Causal inference, week 4
Stratification, Regression Adjustment, Matching

Daniel Stegmueller

Duke University

Observational studies

- Do not control (or do not know) the assignment mechanism
- Presence of measured and unmeasured confounders is unbalanced between groups
- Measured confounders ["covariates", "controls", "pre-treatment variables"]: imbalance is the norm than exception!
- Structural assumptions have to be made (often untestable) to identify causal effects
- Also: model / functional form assumptions

Some estimands

- *Average treatment effect (ATE):*

$$\tau^{ATE} = E[Y_{i1} - Y_{i0}].$$

- *Conditional Average treatment effect (CATE):*

$$\tau(x) = E[Y_{i1} - Y_{i0} \mid X_i = x].$$

- *Average treatment effect for the treated (ATT) or for the control (ATC):*

$$\tau^{ATT} = E[Y_{i1} - Y_{i0} \mid D_i = 1], \quad \tau^{ATC} = E[Y_{i1} - Y_{i0} \mid D_i = 0]$$

- *Causal odds ratio (for dichotomous outcomes)*

$$\tau^{OR} = \frac{\Pr(Y_{i1} = 1)/\Pr(Y_{i1} = 0)}{\Pr(Y_{i0} = 1)/\Pr(Y_{i0} = 0)}.$$

- To identify the causal estimands, two important assumptions
  - Unconfoundedness
  - Overlap

Unconfoundedness

- Unconfoundedness or ignorability:

$$(Y_{i0}, Y_{i1}) \perp D_i \mid X_i$$

  or, eqeuivalently:

$$\Pr(D_i \mid Y_{i0}, Y_{i1}, X_i) = \Pr(D_i \mid X_i).$$

- Within subpopulations defined by values of observed covariates, treatment assignment is random
- Rules out unobserved confounders
- Given by design in randomized experiment
- Untestable in most observational analyses
  - Sometimes testable indirectly
  - Sensitivity analyses!

Overlap

- The other condition (less well understood and more often ignored): overlap

$$0 < \Pr(D_i = 1 \mid X_i) < 1, \quad \text{for all } i$$

- For all possible values of the covariates there are both treated and control units
- Sometimes referred to as "positivity" or "probabilistic assignmemt"
- Can be examined from the data!
- Little overlap in covariates between treatment groups:
  - reliable causal inference practically impossible
  - relies on extrapolation (large bias and variance, sensitive)
- Unconfoundedness and overlap jointly define the "strong ignorability" assumption (Rosenbaum and Rubin, 1983)

Note: $e(x) = \Pr(D_i = 1 \mid X_i = x)$ is called the propensity score

Identification

- Under unconfoundedness and overlap, we have

$$\Pr(Y_d \mid X) = \Pr(Y \mid X, D = d).$$

- The observed distribution of Y in treatment group $D = d$ is equal to the distribution of the potential outcome Y_d.

- Thus we have the following two identification formulas

$$\begin{aligned}\tau^{ATE} &= E[\mu_1(X) - \mu_0(X)] \\ &= E\left[\frac{DY}{e(X)} - \frac{(1-D)Y}{1-e(X)}\right],\end{aligned}$$

where

$$\mu_d(X) = E(Y_d \mid X) = E(Y \mid D = d, X)$$

is the *outcome model* under treatment d ($d = 0, 1$) and

$$e(x) = \Pr(D_i = 1 \mid X_i = x)$$

is the *propensity score*.

Two (or 2.5) estimation strategies

- *Outcome regression* :

$$\tau = N^{-1} \sum_{i=1}^{N} \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}.$$

- *Inverse probability weighting (IPW)*:

$$\tau = \frac{\sum_{i=1}^{N} D_i Y_i / \hat{e}(X_i)}{\sum_{i=1}^{N} D_i / \hat{e}(X_i)} - \frac{\sum_{i=1}^{N} (1 - D_i) Y_i / \{1 - \hat{e}(X_i)\}}{\sum_{i=1}^{N} (1 - D_i) / \{1 - \hat{e}(X_i)\}}.$$

- Combination of both, so-called “Doubly Robust” estimators

$$\tau = \tau^{reg} + N^{-1} \sum_{i=1}^{N} \left\{ \frac{D_i R_i}{\hat{e}(X_i)} - \frac{(1 - D_i) R_i}{1 - \hat{e}(X_i)} \right\},$$

Here $R_i = Y_i - \mu_{D_i}(X_i)$ is the residual from outcome modeling

Outcome models

- Specify regression model for the (potential) outcome on treatment and covariates: $\mu_d(X)$
- Fit $\mu_d(X)$ to observed data and obtain fitted potential outcomes $\hat{\mu}_d(X_i)$ with $d = 0, 1$ This is akin to imputing the missing potential outcomes
- All estimands can be estimated based on $\hat{\mu}_d(X_i)$:
  - ATE:

$$\hat{\tau}^{ATE} = \frac{1}{N} \sum_{i=1}^{N} \Big[D_i \big(Y_i - \hat{\mu}_0(X_i) \big) + (1 - D_i) \big(\hat{\mu}_1(X_i) - Y_i \big) \Big]$$

  - ATT:

$$\hat{\tau}^{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \big\{ Y_i - \hat{\mu}_0(X_i) \big\}$$

  - CATE $\tau(x)$:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

The most common outcome model of all

- Linear regression:

$$\mu_d(x) = \alpha_d + \beta_d X_i + \epsilon_{d,i}.$$

- Fitting one model for each d is equivalent to adding an interaction term Xd This is not the case for nonlinear models!

- The outcome-modeling estimator of the ATE has the same form as the ANCOVA estimator in experiments:

$$\hat{\tau}^{ATE} = \left\{\bar{Y}_1 - \hat{\beta}_1(\bar{X}_1 - \bar{X})\right\} - \left\{\bar{Y}_0 - \hat{\beta}_0(\bar{X}_0 - \bar{X})\right\},$$

  where $\hat{\beta}_d$ is the OLS estimate of the coefficient of D in the regression $\mu_d(x)$.

- NOTE! Estimator is not consistent if the linear model is misspecified – unlike randomized experiments

- S.E. via or bootstrap or delta method

The importance of overlap

- Can use a wide range of outcome models besides linear models (e.g., from machine learning, Bayesian stats)
- Key issue is model specification:
  - separate models for each treatment group ?
  - unified model with treatment indicator (important to include treatment covariate interaction)? Equiv. for simple linear models, not for nonlinear ones
- If imbalance of covariates between treatment groups is small, model specification is less important
- If imbalance is large, results rely on extrapolation in the region with little overlap
- High model sensitivity

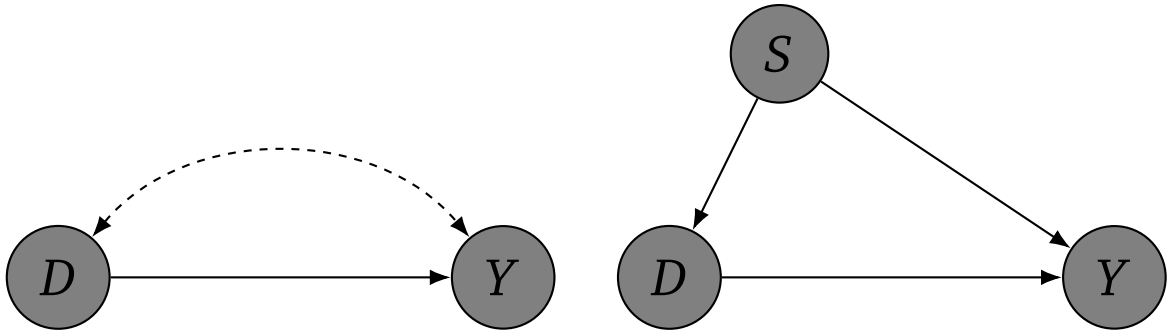
How to reduce model sensitivity

- The two well-known choices
  - DESIGN
  - FLEXIBLE MODEL
- Best of course to have both (double robust / double learning estimators): balance covariates in design stage; use flexible models in analysis stage
- Balance covariates (often involve propensity scores):
  - Matching
  - Stratification
  - Weighting
- Flexible models [more on that later in course on request...]
  - Semi-parametric /nonparametric models
  - Machine Learning methods (tree-based methods [CART, RF], boosting)
  - Bayesian nonparametrics (BART, GP, DPM)

A graphical look

The basic problem of confounding / selection

- Consider (again) the following structures



where S denoted all variables that systematically determine treatment assignment

- In other words, we have a conditional probability distribution

$$P(D = 1|S)$$

- For example (where S is gender)

$$P(D = 1|female) = 0.3$$
$$P(D = 1|male) = 0.5$$

The basic problem of confounding / selection

- Completely observing S makes treatment assignment "**ignorable**"

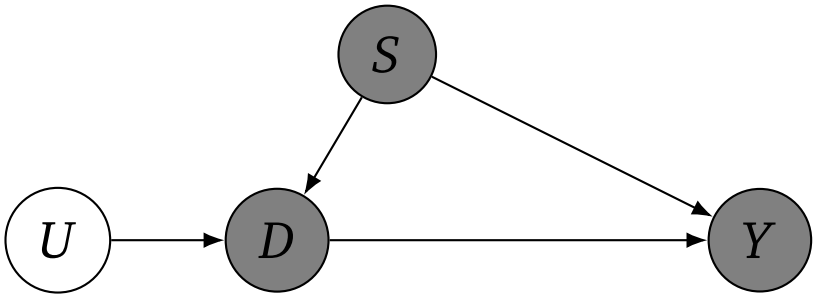
$$(Y_0, Y_1) \perp D|S$$

- In words, the potential outcomes (and functions of them) are independent of D with the strata (think of filtering) defined by S
- ATE can be simply obtained by filtering/conditioning/stratification:

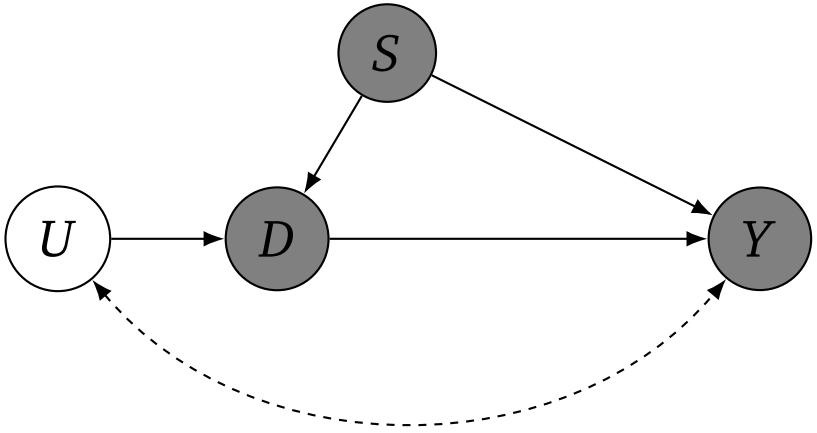
Effect	Strata	weights
$E(Y\|D)$	$S = 0$ [male]	$P(S = 0)$
$E(Y\|D)$	$S = 1$ [female]	$P(S = 1)$

A note on commonly used language

- The assumption underlying the strategy discussed above is often referred to as **selection on observables**. Example DAG:

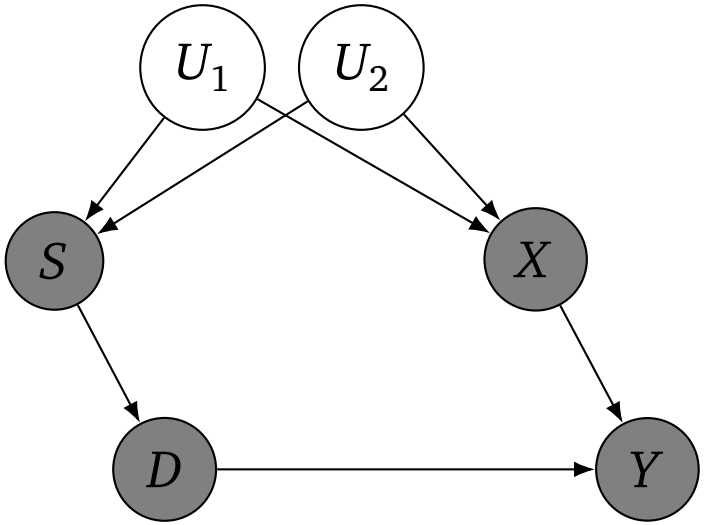


- The logical complement to this situation is **selection on unobservables**



In other words, treatment selection/assignment is non-ignorable.

Confounding, conditioning and adjustment



- Task: block backdoor paths from D to Y Note, neither X nor S is a collider
- Two strategies
  - *Balancing:* Identify and condition on variables S
  - *Adjustment:* Identify and condition on variables X

Note: distinction between X and S is somewhat arbitrary, but helpful to us conceptually...

Confounding, conditioning and adjustment

- Why look for S-type variables?
  - Better substantive intuitions:
    - If individuals select into D then it is often easier to theorize a selection process
  - Fewer estimation problems:
    - Searching for X to predict Y risks 'data-mining' which compromises significance tests etc...
    - Data mining for D does not (directly) compromise inference about D $\rightarrow$ Y

Balance

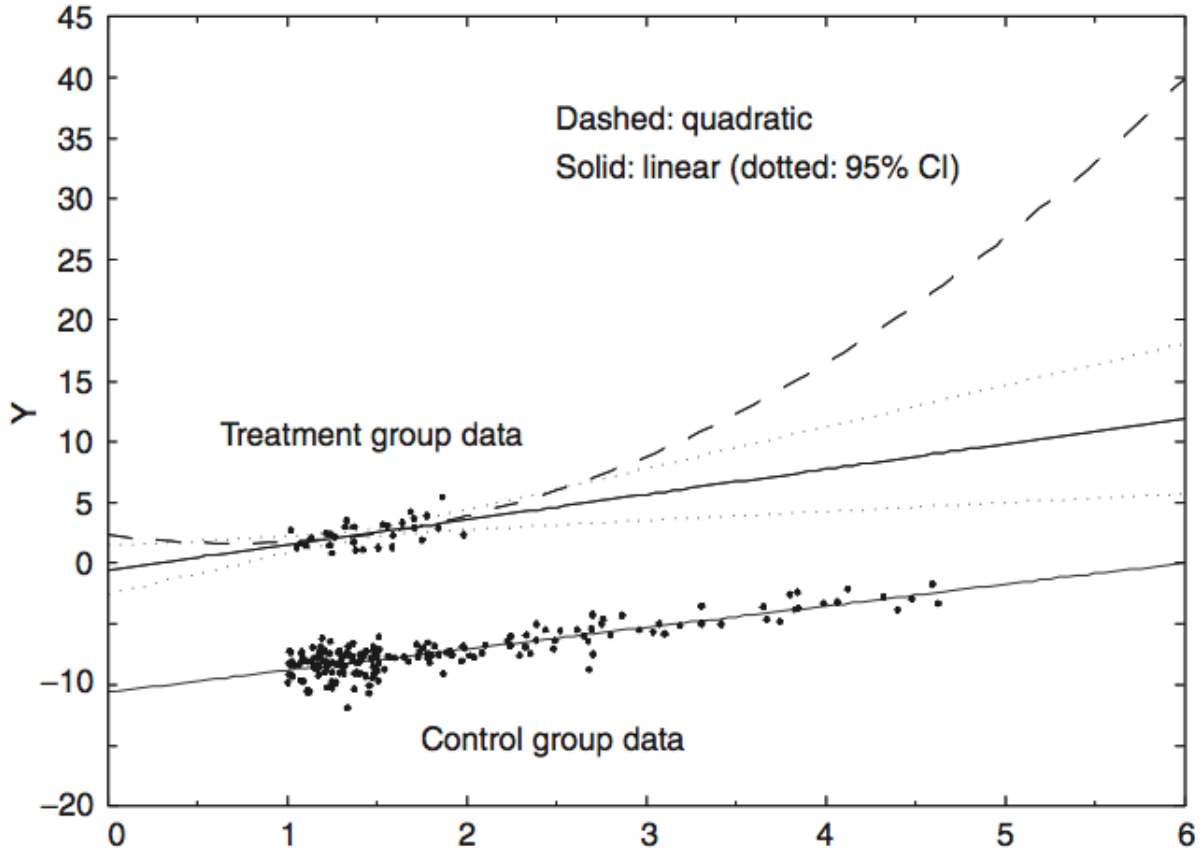
- To identify D → Y we need comparable cases in treatment and control groups
  - Comparable means: balanced confounding variables
  - Consider an experiment with two groups created using simple random assignment
  - Here, balance means:

    $$P(S|D = 1) = P(S|D = 0)$$

    i.e., the probability distribution of S is the same in the treatment and control group

Balance

- Remember our discussion of this example...



Matching to achieve balance

- For ATT, we can achieve balance by matching control cases to individual treatment cases
- The core idea here is *stratification*: treatment and control cases are matched = put together into strata

Stratification

- Assume that we have selection observables S
- We can identify the ATE assuming

$$E(Y_1|D=1,S) = E(Y_1|D=0,S)$$
$$E(Y_0|D=1,S) = E(Y_0|D=0,S)$$

- We get unbiased estimates of the treatment effect given strata $S=s$:

$$[E_N(y_i|d_i=1,s_i=s) - E_N(y_i|d_i=0,s_i=s)] \xrightarrow{p} E(\delta|S=s)$$

- The corresponding unconditional quantities are obtained by weighting with $P_N(s_i=s)$

$$\sum_s [E_N(y_i|d_i=1,s_i=s) - E_N(y_i|d_i=0,s_i=s)]\, P_N(s_i=s) \xrightarrow{p} E(\delta)$$

Stratification

- Consider an infinitely large data set (= don't worry about sampling).
- Our data yields 4.4 in the control and 10.2 in the treatment group.
- Our population has the following joint distribution of S and D

	Joint probability distribution, $P(S,D)$		
	$D=0$	$D=1$	
$S=1$	$P[S=1,D=0]=0.36$	$P[S=1,D=1]=0.08$	$P[S=1]=0.44$
$S=2$	$P[S=2,D=0]=0.12$	$P[S=2,D=1]=0.12$	$P[S=2]=0.24$
$S=3$	$P[S=3,D=0]=0.12$	$P[S=3,D=1]=0.20$	$P[S=3]=0.32$
	$P[D=0]=0.6$	$P[D=1]=0.4$	

- with potential outcomes

	Potential outcomes		
	Control state	**Treatment state**	
$S=1$	$E[Y^0\|S=1]=2$	$E[Y^1\|S=1]=4$	$E[Y^1-Y^0\|S=1]=2$
$S=2$	$E[Y^0\|S=2]=6$	$E[Y^1\|S=2]=8$	$E[Y^1-Y^0\|S=2]=2$
$S=3$	$E[Y^0\|S=3]=10$	$E[Y^1\|S=3]=14$	$E[Y^1-Y^0\|S=3]=4$
	$E[Y^0\|D=0]=4.4$	$E[Y^1\|D=1]=10.2$	

Stratification

- The empirical counterparts are the mean observed outcomes

	Estimated mean observed outcome given s_i and d_i	
	Control group	**Treatment group**
$s_i = 1$	$E_N[y_i \mid s_i = 1, d_i = 0] = 2$	$E_N[y_i \mid s_i = 1, d_i = 1] = 4$
$s_i = 2$	$E_N[y_i \mid s_i = 2, d_i = 0] = 6$	$E_N[y_i \mid s_i = 2, d_i = 1] = 8$
$s_i = 3$	$E_N[y_i \mid s_i = 3, d_i = 0] = 10$	$E_N[y_i \mid s_i = 3, d_i = 1] = 14$

- The conditional distribution of S given D

	$P(S \mid D)$	
	$d_i = 0$	$d_i = 1$
$s_i = 1$	$P_N[s_i = 1 \mid d_i = 0] = 0.6$	$P_N[s_i = 1 \mid d_i = 1] = 0.2$
$s_i = 2$	$P_N[s_i = 2 \mid d_i = 0] = 0.2$	$P_N[s_i = 2 \mid d_i = 1] = 0.3$
$s_i = 3$	$P_N[s_i = 3 \mid d_i = 0] = 0.2$	$P_N[s_i = 3 \mid d_i = 1] = 0.5$

- Thus, we can now calculate ATE, ATC, ATT as weighted average of within-strata estimates
- ATT: $(4-2)(.2)+(8-6)(.3)+(14-10)(.5) = 3$
- ATC: $(4-2)(.6)+(8-6)(.2)+(14-10)(.2) = 2.4$
- ATE: $(4-2)(.44)+(8-6)(.24)+(14-10)(.32) = 2.64$

The importance of overlap

- Imagine a strata (e.g., $S = 1$) where there are no treated cases
- In terms of the joint probability distribution, $P(D = 1, S = 1) = 0$
- No value for $E(y_i|s_i = 1, d_i = 1)$ exists
- Makes it impossible to identify ATE
- Still possible to identify ATT

Note: instead of structural zeros (as above), often empty treatment / control cases in some strata simply due to empirical sparsity. Balance problems increase with number of covariates

Propensity score

- Summarize the effects of all confounders in a single index
- Propensity score [e(S)]

$$P(D|S_1, S_2, \dots)$$

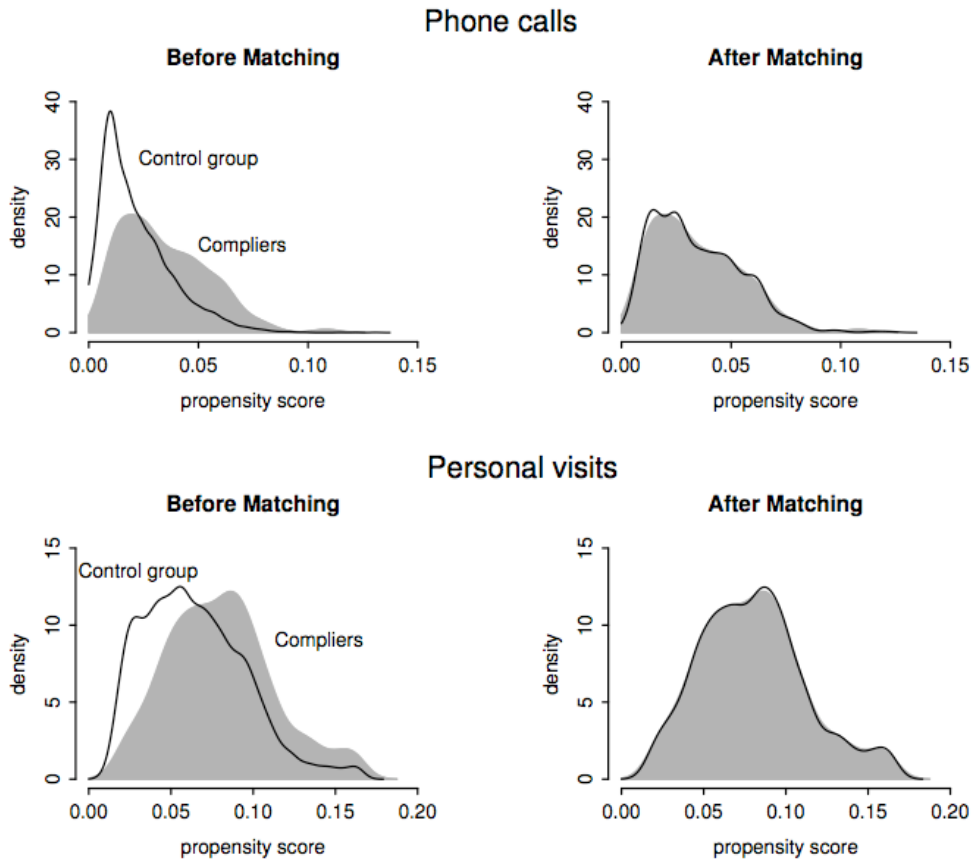
- Variables that don't affect D are irrelevant
- Even variables that do affect D are only relevant to the extent that they change $P(D)$
- A simple propensity score model (usually estimated as logit/probit)

$$D = \alpha + S_1\beta_1 + S_2\beta_2 + \cdots + \epsilon$$

- PS takes the slice of variation in S_1, S_2,... relevant to D:
  - if $\beta_1 = 0$ then cases with $D = 0$ and $D = 1$ can still have different values of S_1, but that doesn't matter

Propensity scores as distance

- Propensity score is now a distance measure that can be used to match and discard cases
- Allows for easier balance checking (in one dimension)



Propensity score problems

- We don't really know the propensity score – we have to rely on model based estimates
- Not straightforward to take into account uncertainty in estimation
- If propensity score for a case is 0 or 1 then weight can be infinite and the relevant counterfactual is undefined
  (this may be an advantage...)

General structure of matching estimators

- Partition the data into sets T of treated and C control cases. Denote by n^T the number of treated cases
- We estimate the ATT as

$$\delta_{ATT} = \frac{1}{n^T} \sum_{i \in T} \left[(y_i | d_i = 1) - \sum_{j \in C} w_{ij} (y_j | d_j = 0) \right]$$

where distance weights w_{ij} are defined differently by different matching estimators

Some algorithms

- Exact matching
  - Treated cases are matched to controls with identical values of S covariates creating k_i treatment–control pairs
  - w_{ij} are set to $1/k_i$ for matched cases, 0 for unmatched ones
  - Variant: randomly pick control from possible matches ($w_{ij} = 1$)
- Nearest neighbor (and variants)
  - Select control cases (usually 1 or n) closest to treatment on a 1-d distance metric calculated from S
  - Various distance metrics are used, e.g., Mahalanobis distance or the estimates propensity score
  - Weights are set to 1 or $1/n$ in the multiple neighbor case, 0 for unmatched cases
  - To prevent badly matched cases (esp. for n-NN) caliper (max. distance) can be used. Might lead to varying number of neighbors (weights are now $1/n_i$)

- Interval/subclassifcation and full matching
  - Sort treatment and control cases on 1-d distance metric into groups of equal size
  - Full matching uses intervals of variable size (minimizing within-interval heterogeneity)
  - Estimator is calculated within each interval with weights set so that equal weight is given to treated and control cases
  - ATT is the sum of interval-specific estimates weighted by number of cases in the interval
- Kernel matching
  - Uses all control cases to match
  - w_{ij} are constructed using kernel estimate of distance between treated case and all control cases (nearest cases get the most weight)

Balance problems

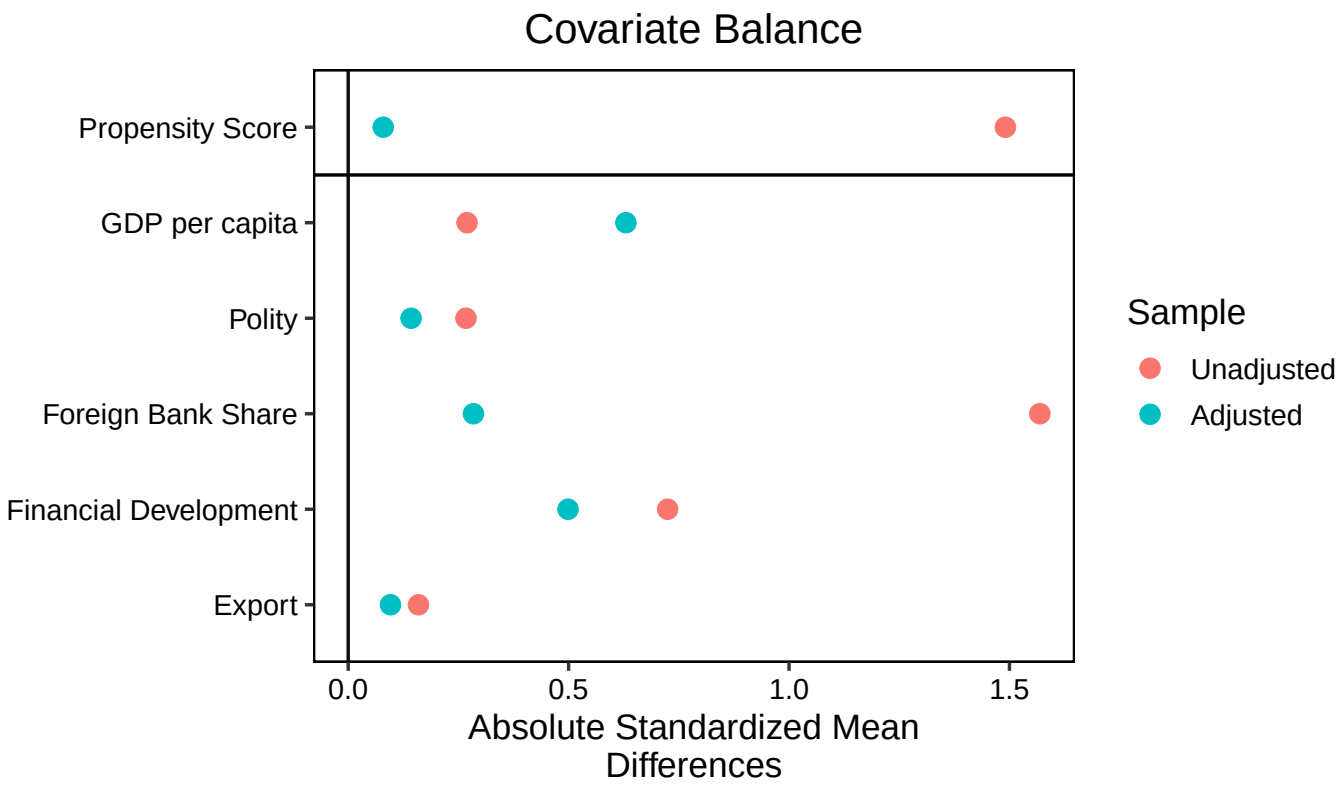
- Curse of dimensionality
  - K binary variables have 2^K possible combinations
  - Most are never observed: the covariate space is empty
  - Covariate space might be weirdly structured
- By discarding cases outside common support we may change the subject
  - ATT $\neq$ average treatment effect on the treated which happen to overlap enough of the control group

Some practical considerations

- Assess covariate balance between treatment and control groups
  - By inspecting the propensity score
  - Graphically (only feasible with limited number of variables)
  - Using hypothesis tests (e.g., diff. in means). Beware of sample size issues and Type I and II errors (!!)
  - Using standardized distance metrics. E.g.,

$$\frac{|E(x_i|d_i=1)-E(x_i|d_i=0)|}{\sqrt{0.5*Var(x_i|d_i=1)+0.5*Var(x_i|d_i=0)}}$$

    before and after matching



Using function `love.plot()` from R package `cobalt`