Matching in practice

PSCI 2301: Quantitative Political Science II

Prof. Brenton Kenkel

brenton.kenkel@gmail.com

Vanderbilt University

February 17, 2025

Recap

1. Confounding variables
    - Affects outcome of interest *and* assignment to treatment
    - Presence of confounding $\rightsquigarrow$ independence condition fails
2. Controlling for confounders
    - Compare observations that are similar except for treatment status
    - Kills selection bias if all confounders are observed
3. The subclassification estimator
    - Divide observations into subgroups based on confounder values
    - Weighted average of within-subgroup differences

Today's agenda

1. Work through philosophy of matching with many confounders
2. See how to implement matching methods in R
3. Briefly discuss Eggers & Hainmueller results

Controlling for many confounders

Recap on subclassification

Can use subclassification when there aren’t many confounders

1. Divide observations into groups based on confounder values
2. Take difference of means within each subgroup:

$$\textbf{avg}[Y_i \mid D_i = 1, X_i = x] - \textbf{avg}[Y_i \mid D_i = 0, X_i = x]$$

3. Estimate ATE by weighted average of within-subgroup differences

Runs into curse of dimensionality with many confounders

- Too few observations per group to accurately estimate differences
- Many groups won’t have both treatment + control observations

Matching

Typical algorithm:

1. For each treatment $(D_i = 1)$ observation, find the control $(D_i = 0)$ observation with the closest confounder values
    - How to define "closest"? Stay tuned!
2. Create a comparison group from the set of matched observations
3. Take average difference in outcome between treatment group and matched controls

ATT versus ATE

Up to now we've focused on estimating the ATE, $\mathbb{E}[Y_{1i} - Y_{0i}]$

→ Difference in potential outcomes for the average population member

Typical matching methods instead estimate the **average treatment effect on the treated**, or ATT:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$$

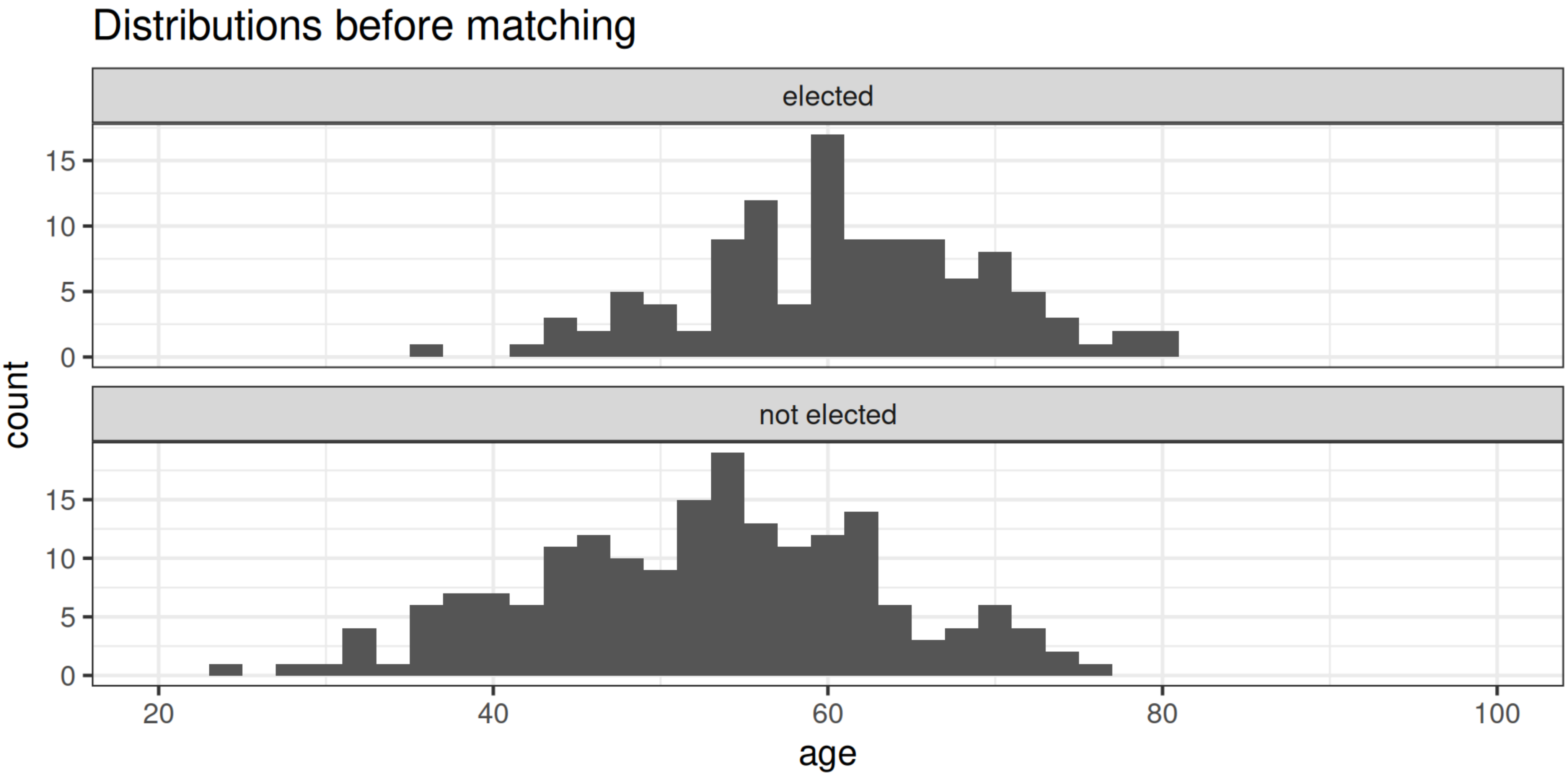
→ Difference in potential outcomes for the average population member who would receive the treatment

Randomly assigned treatment $\rightsquigarrow$ ATE $\approx$ ATT

Self selection $\rightsquigarrow$ ATE $\not\approx$ ATT (except in special situations)

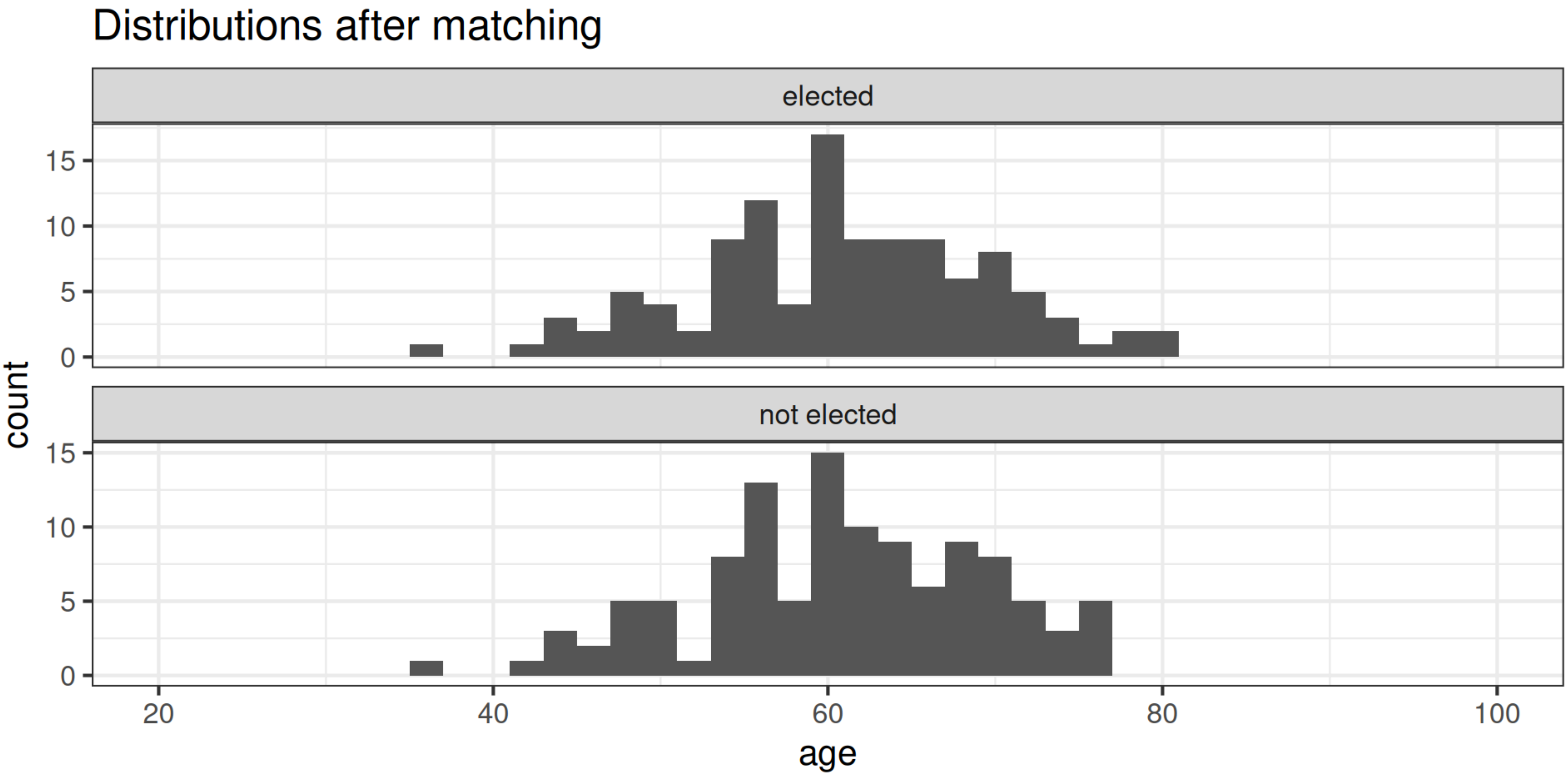
ATT versus ATE

Hypothetical example: Matching on candidate age



ATT versus ATE

Hypothetical example: Matching on candidate age



Measuring closeness

Which control observation is the best match for the treated one?

	Treated	Control 1	Control 2
Female	1	1	0
Years of education	14	12	14
Aristocrat	0	1	0
Year of birth	1928	1946	1931
Year of death	2003	2005	1989

Distance between observations

Could just sum up component-by-component differences

Treated distance to Control 1:

$$|1-1|+|14-12|+|0-1|+|1928-1946|+|2003-2005|=23$$

Treated distance to Control 2:

$$|1-0|+|14-14|+|0-0|+|1928-1931|+|2003-1989|=18$$

Do you see a problem with doing it this way?

The Mahalanobis distance

Variables might be measured on very different scales

Even when units are the same, normal variation might differ

→ 4-year diff in education means more than 4-year diff in birth year

To correct for this, Mahalanobis distance normalizes by standard deviations

ⓘ **The Mahalanobis distance**

When the confounding variables $X_{1i}, \ldots, X_{Ki}$ are uncorrelated with each other, the Mahalanobis distance between two observations is

$$d(X_i, X_j) = \sqrt{\frac{(X_{1i} - X_{1j})^2}{\text{sd}[X_1]^2} + \cdots + \frac{(X_{Ki} - X_{Kj})^2}{\text{sd}[X_K]^2}}.$$

Propensity score

Other most common way to match observations with many confounders

1. Create statistical model of selection into treatment
    - e.g., logistic regression
2. Using model, calculate the propensity scores $\Pr(D_i = 1 \mid X_i)$
3. Match observations with closest propensity scores

Advantage: Easier to find close matches than with Mahalanobis distance

Disadvantage: Everything hinges on having a good propensity model

- Can be especially challenging with many confounders/few observations
- … exactly the circumstances when you most need matching!

Don’t match on post-treatment variables

Whether using subclassification, Mahalanobis distance, or propensity scores…

Never control for post-treatment variables whose value may be affected by treatment assignment

ⓘ Matching on a post-treatment variable: Lung tar

Imagine you want to study the effects of smoking on lung cancer.

For each patient in your study, you have a measure of the amount of tar in their lungs.

Why will your study be less accurate if you control for this?

Matching in R

Why we’re not using the MPs data

Public data just contains the raw bios, not the outcome or confounders

```r
library("archive")
df_eh <-
  archive_read("https://andy.egge.rs/data/THC_candidates.csv.zip",
               file = "THC_candidates.csv") |>
  read_csv()
print(df_eh)
```

```
# A tibble: 11,485 × 8
  election_id date       constituency.name sname      party   votes winner bio
        <dbl> <date>     <chr>             <chr>      <chr>   <dbl>  <dbl> <chr>
1       35779 1950-02-23 Battersea North   Jay        Lab.    24762      1 "Mr....
2       35779 1950-02-23 Battersea North   Maddan     C        9084      0 "Mr....
3       35779 1950-02-23 Battersea North   Handscombe L.       1090      0 "Mr....
4       35779 1950-02-23 Battersea North   Mahon      Comm.     655      0 "Mr....
5       35780 1950-02-23 Battersea South   Ganley     Co-op.... 16142      1 "Mrs...
# i 11,480 more rows
```

Gilligan & Sergenti data

```r
df_gs
```

```
# A tibble: 87 × 11
  id    ethfrac country       intervention ln_peace_duration ln_deaths ln_wardur
  <chr>   <dbl> <chr>                <dbl>             <dbl>     <dbl>     <dbl>
1 41_2     1.36 Haiti                    0              2.40      0            9
2 41_3     1.36 Haiti                    1              4.96      5.52        12
3 52_2    55.8  Trinidad and…             0              5.07      3.40         1
4 70_2    30.5  Mexico                   0              3.40      4.98         1
5 70_3    30.5  Mexico                   0              4.42      0            4
# ℹ 82 more rows
# ℹ 4 more variables: ln_population <dbl>, ln_military <dbl>, ln_gdppc <dbl>,
#   polity <dbl>
```

Mahalanobis distance matching

```r
library("MatchIt")

match_gs_md <- matchit(
    intervention ~ ethfrac + ln_deaths + ln_wardur
        ln_military + ln_gdppc + polity,
    data = df_gs,
    method = "nearest",
    distance = "mahalanobis",
    ratio = 1,
    estimand = "ATT"
)
summary(match_gs_md)
```

```
Call:
matchit(formula = intervention ~ ethfrac +
ln_deaths + ln_wardur +
    ln_population + ln_military + ln_gdppc +
polity, data = df_gs,
    method = "nearest", distance = "mahalanobis",
estimand = "ATT",
    ratio = 1)

Summary of Balance for All Data:
              Means Treated Means Control
ethfrac             49.2130       56.5038
ln_deaths            8.9815        6.6473
ln_wardur           80.5263       50.2794
ln_population        8.7539        9.5094
```

Propensity score step 1: Model treatment assignment

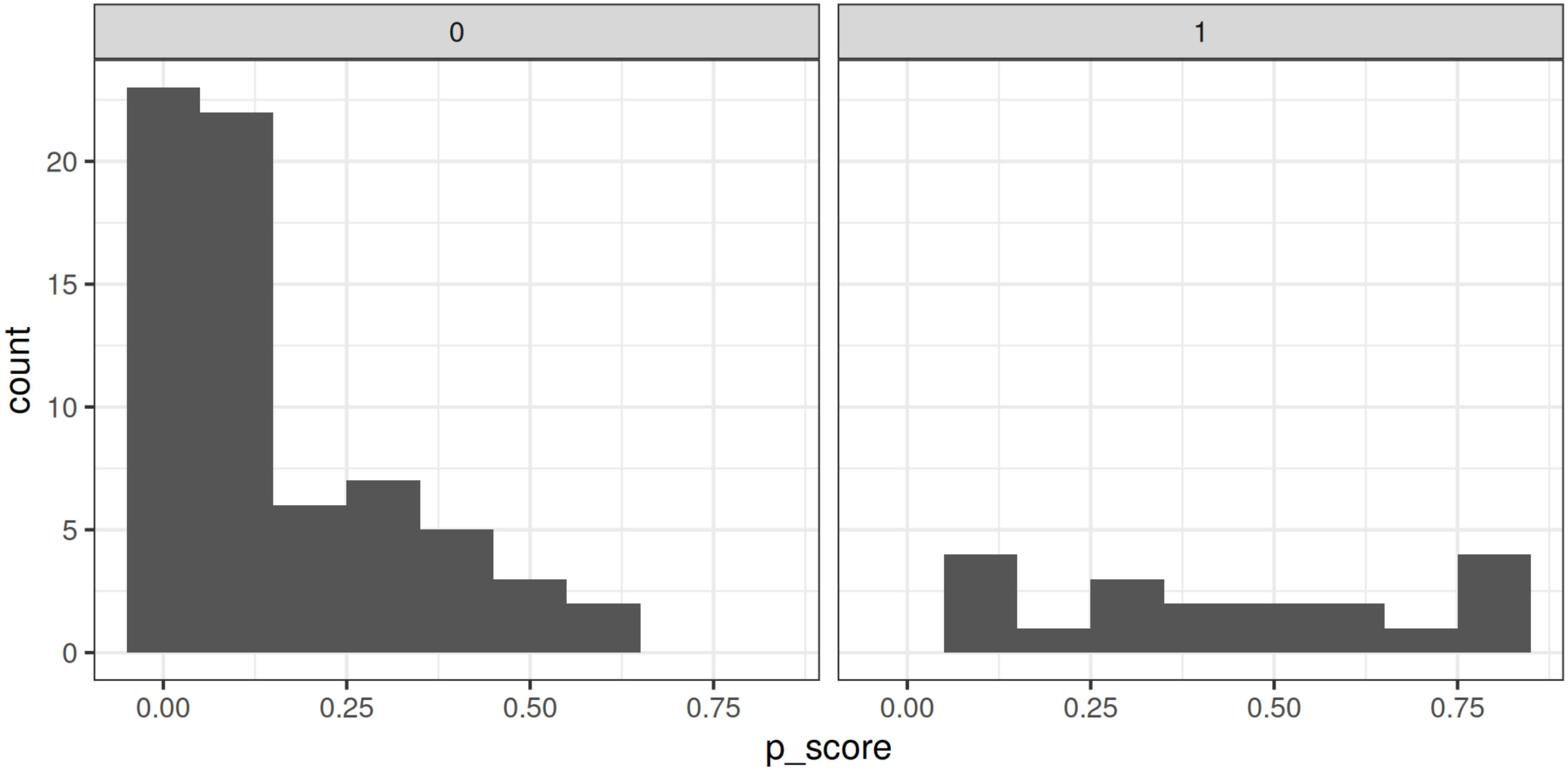
```r
library("broom")
fit_gs_prop <- glm(
  intervention ~ ethfrac + ln_deaths + ln_wardur + ln_population +
    ln_military + ln_gdppc + polity,
  data = df_gs,
  family = binomial()
)
tidy(fit_gs_prop)
```

```
# A tibble: 8 × 5
  term          estimate std.error statistic p.value
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)   -5.33      5.46       -0.977 0.329
2 ethfrac       -0.0108    0.0118     -0.912 0.362
3 ln_deaths      0.629     0.210       3.00  0.00271
4 ln_wardur     -0.00615   0.00466    -1.32  0.186
5 ln_population -0.142     0.473      -0.300 0.764
6 ln_military   -0.852     0.508      -1.68  0.0932
7 ln_gdppc       0.627     0.423       1.48  0.138
8 polity        -0.0912    0.0752     -1.21  0.225
```

Propensity score step 2: Extract propensity scores

```r
df_gs$p_score <- predict(fit_gs_prop, type = "response")

ggplot(df_gs, aes(x = p_score)) + geom_histogram(binwidth = 0.1) + facet_wrap(~ intervention)
```



Propensity score step 3: Matching

```r
match_gs_ps <- matchit(
    intervention ~ ethfrac + ln_deaths + ln_wardur
        ln_military + ln_gdppc + polity,
    data = df_gs,
    method = "nearest",
    distance = df_gs$p_score,
    ratio = 1,
    estimand = "ATT"
)
summary(match_gs_ps)
```

```

Call:
matchit(formula = intervention ~ ethfrac +
ln_deaths + ln_wardur +
    ln_population + ln_military + ln_gdppc +
polity, data = df_gs,
    method = "nearest", distance = df_gs$p_score,
estimand = "ATT",
    ratio = 1)

Summary of Balance for All Data:
              Means Treated Means Control
distance             0.4375        0.1572
ethfrac             49.2130       56.5038
ln_deaths            8.9815        6.6473
ln_wardur           80.5263       50.2794
```

Estimating the ATT from matched samples

```r
fit_unmatched <- lm(
  ln_peace_duration ~ intervention,
  data = df_gs
)
tidy(fit_unmatched)
```

```
# A tibble: 2 × 5
  term   estimate std.error statistic  p.value
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>
1 (Int…     3.36      0.136     24.8  1.27e-40
2 inte…     0.872     0.290      3.00 3.52e- 3
```

```r
fit_md <- lm(
  ln_peace_duration ~ intervention,
  data = match.data(match_gs_md)
)
tidy(fit_md)
```

```
# A tibble: 2 × 5
  term   estimate std.error statistic  p.value
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>
1 (Int…     3.51      0.225     15.6  1.33e-17
2 inte…     0.723     0.318      2.28 2.89e- 2
```

```r
fit_ps <- lm(
  ln_peace_duration ~ intervention,
  data = match.data(match_gs_ps)
)
tidy(fit_ps)
```

```
# A tibble: 2 × 5
  term   estimate std.error statistic  p.value
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>
1 (Int…     3.31      0.259     12.8  6.16e-15
2 inte…     0.926     0.366      2.53 1.60e- 2
```

“MPs for Sale?”: The results

Eggers & Hainmueller research design

Population: British candidates for Parliament elected 1950–1970

Outcome: Total wealth at death

Treatment: Being elected to Parliament

Comparison: Not being elected to Parliament

Controls: Age, gender, aristocrat status, educational history, career history

→ They match on these variables to estimate treatment effects

Eggers & Hainmueller results

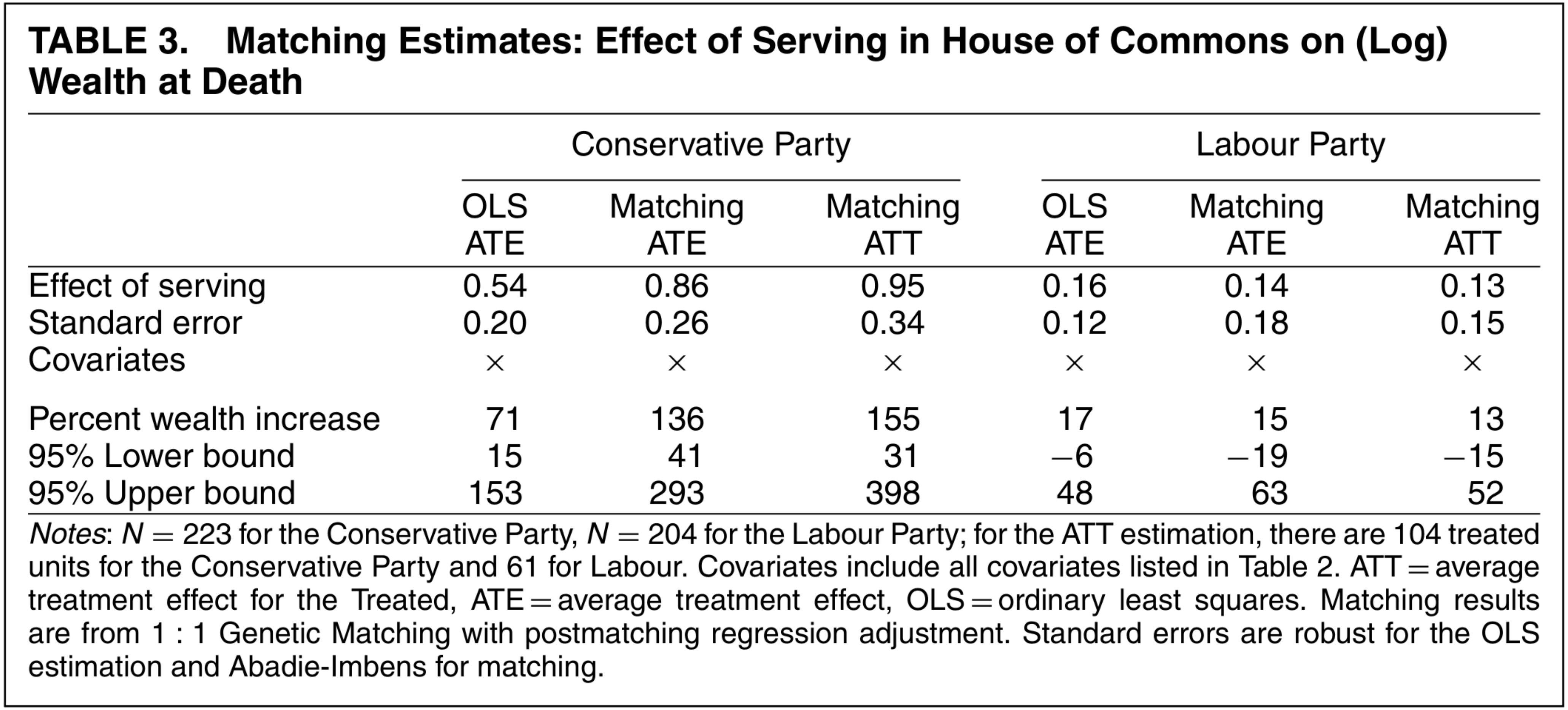


TABLE 3. Matching Estimates: Effect of Serving in House of Commons on (Log) Wealth at Death

	Conservative Party			Labour Party		
	OLS ATE	Matching ATE	Matching ATT	OLS ATE	Matching ATE	Matching ATT
Effect of serving	0.54	0.86	0.95	0.16	0.14	0.13
Standard error	0.20	0.26	0.34	0.12	0.18	0.15
Covariates	×	×	×	×	×	×
Percent wealth increase	71	136	155	17	15	13
95% Lower bound	15	41	31	−6	−19	−15
95% Upper bound	153	293	398	48	63	52

Notes: $N = 223$ for the Conservative Party, $N = 204$ for the Labour Party; for the ATT estimation, there are 104 treated units for the Conservative Party and 61 for Labour. Covariates include all covariates listed in Table 2. ATT = average treatment effect for the Treated, ATE = average treatment effect, OLS = ordinary least squares. Matching results are from 1 : 1 Genetic Matching with postmatching regression adjustment. Standard errors are robust for the OLS estimation and Abadie-Imbens for matching.

Concerns about the matching strategy

Controls: Age, gender, aristocrat status, educational history, career history

These probably don't fully capture all sources of confounding bias

---

E&H follow-up analysis: Regression discontinuity design

Reduce unobserved confounding by comparing close winners to close losers

Key assumption: In close elections, who wins is close to random

Wrapping up

What we did today

1. Matching methods with many confounders
    - Mahalanobis distance — variance-adjusted differences
    - Propensity score matching — match on likelihood of being treated
    - Typically obtain ATT instead of ATE
    - Don’t control for post-treatment variables
2. Implementation with `MatchIt` in R
3. Eggers & Hainmueller results
    - Officeholding appears lucrative, especially for Tories
    - …but lingering worries about unobserved confounding

Next time

Regression for treatment effect estimation with observed confounders

1. Read Bartels research paper, “Beyond the Running Tally”
2. Read *Mastering ’Metrics*, chapter 2, pages 56–81
3. Remember that Problem Set 3 is due Friday
4. Project proposals due at end of the month — find data!