# Econ 2110, fall 2016, Part II Causality and Identification 

Maximilian Kasy<br>Department of Economics, Harvard University

## Takeaways for this part of class

1. fundamental notions of causal inference:

- causality
- structural objects
- identification

2. identification approaches:

- randomized experiments
- instrumental variables
- approaches based on conditional independence


## Roadmap

Basic concepts

Origins: systems of structural equations

Treatment effects

Instrumental variables

Conditional independence

## Causal and structural objects in economics

- Returns to schooling
- Elasticity of the tax base with respect to tax rates
- Effect of minimum wage on employment
- Effect of deworming pills on school attendance
- Price elasticity of demand for gasoline
- ...


## Correlation and causality

Do observable distributions tell us something about causality?

- College graduates earn $x \%$ more than high school graduates
- Countries with higher GDP have higher tax rates on average
- Minimum wage levels seem uncorrelated with unemployment levels across time and space
- Gasoline consumption and gasoline price are negatively correlated over time
- ...

Figure: correlation and causation


## Identification vs inference



## Identification vs inference

- goal of econometrics:
learning interesting things (hopefully) about economic phenomena from observations.
- two separate components of econometrics:

1. identification
2. estimation and inference

## Estimation and Inference

1. learning about a population distribution
2. from a finite number of observations.
3. examples:

- estimate a difference in expectations using a difference in means
- perform inference using a t-test.


## Identification

1. learning about underlying structures, causal mechanisms
2. from a population distribution.
3. example:
identify a causal effect
by a difference in expectations
if we have a randomized experiment.

- identification inverts the mapping
- from underlying structures to a population distribution
- implied by a model and identifying assumptions.


## Causality

## Practice problem

How would you define causality?

## "Pure" statistics

- causality is meaningless
- observations only tell us about correlations
- more generally, joint distributions
- disclaimer: few statisticians today would say this!


## Sciences

- Galileo Galilei: one of the first to follow experimental ideal
- full control of experimental circumstances.
- do the same thing
$\Rightarrow$ same thing happens to the outcomes you measure
- variation in experimental circumstances
$\Rightarrow$ difference in observed outcomes $\approx$ causal effect
- example:
- dropping a ball from different floors of the tower of Pisa
- different time till the ball hits the ground
- crucial component:
external intervention ("exogenous variation")
$\Rightarrow$ allows to interpret correlation as causation


## Social and biological sciences

- economics is not physics
- this version of the experimental ideal is not very useful
- reason: many unobserved, and unknown, factors which we cannot hope to control
- $\Rightarrow$ can never replicate experimental circumstances fully
- there is "unobserved heterogeneity."
- Aside:

I also think we should be careful with other concepts from the sciences; for instance "theory testing"

## Social and biological sciences ctd

- not all is lost, however
- can still hope create experimental circumstances which are the same on average
- this is the idea of a randomized experiment!
- randomly pick treatment and control groups
$\Rightarrow$ they are identical on average.
- "compare apples with apples."
- many settings of interest in economics: not possible to run experiments
- but: definition of causality is intimately tied to randomized experiment, hypothetical or actual


## Recap

- Framework discussed in part I of class
- Does not allow to talk about causality,
- only joint distributions.
- Causality in the sciences:
- Additional concept:
- External intervention / exogeneous variation
- $\Rightarrow$ experiments.
- Causality in econometrics, biostatistics,...:
- Additional concept:
- Unobserved heterogeneity.
- $\Rightarrow$ randomized experiments
- (or "quasi-experiments")


## Structural objects

## Practice problem

How would you define "structural?"

## My preferred definition

- An object is structural, if it is invariant across relevant counterfactuals.
- Example: dropping a ball from tower of Pisa
- acceleration is the same, no matter which floor you drop it from
- also the same if you do this on the Eiffel tower
- time to ground would not be the same
- acceleration is not the same if you do this on the moon


## Possibly structural objects

- "economic primitives"
- preferences
- technologies
- assumed to be invariant across policy changes
- derived objects
- demand function, as supply, and thus price, varies
- more generally causal effects, as treatment varies


## Another common use for "structural"

- full specification of parametric forms for economic primitives
- assumptions strong enough to identify these primitives
- somewhat misleading terminology:

1. less ambitious approaches with weaker assumptions can also identify structural objects
2. if assumptions for full identification of economic primitives are violated
then the objects identified are not structural.

## Origins: systems of structural equations

- econometrics pioneered by "Cowles commission" starting in the 1930s
- they were interested in demand (elasticities) for agricultural goods
- introduced systems of simultaneous equations
- outcomes as equilibria of some structural relationships
- goal: recover the slopes of structural relationships
- from observations of equilibrium outcomes and exogenous shifters


## System of structural equations

$$
Y=A \cdot Y+B \cdot Z+\varepsilon
$$

- $Y$ : $k$-dimensional vector of equilibrium outcomes
- Z: I-dimensional vector of exogenous variables
- A: unknown $k \times k$ matrix of coefficients of interest
- B: unknown $k \times I$ matrix
- $\varepsilon$ : further unobserved factors affecting outcomes


## Example: supply and demand

$$
\begin{aligned}
& Y=(P, Q) \\
& P=A_{12} \cdot Q+B_{1} \cdot Z+\varepsilon_{1} \text { demand } \\
& Q=A_{21} \cdot P+B_{2} \cdot Z+\varepsilon_{2} \text { supply }
\end{aligned}
$$

- demand function: relates prices to quantity supplied and shifters $Z$ and $\varepsilon_{1}$ of demand
- supply function relates quantities supplied to prices and shifters $Z$ and $\varepsilon_{2}$ of supply.
- does not really matter which of the equations puts prices on the "left hand side.'
- price and quantity in market equilibrium: solution of this system of equations.


## Reduced form

- solve equation $Y=A \cdot Y+B \cdot Z+\varepsilon$ for $Y$ as a function of $Z$ and $\varepsilon$
- bring $A \cdot Y$ to the left hand side, pre-multiply by $(I-A)^{-1} \Rightarrow$

$$
\begin{aligned}
& Y=C \cdot Z+\eta \text { "reduced form" } \\
& C:=(I-A)^{-1} \cdot B \text { reduced form coefficients } \\
& \eta:=(I-A)^{-1} \cdot \varepsilon
\end{aligned}
$$

- suppose $E[\varepsilon \mid Z]=0$ (ie., $Z$ is randomly assigned)
- then we can identify $C$ from

$$
E[Y \mid Z]=C \cdot Z
$$

## Exclusion restrictions

- suppose we know $C$
- what we want is $A$, possibly $B$
- problem: $k \times I$ coefficients in $C=(I-A)^{-1} \cdot B$ $k \times(k+l)$ coefficients in $A$ and $B$
- $\Rightarrow$ further assumptions needed
- exclusion restrictions: assume that some of the coefficients in $B$ or $A$ are $=0$.
- Example: rainfall affects grain supply but not grain demand


## Supply and demand continued

- suppose $Z$ is (i) random, $E[\varepsilon \mid Z]=0$
- and (ii) "excluded" from the demand equation
$\Rightarrow B_{11}=0$
- by construction, $\operatorname{diag}(A)=0$
- therefore

$$
\operatorname{Cov}(Z, P)=\operatorname{Cov}\left(Z, A_{12} \cdot Q+B_{1} \cdot Z+\varepsilon_{1}\right)=A_{12} \cdot \operatorname{Cov}(Z, Q),
$$

- $\Rightarrow$ the slope of demand is identified by

$$
A_{12}=\frac{\operatorname{Cov}(Z, P)}{\operatorname{Cov}(Z, Q)}
$$

- $Z$ is an instrumental variable


## Remarks

- historically, applied researchers have not been very careful about choosing $Z$ for which
(i) randomization and (ii) exclusion restriction are well justified.
- since the 1980s, more emphasis on credibility of identifying assumptions
- some additional problematic restrictions we imposed:

1. linearity
2. constant (non-random) slopes
3. heterogeneity is $k$ dimensional and enters additively

- $\Rightarrow$ causal effects assumed to be the same for everyone
- next section: framework which does not impose this


## Treatment effects and potential outcomes

- coming from biostatistics / medical trials
- potential outcome framework: answer to "what if" questions
- two "treatments:" $D=0$ or $D=1$
- eg. placebo vs. actual treatment in a medical trial
- $Y_{i}$ person i's outcome
eg. survival after 2 years
- potential outcome $Y_{i}^{0}$ : what if person $i$ would have gotten treatment 0
- potential outcome $Y_{i}^{1}$ : what if person $i$ would have gotten treatment 1
- question to you: is this even meaningful?
- causal effect / treatment effect for person $i$ : $Y_{i}^{1}-Y_{i}^{0}$.
- average causal effect / average treatment effect:

$$
A T E=E\left[Y^{1}-Y^{0}\right]
$$

- expectation averages over the population of interest


## The fundamental problem of causal inference

- we never observe both $Y^{0}$ and $Y^{1}$ at the same time
- one of the potential outcomes is always missing from the data
- treatment $D$ determines which of the two we observe
- formally:

$$
Y=D \cdot Y^{1}+(1-D) \cdot Y^{0}
$$

## Selection problem

- distribution of $Y^{1}$ among those with $D=1$ need not be the same as the distribution of $Y^{1}$ among everyone.
- in particular

$$
\begin{aligned}
& E[Y \mid D=1]=E\left[Y^{1} \mid D=1\right] \neq E\left[Y^{1}\right] \\
& E[Y \mid D=0]=E\left[Y^{0} \mid D=0\right] \neq E\left[Y^{0}\right] \\
& E[Y \mid D=1]-E[Y \mid D=0] \neq E\left[Y^{1}-Y^{0}\right]=A T E .
\end{aligned}
$$

## Randomization

- no selection $\Leftrightarrow D$ is random

$$
\left(Y^{0}, Y^{1}\right) \perp D
$$

- in this case,

$$
\begin{aligned}
& E[Y \mid D=1]=E\left[Y^{1} \mid D=1\right]=E\left[Y^{1}\right] \\
& E[Y \mid D=0]=E\left[Y^{0} \mid D=0\right]=E\left[Y^{0}\right] \\
& E[Y \mid D=1]-E[Y \mid D=0]=E\left[Y^{1}-Y^{0}\right]=A T E .
\end{aligned}
$$

- can ensure this by actually randomly assigning $D$
- independence $\Rightarrow$ comparing treatment and control actually compares "apples with apples"
- this gives empirical content to the "metaphysical" notion of potential outcomes!


## Empirical example

- Krueger, A. (1999). Experimental estimates of education production functions. The Quarterly Journal of Economics, 114(2):497-532.
- project STAR experiment
- randomly assign primarily school students (and their teachers) to large and small classes, within schools
- estimate causal effect on test-scores, later life outcomes


## Empirical example

- Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review, 94(4):991-1013.
- randomly assign names which are statistically "white" or "black" to resumes which are sent out as job applications
- estimate causal effect on likelihood of getting invited to a job interview


## Relation to linear structural equations?

- Linear structural equations are a special case of treatment effect framework.
- Suppose

$$
Y=\alpha+\beta D+\varepsilon
$$

- where this equation is "structural."
$\Leftrightarrow$ Coefficients and the residual stay the same if treatment is changed.
- Then

$$
\begin{aligned}
& Y^{0}=\alpha+\varepsilon \\
& Y^{1}=\alpha+\beta+\varepsilon,
\end{aligned}
$$

- and $A T E=\beta$.
- Note: Causal effects are the same for everyone in the linear framework.


## Nonlinear structural equations

- without imposing linearity / restrictions on the dimension of heterogeneity:

$$
Y=g(D, \varepsilon)
$$

- equivalent to potential outcomes
- potential outcomes notation: more compact
- structural function notation: more explicit which determinants are held constant to define causal effects in more complex settings? Eg.

$$
Y=g(D, X, W, \varepsilon)
$$

## Instrumental variables

- recall: simultaneous equations models with exclusion restrictions
- $\Rightarrow$ instrumental variables

$$
\beta=\frac{\operatorname{Cov}(Z, Y)}{\operatorname{Cov}(Z, D)}
$$

- we will now give a new interpretation to $\beta$
- using the potential outcomes framework, allowing for heterogeneity of treatment effects
- "Local Average Treatment Effect" (LATE)


## Empirical example

- Angrist, J.D. and Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4):979-1014.
- compare individuals born in different quarters of the year
- school start age and structure compulsory schooling laws
- $\Rightarrow$ people born late in the year have to stay in school longer
- quarter of birth as an instrument for educational attainment in estimates of returns to schooling
- estimates effect for those affected by compulsory schooling laws


## 6 assumptions

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434):444-455.

$$
\begin{aligned}
& \text { 1. } Z \in\{0,1\}, D \in\{0,1\} \\
& \text { 2. } Y=D \cdot Y^{1}+(1-D) \cdot Y^{0} \\
& \text { 3. } D=Z \cdot D^{1}+(1-Z) \cdot D^{0} \\
& \text { 4. } D^{1} \geq D^{0} \\
& \text { 5. } Z \perp\left(Y^{0}, Y^{1}, D^{0}, D^{1}\right) \\
& \text { 6. } \operatorname{Cov}(Z, D) \neq 0
\end{aligned}
$$

## Discussion of assumptions

Generalization of randomized experiment

- $D$ is "partially randomized"
- instrument $Z$ is randomized
- $D$ depends on $Z$, but is not fully determined by it

1. Binary treatment and instrument: both $D$ and $Z$ can only take two values results generalize, but things get messier without this
2. Potential outcome equation for $Y: Y=D \cdot Y^{1}+(1-D) \cdot Y^{0}$

- exclusion restriction: $Z$ does not show up in the equation determining the outcome.
- "stable unit treatment values assumption" (SUTVA): outcomes are not affected by the treatment received by other units. excludes general equilibrium effects or externalities.

3. Potential outcome equation for $D: D=Z \cdot D^{1}+(1-Z) \cdot D^{0}$ SUTVA; treatment is not affected by the instrument values of other units
4. No defiers: $D^{1} \geq D^{0}$

- four possible combinations for the potential treatments ( $D^{0}, D^{1}$ ) in the binary setting
- $D^{1}=0, D^{0}=1$, is excluded
- $\Leftrightarrow$ monotonicity

Table: No defiers

|  | $D^{0}$ | $D^{1}$ |
| :--- | :---: | :---: |
| Never takers (NT) | 0 | 0 |
| Compliers (C) | 0 | 1 |
| Always takers (AT) | 1 | 1 |
| Defiers | 1 | 0 |

## Practice problem

Who do you think are the compliers for the quarter of birth instrument?
5. Randomization: $Z \perp\left(Y^{0}, Y^{1}, D^{0}, D^{1}\right)$

- $Z$ is (as if) randomized.
- in applications, have to justify both exclusion and randomization
- no reverse causality, common cause!

6. Instrument relevance: $\operatorname{Cov}(Z, D) \neq 0$

- guarantees that the IV estimand is well defined
- there are at least some compliers
- testable
- near-violation: weak instruments


## Graphical illustration



## Illustration explained

- 3 groups, never takers, compliers, and always takers
- by randomization of $Z$ :
each group represented equally given $Z=0 / Z=1$
- depending on group: observe different treatment values and potential outcomes.
- will now take the IV estimand

$$
\frac{\operatorname{Cov}(Z, Y)}{\operatorname{Cov}(Z, D)}
$$

- interpret it in terms of potential outcomes: average causal effects for the subgroup of compliers
- idea of proof: take the "top part" of figure 49, and subtract the "bottom part."


## Preliminary result:

If $Z$ is binary, then

$$
\frac{\operatorname{Cov}(Z, Y)}{\operatorname{Cov}(Z, D)}=\frac{E[Y \mid Z=1]-E[Y \mid Z=0]}{E[D \mid Z=1]-E[D \mid Z=0]} .
$$

## Practice problem

Try to prove this!

## Proof

- Consider the covariance in the numerator:

$$
\begin{aligned}
& \operatorname{Cov}(Z, Y)=E[Y Z]-E[Y] \cdot E[Z] \\
&= E[Y \mid Z=1] \cdot E[Z]-(E[Y \mid Z=1] \cdot E[Z]-E[Y \mid Z=0] \cdot E[1-Z]) \cdot E[Z] \\
&=(E[Y \mid Z=1]-E[Y \mid Z=0]) \cdot E[Z] \cdot E[1-Z] .
\end{aligned}
$$

- Similarly for the denominator:

$$
\operatorname{Cov}(Z, D)=(E[D \mid Z=1]-E[D \mid Z=0]) \cdot E[Z] \cdot E[1-Z] .
$$

- The $E[Z] \cdot E[1-Z]$ terms cancel when taking a ratio


## The "LATE" result

$$
\frac{E[Y \mid Z=1]-E[Y \mid Z=0]}{E[D \mid Z=1]-E[D \mid Z=0]}=E\left[Y^{1}-Y^{0} \mid D^{1}>D^{0}\right]
$$

## Practice problem

Try to prove this!
Hint: decompose $E[Y \mid Z=1]-E[Y \mid Z=0]$ in 3 parts corresponding to our illustration


## Proof

- "top part" of figure:

$$
\begin{aligned}
E[Y \mid Z=1]= & E[Y \mid Z=1, N T] \cdot P(N T \mid Z=1) \\
& +E[Y \mid Z=1, C] \cdot P(C \mid Z=1) \\
& +E[Y \mid Z=1, A T] \cdot P(A T \mid Z=1) \\
= & E\left[Y^{0} \mid N T\right] \cdot P(N T)+E\left[Y^{1} \mid C\right] \cdot P(C)+E\left[Y^{1} \mid A T\right] \cdot P(A T) .
\end{aligned}
$$

- first equation relies on the no defiers assumption
- second equation uses the exclusion restriction and randomization assumptions.
- Similarly

$$
\begin{aligned}
& E[Y \mid Z=0]=E\left[Y^{0} \mid N T\right] \cdot P(N T)+ \\
& E\left[Y^{0} \mid C\right] \cdot P(C)+E\left[Y^{1} \mid A T\right] \cdot P(A T) .
\end{aligned}
$$

proof continued:

- Taking the difference, only the complier terms remain, the others drop out:

$$
E[Y \mid Z=1]-E[Y \mid Z=0]=\left(E\left[Y^{1} \mid C\right]-E\left[Y^{0} \mid C\right]\right) \cdot P(C)
$$

- denominator:

$$
\begin{aligned}
E[D \mid Z=1]-E[D \mid Z= & 0]=E\left[D^{1}\right]-E\left[D^{0}\right] \\
& =(P(C)+P(A T))-P(A T)=P(C) .
\end{aligned}
$$

- taking the ratio, the claim follows.


## Recap

LATE result:

- take the same statistical object economists estimate a lot
- which used to be interpreted as average treatment effect
- new interpretation in a more general framework
- allowing for heterogeneity of treatment effects
- $\Rightarrow$ treatment effect for a subgroup


## Practice problem

Is the LATE, $E\left[Y^{1}-Y^{0} \mid D^{1}>D^{0}\right]$, a structural object?

## An alternative approach: Bounds

- keep the old structural object of interest: average treatment effect
- but analyze its identification in the more general framework with heterogeneous treatment effects
- in general: we can learn something, not everything
- $\Rightarrow$ bounds / "partial identification"

Manski, C. (2003). Partial identification of probability distributions.
Springer Verlag.

## Same assumptions as before

1. $Z \in\{0,1\}, D \in\{0,1\}$
2. $Y=D \cdot Y^{1}+(1-D) \cdot Y^{0}$
3. $D=Z \cdot D^{1}+(1-Z) \cdot D^{0}$
4. $D^{1} \geq D^{0}$
5. $Z \perp\left(Y^{0}, Y^{1}, D^{0}, D^{1}\right)$
6. $\operatorname{Cov}(Z, D) \neq 0$
additionally:
7. $Y$ is bounded, $Y \in[0,1]$

## Decomposing ATE in known and unknown components

- decompose $E\left[Y^{1}\right]$ :

$$
E\left[Y^{1}\right]=E\left[Y^{1} \mid N T\right] \cdot P(N T)+E\left[Y^{1} \mid C \vee A T\right] \cdot P(C \vee A T) .
$$

- terms that are identified:

$$
\begin{aligned}
E\left[Y^{1} \mid C \vee A T\right] & =E[Y \mid Z=1, D=1] \\
P(C \vee A T) & =E[D \mid Z=1] \\
P(N T) & =E[1-D \mid Z=1]
\end{aligned}
$$

and thus

$$
E\left[Y^{1} \mid C \vee A T\right] \cdot P(C \vee A T)=E[Y D \mid Z=1] .
$$

- Data tell us nothing about is $E\left[Y^{1} \mid N T\right]$. $Y^{1}$ is never observed for never takers.
- but we know, since $Y$ is bounded, that

$$
E\left[Y^{1} \mid N T\right] \in[0,1]
$$

- Combining these pieces, get upper and lower bounds on $E\left[Y^{1}\right]$ :

$$
\begin{aligned}
& E\left[Y^{1}\right] \in[E[Y D \mid Z=1], \\
& E[Y D \mid Z=1]+E[1-D \mid Z=1]] .
\end{aligned}
$$

- For $Y^{0}$, similarly

$$
\begin{aligned}
E\left[Y^{0}\right] \in[E[Y(1-D) \mid Z & =0], \\
E[Y(1-D) \mid Z & =0]+E[D \mid Z=0]] .
\end{aligned}
$$

- Data are uninformative about $E\left[Y^{0} \mid A T\right]$.


## Practice problem

Show this.

## Combining to get bounds on ATE

- lower bound for $E\left[Y^{1}\right]$, upper bound for $E\left[Y^{0}\right] \Rightarrow$ lower bound on $E\left[Y^{1}-Y^{0}\right]$

$$
E\left[Y^{1}-Y^{0}\right] \geq E[Y D \mid Z=1]-E[Y(1-D) \mid Z=0]-E[D \mid Z=0]
$$

- upper bound for $E\left[Y^{1}\right]$, lower bound for $E\left[Y^{0}\right]$
$\Rightarrow$ upper bound on $E\left[Y^{1}-Y^{0}\right]$

$$
E\left[Y^{1}-Y^{0}\right] \leq E[Y D \mid Z=1]-E[Y(1-D) \mid Z=0]+E[1-D \mid Z=1]
$$

## Between randomized experiments and nothing

- bounds on ATE:

$$
\left.\left.\left.\begin{array}{rl}
E\left[Y^{1}-Y^{0}\right] \in[E[Y D \mid Z & =1]-E[Y(1-D) \mid Z
\end{array}\right)=0\right]-E[D \mid Z=0], ~ 子[Y D \mid Z=1]-E[Y(1-D) \mid Z=0]+E[1-D \mid Z=1]\right] .
$$

- length of this interval:

$$
E[1-D \mid Z=1]+E[D \mid Z=0]=P(N T)+P(A T)=1-P(C)
$$

- Share of compliers $\rightarrow 1$
- interval ("identified set") shrinks to a point
- In the limit, $D=Z$
- thus $\left(Y^{1}, Y^{0}\right) \perp D$ - randomized experiment
- Share of compliers $\rightarrow 0$
- length of the interval goes to 1
- In the limit the identified set is the same as without instrument


## Conditional independence

- Two generalizations of randomized experiments:
- last section:

Generalizing from $D$ random
to $Z$ random.

- this section:

Generalizing from $D$ random
to $D$ random conditional on covariates $X$.

- If that's the case,
"comparing apples with apples"
only requires comparing units with the same values of $X$.


## Assumptions

1. $D \in\{0,1\}$
2. $Y=D \cdot Y^{1}+(1-D) \cdot Y^{0}$
3. $\left(Y^{1}, Y^{0}\right) \perp D \mid X$
4. $0<p(X)<1$ for almost all $X$, where $p(X)=P(D=1 \mid X)$

## Discussion of assumptions

1. Binary treatment: easy to generalize the following to arbitrary support of $D$.
2. Potential outcome equation for $Y: Y=D \cdot Y^{1}+(1-D) \cdot Y^{0}$ same as before (SUTVA!).
3. Conditional independence: $\left(Y^{1}, Y^{0}\right) \perp D \mid X$

- within subgroups defined by $X$, we essentially have a randomized experiment.
- somewhat hard to justify in practice
- best thought of as a plausible approximation, if $X$ captures the main components of heterogeneity which might drive dependence between potential outcomes and treatment

4. Overlapping support: $0<p(X)<1$

- there are no groups, defined by $X$, for which everybody (or nobody) was treated.
- can be checked directly in the data (like instrument relevance)

Under these assumptions, the average treatment effect is identified

## Practice problem

Try to prove this!

## Identification of ATE,

 proof using regression- conditional expectation of outcomes given covariates and treatment:

$$
E[Y \mid X, D=1]=E\left[Y^{1} \mid X, D=1\right]=E\left[Y^{1} \mid X\right] .
$$

- first equality holds by the potential outcomes equation.
- second equality uses conditional independence.
- $E[Y \mid X, D=1]$ is identified as long as $p(X)>0$.
- law of iterated expectations, averaging across the distribution of $X \Rightarrow$

$$
E\left[Y^{1}\right]=E_{X}\left[E\left[Y^{1} \mid X\right]\right]=E_{X}[E[Y \mid X, D=1]] .
$$

- note that

$$
E_{X}[E[Y \mid X, D=1]] \neq E[Y \mid D=1]!
$$

- left hand term averages $E[Y \mid X, D=1]$ over the marginal distribution of $X$
- right hand term averages $E[Y \mid X, D=1]$ over the conditional distribution of $X$ given $D$.
- similarly for $D=0$

$$
E\left[Y^{0}\right]=E_{X}\left[E\left[Y^{0} \mid X\right]\right]=E_{X}[E[Y \mid X, D=0]] .
$$

- average treatment effect is identified by

$$
E\left[Y^{1}-Y^{0}\right]=E_{X}[E[Y \mid X, D=1]-E[Y \mid X, D=0]] .
$$

## Identification of ATE, proof using reweighting

- intuition: units for which $p(X)$ is small are under-represented among the observations such that $D=1$
- need to upweight them in order to get the distribution of $Y^{1}$
- upweight by factor $1 / p(x)$ - "inverse probability weighting"
- Law of iterated expectations, $D$ binary $\Rightarrow$

$$
\begin{aligned}
E[Y D \mid X] & =E[Y D \mid X, D=1] \cdot p(X)+E[Y D \mid X, D=0] \cdot(1-p(X)) \\
& =E[Y \mid X, D=1] \cdot p(X)
\end{aligned}
$$

- Potential outcome equation, conditional independence $\Rightarrow$

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

- Rearranging $\Rightarrow$

$$
E\left[\left.\frac{D}{p(X)} \cdot Y \right\rvert\, X\right]=E\left[Y^{1} \mid X\right]
$$

- Iterated expectations, again $\Rightarrow$

$$
E\left[\frac{D}{p(X)} \cdot Y\right]=E\left[Y^{1}\right]
$$

- Similar argument for $D=0 \Rightarrow$

$$
E\left[\frac{1-D}{1-p(X)} \cdot Y\right]=E\left[Y^{0}\right]
$$

## Practice problem

## Show this.

- $\Rightarrow$ Average treatment effect is identified by

$$
E\left[Y^{1}-Y^{0}\right]=E\left[\left(\frac{D}{p(X)}-\frac{1-D}{1-p(X)}\right) \cdot Y\right]
$$

## Brief remark on estimation

- This entire part of class was about identification.
- Mapping features of the population distribution into features of the underlying structure.
- Suggests analogous approaches for estimation.
- $E[Y \mid X, D]$ can be estimated as the predicted value of a regression of $Y$ on $D$, controlling for $X$ then average across $X$.
- "Propensity score" $p(X)$ can be estimated, using flexible (logit) regression of $D$ on $X$ then average reweighted $Y$.


## Supplementary readings

- Applied microeconomics perspective: Angrist, J. D. and Pischke, J. S. (2009). Mostly harmless econometrics: an empiricist's companion. Princeton Univ Press.
- Principled treatment of (partial) identification:

Manski, C. (2003). Partial identification of probability distributions. Springer Verlag.

- Theoretical computer scientist on the notion of causality:

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

- Forthcoming book by Imbens and Rubin themselves:

Imbens, G. and D. Rubin (2014). Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

