Matching

EDLD 650: Week 8

David D. Liebowitz

Agenda

1. Roadmap and Goals (9:00-9:10)

2. Discussion Questions (9:10-10:20)

- Diaz & Handa
- Murnane & Willett, Ch. 12

3. Break (10:20-10:30)

4. Applied matching (10:30-11:40)

- PSM and CEM

5. Wrap-up (11:40-11:50)

Roadmap

Causal identification problem:

$Y_i(D_i)$: outcome for individual i, where D_i = 1 if treated, 0 if control

$ITE_i = Y_i(1) - Y_i(0)$; but unobserved so:

$$\widehat{ATE} = \left(\sum_{i=1}^{n_1} Y_i \Big/ n_1\right) - \left(\sum_{i=1}^{n_0} Y_i \Big/ n_0\right)$$

Is assignment randomized?

YES

$E[\varepsilon_i|X_i] = E[\varepsilon_i|D_i] = 0$; $\varepsilon_i \perp D_i$

- -Check balance: $\bar{X}_{D=1} \cong \bar{X}_{D=0}$
- -Estimate effects:

  $\bar{Y}_{D=1} - \bar{Y}_{D=0}$; **or**

  $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$

  $\widehat{\beta_1} = \widehat{ATE} = \widehat{ATT}$
- -Sampling approach/External validity
- -Power?
- -Internal validity:

  Spillover (SUTVA)? Hawthorne/John Henry? Non-compliance? Attrition?

ITT v. TOT

NO, exogenous variation in D_i

$E[\varepsilon_i| D_i] \neq 0$

-Difference-in-Differences (DD)

$\widehat{ATE} = (\bar{Y}(1)_{t=1} - \bar{Y}(1)_{t=0}) - (\bar{Y}(0)_{t=1} - \bar{Y}(0)_{t=0})$;

or $Y_{it} = \beta_1 D_{it} + \Gamma_i + \Pi_t + \varepsilon_{it}$

→ D_{it}: 1 if unit under treatment in time t, otherwise 0

Common trends? Simultaneous shocks? Time-varying effects? Weighting?

-Regression Discontinuity (RD)

given forcing var $D_i = (S_i \geq C)$

$Y_i = \beta_0 + \beta_1 S_i + 1(S_i \geq C)\beta_2 + \varepsilon_i$

Smooth and non-manipulated forcing variable? Bandwidth? Functional form? Generalizability/LATE? Sharp or fuzzy?

-Instrumental Variables (IV)

Z_i correl w/ D_i but $Z_i \perp \varepsilon_i$

1st: $D_i = \alpha_0 + \alpha_1 Z_i + \delta_i$

2nd: $Y_i = \beta_0 + \beta_1 \widehat{D_i} + \varepsilon_i$

NO, D_i based on observables

$E[\varepsilon_i| D_i, X_i] = 0$

- -OLS
- -Multi-level modeling
  - Random: $Y_{ij} = \gamma_0 + \gamma_1 D_{ij} + (\varepsilon_{ij} + \upsilon_j)$
  - Fixed: $Y_{ij} = \Sigma_1^J \alpha_j S_{ij} + \gamma_1 D_{ij} + \varepsilon_{ij}$
- -Matching estimators
  - -PSM
  - -CEM
  - -IPW

Goals

1. Describe conceptual approach to matching analysis

2. Assess validity of matching approach and what selection on observable assumptions implies

3. Conduct matching analysis in simplified data using both propensity-score matching and coarsened-exact matching (CEM)

So random…

Break

Matching:

Propensity scores

Recall the Catholic school data

Show 5 entries   Search:

	id	math12	catholic	math8	faminc8
1	124902	49.7700004577637	1	50.2700004577637	10
2	180625	51.5099983215332	1	41.310001373291	11
3	702949	48.2799987792969	0	45.75	11
4	710976	53.0099983215332	0	46.0499992370605	9
5	1425490	65.3499984741211	1	66.6900024414062	10

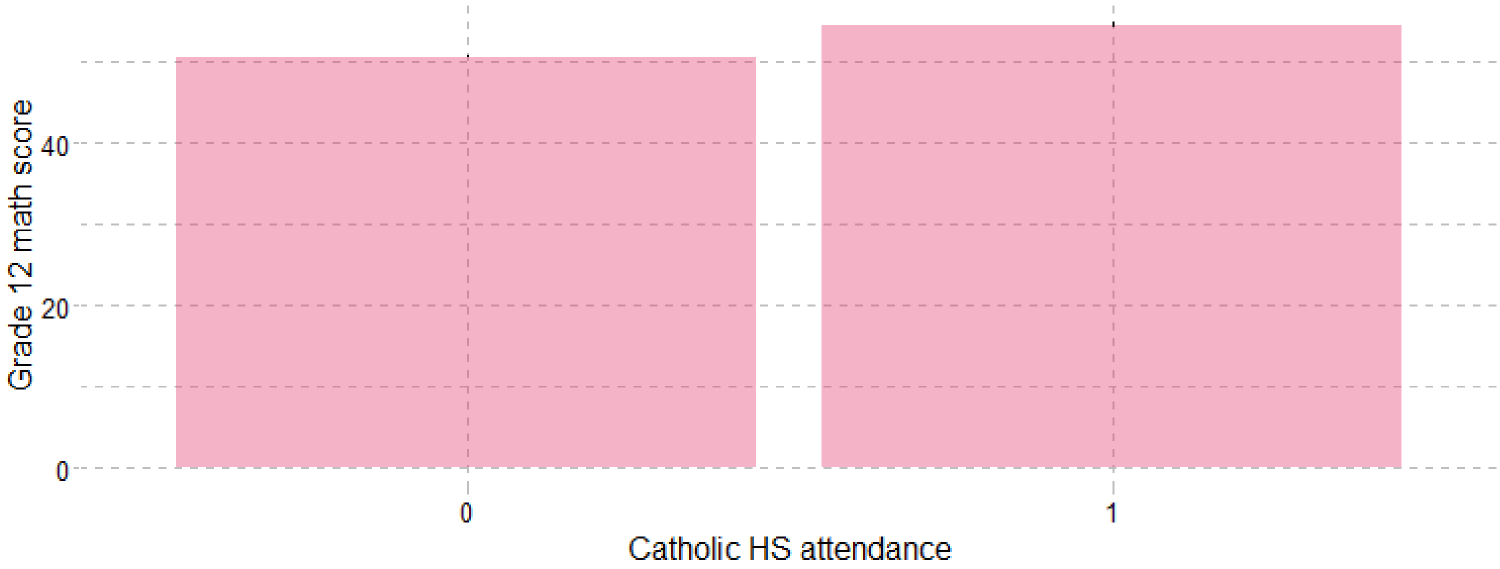
Showing 1 to 5 of 5 entries   Previous 1 Next

Are Catholic HSers higher-performing?

```
catholic %>% group_by(catholic) %>%
  summarise(n_students = n(),
   mean_math = mean(math12), SD_math = sd(math12))
```

```
#> # A tibble: 2 x 4
#>    catholic n_students mean_math SD_math
#>   <dbl+lbl>      <int>     <dbl>   <dbl>
#> 1   0 [no]        5079      50.6    9.53
#> 2   1 [yes]        592      54.5    8.46
```

Are Catholic HSers higher-performing?



Are Catholic HSers higher-performing?

```
ols1 ← lm(math12 ~ catholic, data=catholic)
summary(ols1)
```

```
...
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  50.6447     0.1323 382.815   <2e-16 ***
#> catholic      3.8949     0.4095   9.512   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 9.428 on 5669 degrees of freedom
#> Multiple R-squared:  0.01571,    Adjusted R-squared:  0.01554
#> F-statistic: 90.48 on 1 and 5669 DF,  p-value: < 2.2e-16
...
```

What is wrong with all of these approaches?

Are Catholic HS attendees different?

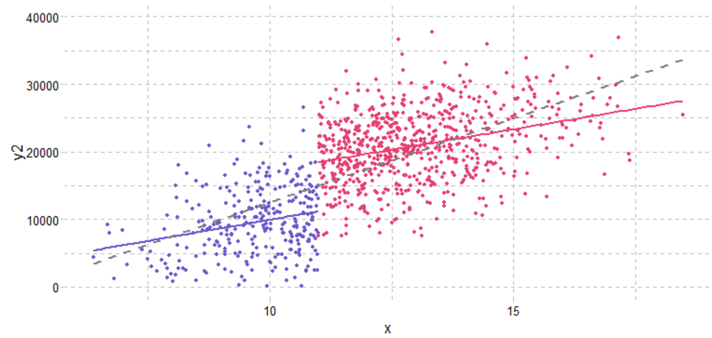
```
table <- tableby(catholic ~ faminc8 + math8 + white + female,
        numeric.stats=c("meansd"), cat.stats=c("N", "countpct"),
        digits=2, data=catholic)
mylabels <- list(faminc8 = "Family income level in 8th grade",
        math8 = "8th grade math score")
summary(table, labelTranslations = mylabels)
```

	0 (N=5079)	1 (N=592)	Total (N=5671)	p value
Family income level in 8th grade				< 0.001
Mean (SD)	9.43 (2.25)	10.36 (1.68)	9.53 (2.22)	
8th grade math score				< 0.001
Mean (SD)	51.24 (9.75)	53.66 (8.83)	51.49 (9.68)	
student is white?				< 0.001
Mean (SD)	0.68 (0.47)	0.80 (0.40)	0.69 (0.46)	
student is female?				0.253
Mean (SD)	0.52 (0.50)	0.54 (0.50)	0.52 (0.50)	

Implementing matching

Reminder of key assumptions/issues:

1. Selection on observables
2. Treatment is as-good-as-random, conditional on known set of observables
3. Tradeoff between bias, variance and generalizability



Practical considerations

Can implement this various ways. Pedagogically, we'll implement matching using a combination of the `MatchIt` package (which is similar to the `cem` package for Coarsened Exact Matching), the `fixest` implementation of logistic regression and data manipulation by hand.[1]

```
# install.packages("MatchIt")
# install.packages("gtools")
```

[1] Most of the coarsening we'll do can be done directly within the `MatchIt` package, but it's good to get your hands into the data to truly understand what it is you're doing!

Phase I: Generate propensities

Step 1: Estimate selection model

```r
pscores <- feglm(catholic ~ inc8 + math8 + mathfam,
                 family=c("logit"), data=catholic)
summary(pscores)
```

```
#> GLM estimation, family = binomial(link = "logit"), Dep. Var.: catholic
#> Observations: 5,671
#> Standard-errors: IID
#>              Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) -5.208846   0.586532 -8.88075  < 2.2e-16 ***
#> inc8         0.061803   0.014058  4.39633 1.1009e-05 ***
#> math8        0.042959   0.011138  3.85707 1.1476e-04 ***
#> mathfam     -0.000734   0.000262 -2.80586 5.0183e-03 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> Log-Likelihood: -1,837.6   Adj. Pseudo R2: 0.030071
#>            BIC:  3,709.8     Squared Cor.: 0.018645
```

Phase I: Generate propensities

Step 2: Predict selection likelihood

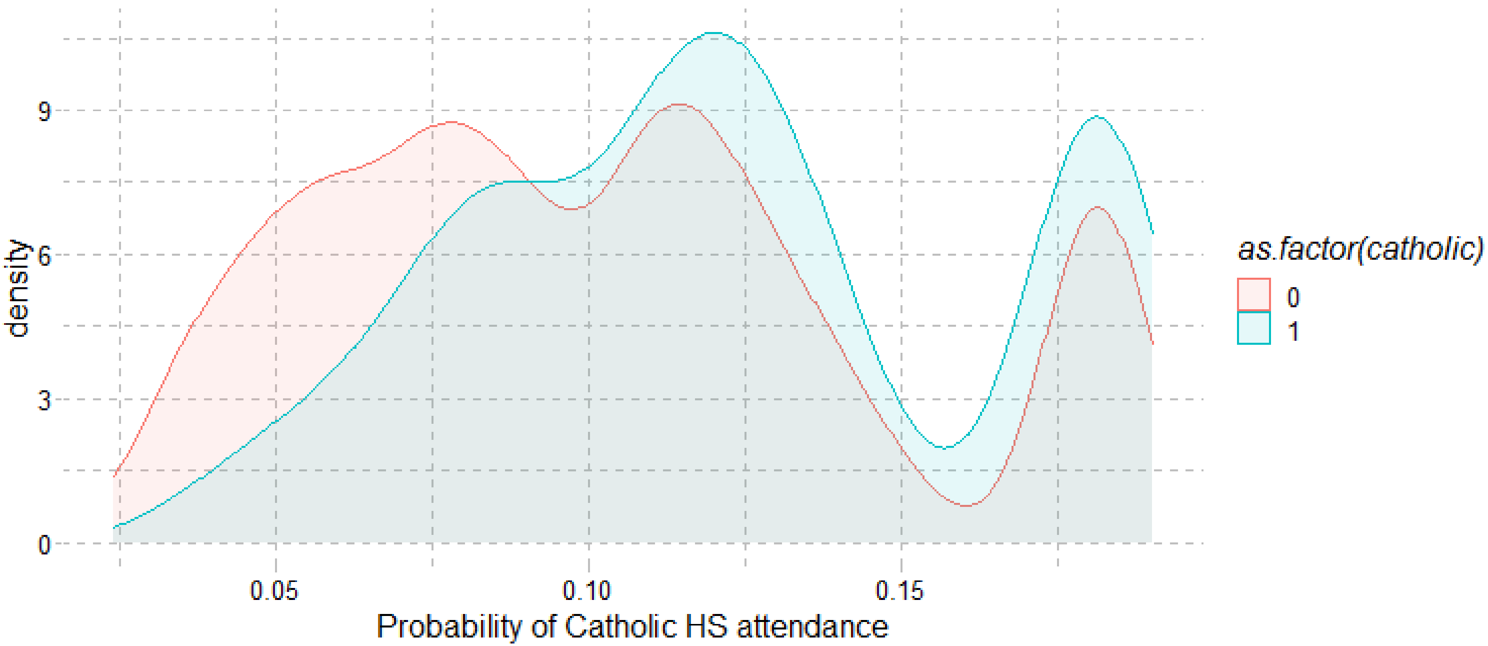
```
pscore_df <- data.frame(p_score = predict(pscores, type="response")
                        catholic = catholic$catholic)
head(pscore_df)
```

```
#>      p_score catholic
#> 1 0.09094085        1
#> 2 0.09312787        1
#> 3 0.08635750        1
#> 4 0.08478468        1
#> 5 0.13309352        1
#> 6 0.07903282        1
```

Note: to apply Inverse-Probability Weights (IPW), you would take these propensities and assign weights of $1/\hat{p}$ to treatment and $1/(1-\hat{p})$ to control units.

Phase I: Generate propensities

Step 3: Common support (pre-match)



Phase 2: PS Matching

Step 1: Assign nearest-neighbor match[1]

```
matched ← matchit(catholic ~ math8 + inc8, method="nearest",
                  replace=T, discard="both", data=catholic)
df_match ← match.data(matched)

# How many rows/columns in resulting dataframe?
dim(df_match)
```

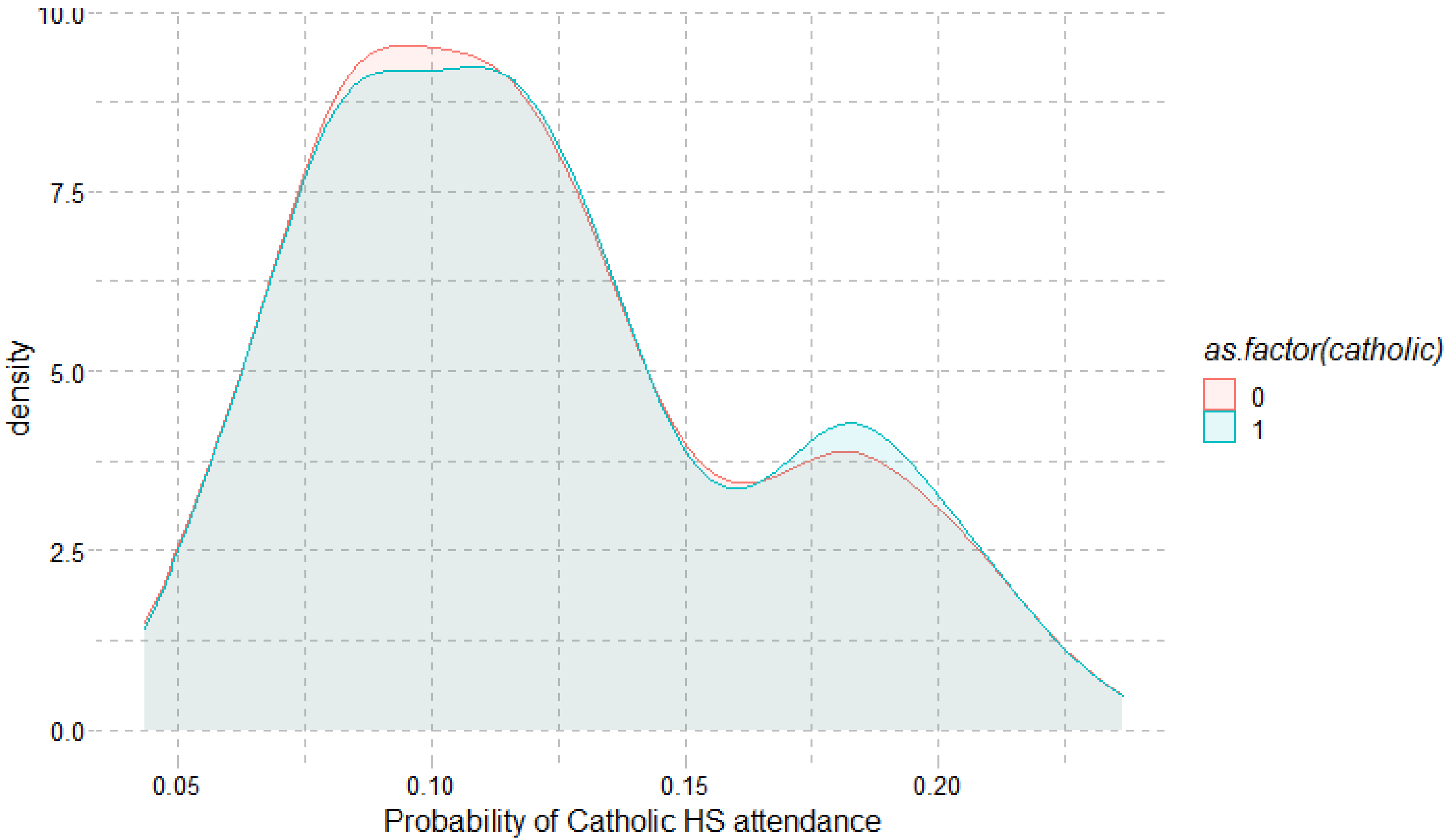
```
#> [1] 1118   30
```

This is the **NOT** same number of observations as were in the original sample... what happened?

[1] As you might anticipate, there are *lots* of different ways besides "nearest-neighbor with replacement" to create these matches.

Phase 2: PS Matching

Step 2: Common support (post-match)



Phase 2: PS Matching

Step 3: Examine balance

(doesn't really fit on screen)

```
summary(matched)
```

```
#>
#> Call:
#> matchit(formula = catholic ~ math8 + inc8, data = catholic, method = "nea
#>     discard = "both", replace = T)
#>
#> Summary of Balance for All Data:
#>          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mea
#> distance        0.1216        0.1024          0.4351     1.0216    0.134
#> math8          53.6604       51.2365          0.2746     0.8201    0.075
#> inc8           39.5346       31.8548          0.4714     0.8886    0.077
#>          eCDF Max
#> distance   0.2142
#> math8      0.1550
```

Phase 2: PS Matching

Step 3: Examine balance

Summary of balance for all data:

Variable	Means Treated	Means Control	Std. Mean Diff
distance	0.1216	0.1024	0.4351
math8	53.6604	51.2365	0.2746
inc8	39.5346	21.8548	0.4714

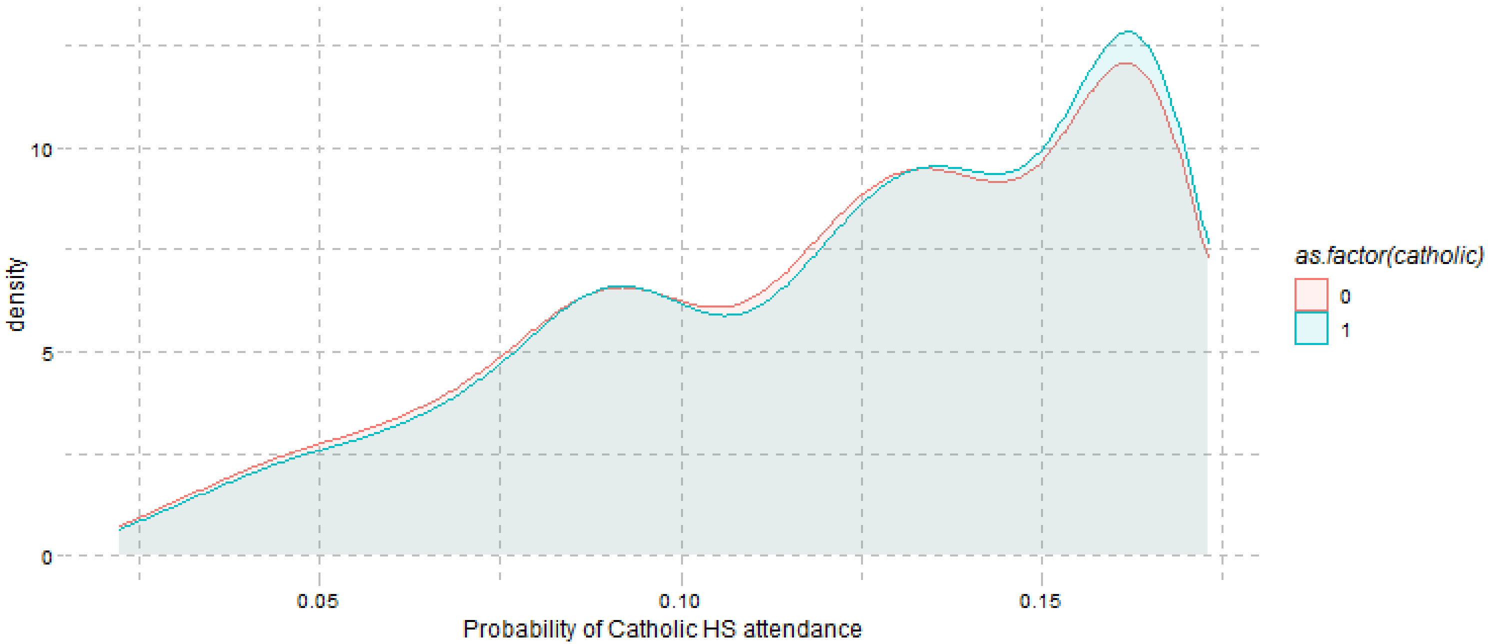
Summary of balance for matched data:

Variable	Means Treated	Means Control	Std. Mean Diff
distance	0.1216	0.1216	0.0000
math8	53.6604	53.4416	0.0248
inc8	39.5346	39.6698	-0.0083

Phase 2: PS Matching

Could get even closer with fuller model:

```
matched2 ← matchit(catholic ~ math8 + inc8 + inc8sq + mathfam,
        method="nearest", replace=T, discard="both", data=catholic)
```



Phase 2: Estimate effects

```
psmatch2 <- lm(math12 ~ catholic + math8 + inc8 + inc8sq + mathfam,
               weights = weights, data=df_match)
#Notice how we have matched on just math8 and inc8 but are now
#   adjusting for more in our estimation. This is fine!
# Very important to include weights!
summary(psmatch2)
```

```
...
#>
#> Coefficients:
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  1.3079490  2.4457425   0.535 0.592905
#> catholic     1.5990422  0.3144335   5.085 4.30e-07 ***
#> math8        0.9065628  0.0468521  19.349  < 2e-16 ***
#> inc8         0.3701132  0.0663303   5.580 3.02e-08 ***
#> inc8sq      -0.0015921  0.0005686  -2.800 0.005194 **
#> mathfam     -0.0040694  0.0010783  -3.774 0.000169 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
```

Can you interpret these results?

> In a matched sample of students who had nearly identical 8th grade math test scores and family income levels and were equally likely to attend private school based on these observable conditions, the effect of attending parochial high school was to increase 12th grade math test scores by 1.59 scale score points [95% CI: 0.98, 2.22]. To the extent that families' selection into Catholic high school is based entirely on their children's 8th grade test scores and their family income, we can interpret this a credibly causal estimate of the effect of Catholic high school attendance, purged of observable variable bias.

Matching:

Coarsened Exact Matching (CEM)

A different approach: CEM

Some concerns with PSM:

- Model (rather than theory) dependent
- Lacks transparency
- Can exclude large portions of data
- Potential for bias
- *We'll return to these at the end!*

$\rightarrow$ more transparent (?) approach ... **Coarsened Exact Matching** ... literally what the words say!

Basic intuition:

- Create bins of observations by covariates and require observation to match exactly within these bins.
- Can require some bins be as fine-grained as original variables (then, it's just exact matching).

Creating bins

```r
table(catholic$faminc8)
```

```
#>
#>     1    2    3    4    5    6    7    8    9   10   11   12
#>    18   42   84   85  144  175  447  441  655 1267 1419  894
```

```r
catholic ← mutate(catholic, coarse_inc=ifelse(faminc8<5,1,faminc8)
catholic$coarse_inc ← as.ordered(catholic$coarse_inc)
levels(catholic$coarse_inc)
```

```
#> [1] "1"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12"
```

```r
summary(catholic$math8)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   34.48   43.45   50.45   51.49   58.55   77.20
```

```r
mathcuts ← c(43.45, 51.49, 58.55)
```

CEM matches

```
cem <- matchit(catholic ~ coarse_inc + math8,
      cutpoints=list(math8=mathcuts), method="cem", data=catholic)
df_cem <- match.data(cem)
table(df_cem$catholic)
```

```
#>
#>    0    1
#> 5079  592
```

This is the same number of observations as were in the original sample. What does this imply?

Quality of matches

```
summary(cem)
```

```
#>
#> Call:
#> matchit(formula = catholic ~ coarse_inc + math8, data = catholic,
#>     method = "cem", cutpoints = list(math8 = mathcuts))
#>
#> Summary of Balance for All Data:
#>              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF
#> coarse_inc1         0.0135        0.0435         -0.2598          .    0
#> coarse_inc5         0.0101        0.0272         -0.1701          .    0
#> coarse_inc6         0.0101        0.0333         -0.2310          .    0
#> coarse_inc7         0.0338        0.0841         -0.2783          .    0
#> coarse_inc8         0.0524        0.0807         -0.1273          .    0
#> coarse_inc9         0.0794        0.1197         -0.1491          .    0
#> coarse_inc10        0.2196        0.2239         -0.0103          .    0
#> coarse_inc11        0.3345        0.2404          0.1994          .    0
#> coarse_inc12        0.2466        0.1473          0.2305          .    0
#> math8              53.6604       51.2365          0.2746     0.8201    0
#>              eCDF Max
#> coarse_inc1    0.0300
```

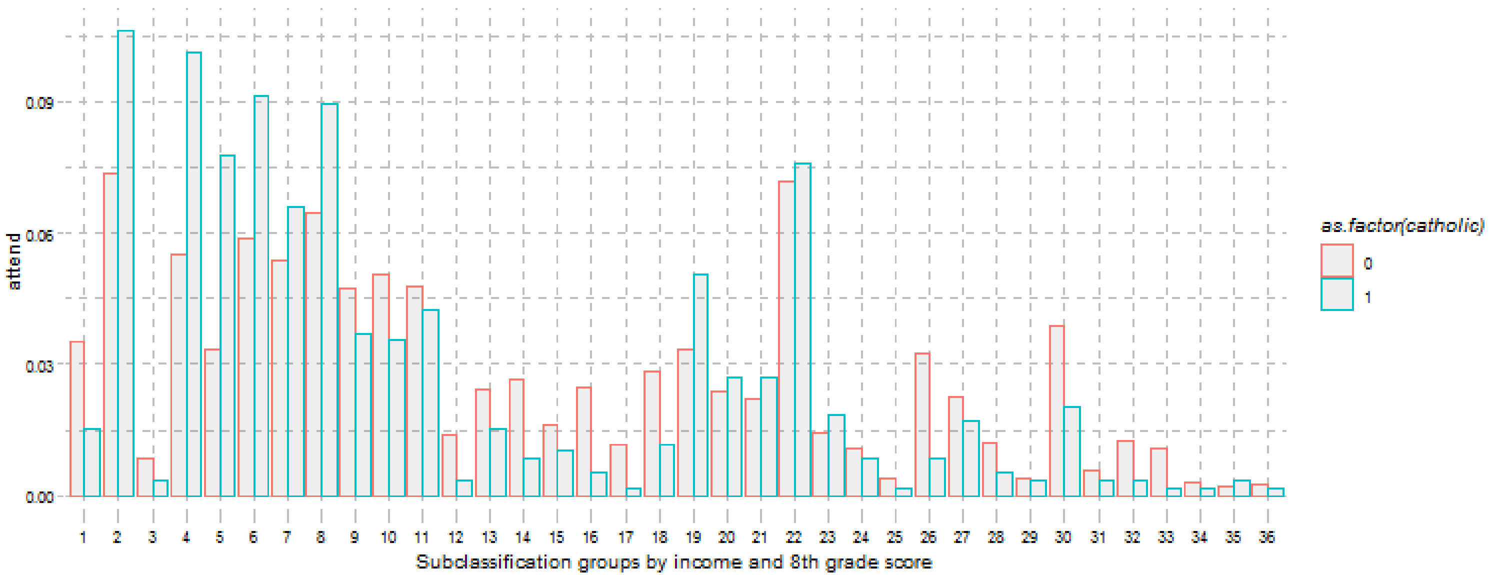
Quality of matches

Summary of balance for all data:

Variable	Means Treated	Means Control	Std. Mean Diff
coarse_inc1	0.0135	0.0435	-0.2598
coarse_inc5	0.0101	0.0272	-0.1701
coarse_inc6	0.101	0.0333	-0.2310
coarse_inc7	0.0338	0.0841	-0.2783
coarse_inc8	0.0524	0.0807	-0.1273
coarse_inc9	0.0794	0.1197	-0.1491
coarse_inc10	0.2196	0.2239	-0.0103
coarse_inc11	0.3345	0.2404	-0.1994
coarse_inc12	0.2466	0.1473	-0.2305
math8	53.6604	51.2365	-0.2746

Common support?

```
df_cem1 ← df_cem %>% group_by(catholic, subclass) %>%
            summarise(count= n())
df_cem1 ← df_cem1 %>%  mutate(attend = count / sum(count))
```



Different cuts?

Can generate different quantiles, e.g., quintiles

```
math8_quints <- gtools::quantcut(catholic$math8, 5)
table(math8_quints)
```

```
#> math8_quints
#> [34.5,42.1] (42.1,47.6] (47.6,53.3] (53.3,60.6] (60.6,77.2]
#>        1136        1133        1134        1134        1134
```

You might also have a substantive reason for the cuts:

```
mathcuts2 <- c(40, 45, 50, 55, 60, 65, 70)
```

Different cuts: Balance

```
#>
#> Call:
#> matchit(formula = catholic ~ coarse_inc + math8, data = catholic,
#>     method = "cem", cutpoints = list(math8 = mathcuts2))
#>
#> Summary of Balance for All Data:
#>              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF
#> coarse_inc1         0.0135        0.0435         -0.2598          .    0
#> coarse_inc5         0.0101        0.0272         -0.1701          .    0
#> coarse_inc6         0.0101        0.0333         -0.2310          .    0
#> coarse_inc7         0.0338        0.0841         -0.2783          .    0
#> coarse_inc8         0.0524        0.0807         -0.1273          .    0
#> coarse_inc9         0.0794        0.1197         -0.1491          .    0
#> coarse_inc10        0.2196        0.2239         -0.0103          .    0
#> coarse_inc11        0.3345        0.2404          0.1994          .    0
#> coarse_inc12        0.2466        0.1473          0.2305          .    0
#> math8              53.6604       51.2365          0.2746     0.8201    0
#>              eCDF Max
#> coarse_inc1    0.0300
#> coarse_inc5    0.0170
#> coarse_inc6    0.0231
```

Big improvements!

Summary of balance for matched data:

Variable	Means Treated	Means Control	Std. Mean Diff
coarse_inc1	0.0269	0.0269	-0.000
coarse_inc5	0.0203	0.0203	0.000
coarse_inc6	0.0251	0.0251	-0.000
coarse_inc7	0.0799	0.0799	0.000
coarse_inc8	0.0744	0.0744	0.000
coarse_inc9	0.1171	0.1171	0.000
coarse_inc10	0.2322	0.2322	0.000
coarse_inc11	0.2601	0.2601	0.000
coarse_inc12	0.1639	0.1639	-0.000
math8	51.6351	51.3938	0.026

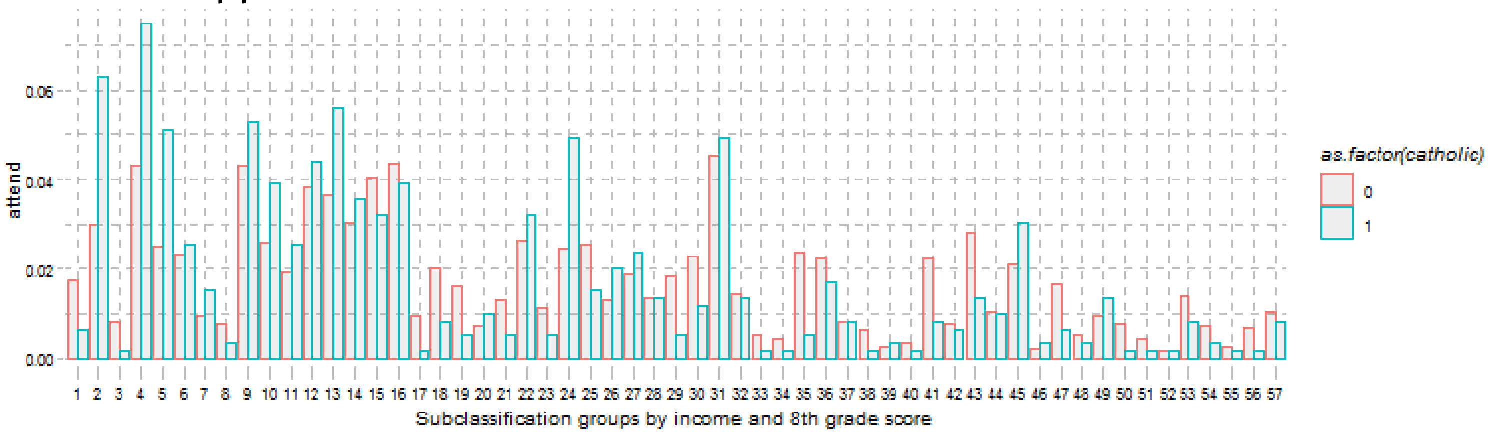
We've forced T/C to be identical within coarsened income bins. The *original* **math8** variable still has some imbalance (but it's much better). Within **mathcuts2**, T/C would be identical.

Minimal sample loss

Sample sizes:

Category	Control	Treated
All	5079	592
Matched	4866	590
Unmatched	213	2

Common support?



Estimating effects

```
att2 <- lm(math12 ~ catholic + coarse_inc + math8,
           data=df_cem2, weights = weights)
summary(att2)
```

```
#>
#> Call:
#> lm(formula = math12 ~ catholic + coarse_inc + math8, data = df_cem2,
#>     weights = weights)
#>
#> Weighted Residuals:
#>     Min      1Q  Median      3Q     Max
#> -28.504  -3.144  -0.064   3.192  26.186
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  10.26504    0.44843  22.891  < 2e-16 ***
#> catholic      1.50497    0.22886   6.576 5.28e-11 ***
#> coarse_inc.L  2.85163    0.50177   5.683 1.39e-08 ***
#> coarse_inc.Q  0.02882    0.43027   0.067    0.947
#> coarse_inc.C -0.19212    0.47399  -0.405    0.685
#> coarse_inc^4 -0.26505    0.48358  -0.548    0.584
```

Let's look across estimates

	OLS	PSM			CEM	
	(1)	(2)	(3)	(4)	(5)	(6)
Attend catholic school	3.895*** (0.409)	1.612*** (0.318)	1.599*** (0.314)	1.688*** (0.306)	1.561*** (0.228)	1.505*** (0.229)
Observations	5,671	1,118	1,118	1,126	5,671	5,456
R^2	0.016	0.623	0.632	0.651	0.656	0.644
Note:	*p<0.05; **p<0.01; ***p<0.001. Models 2-3 and 5-6 match on income and math score. Model 3 adjusts for higher-order terms and interactions post matching; Model 4 includes them in matching algorithm. Model 6 uses narrower bins to match than Model 5. All CEM and PSM estimates are doubly-robust. Outcome mean (SD) for treated = 54.5 (8.5)					

How would you describe results?

> A naïve estimate of 12th grade math test score outcomes suggests that students who attend Catholic high school scored nearly 4 scale score points higher than those who attended public school (almost half a standard deviation). However, the characteristics of students who attended Catholic high school were substantially different. They had higher family income, scored higher on 8th grade math tests and were more likely to be White, among other distinguishing characteristics. We theorize that the primary driver of Catholic school attendance is student 8th grade performance and family income. Conditional on these two characteristics, we implement two separate matching algorithms: Propensity Score Matching and Coarsened Exact Matching. Both sets of estimates indicate that the benefits of Catholic school are overstated in the full sample, but the attenuated results are still large in magnitude (just under one-fifth of a SD) and statistically significant.

Strengths/limitations of approaches

Approach	Strengths	Limitations
PS nearest neighbor matching w/ calipers and replacement	- Simulates ideal randomized experiment - Limits dimensionality problem - Calipers restrict poor matches - Replacement takes maximal advantage of available data	- May generate poor matches - Model dependent Lacks transparency; PS in aribtrary units - Potential for bias (King & Nielsen, 2019)
PS stratification	- Simulates block-randomized experiment - Limits dimensionality problem	- May produce worse matches than nearest neighbor - Lacks transparency; stratum arbitrary
Inverse probablity (PS) matching	- Retains all original sample data - Corrects bias of estimate with greater precision than matching/stratification	- Non-transparent/a-theoretical
Coarsened Exact Matching	- Matching variables can be pre-specified (and pre-registered) - Matching substantively driven Transparent matching process - Eliminates same bias as propensity score if SOO occurs	- May generate poor matches depending on how coarsened variables are - May lead to disgarding large portions of sample

Synthesis and wrap-up

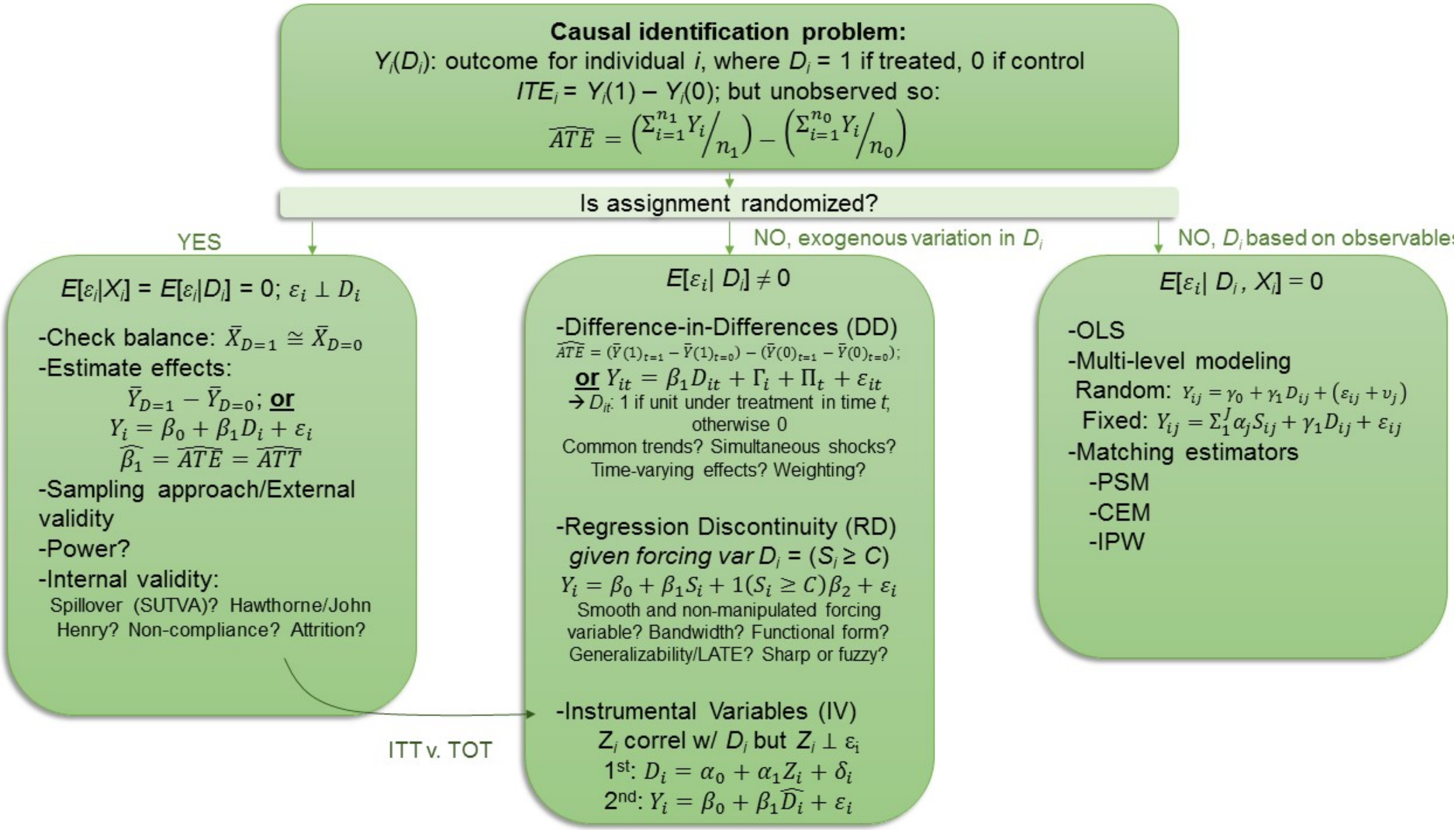
Goals

1. Describe conceptual approach to matching analysis

2. Assess validity of matching approach

3. Conduct matching analysis in simplified data using both propensity-score matching and CEM

Can you explain this figure?



To-Dos

Week 9: Matching and presenting

Readings:

- Umansky & Dumont (2021)

Assignments Due

DARE 4

- Due 9:00am, February 28

Final Research Project

- Presentation, March 8
- Paper, March 17 (submit March 10 for feedback)

Feedback

Plus/Deltas

Front side of index card

Clear/Murky

On back