Causal inference with observational data: matching and weighting methods

Ben Ackerman and Rebecca Johnson [1]

July 30, 2018

[1] Acknowledgments to Elizabeth Stuart, Brandon Stewart, and others for sharing slides

Motivating example: want to measure the effect of mental health services on future jail bookings (some notation)

- $Y_i(t)$: potential outcome for observation i under treatment t
- T_i: binary indicator for treatment status for observation i
  - $t \in \{0, 1\}$, $0 =$ control, $1 =$ treatment
- In this example...
  - $Y_i = \begin{cases} 0 & \text{if person } i \text{ not booked in jail} \\ 1 & \text{if person } i \text{ booked in jail} \end{cases}$
  - $T_i = \begin{cases} 0 & \text{if person } i \text{ doesn't receive tx from social worker} \\ 1 & \text{if person } i \text{ receives tx from social worker} \end{cases}$
- X: observed covariate correlated with T and Y
- U: unobserved covariate correlated with T and Y
- *Note:* we know this is not what the JoCo team is working on! But just wanted to include a relevant example

Ideal Setting for Estimating Causal Effect

Randomized Controlled Trial (RCT)

- Can we randomize people to receive mental health treatment?

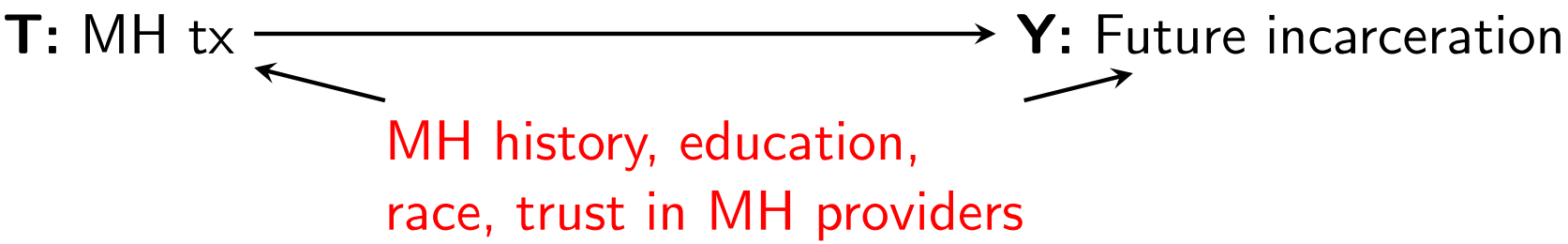
If so, can estimate the average treatment effect (ATE):

$$E[Y(1) - Y(0)]$$

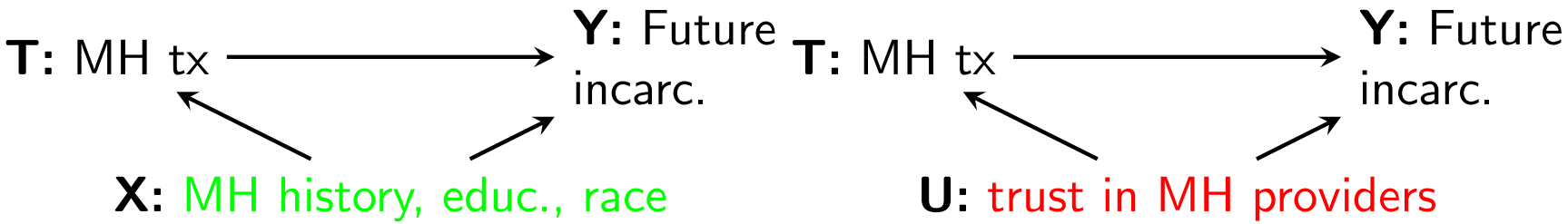
If not, we must try to use *non-experimental data* to estimate the ATE.
But, one big challenge...

Confounding

The problem of confounding when measuring the effect of T on Y



Where Aniket and Ancil left off: two classes of confounders



- Unobserved confounders: e.g., trust in MH providers; some latent measure of severe mental health issues not captured in the observed mental health history
- Observed confounders: race, sex, etc.
- The strategies we'll cover today help make an assumption–that treatment units' potential outcomes are similar to control units once we condition on the correct set of observed confounders (but may still vary substantially along unobserved dimensions)– more plausible
  - Sensitivity analyses exist showing how results change with different magnitudes of unobserved confounding
    - E.g.: Bounding (Rosenbaum, 2002); confounding-adjusted outcomes (Blackwell, 2014); E-value (VanderWeele and Ding, 2017)

Comparing MH tx and control: What if randomly assigned?

	MH tx	Controls
Age (mean)	30	
% Male	67.2	
% Black	53	
% High School Educated or more	45	
Num. of co-occurring conditions (median)	5	
N	137	

Comparing MH tx and control: In reality

	MH tx	Controls
Age (mean)	30	45
% Male	67.2	39.9
% Black	53	40
% High School Educated or more	45	44
Num. of co-occurring conditions (median)	5	1
N	137	393

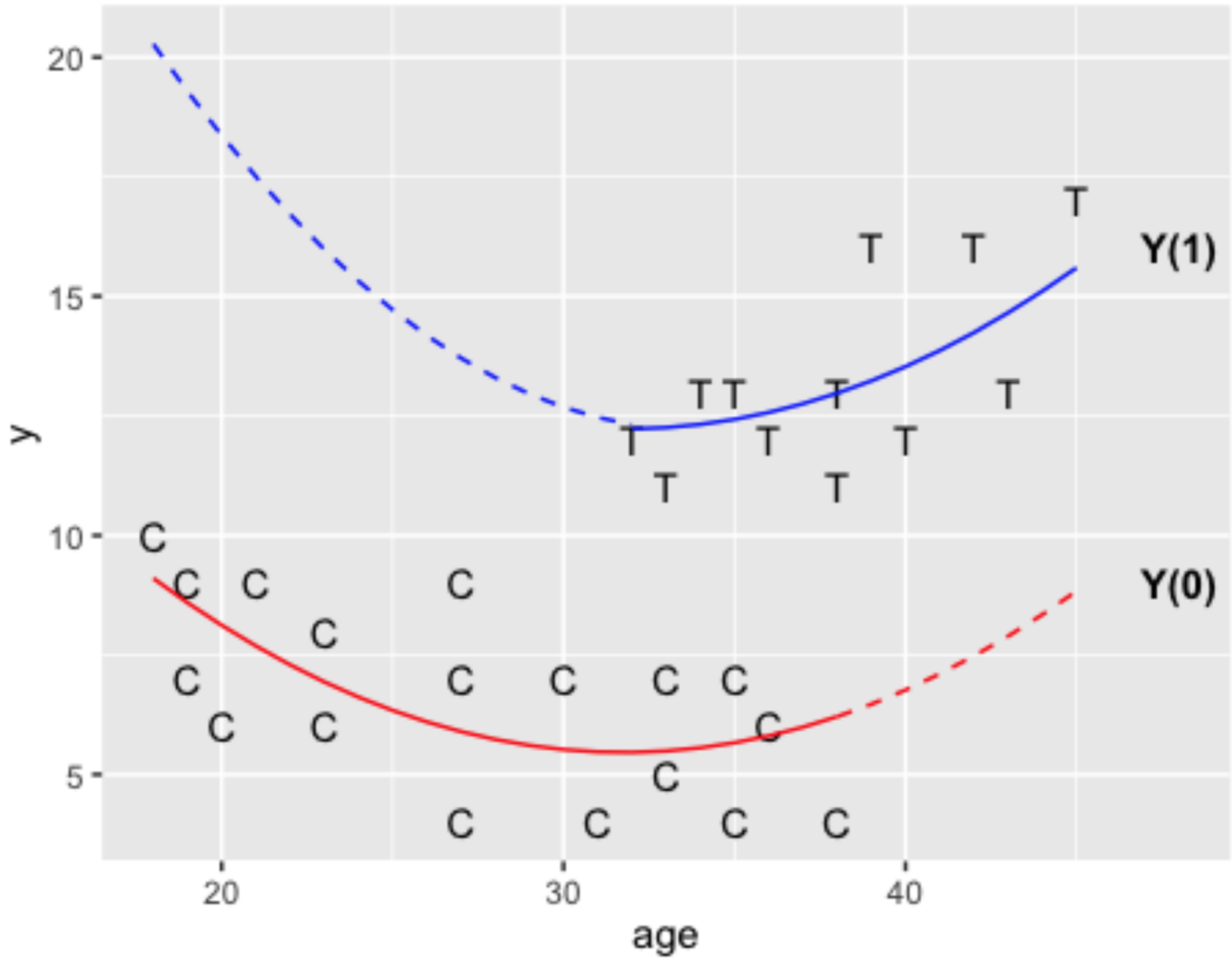
Traditional non-experimental design options

- Stratification
  - Put people into groups with same values of covariates (e.g., analyze individuals in poverty and not in poverty separately)
  - But lots of variables to stratify on, limited sample size
  - Hard to adjust for many covariates this way
- Regression analysis
  - e.g., normal linear regression of outcome given treatment and covariates
  - Predict incarceration given covariates and mental health services use; look at coefficient on mental health services

Dangers of regression adjustment on full samples

- When the treated and control groups have very different distributions of the confounders, can lead to bias if model misspecified
  - Observe $Y(0)$ in the control group, $Y(1)$ in the treatment group
  - Predicting $Y(0)$ for the treatment group may involve extrapolation and pure reliance on functional form
- Regression is only "trustworthyïf the treatment and comparison groups look similar on covariates (Rubin, 2001; Ho et al., 2007)

Visualizing Extrapolation



So can we never use regression adjustment?

- No, not at all!
  - If differences between groups not large (e.g., $<$ .1 or .2 standard deviations) regression works fine (Rubin, 2001)
  - And in fact, matching methods (like propensity scores) work best in conjunction with regression

So what should we do instead?

Some methods attempt to remove both forms of confounding:

- *Regression discontinuity*
  - Org has limited MH providers. Do they score people and only offer treatment to those who score above some threshold? If so, assume people around threshold are similar on both observed and unobserved
- *Instrumental variables*
  - Amount of rainfall meaningfully predicts whether people receive mental health treatment (no weak instrument) and is uncorrelated with their likelihood of future incarceration (exclusion restriction)

When those don't work for a question, make assumption covered on the next slides[2] and use matching/weighting to make assumption more plausible. Cover two methods:

1. Propensity Scores
2. Alternative methods (CEM)

[2] Goes by different names: selection on observables (SOO); no unobserved confounders; strong ignorability

So what should we do instead?

- **Matching/Weighting**
  1. **Propensity Scores**
  2. Alternative methods (CEM)

Propensity Score Methods

- Propensity score methods attempt to replicate two features of randomized experiments
  - Create groups that look only randomly different from one another (at least on observed variables)
  - Don't use outcome when setting up the design
- Idea is to find treated and control individuals with similar covariate values
  - Increase "balance"

Propensity scores

- Probability of receiving the treatment (T), given the covariates (X)

$$e_i = P(T_i = 1|X_i)$$

- Two key features:
  1. Balancing score: At each value of the propensity score, the distribution of observed covariates (that went into the propensity score) the same in the treated and control groups
  2. If treatment assignment independent of potential outcomes given covariates, then also independent of potential outcomes given the propensity score (no unmeasured confounders)
- Facilitate matching because can match just on propensity score, rather than all of the covariates individually
- Can estimate propensity scores using a simple logistic regression model

Feature 1: Propensity scores as balancing scores

- At each value of the propensity score, the distribution of observed covariates (that went into the propensity score) the same in the treated and control groups
  - Intuitively, if two people had the same probability of receiving the treatment (e.g., mental health services) and one did and one didn't, it must have been random as to who did and who didn't
  - Within small range of propensity score values, treated and comparison individuals should look only randomly different on the observed covariates
  - Difference in outcomes within groups with same/similar propensity scores gives unbiased estimate of treatment effect

Feature 2: Unconfoundedness

- If unconfoundedness holds given the full set of observed covariates, also holds given the propensity score
  - $P(T|X, Y(0), Y(1)) = P(T|X)$ implies $P(T|X, Y(0), Y(1)) = P(T|e(X))$
- This is what allows us to match just on propensity score; don't need to deal with all the covariates individually

Common support/overlap

- Additional requirement for causal inference: 'common support'
- Everyone has a positive probability of receiving any of the conditions
  - Known as 'positivity' in epidemiology
  - Part of 'strong ignorability' in statistical literature
- Fundamentally, if someone has 0 probability of receiving one of the treatments, we have no way of learning what their outcomes would be under that treatment, so can't learn about their causal effects
- In practice this examined by looking at the common support/overlap of the propensity score distribution: do the groups overlap (e.g., histograms)

Using propensity scores/types of "matching"

- k to 1 nearest neighbor matching
  - For each treated unit, select k controls with closest propensity scores
  - Will discussion variations on this later
- Subclassification/stratification
  - Group individuals into groups with similar propensity score values
  - Often 5 subclasses used (Cochran 1968)
- Weighting adjustments
  - Inverse probability of exposure weights (IPTW)
  - Weighting by the odds

So what should we do instead?

- **Matching/Weighting**
  1. Propensity Scores
  2. **Alternative methods (CEM)**

Returning to the rationale behind propensity scores: difficult to perform exact matching of treatment units with controls with continuous covariates / as k increases

T = 1; k predictors

T	**age**	**sex**	**race**	**income(k)**	**yrseduc**	...	k
1	27	*M*	*W*	10.5	12	...	
1	24.5	*M*	*W*	30	10	...	
⋮							
1	21	*F*	*B*	60	16	...	

T = 0; k predictors

T	**age**	**sex**	**race**	**income(k)**	**yrseduc**	...	k
0	21	*M*	*W*	25.5	16	...	
0	23	*M*	*B*	60	14	...	
⋮							
0	28	*F*	*H*	42	20	...	

Propensity score approach: address inability to exact match by compressing k into a single scalar and then use that scalar to reweight data

T	**age**	**yrseduc**	...	k	$weight(w)$
1	27	12	...		1
1	24.5	10	...		1
⋮					
1	21	16	...		1
0	21	16	...		1
0	23	14	...		1
⋮					
0	28	20	...		1

$$\xrightarrow{\pi(X) = \Pr(T = 1|X)}$$

T	$\hat{\pi}$	$w(IPT)$
1	0.8	$\frac{1}{0.8} = 1.25$
1	0.2	$\frac{1}{0.2} = 5$
⋮		
1	0.4	$\frac{1}{0.4} = 2.5$
0	0.4	$\frac{1}{(1-0.4)} = 1.666$
0	0.6	$\frac{1}{(1=0.6)} = 2.5$
⋮		
0	0.1	$\frac{1}{(1-0.1)} = 1.111$

Two classes of extensions

1. Use methods other than logistic regression to estimate $\pi(X) = \Pr(T = 1|X)$
   - *Why?* Often optimize logit to find $\hat{\beta}_{MLE}$ (vector of weights on coefficients that maximize the likelihood function); in reality, we care less about weights on X that highly predict treatment and more about weights that optimize balance between treatment and control on X
   - *How:* covariate-balancing propensity scores (CBPS; Imai and Ratkovic, 2014)
2. Return to original goal–exact matching of each treatment unit with a control unit–and perform that matching based on original covariate values (X) rather than a compression of those covariate values into $\pi(X)$
   - *Why?* simulations show (King and Nielsen, 2016):
     - When PSM is most likely to reduce imbalance: larger divergence between $\pi(T = 1)$ and $\pi(T = 0)$; PSM useful but common support condition most likely to be violated
     - When PSM increases imbalance: as $\pi(T = 1) = \pi(T = 0) \rightarrow 0.5$, PSM least useful due to random pruning but common support condition most likely to be met
   - *How?* CEM (Iacus, King, and Porro, 2011)

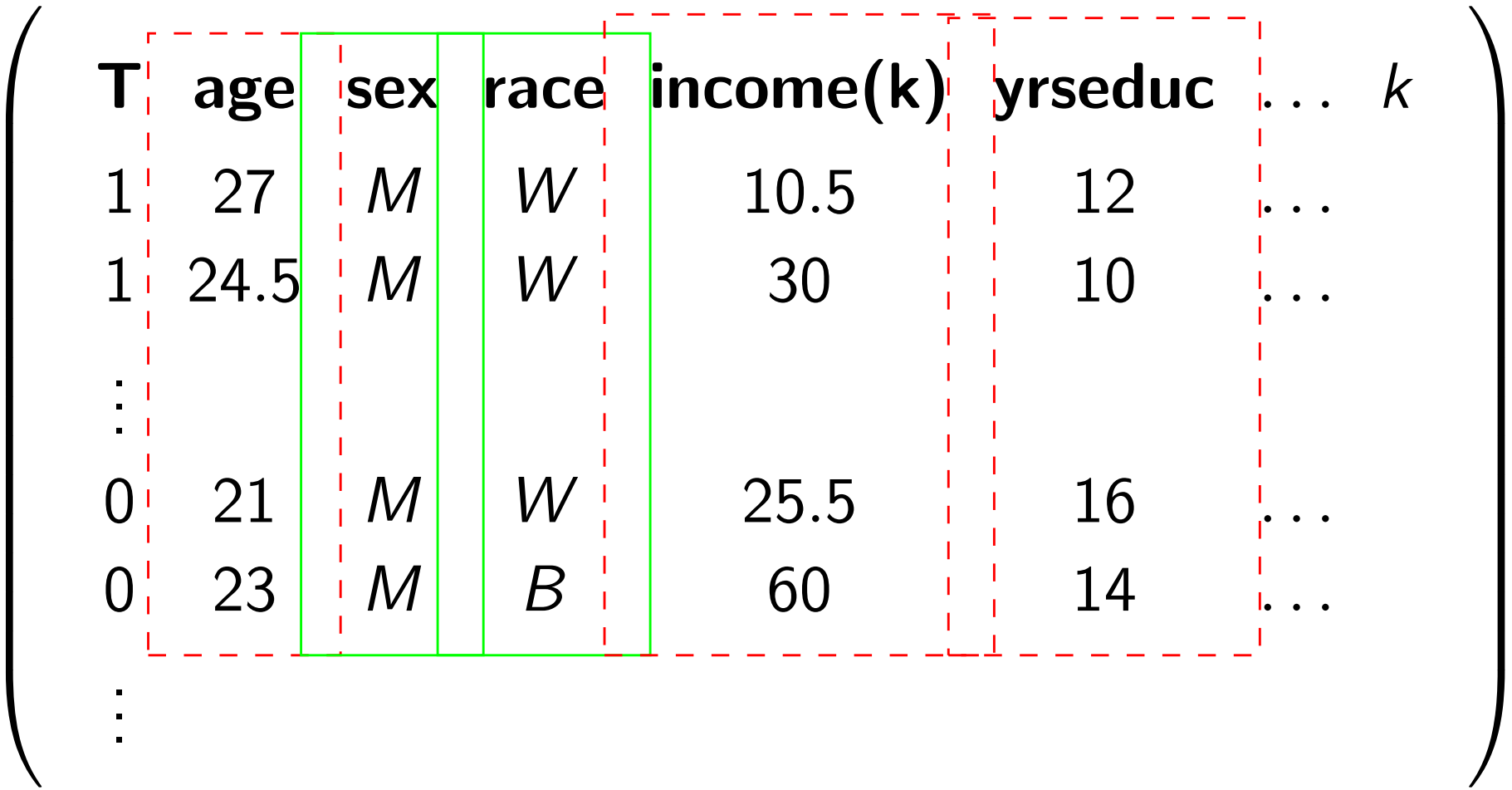
Coarsened exact matching (CEM): stick with original covariates rather than π; get as close as possible to exact matching in two ways

T	**age**	**sex**	**race**	**income(k)**	**yrseduc**	...	k
1	27	*M*	*W*	10.5	12	...	
1	24.5	*M*	*W*	30	10	...	
⋮							
0	21	*M*	*W*	25.5	16	...	
0	23	*M*	*B*	60	14	...	
⋮							

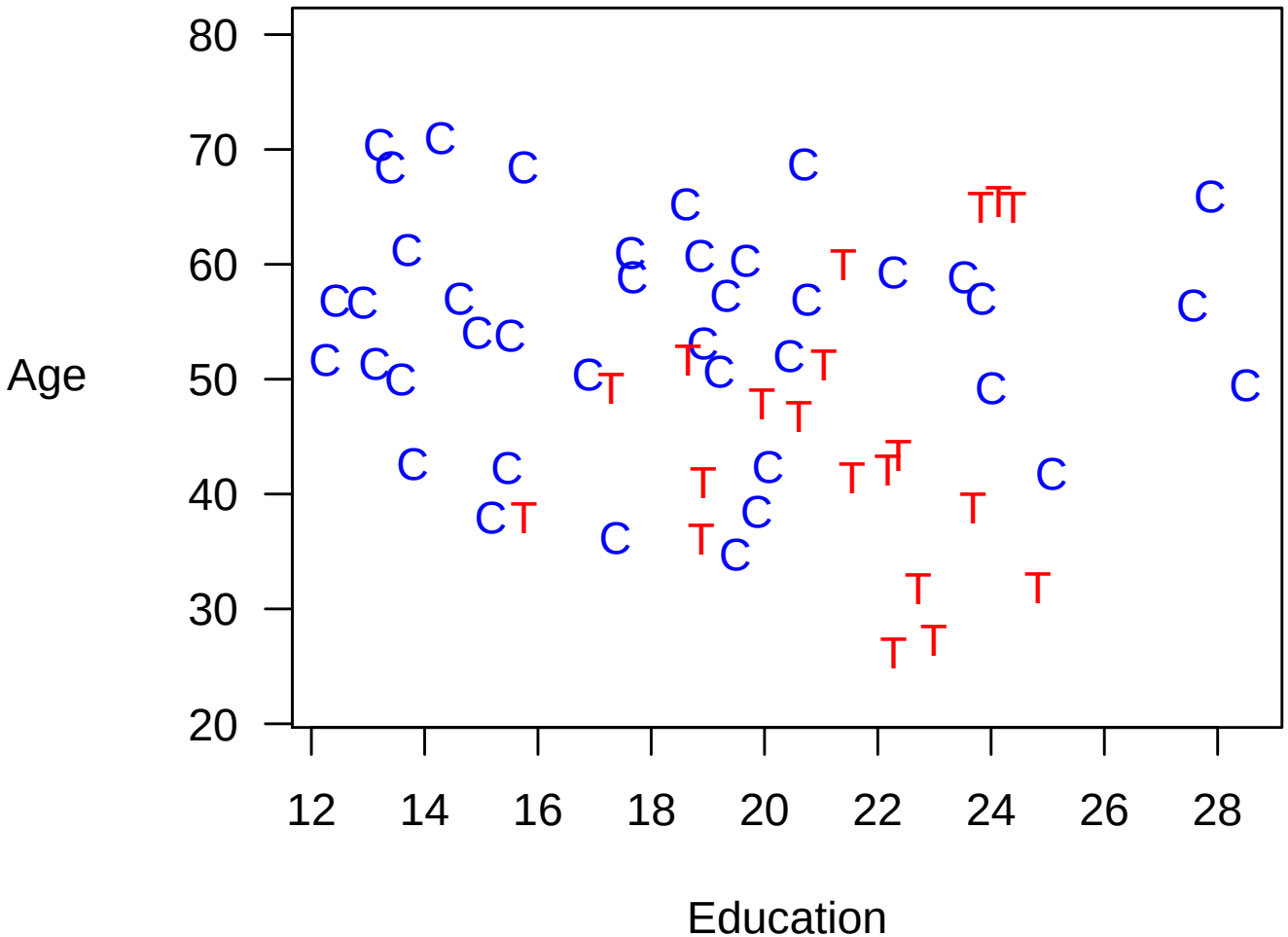
Match treatment and control units exactly on nominal k

T	**age**	**sex**	**race**	**income(k)**	**yrseduc**	...	k
1	27	*M*	*W*	10.5	12	...	
1	24.5	*M*	*W*	30	10	...	
⋮							
0	21	*M*	*W*	25.5	16	...	
0	23	*M*	*B*	60	14	...	
⋮							

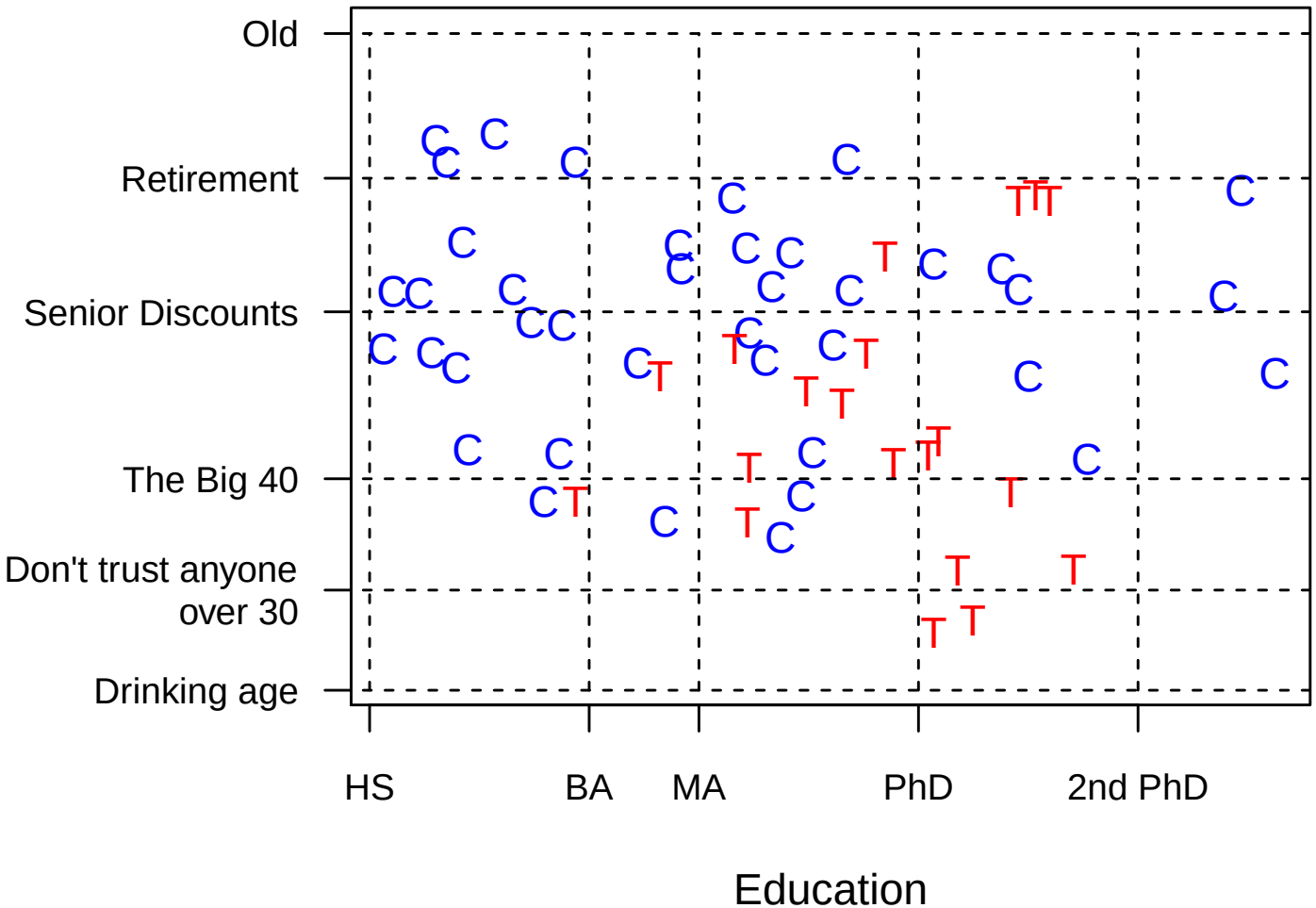
Coarsen continuous k and perform exact matching within strata



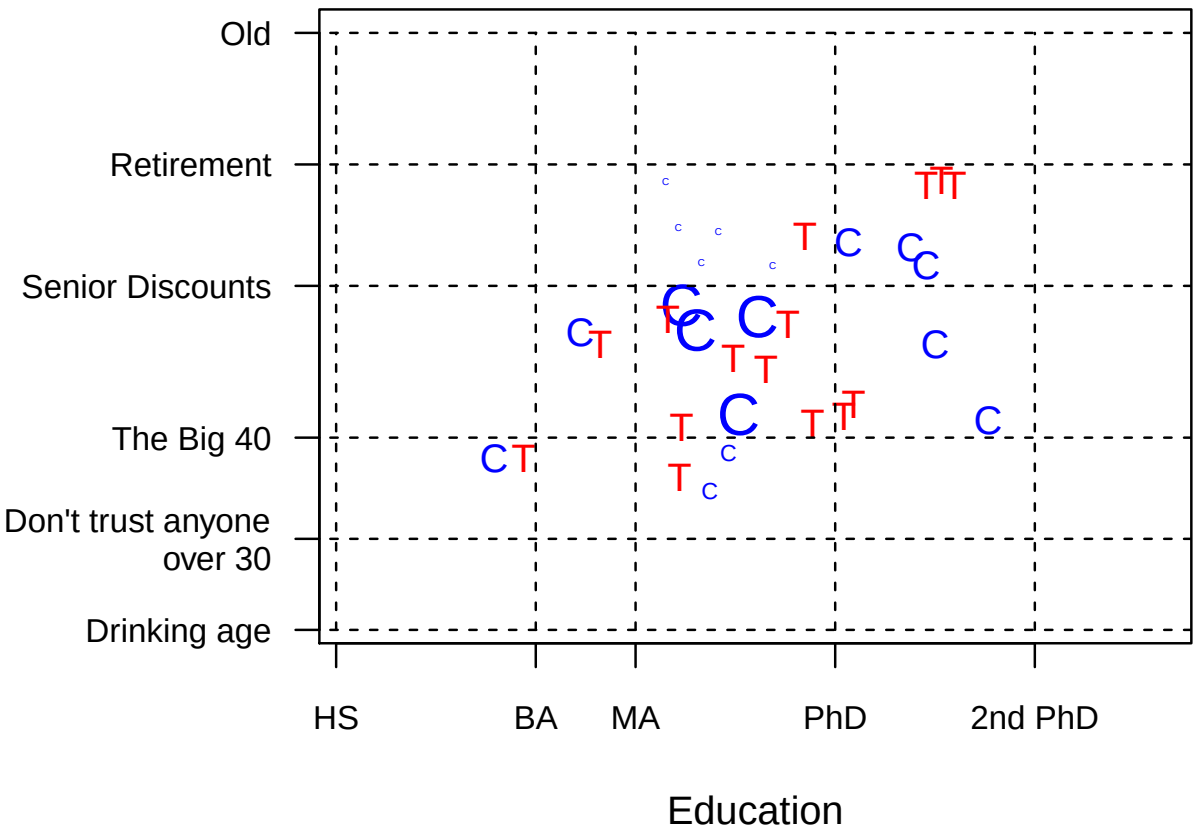
Visual illustration. Start with un-coarsened continuous values...



Coarsen based on meaningful categories and construct strata



Drop all units in strata where matching can't occur and match/reweight (note that if tx units are dropped, leads to new estimand)



Matching: implementation

- `MatchIt`: can specify distance metric to calculate (e.g., distance between π; mahalanobis distance on raw covar values); method to match on (e.g., nearest neighbor); how many control units to match with each treatment (e.g., 1:1, 2:1)
- `CEM`
- `WeightIt`
- `Cobalt`

- `causalinference`: less familiar but seems to have propensity score
- Or...estimate prop score using preferred method and write own matching function (written in R but could translate)

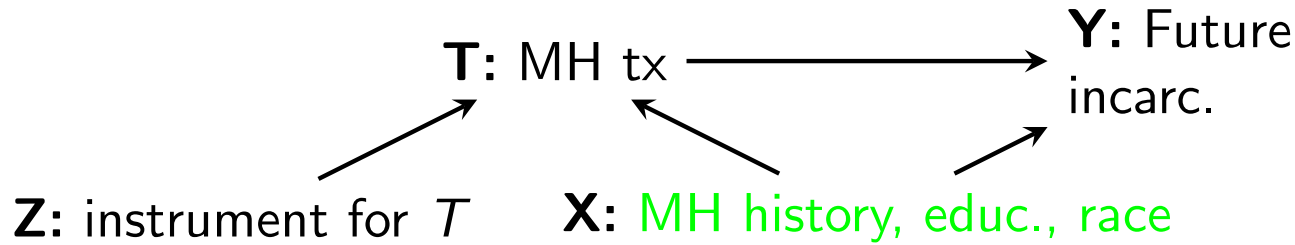
```
##matching function with tie break
custom_match_tiebreak <- function(controlpool, txpool, num2match){

        #sort control by propensity score:
        #highest to lowest
        control_sorted <- controlpool %>%
              arrange(desc(logit_fit))

        match_store <- matrix(NA, nrow = num2match, ncol = 2)
        colnames(match_store) <- c("controlID", "treatedID")
        for(i in 1:num2match) {
          ##subset to unit i of controlpool
          control_unit <- control_sorted[i, ]

          #calculate absolute value distance between that control and remainin
          #and add that as variable to treatment
          txpool <- txpool %>%
              mutate(dist_2_control= abs(control_unit$logit_fit -
                    logit_fit))
```

Matching and DSSG projects

- Three general quantities of interest with our motivating example
  1. $\hat{\beta}$: estimate effect of mental health treatment ($T = 1$) on jail (re-)booking
  2. $\hat{y}$: predict jail (re-)booking
  3. Some combination, e.g.: $\hat{\tau} = \hat{Y}(T = 1) - \hat{Y}(T = 0)$
- Goal of DSSG projects is often #2; most work on matching + ML focuses on potential of latter for goal #1

How ML can help with matching if we care about $\hat{\beta}$

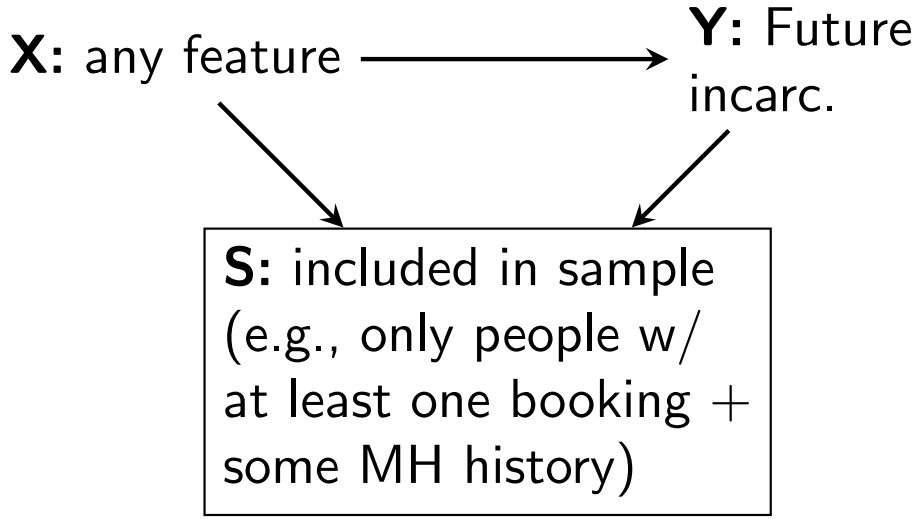
- Estimating $\pi(X) = \Pr(T = 1|X)$
  - Properties reviewed earlier depend on conditioning on the 'correct' set of confounders. E.g., X but not Z (Bhattacharya and Vogt, 2007; Pearl, 2010)



  - Some evidence that using ML methods like random forest makes the estimation of $\pi(X)$ more robust to accidental inclusion of Z
- Inadvertant byproduct of introducing ML methods to predict a unit's treatment status is that the same mark of a 'good' ML model–units where observed $T_i = 1$ have high $\hat{\pi}_i$; observed $T_i = 0$ have low $\hat{\pi}_i$–makes the 'common support'/overlap assumption $(0 < P(T_i = 1|X_i) < 1)$ more difficult to defend
  - As assumption, it's untestable; in practice, people look at overlap in empirical distributions of $\hat{\pi}_i$

How matching can help ML if we care about $\hat{y}$

- Many of our training (and test) sets are neither a full census nor a random sample of units (people; building; workplaces)
- Problematic if T or Y is correlated with a unit's probability of being in the sample (collider bias)



- Think of S as the thing to match on and model that process

Conclusion

- Other strategies for causal inference are aimed at addressing both observed features of a unit that influence its likelihood of treatment and unobserved features
- Matching makes a strong assumption–that once we condition on the correct set of observed features, the potential outcomes of a unit are independent from its treatment status. But it allows us to work with a broader range of observational data
- Once we decide *whether* 'to match', there are many choices about *how* to match with non-trivial implications for parameter estimates