Quant II
Lab 3: Conditioning on observables

Giacomo Lemoli

February 14, 2022

Today's plan

- Matching and weighting
- Sensitivity analysis

Conditioning on observables: Recap

- We started with the goal standard: a randomized experiment. There the treatment assignment is independent of any pre-existing characteristic by construction
- Then we have moved to observational settings, where we need assumptions to identify causal effects
- The assumption we have been working with is that the treatment assignment is independent of the potential outcomes once we account for/condition on a set of observable characteristics
- Intuition: if the same characteristics give us an equal probability of being treated, then differences in treatment status between us are just the outcome of a random draw
  - Think of drawing the treatment status from a common Bernoulli distribution with a unique probability parameter q
- In empirical applications, this is also informally called an assumption of "selection on observables". It is usually a strong assumption

CIA and causal effects

Recall the CIA assumption.

$$D_i \perp (Y_{1i}, Y_{0i})|X_i, 0 < P(D_i = 1|X_i) < 1$$

This assumption guarantees that group/covariate stratum-specific ATEs are identified in the data:

$$\tau(x) = E[Y_{1i} - Y_{0i}|X_i = x] = E[Y_{1i}|D_i = 1, X_i = x] - E[Y_{0i}|D_i = 0, X_i = x]$$

And then τ_{PATE} is identified by averaging over the covariates distribution (cf. Imbens and Wooldridge 2009, p.26-27)

$$\tau_{PATE} = E[\tau(x)] = \int_X \tau(x) dF(x)$$

Same holds for τ_{PATT}, changing the distribution to be the one of the population we care about: $F(x|D_i = 1)$

Conditioning on observables in empirical research

When can CIA be plausibly invoked?

Geographic factors influence treatment assignment. E.g. much of the comparative development literature.

Conditioning on observables in empirical research

When can CIA be plausibly invoked?

Geographic factors influence treatment assignment. E.g. much of the comparative development literature.

ON THE ORIGINS OF GENDER ROLES: WOMEN AND THE PLOUGH*

ALBERTO ALESINA
PAOLA GIULIANO
NATHAN NUNN

The study examines the historical origins of existing cross-cultural differences in beliefs and values regarding the appropriate role of women in society. We test the hypothesis that traditional agricultural practices influenced the historical gender division of labor and the evolution of gender norms. We find that, consistent with existing hypotheses, the descendants of societies that traditionally practiced plough agriculture today have less equal gender norms, measured using reported gender-role attitudes and female participation in the workplace, politics, and entrepreneurial activities. Our results hold looking across countries, across districts within countries, and across ethnicities within districts. To test for the importance of cultural persistence, we examine the children of immigrants living in Europe and the United States. We find that even among these individuals, all born and raised in the same country, those with a heritage of traditional plough use exhibit less equal beliefs about gender roles today. *JEL* Codes: D03, J16, N30.

Conditioning on observables in empirical research

Treatment assignment function is observed (but cannot be deterministic. Why?). E.g. media effects literature.

Treatment assignment function is observed (but cannot be deterministic. Why?). E.g. media effects literature.

Panel A. Predicted signal strength

Panel B. Signal strength in the free space

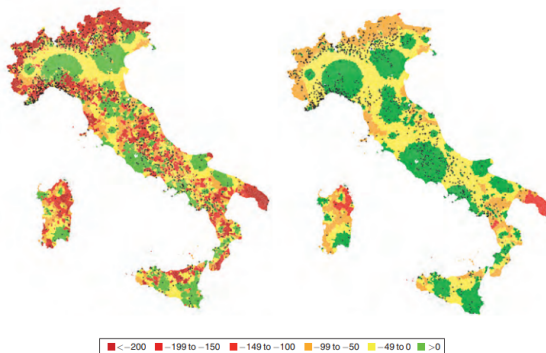<small>■ < −200 ■ −199 to −150 ■ −149 to −100 ■ −99 to −50 ■ −49 to 0 ■ >0</small>

FIGURE 4. MEDIASET SIGNAL STRENGTH AND HYPOTHETICAL SIGNAL STRENGTH IN THE FREE SPACE IN 1985

Notes: Panel A shows Mediaset actual signal strength, in decibels (dB), across municipalities in 1985. Panel B shows the hypothetical signal strength in the absence of geomorphological obstacles. The black dots represent the location of transmitters.

Conditionining on observables in empirical research

- CIA is a strong assumption and needs to be justified on substantive grounds
- Usually authors are required to reduce the number of units in the conditioning set (e.g. by adding fixed effects for smaller groups or creating artificial groups)
- This may be demanding of the data in terms of uncertainty, so there may be trade-offs in the assumptions that are required
- It is common practice to check for CIA plausibility by assessing covariate balance

Regression and CIA

- Linear regression is the most obvious tool to estimate causal effects under CIA
- Estimates linear approximations of the conditional means of the potential outcomes
- Problematic approximation if the treated and control units are very different in terms of their covariates
- We are using the linear relationship estimated in populated covariate cells to extrapolate outcomes in unpopulated covariate cells

Suppose we have

$$E[Y_{0i}|X_i] = \alpha_0 + \beta_0(X_i - \bar{X})$$
$$E[Y_{1i}|X_i] = \alpha_1 + \beta_1(X_i - \bar{X})$$

We can implement the estimation of τ_{PATE} by running an OLS regression on D_i, $(X_i - \bar{X})$ and $D_i * (X_i - \bar{X})$. Then the coefficient of D_i is an estimate of the PATE:

$$\hat{\tau}_{reg} = \hat{\alpha}_1 - \hat{\alpha}_0$$

Extrapolation

To see where problems with covariates distribution kick in, let's use the decomposition used by Imbens and Rubin (Ch. 12) and Imbens and Woldridge (2009):

Extrapolation

To see where problems with covariates distribution kick in, let's use the decomposition used by Imbens and Rubin (Ch. 12) and Imbens and Woldridge (2009):

$$\hat{\tau}_{reg} = \underbrace{\bar{Y}_1 - \bar{Y}_0}_{\text{Difference in means}} - \underbrace{(\frac{N_0}{N_1 + N_0}\hat{\beta}_1 + \frac{N_1}{N_1 + N_0}\hat{\beta}_0)(\bar{X}_1 - \bar{X}_0)}_{\text{Regression adjustment}}$$

Extrapolation

To see where problems with covariates distribution kick in, let's use the decomposition used by Imbens and Rubin (Ch. 12) and Imbens and Woldridge (2009):

$$\hat{\tau}_{reg} = \underbrace{\bar{Y}_1 - \bar{Y}_0}_{\text{Difference in means}} - \underbrace{(\frac{N_0}{N_1 + N_0}\hat{\beta}_1 + \frac{N_1}{N_1 + N_0}\hat{\beta}_0)(\bar{X}_1 - \bar{X}_0)}_{\text{Regression adjustment}}$$

Another way to see it:

$$\hat{\tau}_{reg} = \frac{N_1}{N_1 + N_0}\hat{\tau}_1 + \frac{N_0}{N_1 + N_0}\hat{\tau}_0$$

$$\hat{\tau}_1 = \bar{Y}_1 - \bar{Y}_0 - (\bar{X}_1 - \bar{X}_0)\hat{\beta}_0$$

$$\hat{\tau}_0 = \bar{Y}_1 - \bar{Y}_0 - (\bar{X}_1 - \bar{X}_0)\hat{\beta}_1$$

Extrapolation

To see where problems with covariates distribution kick in, let's use the decomposition used by Imbens and Rubin (Ch. 12) and Imbens and Woldridge (2009):

$$\hat{\tau}_{reg} = \underbrace{\bar{Y}_1 - \bar{Y}_0}_{\text{Difference in means}} - \underbrace{(\frac{N_0}{N_1 + N_0}\hat{\beta}_1 + \frac{N_1}{N_1 + N_0}\hat{\beta}_0)(\bar{X}_1 - \bar{X}_0)}_{\text{Regression adjustment}}$$

Another way to see it:

$$\hat{\tau}_{reg} = \frac{N_1}{N_1 + N_0}\hat{\tau}_1 + \frac{N_0}{N_1 + N_0}\hat{\tau}_0$$
$$\hat{\tau}_1 = \bar{Y}_1 - \bar{Y}_0 - (\bar{X}_1 - \bar{X}_0)\hat{\beta}_0$$
$$\hat{\tau}_0 = \bar{Y}_1 - \bar{Y}_0 - (\bar{X}_1 - \bar{X}_0)\hat{\beta}_1$$

Predict (unobserved) counterfactuals with the coefficients estimated on the other (observed) group. Extrapolation is needed if $\bar{X}_1 \neq \bar{X}_0$.

Solutions

- Achieve covariate balance by trimming extreme values is an option
- But it turns out that regression is even messier than that under effect heterogeneity (cf. Słoczińsky 2022)
- We can use alternative estimators that don't suffer from the same problems

Conditioning on the propensity score

- Let's first introduce the concept of **balancing score**: a function of the covariates that is a sufficient statistic for the covariates values in the treatment distribution
    - A statistic $T(X)$ is sufficient for a parameter θ if $f(X|T(X), \theta) = f(X|T(X))$ (e.g. Mukhopadhyay 2000)

$$D_i \perp X_i | b(X_i)$$

- This is a definition: there are several possible balancing scores that make the treatment orthogonal to covariates. The covariates themselves are a balancing score
- What's appealing about the balancing score: if treatment is independent of the POs conditional on X_i, it is also independent conditional on $b(X_i)$ (proof in Imbens and Rubin, p.267)
- We are interested in scores that reduce the dimensionality of the conditioning set X_i

- Now define the **propensity score**

$$e(X_i) = P(D_i = 1 | X_i) = E(D_i | X_i)$$

- The propensity score is a balancing score (Imbens and Rubin, p.266)

$$D_i \perp X_i | e(X_i)$$

Weighting units by their propensity score is theoretically appealing, because it gives an unbiased estimate for the PATE under CIA.

Weighting

Weighting units by their propensity score is theoretically appealing, because it gives an unbiased estimate for the PATE under CIA.

$$E\left[\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right] = E\left[\frac{D_i Y_{1i}}{e(X_i)} - \frac{(1-D_i)Y_{0i}}{1-e(X_i)}\right] =$$

$$= E\left[E\left[\frac{D_i Y_{1i}}{e(X_i)} - \frac{(1-D_i)Y_{0i}}{1-e(X_i)}\right]|X_i\right] = E\left[\frac{E[D_i|X_i]E[Y_{1i}|X_i]}{e(X_i)} - \frac{E[1-D_i|X_i]E[Y_{0i}|X_i]}{1-e(X_i)}\right] =$$

$$= E\left[\frac{e(X_i)E[Y_{1i}|X_i]}{e(X_i)} - \frac{(1-e(X_i))E[Y_{0i}|X_i]}{1-e(X_i)}\right] =$$

$$= E[E[Y_{1i} - Y_{0i}|X_i]] = E[Y_{1i} - Y_{0i}] = \tau_{PATE}$$

Weighting

A natural weighting estimator is the IPW estimator.

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

We usually use a "normalized" version of it.

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

In any case, relying on the PS shifts modeling issues from estimating $E[Y_i|X_i]$ to estimating $e(X_i)$.

Matching

- Non-parametric methods for causal effects under "selection on observables"
- Approximates an experiment with block-randomization
- Let's focus on the ATT as an estimand, because it requires weaker assumptions than CIA to be identified

Matching

- Non-parametric methods for causal effects under "selection on observables"
- Approximates an experiment with block-randomization
- Let's focus on the ATT as an estimand, because it requires weaker assumptions than CIA to be identified
- CMI: $E[Y_{0i}|D_i = 1, X_i] = E[Y_{0i}|D_i = 0, X_i], P(D_i = 1|X_i) < 1$
    - Weaker overlap conditions as we just need to care about counterfactuals for treated units and not for control units

- The general intuition is a process like the following:

1. For each treated unit, find control units with $\sim$ values of X_i
2. Compute the difference in means within these strata and compute a weighted average using the distribution of X of the treated group

- In practice, this is done by targeting balance wrt the covariate distribution of the treated group

Matching

- Several matching algorithms available
- Select close units by minimizing some function of the covariates
- Generally create a matched sample of units, on which one can estimate ATT/ATE
- Researchers can try different matching algorithms until an acceptable level of balance is achieved
- Useful diagnostics: Kolmogorov-Smirnov tests for equality of distributions, quantile-quantile plots

Matching to improve balance

Let's see how to improve balance with matching using data from Ruggeri, Dorussen, and Gizelis (2016)

```r
# Import the data
library(haven)
data <- read_dta("matchingdata.dta")

# Keep non-missing values
data <- na.omit(data)

# Treatment distribution
library(janitor)
tabyl(data$PKO)
```

```
## data$PKO     n      percent
##        0 13051 0.94799157
##        1   716 0.05200843
```

Matching to improve balance

```r
# Use MatchIt package
library(MatchIt)

# Nearest-neighbor matching on the propensity score
nn_match <- matchit(PKO ~ avgttime + avgadjimr + popgpw2000_40 + avgmnt + borddist + capdist + prec_new, metho

nn_match

## A matchit object
##  - method: 1:1 nearest neighbor matching without replacement
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 13767 (original), 1432 (matched)
##  - target estimand: ATT
##  - covariates: avgttime, avgadjimr, popgpw2000_40, avgmnt, borddist, capdist, prec_new
# Matched sample
data_nn <- match.data(nn_match)

# Matched data
tabyl(data_nn$PKO)

## data_nn$PKO    n percent
##           0  716     0.5
##           1  716     0.5
```

Matching to improve balance

```
# Compare balance in raw and matched data
summary(nn_match)$sum.all[,1:3]
```

```
##                Means Treated Means Control Std. Mean Diff.
## distance        1.886561e-01  4.451170e-02      0.9039073
## avgttime        3.179041e+02  6.365755e+02     -1.4044072
## avgadjimr       1.402006e+02  1.309813e+02      0.3937550
## popgpw2000_40   5.282647e+05  1.842708e+05      0.6475519
## avgmnt          1.908064e-01  1.039625e-01      0.2943259
## borddist        1.417071e+02  1.988932e+02     -0.3827070
## capdist         5.797424e+02  8.136026e+02     -0.4359644
## prec_new        1.769303e+03  1.356781e+03      0.4541444
```

```
summary(nn_match)$sum.matched[,1:3]
```

```
##                Means Treated Means Control Std. Mean Diff.
## distance        1.886561e-01  1.820085e-01      0.04168650
## avgttime        3.179041e+02  3.352170e+02     -0.07629930
## avgadjimr       1.402006e+02  1.454425e+02     -0.22388296
## popgpw2000_40   5.282647e+05  4.773519e+05      0.09584090
## avgmnt          1.908064e-01  1.731964e-01      0.05968288
## borddist        1.417071e+02  1.571460e+02     -0.10332192
## capdist         5.797424e+02  5.044745e+02      0.14031509
## prec_new        1.769303e+03  1.822873e+03     -0.05897540
```
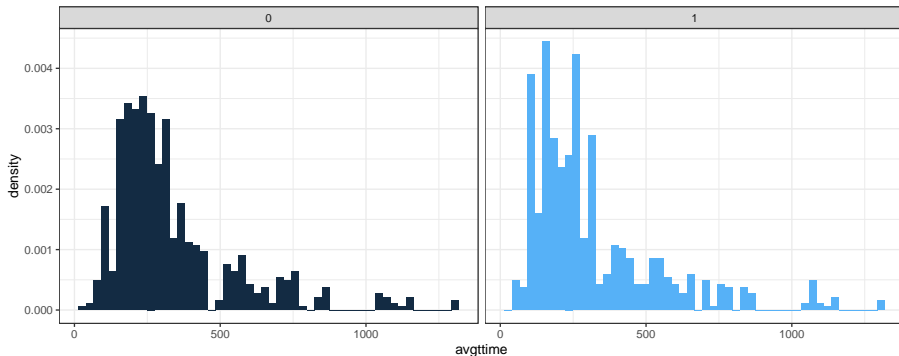
Matching to improve balance

```
library(ggplot2)
# Raw
ggplot(data, aes(x=avgttime, y = ..density.., fill=PKO)) +
  geom_histogram(bins=50) + facet_wrap(~PKO)+ theme_bw() +
  theme(legend.position="none")
```
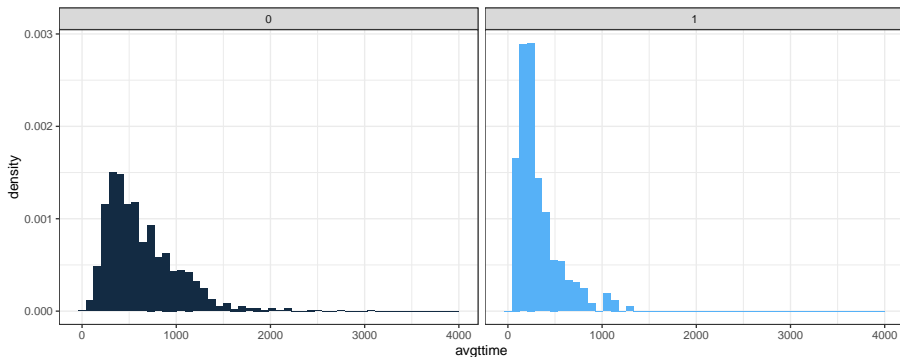
Matching to improve balance

```
# Matched
ggplot(data_nn, aes(x=avgttime, y = ..density.., fill=PKO)) +
  geom_histogram(bins=50) + facet_wrap(~PKO)+ theme_bw() +
  theme(legend.position="none")
```



```
# Larger set of diagnostic plots provided:
# plot(nn_match)
```

Matching algorithms

- Exact matching
- Genetic matching
- Coarsened Exact Matching
- Covariate Balancing Propensity Score

Combining weighting and regression

Doubly-robust estimators combine estimation of the conditional assignment probability (PS) and conditional outcome (regression), and are consistent if any of the two is misspecified (not both).

Combining weighting and regression

Doubly-robust estimators combine estimation of the conditional assignment probability (PS) and conditional outcome (regression), and are consistent if any of the two is misspecified (not both).

There are many, a standard one is the Augmented Inverse Probability Weighting estimator: augments IPW with predicted outcomes from separate regressions on treated and control group.

$$\hat{\tau}_{aipw} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{\hat{Y}_i(D_i - \hat{e}(X_i))}{\hat{e}(X_i)} \right) -$$
$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{(1-D_i)Y_i}{1 - \hat{e}(X_i)} + \frac{\hat{Y}_i(D_i - \hat{e}(X_i))}{1 - \hat{e}(X_i)} \right)$$

Implentation

There are several packages for matching and weighting methods

In R:

- `MatchIt, Matching, GenMatch`
- `CBPS`
- `PSweight`
- `Causalweight`
- `cem`

In Stata:

- `teffects`
- `cem`

Sensitivity

- All these methods rely on some form of the CIA
- As any assumption, "selection on observables" is untestable
- We can however study how results could be affected by hypothetical departures from it, or how *sensitive* they are
- There are several proposed strategies for sensitivity analysis, we will focus on the one by Cinelli and Hazlett (2020), implemented through `sensemakr` (there should now be a Stata version)

Sensitivity as an omitted variable problem

Basic framework:

- A linear model $Y = \tau D + X'\beta + \gamma Z + \epsilon_{full}$, where D is the treatment, X are observed controls and Z are unobserved controls
- The researcher can only estimate $Y = \tau_{res} D + X'\beta_{res} + \epsilon_{res}$
- We know that $\hat{\tau}_{res} = \hat{\tau} + \hat{\gamma}\hat{\delta} = \hat{\tau} + Bias$, where $\hat{\delta} = \frac{Cov(\tilde{D}, \tilde{Z})}{V(\tilde{D})}$
- In essence, what we do is to give values to the $Bias$ and study how much the estimate of $\hat{\tau}_{res}$ change
- Now, note that the Bias has two components:
    - $\hat{\gamma}$: the *impact* of Z on the outcome
    - $\hat{\delta}$: the *imbalance* of Z across treated/control groups

- It is convenient to rewrite OVB in terms of partial R-squared/correlations
- Allows for non-linearities in the effects of confounders and in assessing the sensitivity of the standard errors
- See the technical details in the journal article and examples in the sensemakr website
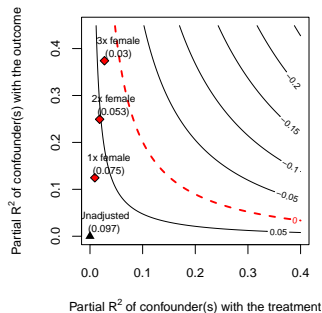
Cinelli and Hazlett (2020)

```
library(sensemakr)
data("darfur")

# Run regression model
model <- lm(peacefactor ~ directlyharmed + age + farmer_dar + herder_dar + pastvoted + hhsize_darfur + female +
            village, data = darfur)

# Run sensitivity analysis
sensitivity <- sensemakr(model, treatment = "directlyharmed", benchmark_covariates = "female", kd = 1:3)

# Results description
# sensitivity

# Plot
plot(sensitivity)
```

Cinelli and Hazlett (2020)

```
# Plot
plot(sensitivity, sensitivity.of = "t-value")
```