i = 1)$$

$$= \int E[Y_i(1)|A_i = 1, X_i = x]dP(X_i = x|A_i = 1)$$

$$- \int E[Y_i(0)|A_i = 1, X_i = x]dP(X_i = x|A_i = 1)$$

$$= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1]$$

Weakening the identification assumptions

- Consistency, no unmeasured confounders, total expectation and exact matching property

  $\Rightarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to conditional mean independence (CMI):
  $$E[Y_i(0)|X_i = x, A_i = 1] = E[Y_i(0)|X_i = x, A_i = 0]$$
- Nice features of CMI:

  1. Only make assumptions about $Y_i(0)$ not $Y_i(1)$.

  2. Only make assumptions on the means, not other aspects of distribution (variance, skewness, kurtosis, etc).

Analyzing exactly matched data

- Simple difference in observed means:

$$\hat{\tau}_{ATT} = \frac{1}{N_T}\sum_{i\in I_t} Y_i - \frac{1}{N_c}\sum_{j\in I_c} Y_j$$

- In simple matching mentioned above (exact, 1-to-1, no replacement):

$$\hat{\tau}_{ATT} = \frac{1}{N_T}\sum_{i=1\ldots N_T}\left(Y_i - Y_{j(i)}\right)$$

$$\widehat{var}\ (\hat{\tau}_{ATT}) = \frac{1}{N_T}\sum\nolimits_{i=1\ldots N_T}\left(Y_i - Y_{j(i)} - \hat{\tau}_{ATT}\right)^2$$

- In practice, such an exact matching scheme is rarely feasible

➢ exact matching is typically impossible

➢ the pool of potential matches is often too small to ignore the conflicts

Inexact Matching without Replacement

- Match the ith treated unit with covariate values X_i to control unit m_i, that is, the control unit that solves
$$m_i^c = argmin_{i' \in I_c} \|X_i - X_{i'}\|$$
- one control unit might be identified as match for more than one unit

> match all units simultaneously

$argmin_{m_1^c,\dots,m_{N_t}^c \in I_c} \sum_{i=1}^{N_t} \left\|X_i - X_{m_i^c}\right\|$ subject to $m_i \neq m_{i'}$

> match units sequentially (“greedy” matching algorithm): the ordering matters

One option: match those difficult ones first: the rank of the estimated propensity scores (high -> low)

Inexact Matching without Replacement

- When multiple matches (equally close) to one treated unit

  > use the average of the outcomes for this set of tied matches as the control potential outcome for treated unit i, $\sum_{i' \in M_i^c} Y_{i'}(0) \,/ M_i$, with M_i be the cardinality of M_i^c

  reduced sampling variance of the resulting estimator

  removing more units from the pool of possible control units available for subsequent matches

  > some selection mechanism, e.g. random selection

Distance metrics

- We need a distance metric to define distance/similarity on X_i and X_j, which might be high dimensional.

  -- Lower value $\longrightarrow$ more similar values

  -- Choice of distance metric will lead to different matches
- Possible choices of distance:

  > Propensity score distance metric

  > Euclidean distance metric

  > Mahalanobis distance metric

  > Caliper metric

  > Hybrid metric

Propensity score distance

- Propensity score: $e(X_i) = P(A_i = 1|X_i)$
- Rubin et al have shown that propensity score matching has good properties if covariates are roughly normal.
- Propensity score distance:

Option 1: $D_{ij} = \left|e(X_i) - e\left(X_j\right)\right|$

Option 2: $D_{ij} = \left|logit(e(X_i)) - logit(e\left(X_j\right))\right|$

Euclidean distance

- Suppose that $X_i = (X_{i1}, \dots, X_{iK})$.
- The Euclidean distance metric is

$$D_{ij} = \sqrt{\sum_{k=1}^{K} \frac{\left(X_{ik} - X_{jk}\right)^2}{\hat{\sigma}_k^2}} \quad ,$$

where $\hat{\sigma}_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ik} - \bar{X}_k)^2$.

Mahalanobis distance

- Intuition: if X_{ik} and $X_{ik'}$ are highly correlated, then their contribution to the distances should be lower.

    - Easy to get close on correlated covariates, then downweight it

    - Harder to get close on uncorrelated covariates, then upweight it

- The Mahalanobis distance is

$$D_{ij} = \sqrt{\left(X_i - X_j\right)^T \hat{\Sigma}^{-1} (X_i - X_j)},$$

where weight matrix $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T$.

Caliper

- To overcome shortcomes of erroneously choosing control, we will only select the control if its distance to case meet condition:

$$D_{ij} < \varepsilon$$

- Rubin (1985) suggested using a caliper size of a quarter of a standard deviation of the sample estimate propensity score ($\varepsilon = \sigma_{ps}/4$).

HEALTH
UNIVERSITY OF UTAH

Hybrid metrics

Example:

Exact on race/gender, Mahalanobis on the rest.

The general matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)
4. Check balance
5. Repeat 1-4 until balance is acceptable
6. Calculate the effect of the treatment on the outcome in the matched dataset.

The Card-Krueger Minimum Wage Data

- Card and Krueger (1995) were interested in evaluating the effect of raising the state minimum wage in New Jersey in 1993.
- data were collected on employment at fast-food restaurants in New Jersey and in the neighboring state of Pennsylvania.
- Covariates are measured on restaurant level prior to the raise;
- The outcome is employment after the raise (final empl).

The Card-Krueger New Jersey and Pennsylvania Minimum Wage Data

	$(N = 347)$		$(N_t = 279)$ (treated)		$(N_c = 68)$ (controls)			
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Nor Dif	Log Ratio of STD
initial empl	17.84	(9.62)	20.17	(11.96)	17.27	(8.89)	−0.28	−0.30
burger king	0.42	(0.49)	0.43	(0.50)	0.42	(0.49)	−0.02	−0.01
kfc	0.19	(0.40)	0.13	(0.34)	0.21	(0.41)	0.20	0.17
roys	0.25	(0.43)	0.25	(0.44)	0.25	(0.43)	0.00	−0.00
wendys	0.14	(0.35)	0.19	(0.40)	0.13	(0.33)	−0.18	−0.18
initial wage	4.61	(0.34)	4.62	(0.35)	4.60	(0.34)	−0.05	−0.02
time until raise	17.96	(11.01)	19.05	(13.46)	17.69	(10.34)	−0.11	−0.26
pscore	0.80	(0.05)	0.79	(0.06)	0.81	(0.04)	0.28	−0.35
final empl	17.37	(8.39)	17.54	(7.73)	17.32	(8.55)		

HEALTH
UNIVERSITY OF UTAH

Excise on the Card-Krueger Data

- For this illustration, we focus on a small subset of 20 restaurants
- 5 from New Jersey and 15 from Pennsylvania
- We use only initial employment (initial empl) and restaurant chain (burger king or kfc) as pre-treatment variables
- Inexact match without replacement

20 Units from the Card-Krueger Dataset

Unit i	State W_i	chain X_{i1}	initial empl X_{i2}	final empl Y_i^{obs}
1	NJ	BK	22.5	40.0
2	NJ	KFC	14.0	12.5
3	NJ	BK	37.5	20.0
4	NJ	KFC	9.0	3.5
5	NJ	KFC	8.0	5.5
6	PA	BK	10.5	15.0
7	PA	KFC	13.8	17.0
8	PA	KFC	8.5	10.5
9	PA	BK	25.5	18.5
10	PA	BK	17.0	12.5
11	PA	BK	20.0	19.5
12	PA	BK	13.5	21.0
13	PA	BK	19.0	11.0
14	PA	BK	12.0	17.0
15	PA	BK	32.5	22.5
16	PA	BK	16.0	20.0
17	PA	KFC	11.0	14.0
18	PA	KFC	4.5	6.5
19	PA	BK	12.5	31.5
20	PA	BK	8.0	8.0

Match Order = 1,2,3,4,5; Metric = $x_1^2 + x_2^2$

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{obs}$	$\hat{\tau}_i^{match}$
1	11	40.0	19.5	20.5
2	7	12.5	17	−4.5
3	15	20.0	22.5	−2.5
4	8	3.5	10.5	−7
5	20	5.5	8.0	−2.5
$\hat{\tau}_t^{match}$				+0.8

Match Order = 1,2,3,5,4; Metric = $x_1^2 + x_2^2$

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{obs}$	$\hat{\tau}_i^{match}$
1	11	40.0	19.5	20.5
2	7	12.5	17.0	−4.5
3	15	20.0	22.5	−2.5
5	8	5.5	10.5	−5
4	20	3.5	8.0	−4.5
$\hat{\tau}_t^{match}$				+0.8

Match Order = 1,2,3,4,5; Metric = $100 \cdot x_1^2 + x_2^2$

i	m_i^c	Y_i^{obs}	$Y_{m_i^c}^{obs}$	$\hat{\tau}_i^{match}$
1	11	40.0	19.5	20.5
2	7	12.5	17.0	−4.5
3	15	20.0	22.5	−2.5
4	8	3.5	10.5	−7
5	17	5.5	14.0	−8.5
$\hat{\tau}_t^{match}$				−0.4

The Bias of Matching Estimator

- The potential bias created by discrepancies between the pre-treatment covariates of the units within a matched pair.

$$E\left[\hat{\tau}_i^{match} | A_i = 1, X_i, X_{m_i^c}\right] = E\left[Y_i^1 - Y_{m_i^c}^0 | X_i, X_{m_i^c}\right]$$
$$= \mu_t(X_i) - \mu_c\left(X_{m_i^c}\right) = \tau(X_i) + \mu_c(X_i) - \mu_c\left(X_{m_i^c}\right)$$

- The unit-level bias is $B_i = \mu_c(X_i) - \mu_c\left(X_{m_i^c}\right)$
- Bias adjustment:

$$\hat{\tau}_t^{adj} = \frac{1}{N_t} \sum_{i:A_i=1} \left(Y_i - Y_{m_i^c} - \widehat{B}_i\right)$$

$\widehat{B}_i$ can be estimated through linear model

Bias Correction Using Linear Model

- If we assume linear models for the group specific means $\mu_c(x) = \alpha_d + x\beta_d$ and $\mu_t(x) = \tau + \alpha_d + x\beta_d$

- Then we can estimate the bias as $\hat{B}_i = \hat{\mu}_c(X_i) - \hat{\mu}_c\left(X_{m_i^c}\right) = (X_i - X_{m_i^c})\hat{\beta}_d$

- Three simple regression based approaches can be considered to obtain $\hat{B}_i$:

1. **Regression on the Matching Discrepancy** $(D_i = X_i - X_{m_i^c})$

$$Y_i^{obs} - Y_{m_i^c}^{obs} = \tau + D_i\beta_d + v_i: \quad Y_i^{obs} - Y_{m_i^c}^{obs} \sim D_i \Rightarrow \hat{\beta}_d$$

2. **Control Regression on Covariates**

$$Y_{m_i^c} = \alpha_c + X_{m_i^c}\beta_c + v_{ci}: \quad Y_{m_i^c} \sim X_{m_i^c} \Rightarrow \hat{\beta}_c$$

3. **Pooled Regression on Covariates**

$$Y_i = \alpha_p + \tau_p A_i + X_i\beta_p + v_i: \quad Y_i \sim A_i + X_i \Rightarrow \hat{\beta}_p$$

Data Illustration of Bias Correction

Matching Discrepancy for the 20 Units from the Card-Krueger Data (Match Order 1,2,3,4,5; Metric $x_1^2 + x_2^2$)

i	m_i	Y_i^{obs}	$Y_{m_i^c}^{obs}$	$\hat{\tau}_i^{match}$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	$D_{i,1}$	$D_{i,2}$
1	11	40.0	19.5	20.5	0	22.5	0	20.0	0	2.5
2	7	12.5	17.0	−4.5	1	14.0	1	13.8	0	0.2
3	15	20.0	22.5	−2.5	0	37.5	0	32.5	0	5.0
4	8	3.5	10.5	−7.0	1	9.0	1	8.5	0	0.5
5	20	5.5	8.0	−2.5	1	8.0	0	8.0	1	0

Bias-Adjustment Regression Coefficients for the 20 Units from the Card-Krueger Data

	Difference Regression (Approach #1)	Control Regression (Approach #2)	Pooled Regression (Approach #3)
Regression coefficients			
Intercept	−1.30	4.21	12.01
Treatment indicator	–	–	1.63
Restaurant chain	−1.20	2.65	−7.32
Initial employment	1.43	0.62	0.39

Data Illustration of Bias Correction

Regression on the Matching Discrepancy (Difference Regression):

First pair $(i, m_i) = (1{,}11)$, $X_1 = (0{,}22.5), X_{m_1} = (0{,}20.0)$

Thus the adjusted control outcome:

$$\hat{Y}_1(0) = Y_{m_1} + D_1\hat{\beta}_d = 19.5 - 1.20 \times D_{1,1} + 1.43 \times D_{1,2}$$
$$= 19.5 - 1.20 \times 0 + 1.43 \times 2.5 = 23.1$$

The adjusted estimate of the unit-level treatment effect

$$\hat{\tau}_1^{adj} = Y_1(1) - \hat{Y}_1(0) = 40.0 - 23.1 = 16.9$$

Similarly, we can obtain the following full set of results:

i	m_i	$Y_i(1)$	$Y_{m_i^c}(0)$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	$D_{i,1}$	$D_{i,2}$	$\hat{\beta}_d^T D_i$	$\hat{Y}_i(0)$
1	11	40.0	19.5	0	22.5	0	20.0	0	2.5	3.6	23.1
2	7	12.5	17.0	1	14.0	1	13.8	0	0.2	0.3	17.3
3	15	20.0	22.5	0	37.5	0	32.5	0	5.0	7.1	29.6
4	8	3.5	10.5	1	9.0	1	8.5	0	0.5	0.7	11.2
5	20	5.5	8.0	1	8.0	0	8.0	1	0	−1.2	6.8
		$\hat{\tau}_t^{\text{match}} = +0.8$				$\hat{\tau}_t^{\text{adj}} = -1.3$					

Data Illustration of Bias Correction

Control Regression on Covariates:

First pair $(i, m_i) = (1{,}11)$, $X_1 = (0{,}22.5), X_{m_1} = (0{,}20.0)$

Thus the adjusted control outcome:

$$\hat{Y}_1(0) = Y_{m_1} + D_1\hat{\beta}_c = 19.5 + 2.65 \times D_{1,1} + 0.62 \times D_{1,2}$$

$$= 19.5\ - 2.65 \times 0 + 0.62 \times 2.5 = 21.1$$

The adjusted estimate of the unit-level treatment effect

$$\hat{\tau}_1^{adj} = Y_1(1) - \hat{Y}_1(0) = 40.0\ - 21.1 =\ 18.9$$

Similarly, we can obtain the following full set of results:

i	m_i	$Y_i(1)$	$Y_{m_i^c}(0)$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	D_{i1}	D_{i2}^*	$\hat{\beta}_c^T D_i$	$\hat{Y}_i(0)$
1	11	40.0	19.5	0	22.5	0	20.0	0	2.5	1.5	21.0
2	7	12.5	17.0	1	14.1	1	13.8	0	0.2	0.1	17.1
3	15	20.0	22.5	0	37.5	0	32.5	0	5.0	3.1	25.6
4	8	3.5	10.5	1	9.0	1	8.5	0	0.5	0.3	10.8
5	20	5.5	8.0	1	8.0	0	8.0	1	0	2.7	10.7

$\hat{\tau}_t^{\text{match}} = +0.8$ $\qquad$ $\hat{\tau}_t^{\text{adj}} = -0.7$

Data Illustration of Bias Correction

Pooled Regression on Covariates:

First pair $(i, m_i) = (1{,}11)$, $X_1 = (0{,}22.5), X_{m_1} = (0{,}20.0)$

Thus the adjustment for the 1st pair:

$$\hat{B}_1 = -7.32 \times D_{1,1} + 0.39 \times D_{1,2}$$
$$= -7.32 \times 0 + 0.39 \times 2.5 = 0.98$$

The adjusted estimate of the unit-level treatment effect

$$\hat{\tau}_1^{adj} = Y_1(1) - Y_{m_1}(0) - \hat{B}_1 = 40.0 - 19.5 - 0.98 = 19.52$$

Similarly, we can obtain the following full set of results:

i	m_i	$Y_i(1)$	$Y_{m_i^c}(0)$	$X_{i,1}$	$X_{i,2}$	$X_{m_i^c,1}$	$X_{m_i^c,2}$	D_{i1}	D_{i2}	$\hat{\beta}_s^T D_i$	$\hat{Y}_i(0)$
1	11	40.0	19.5	0	22.5	0	20.0	0	2.5	1.0	20.5
2	7	12.5	17.0	1	14.0	1	13.8	0	0.2	0.1	17.1
3	15	20.0	22.5	0	37.5	0	32.5	0	5.0	1.9	24.4
4	8	3.5	10.5	1	9.0	1	8.5	0	0.5	0.2	10.7
5	20	5.5	8.0	1	8.0	0	8.0	1	0	−7.3	0.7
		$\hat{\tau}_t^{\text{match}} = +0.8$				$\hat{\tau}_t^{\text{adj}} = +1.6$					

Matching with Replacement

- The set of controls selected does not depend on the ordering of treated units.
- Let $L(i)$ be the number of times each control unit id used as a match

$$L(i) = \sum_{j=i}^{N_t} 1_{j \in M_i^c}$$

When matching without replacement, $L(i) \in \{0,1\}$ for all units.

$$\hat{\tau}_t^{repl} = \frac{1}{N_t} \sum_{i=1}^{N} \left(A_i \cdot Y_i^{obs} - (1 - A_i) \cdot L(i) \cdot Y_i^{obs} \right)$$

With or without replacement

- Matching with replacement: a single control unit could be matched repeatedly with multiple treated units.
- Pro:
  1. Better matches!
  2. Order of matching does not matter.
- Con:
  1. need more complicated inference.
  2. need to account for multiple appearances with weights.
  3. potentially higher uncertainty (using the same data multiple times=relying on less data)

The Number of Matches

- Let σ_c^2 and σ_t^2 be the super-population variances of Y_i^0 and Y_i^1 conditional on the covaraites.
- If we use M matches, the estimator then

$$\hat{\tau}_t^{match,M} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(Y_i^1 - \frac{1}{M}\sum_{j\in M_i^c} Y_j^0\right)$$

- The sample variance is then

$$Var\left(\hat{\tau}_t^{match,M}\right) = \frac{1}{N_t}\left(\sigma_t^2 + \frac{\sigma_c^2}{M}\right)$$

- Assume equal variance σ_c^2=σ_t^2

$$\frac{Var\left(\hat{\tau}_t^{match,1}\right) - Var\left(\hat{\tau}_t^{match,M}\right)}{Var\left(\hat{\tau}_t^{match,1}\right)} = \frac{M-1}{2M}$$

M=2 reduces the sample variance by 25% relative to using a single match

Assessing balance

- All matching methods seek to find the balance:

$$P(X_i = x | A_i = 1, \mathcal{S}) = P(X_i = x | A_i = 0, \mathcal{S})$$

- Choice of balance metric will determine which matching method does better.
- Options: estimation of matching performance

1. Differences in mean/medians, standardized.
2. QQ plot/K-S statistics for comparing the entire distribution.

…

Estimand

- Matching easiest to justify for the average treatment effect for the treated (ATT).

$$\hat{\tau}_t = \frac{1}{N_t} \sum\nolimits_{i:A_i=1} \left(Y_i^{obs} - Y_{m_i^c}^{obs} \right)$$

- Can also justify the average treatment effect for the controls (ATU) by finding matched treated units for the controls.

$$\hat{\tau}_c = \frac{1}{N_c} \sum\nolimits_{i:A_i=0} \left(Y_{m_i^t}^{obs} - Y_i^{obs} \right)$$

- Combined the two to obtain the average treatment effect for the entire sample (ATE):

$$\hat{\tau} = \frac{N_c}{N_c + N_t} \hat{\tau}_c + \frac{N_t}{N_c + N_t} \hat{\tau}_t$$