ECON42720 Causal Inference and Policy Evaluation

3 Matching and Re-weighting

Ben Elsner (UCD)

About this Lecture

Resources

As an **introduction**, I recommend Chapter 5 in Scott Cunningham's Mixtape

Slightly **more detailed coverage** can be found in

- Huntington-Klein's The Effect, Chapter 14
- Huber's Causal Analysis, Chapter 4

Many examples in this chapter, in particular the R codes, have been taken from The Effect or inspired by it.

Starting Point: Conditional Independence

$$(Y^1, Y^0) \perp\!\!\!\perp D \mid X$$

For **causal identification**, we require the assumption that the **treatment** D is as good as **randomly assigned conditional on the covariates** X

Formally, this means that the potential outcomes are **conditionally independent** of the treatment assignment given the covariates

$$E\left[Y^1 \mid D = 1, X\right] = E\left[Y^1 \mid D = 0, X\right]$$
$$E\left[Y^0 \mid D = 1, X\right] = E\left[Y^0 \mid D = 0, X\right]$$

Conditional Independence and Selection on Observables

If CIA holds, we speak of **selection on observables**

- **Independence does not hold** in general
- But it holds in the **subpopulations** defined by the covariates X

The **groups defined by X** (think age, gender, neighbourhood, etc) determine the **treatment assignment**

- But **within each group**, who gets treated is **as good as random**

This is a **strong assumption!**

Example: Smoking and Lung Cancer

Does smoking cause lung cancer?

- Today we would say “yes, of course”
- But answering this question was far from clear in the 1950s
- There is a **strong correlation** between smoking and lung cancer, but is it causal?

(Potential) problem: confounders

- There could be genetic determinants of smoking and lung cancer
- There could be environmental factors that cause both smoking and lung cancer

We don’t have **experimental evidence**

Example: Death Rates per 1,000

The following example from Cochran (1968) will illustrate what **selection on observables** and do for us

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

In all countries, the **highest death rates are for cigar and pipe smokers**

- Does this mean that smoking pipes and cigars is more dangerous than smoking cigarettes?

Smoking and Lung Cancer: Independence?

The **independence assumption** would imply that

$$E[Y^1 \mid \text{Cigarette}] = E[Y^1 \mid \text{Pipe}] = E[Y^1 \mid \text{Cigar}]$$
$$E[Y^0 \mid \text{Cigarette}] = E[Y^0 \mid \text{Pipe}] = E[Y^0 \mid \text{Cigar}]$$

Suppose that the **independence assumption** holds

- This would/should also mean that observable characteristics X are similar between the groups
- I.e. the **covariates should be balanced** between groups

Are cigarette smokers similar to pipe and cigar smokers?

Let's ask Dall-E: show me a picture of a cigarette smoker and a cigar smoker

Age as a Confounder?

Smoking group	Canada	UK	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Clearly, **age affects what people smoke and also their death rates**

- Independence is violated: the **distribution of age** is different between the groups
- There may be other confounders, but let's focus on age for now

We have **covariate imbalance!**. Potential remedy: condition on age (**subclassification**)

Subclassification: Divide Age into Strata

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

The **death rate of cigarette smokers in the population** is:

$$20 \times \frac{65}{100} + 40 \times \frac{25}{100} + 60 \times \frac{10}{100} = 29$$

But: the **age distribution is (heavily) imbalanced** between the groups

Re-weighting: Age-Adjusted Death Rates

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

The **age-adjusted death rate of cigarette smokers** is:

$$20 \times \frac{10}{100} + 40 \times \frac{25}{100} + 60 \times \frac{65}{100} = 51$$

Age-Adjusted Death Rates

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	29.5	14.8	21.2
Cigars/pipes	19.8	11.0	13.7

Here we **achieved balance on one covariate: age**

- The **age-adjusted death rates** are now more similar between the groups
- But there may be an **imbalance on other covariates** (SES, income, health, etc)

We need to **use a DAG** to identify **all confounders** and adjust for them

Identifying Assumptions

In presence of confounders X, we can **identify a causal effect under two assumptions**

1. **Conditional Independence**: $Y^0, Y^1 \perp D \mid X$
2. **Common Support**: $0 < P(D = 1 \mid X) < 1$ with probability one

Common support: for each stratum, we need some units that are treated and others that are control units

- We need **common support** to calculate the **weights for the adjustment**

Causal Identification with Selection on Observables

Under **conditional independence and common support**, the following holds:

$$\begin{aligned}
E\big[Y^1 - Y^0 \mid X\big] &= E\big[Y^1 - Y^0 \mid X, D = 1\big] \\
&= E\big[Y^1 \mid X, D = 1\big] - E\big[Y^0 \mid X, D = 0\big] \\
&= E\big[Y \mid X, D = 1\big] - E\big[Y \mid X, D = 0\big]
\end{aligned}$$

The **estimator for the ATE** is as follows:

$$\widehat{\delta_{ATE}} = \int \Big(E\big[Y \mid X, D = 1\big] - E\big[Y \mid X, D = 0\big]\Big) d\,\Pr(X)$$

The Limits of Subclassification: The Curse of Dimensionality

In the example of smoking and death rates, we adjusted for just one confounder

- The hope was that, by slicing up age into three groups, achieve balance in treated and control groups
- We did achieve balance on age, but what about other confounders?
- Also, are three age groups enough or do we need more?

In practice, we have the **problem of a finite sample size**

- There are **limits to how many strata we can create**
- We cannot have an infinite number of groups defined by one variable (such as age)
- We cannot have an infinite number of variables to adjust for

This problem is known as the **curse of dimensionality**

References

Broockman, David E. 2013. Black Politicians Are More Intrinsically Motivated to Advance Blacks' Interests: A Field Experiment Manipulating Political Incentives. *American Journal of Political Science*, **57**(3), 521–536.

Cochran, W. G. 1968. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, **24**(2).

APPENDIX

Example: Using Broockman (2013) for R examples

Research question: are black politicians more likely to help black citizens even if the incentives are low?

Methodology: audit study; sent emails to U.S. state legislators; asking them to help them sign up for unemployment benefits

Experimental variation:

- Sender with black vs. white name
- Sender lives in same district as legislator or far away

Matching: white and black legislators with similar characteristics

Broockman (2013) data preparation

We use the excellent `Matching` package in R. A great alternative is `MatchIt`

```r
library(Matching)
library(causaldata)
library(tidyverse)

br <- causaldata::black_politicians

# Outcome
Y <- br %>%
    pull(responded)
# Treatment
D <- br %>%
    pull(leg_black)
# Matching variables
# Note select() is also in the Matching package, so we specify dplyr
X <- br %>%
    dplyr::select(medianhhincom, blackpercent, leg_democrat) %>%
    as.matrix()
```

Mahalanobis distance matching in R

```r
# Set weight=2 for Mahalanobis distance
M <- Match(Y, D, X, Weight = 2, caliper = 1)

# See treatment effect estimate
summary(M)
```

```
##
## Estimate...  -0.0073462
## AI SE......  0.072683
## T-stat.....  -0.10107
## p.val......  0.91949
##
## Original number of observations..............  5593
## Original number of treated obs...............  364
## Matched number of observations...............  363
## Matched number of observations  (unweighted).  405
##
## Caliper (SDs).......................................   1 1 1
## Number of obs dropped by ’exact’ or ’caliper’  1
```

Mahalanobis distance matching in R

Previous slide: the estimate -0.007346 means that black legislators were 0.7 percentage points less likely to respond to emails

This effect is not statistically significant

Mahalanobis distance matching in R

```
# Get matched data for use elsewhere. Note that this approach will
# duplicate each observation for each time it was matched
matched_treated <- tibble(id = M$index.treated,
                          weight = M$weights)
matched_control <- tibble(id = M$index.control,
                          weight = M$weights)
matched_sets <- bind_rows(matched_treated,
                          matched_control)
# Simplify to one row per observation
matched_sets <- matched_sets %>%
                    group_by(id) %>%
                    summarize(weight = sum(weight))
# And bring back to data
matched_br <- br %>%
    mutate(id = row_number()) %>%
    left_join(matched_sets, by = 'id')
```

Mahalanobis distance matching in R

```r
# OLS estimation based on matched sample
lm(responded~leg_black, data = matched_br, weights = weight)
```

```
##
## Call:
## lm(formula = responded ~ leg_black, data = matched_br, weights = weight)
##
## Coefficients:
## (Intercept)    leg_black
##    0.398531    -0.007346
```

We can see that the estimate is the same as with matching

Coarsened Exact Matching in R

Broockman performs CEM to make black and white legislators more comparable

We use the `cem` package here. Alternatively we could use the `method_cem` command of the `MatchIt` package.

```r
library(cem); library(tidyverse)
br <- causaldata::black_politicians
```

Coarsened Exact Matching in R

```r
# Limit to just the relevant variables and omit missings
brcem <- br %>%
    dplyr::select(responded, leg_black, medianhhincom,
    blackpercent, leg_democrat) %>%
    na.omit() %>%
    as.data.frame() # Must be a data.frame, not a tibble

# Two ways to create breaks. Use quantiles to create quantile cuts or manually for evenly
# although you MUST do it yourself for binary variables). Be sure
# to include the "edges" (max and min values).
inc_bins <- quantile(brcem$medianhhincom, (0:6)/6)

create_even_breaks <- function(x, n) {
    minx <- min(x)
    maxx <- max(x)

    return(minx + ((0:n)/n)*(maxx-minx))
}
```

Coarsened Exact Matching in R

```r
bp_bins <- create_even_breaks(brcem$blackpercent, 6)

# For binary, we specifically need two even bins
ld_bins <- create_even_breaks(brcem$leg_democrat,2)

# Make a list of bins
allbreaks <- list('medianhhincom' = inc_bins,
                  'blackpercent' = bp_bins,
                  'leg_democrat' = ld_bins)
```

Coarsened Exact Matching in R

```
# Match, being sure not to match on the outcome
# Note the baseline.group is the *treated* group
c <- cem(treatment = 'leg_black', data = brcem,
         baseline.group =  '1',
         drop = 'responded',
         cutpoints = allbreaks,
         keep.all = TRUE)
```

```
##
## Using 'leg_black'='1' as baseline group
```

```
# Get weights for other purposes
brcem <- brcem %>%
    mutate(cem_weight = c$w)
```

Coarsened Exact Matching in R

```r
# OLS estimation with weighted dataset
lm(responded~leg_black, data = brcem, weights = cem_weight)
```

```
##
## Call:
## lm(formula = responded ~ leg_black, data = brcem, weights = cem_weight)
##
## Coefficients:
## (Intercept)    leg_black
##     0.34680      0.02302
```

Coarsened Exact Matching in R

```r
# Use the inbuilt ATT estimation command from cem
att(c, responded ~ leg_black, data = brcem)
```

```
##
##             G0  G1
## All       5229 364
## Matched   4491 338
## Unmatched  738  26
##
## Linear regression model on CEM matched data:
##
## SATT point estimate: 0.023020 (p.value=0.391783)
## 95% conf. interval: [-0.029659, 0.075699]
```

PSM in R

To perform PSM, we can use the `MatchIt` package. Here we estimate the propensity score for the LaLonde data

```
library("MatchIt")
library('marginaleffects')
data("lalonde")

# 1:1 NN PS matching w/o replacement
m.out1 <- matchit(treat ~ age + educ + race + married +
                   nodegree + re74 + re75, data = lalonde,
                 method = "nearest", distance = "glm")
```

PSM in R

Checking balance after nearest neighbor matching

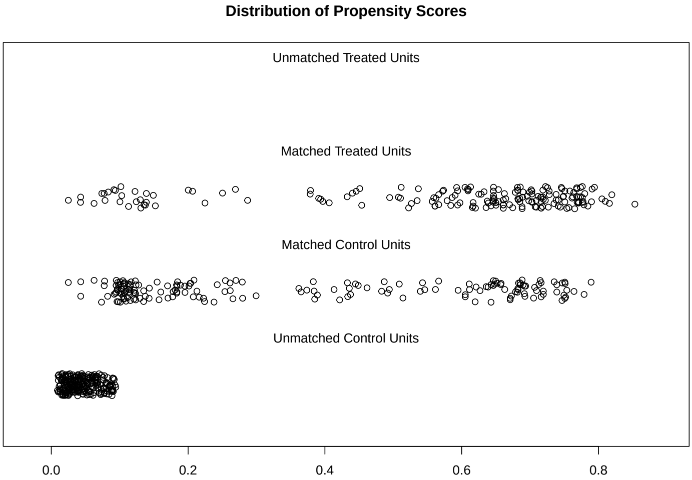
```r
summary(m.out1, un = FALSE)
```

```
##
## Call:
## matchit(formula = treat ~ age + educ + race + married + nodegree + 
##     re74 + re75, data = lalonde, method = "nearest", distance = "glm")
##
## Summary of Balance for Matched Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance          0.5774        0.3629          0.9739     0.7566    0.1321
## age              25.8162       25.3027          0.0718     0.4568    0.0847
## educ             10.3459       10.6054         -0.1290     0.5721    0.0239
## raceblack         0.8432        0.4703          1.0259          .    0.3730
## racehispan        0.0595        0.2162         -0.6629          .    0.1568
## racewhite         0.0973        0.3135         -0.7296          .    0.2162
## married           0.1892        0.2108         -0.0552          .    0.0216
## nodegree          0.7081        0.6378          0.1546          .    0.0703
## re74           2095.5737     2342.1076         -0.0505     1.3289    0.0469
## re75           1532.0553     1614.7451         -0.0257     1.4956    0.0452
##            eCDF Max Std. Pair Dist.
## distance     0.4216          0.9740
## age          0.2541          1.3938
```

PSM in R

We can also plot the distribution of propensity scores

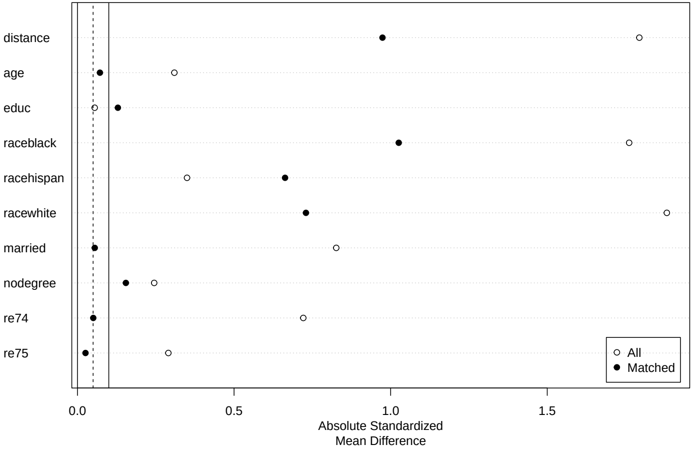
```
plot(m.out1, type = "jitter", interactive = FALSE)
```



PSM in R

Or how about this one. . .

```
plot(summary(m.out1))
```



PSM in R

```r
# Generate matched dataset
m.data <- match.data(m.out1)
# Run a regression on the matched dataset
fit <- lm(re78 ~ treat + age + educ + race + married + nodegree +
             re74 + re75, data = m.data, weights = weights)
summary(fit)
```

```
##
## Call:
## lm(formula = re78 ~ treat + age + educ + race + married + nodegree +
##     re74 + re75, data = m.data, weights = weights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8891  -5063  -1703   3422  53495
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.582e+03  3.296e+03  -0.783  0.43394
## treat        1.345e+03  7.898e+02   1.703  0.08945 .
## age          7.804e+00  4.292e+01   0.182  0.85581
```

PSM in R

```r
# Can also compute the ATT based on the interactions of the treatment
fit <- lm(re78 ~ treat * (age + educ + race + married + nodegree +
             re74 + re75), data = m.data, weights = weights)

avg_comparisons(fit,
                variables = "treat",
                vcov = ~subclass,
                newdata = subset(m.data, treat == 1),
                wts = "weights")
```

```
##
##   Term Contrast Estimate Std. Error    z Pr(>|z|)   S 2.5 % 97.5 %
##  treat    1 - 0     1121        752 1.49    0.136 2.9  -354   2596
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, co
```

PSM in R

Another option based on the `Matching` package; PSM done directly here

```r
library("Matching")
attach (lalonde)
D <- treat
Y <- re78 # define outcome
X <- cbind( age , educ , nodegree , married , re74 , re75)
ps<- glm(D ~ X, family=binomial)$fitted
psmatching <- Match(Y=Y, Tr=D, X=ps , BiasAdjust = TRUE)
```

PSM in R

Another option based on the `Matching` package; PSM done directly here

```r
summary(psmatching)
```

```
##
## Estimate...  597.88
## AI SE......  913.53
## T-stat.....  0.65447
## p.val......  0.51281
##
## Original number of observations..............  614
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  289
```

benjamin.elsner@ucd.ie

www.benjaminelsner.com

Sign up for office hours

YouTube Channel

@ben_elsner

LinkedIn

Contact

Prof. Benjamin Elsner
University College Dublin
School of Economics
Newman Building, Office G206
benjamin.elsner@ucd.ie

Office hours: book on Calendly