Empirical Analysis III

The University of Chicago

April 3, 2019

Content: Week 1

- Defining parameters and arguing their (policy) relevance
- What we can (and cannot) learn from randomized controlled trials

- But first, let's discuss the **three steps** of (good) empirical work
- These steps will be an organizing principle in this class (and hopefully in your work)

Step 1: Define the target parameter(s)

Target parameters, though experiments and counterfactuals

- Defining the causal effect of interest – the target parameter – amounts to specifying precisely a counterfactual question
- Thinking about a counterfactual requires asking "What if..."
- In other words, defining the **target parameter(s)** requires a though experiment; neither data nor actual experimentation needed
- Contrary to slogan "No causation without manipulation"

Examples

- What would happen to unemployment if government increased minimum wages?
- What would happen to prices if two firms merged?
- What would your life been like if you didn't accept U Chicago?

Step 2: Identification of target parameter

Identification links though experiment and data

- The **target parameter**, as defined by the counterfactual question, is a function of the **unobservables**
- Question of identification: What can we learn about this function from the observed data?

- Identification maps assumptions (model) and data to information about target parameter
- A parameter is identified if, under the stated assumptions, alternative values of the parameter implies different distributions of observable data
- Identification is a binary property – target parameter is either (point or partially) identified or not identified

Step 3: Statistical Inference

Statistical inference links population and sample

- In practice, we only see a finite **sample** of the observables
- From this we know the sample distribution
- However, we don't know the **population distribution** of data
- Statistical inference is using the sample to learn about the population

- It is useful to separate identification from statistical inference:

$$\text{sample} \underbrace{\rightarrow}_{\text{statistical inference}} \text{population} \underbrace{\rightarrow}_{\text{identification}} \text{unobserved parameters}$$

- The second arrow is logically the first thing to consider
- Can't recover a parameter when we know the population distribution? Then you also couldn't recover it with the sample distribution!
- Identification: What I can learn if I had "infinite data"....

- We will keep returning to these three steps over the next few weeks

- But first, it is time to introduce some notation and define target parameters

Models and notation

Why use formal models?

- Formal models are useful to be precise about target parameters, identification and inference

Notation and models

- Different researchers use different notation and models, including
  1) **Potential outcome model**: Neyman-Fisher-Quandt-Rubin model
  2) Economic choice models including the **Roy model** and, more generally, **latent variables models**
- As we will see, 1) and 2) are not necessarily different animals
- But explicit choice models can be useful to define counterfactuals and economically interpret assumptions and results

Potential Outcome Notation

- $\mathcal{D}$ is a mutually exclusive and exhaustive set of states ("treatments") e.g. training/no training $\mathcal{D} = \{0, 1\}$
- For each $d \in \mathcal{D}$ there is a **potential outcome** Y_d (a random variable)
- Y_d is what would have happened if the state were endogenously set d

- We observe the actual state, a random variable $D \in \mathcal{D}$
- We also observe an outcome Y, related to potential outcomes as:

$$Y = \sum_{d \in \mathcal{D}} Y_d \mathbb{1}[D = d] = Y_D$$

- $Y = Y_D$ is observed, but Y_d for $d \neq D$ are unobserved

Potential Outcomes and Choices

Binary treatment

- **Switching regression**: $Y = DY_1 + (1 - D)Y_0$
- Without further restrictions, $Y_1 - Y_0$ may vary freely across individuals
- The treatment D may be dependent with Y_0, indicating **selection bias**, or $Y_1 - Y_0$, indicating **selection on the gains**, or both
- Model does not specify why individuals make the treatment choice that they do, in contrast to a an outcome max. model $D = \mathbb{1}[Y_1 > Y_0]$
- Yet model does not preclude possibility that individuals choose treatment with knowledge of (Y_0, Y_1)
- Possible to add potential outcome representation of choices
- E.g., with a binary instrument Z, choice equation can be written:

$$D = ZD_1 + (1 - Z)D_0$$

D_z being value of treatment realized if Z was endogenously set z

Defining target parameter

- We are interested in counterfactuals, Y_d for $d \neq D$
- These variables capture the "what if" aspect of causality
- There are many possible **target parameters**

Example: Program evaluation

- Suppose $d \in \{0, 1\}$ indicates participation in a job training program
- Y is a scalar labor market outcome such as earnings
- If $D = 1$ we observe Y_1 (but not Y_0) and if $D = 0$ we observe Y_0
- There are many possible questions one could ask:
  - What would be average earnings if everyone were trained, i.e. $\mathbb{E}[Y_1]$?
  - What is the average effect of the program, i.e. $\mathbb{E}[Y_1 - Y_0]$?
  - What about only for those who are trained, i.e. $\mathbb{E}[Y_1 - Y_0 | D = 1]$?
- What is useful depends on what question we want to answer!

Latent Variable Notation

Researchers sometimes replace potential outcome notation with **latent variable notation** in the outcome equation, the choice equation, or both.

Latent variable notation to describe outcomes

- Many empirical models in economics look like a special case of:

$$Y = g(D, V)$$

  where g is a function and V are unobservable variables
- A causal interpretation of this model is implicitly saying:

$$Y_d = g(d, V) \text{ for every } d \in \mathcal{D}$$

- This could impose assumptions, depending on what g and V are

Latent Variable Notation

Latent variable notation to describe choices

- An alternative to potential outcomes for D is a latent variable model
- Leading case is binary with **separable latent variable** choice equation:

$$D = \mathbb{1}\left[\underbrace{U}_{\text{latent variable}} \leq \underbrace{\nu(W)}_{\text{unknown function}} \right]$$

- $W \equiv (X, Z)$ are observable with Z being the instrument(s)
- U continuously distributed, normalized to be uniform [0, 1]
  $\rightarrow$ Implies that $\nu(W) = p(D|W) \equiv p(W)$ (can you prove this?)
- Combined with $Y = Y_1 D + Y_0(1 - D)$, this is called the (generalized) **Roy Model** – we will use this model a lot through the course!

- Apply the usual translation: $D_z = \mathbb{1}[U \leq \nu(X, z)]$
- Some advantages - explicitly model D as a choice problem
  $\rightarrow \nu(X, Z)$ - U is the utility of $D = 1$ vs $D = 0$
- Will be useful to think about parameters other than ATE, ATT, ATUT

Roy model and heterogeneity

- A common version of the Roy model:

$$Y_0 = X'\beta_0 + V_0 \qquad \underbrace{D = \mathbb{1}[U \leq W'\gamma]}_{\text{(selection equation)}}$$

$$Y_1 = X'\beta_1 + V_1$$

  Where (V_0, V_1, U) are unobservable and $W \equiv (X, Z)$ are observable
- This model allows for both observed and unobserved heterogeneity:

$$Y_1 - Y_0 = \underbrace{X'(\beta_1 - \beta_0)}_{\text{observed}} + \underbrace{V_1 - V_0}_{\text{unobserved}}$$

- Implies a random coefficient specification for the observed outcome:

$$Y = DY_1 + (1 - D)Y_0 = \underbrace{(V_1 - V_0)}_{\text{random coefficient}} D + X'\beta_0 + DX'(\beta_1 - \beta_0) + V_0$$

- **Selection on unobservables** if U and (V_0, V_1) are dependent

Using latent variables to define target parameters

Definition

- Abstracting from X, define the **marginal treatment effect (MTE)** as:

$$MTE(u) \equiv \mathbb{E}[Y_1 - Y_0 | U = u]$$

- MTE(u) is the ATE for those agents with first stage unobservable u
  $\rightarrow$ Those with small u (close to 0) often choose $D = 1$
  $\rightarrow$ Those with large u (close to 1) infrequently choose $D = 1$
- Unobserved treatment heterogeneity if and only if non-constant MTE

An organizing principle

- U provides a single dimension on which we can organize heterogeneity
- Many quantities can be written as weighted averages of the MTE
- For example, the ATE is the unweighted average of the MTEs:

$$ATE = \mathbb{E}[\mathbb{E}[Y_1 - Y_0 | U]] = \int_0^1 MTE(u) \times 1 \underbrace{du}_{U \text{ uniform}}$$

ATT/ATU as a Weighted Average MTE

ATT

- The ATT can be written as (see problem set)

$$ATT = \int_0^1 MTE(u) \frac{\mathbb{P}[p(Z) \geq u]}{\mathbb{P}[D = 1]} du \equiv \int_0^1 MTE(u) \omega_{ATT}(u) du$$

- Those with low values of u are more highly weighted
  $\rightarrow$ These are the most likely to take treatment
- The weights are known or identifiable and integrate to 1

ATU

- Analogous argument for the ATU:

$$ATU = \int_0^1 MTE(u) \frac{\mathbb{P}[p(Z) < u]}{\mathbb{P}[D = 0]} du \equiv \int_0^1 MTE(u) \omega_{ATU}(u) du$$

- High values of u are more highly weighted (least likely to take treatment)

What is a "Policy Relevant" Parameter?

- The MTE framework partitions all agents in a clear way
- Provides a foundation for thinking about "ideal" treatment effects
- The "ideal" treatment effect clearly depends on the question

- The ATE receives a lot of attention in the literature
  $\rightarrow$ But not very useful for policy - can agents still choose D?
- The ATT is somewhat clearer in this regard
  $\rightarrow$ Loss in benefit to treated group from discontinuing $D = 1$

- Perhaps more relevant is changing the agent's choice problem
- For example, $D \in \{0, 1\}$ is attending a four-year college
- Average effect of forcing college/no college (ATE) is not interesting
- Nor is the effect on college-goers of shutting down college (ATT)
- More interesting are the effects via D of adjusting tuition Z

Policy Relevant Treatment Effects

- Heckman and Vytlacil (2011) formalize this idea as **policy relevant treatment effects (PRTE)**
- Aggregate effect on Y of a change in the propensity score/instrument
- Change corresponds to a policy that affects treatment choice

- Let $p^*(Z^*), Z^*$ be the propensity score/instrument under a new policy
- Let D^* denote the treatment choice under the new policy:

$$D^* = \mathbb{1}[U \leq p^*(Z^*)]$$

- Letting $Y^* = D^* Y_1 + (1 - D^*) Y_0$ be the outcome under the new policy,

$$\text{HV define the PRTE as: } \beta_{PRTE} \equiv \frac{\mathbb{E}(Y^*) - \mathbb{E}(Y)}{\mathbb{E}(D^*) - \mathbb{E}(D)}$$

- The mean effect (per net person) of the policy change
- Implicit assumption is that the policy does not affect (Y_0, Y_1, U)
  $\rightarrow$ Intuitively necessary - see HV for a formalization

The PRTE as a Weighted MTE

- One can show (see problem set) that

$$\beta_{PRTE} \equiv \frac{\mathbb{E}[Y^*] - \mathbb{E}[Y]}{\mathbb{E}[D^*] - \mathbb{E}[D]} = \int_0^1 MTE(u)\omega_{PRTE}(u)du$$

$$\text{with } \omega_{PRTE}(u) \equiv \frac{F_P^-(u) - F_{P^*}^-(u)}{\mathbb{E}[P^*] - \mathbb{E}[P]}$$

PRTEs Between Two Counterfactual Policies

- Instead of contrasting with status quo, could have two policies:

$$D^a \equiv \mathbb{1}[U \leq p^a(X, Z^a)] \qquad \text{and } D^b \equiv \mathbb{1}[U \leq p^b(X, Z^b)]$$

$$Y^a \equiv D^a Y_1 + (1 - D^a) Y_0 \qquad \text{and } Y^b \equiv D^b Y_1 + (1 - D^b) Y_0$$

- Then define the PRTE for b relative to a as

$$PRTE_a^b \equiv \frac{\mathbb{E}[Y^b] - \mathbb{E}[Y^a]}{\mathbb{E}[D^b] - \mathbb{E}[D^a]}$$

- Derivation of the weights just requires relabeling the previous argument

The basic Roy model and selection

Model of College Education

- Suppose you are interested in the benefit of College Education ($D = 1$) relative not having College Education ($D = 0$)
- For each individual you observe realised wage:

$$Y = DY_1 + (1 - D)Y_0$$

- Where:

$$\begin{aligned} Y_1 &= X\beta_1 + U_1 \\ Y_0 &= X\beta_0 + U_0 \\ D &= \mathbb{1}(Y_1 > Y_0) \end{aligned}$$

$$\begin{pmatrix} U_1 \\ U_0 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)$$

The basic Roy model and selection

Model of College Education

- Note:

$$\begin{aligned} U_0 - U_1 \sim & \mathcal{N}(0, \sigma^2 + 1 - 2\rho\sigma) \\ \text{Cov}(U_1, U_0 - U_1) = & \rho\sigma - \sigma^2 \\ \text{Cov}(U_0, U_0 - U_1) = & 1 - 2\rho\sigma \end{aligned}$$

- Decision rule:

$$\begin{aligned} D &= \mathbb{1}(Y_1 > Y_0) \\ &= \mathbb{1}(X\beta_1 + U_1 > X\beta_0 + U_0) \\ &= \mathbb{1}(X(\beta_1 - \beta_0) > U_0 - U_1) \end{aligned}$$

- Implies:

$$\begin{aligned} \mathbb{P}(D = 1|X) &= \mathbb{P}(X(\beta_1 - \beta_0) > U_0 - U_1) \\ &= \Phi\left(\frac{X(\beta_1 - \beta_0)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right) \end{aligned}$$

The basic Roy model, selection and target parameters

College Education: Treatment parameters conditional on X

$$
\begin{aligned}
\text{ATE} &= \mathbb{E}(Y_1 - Y_0|X) = X(\beta_1 - \beta_0) \\
\text{ATT} &= \mathbb{E}(Y_1 - Y_0|X, D = 1) \\
&= \mathbb{E}(X\beta_1 + U_1 - X\beta_0 - U_0|X, X(\beta_1 - \beta_0) > U_0 - U_1) \\
&= X(\beta_1 - \beta_0) - \mathbb{E}(U_0 - U_1|X, U_0 - U_1 < X(\beta_1 - \beta_0)) \\
&= X(\beta_1 - \beta_0) \underbrace{+\sqrt{\sigma^2 + 1 - 2\rho\sigma}\frac{\phi\left(\frac{X(\beta_1-\beta_0)}{\sqrt{\sigma^2+1-2\rho\sigma}}\right)}{\Phi\left(\frac{X(\beta_1-\beta_0)}{\sqrt{\sigma^2+1-2\rho\sigma}}\right)}}_{>0}
\end{aligned}
$$

- Intuition: those who select into college benefit from it

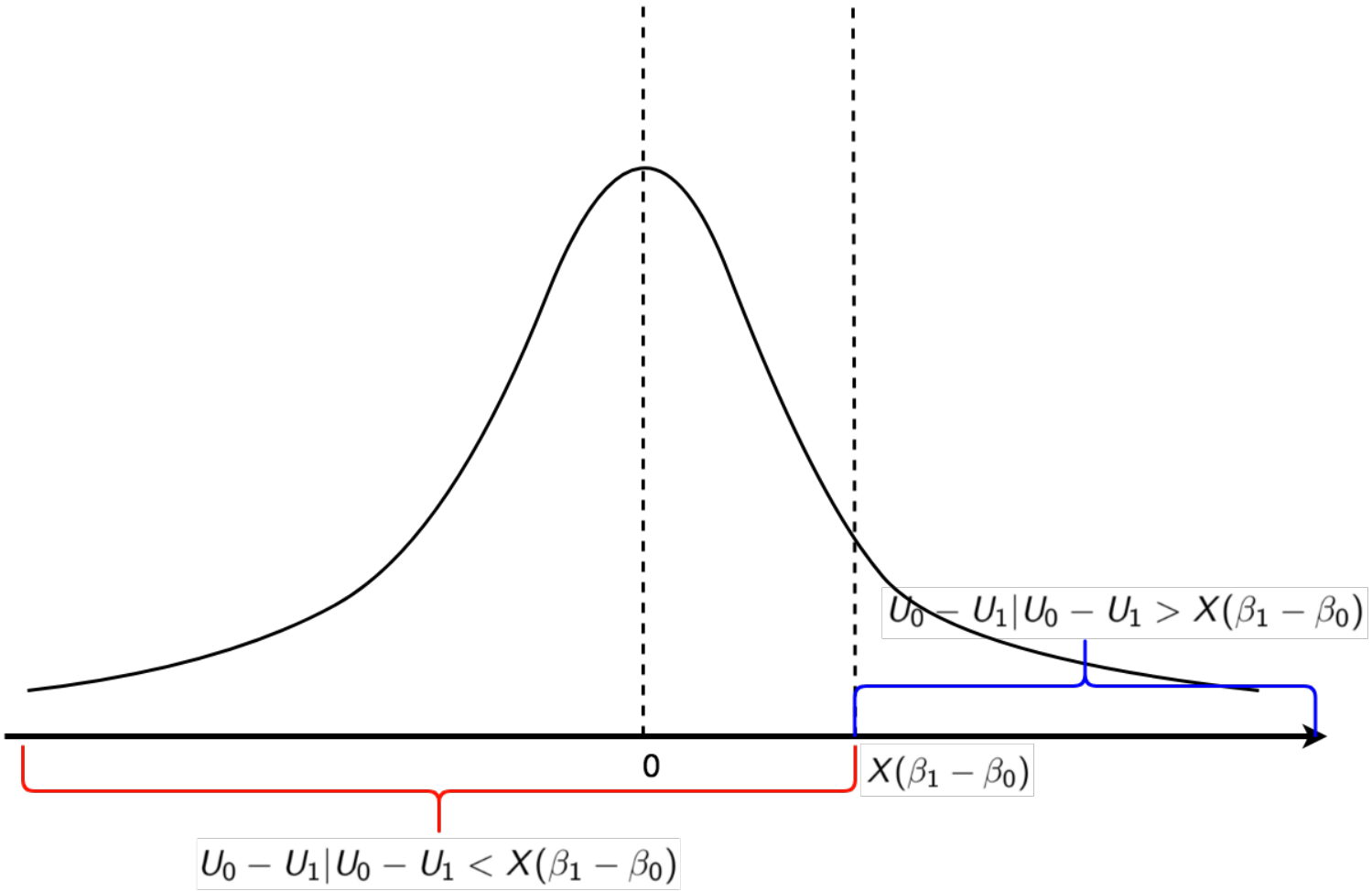
The basic Roy model, selection and target parameters

$$\begin{aligned}\text{ATU} &= X(\beta_1 - \beta_0) - \mathbb{E}(U_0 - U_1 | X, U_0 - U_1 \geq X(\beta_1 - \beta_0)) \\ &= X(\beta_1 - \beta_0) \underbrace{- \sqrt{\sigma^2 + 1 - 2\rho\sigma} \frac{\phi\left(\frac{X(\beta_1 - \beta_0)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right)}{1 - \Phi\left(\frac{X(\beta_1 - \beta_0)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right)}}_{<0}\end{aligned}$$

- Intuition: individuals do not select into college because they do not benefit from it

The basic Roy model, selection and target parameters

Graphical intuition for the sign of the selection bias (the expectation of truncated normal):



- Having defined target parameters, the next step is to think of identification

- To fix ideas, let's begin with a randomized controlled trial
- After all, it is the gold standard, isn't it.....?
  – Banerjee (2006): 'Randomized trials like these - that is, trials in which the intervention is assigned randomly - are the simplest and best way of assessing the impact of a program'
  – Imbens (2010): 'Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top'
  – Duflo (2017) refers to RCTs the 'tool of choice'

Identification, selection and missing data

- The parameter of interest is a function of the unobservables $\{Y_d\}_{d \in \mathcal{D}}$
- What could we learn about this function from the observables, (Y, D)?

- Return to the example of job training and earnings
- Suppose we care about the average effect of the program on participants:

$$\text{ATT} \equiv \mathbb{E}[Y_1 - Y_0 | D = 1] = \underbrace{\mathbb{E}[Y|D=1]}_{\text{fnc. of pop. dist.}} - \underbrace{\mathbb{E}[Y_0|D=1]}_{\text{fnc. of unobs.}}$$

- An important ingredient in a decision to continue or end the program

- The first term is a function of the population distribution
- Using the sample to understand this from data is the domain of statistics

- The question of identification is about the second term
- What can we say about $\mathbb{E}[Y_0|D=1]$ under different assumptions?
- Must answer this question before we can construct an estimate of ATT

Random assignment

- One way to learn about $\mathbb{E}[Y_0|D=1]$ is to perform a RCT
- Recall the potential outcomes model with

$$Y = \sum_{d \in \mathcal{D}} \mathbb{1}[D=d] Y_d$$

- **Random assignment** is the assumption that $\{Y_d\}_{d \in \mathcal{D}} \perp\!\!\!\perp D$
- That is, treatment state D is independent of potential outcomes

- Under random assignment, the distribution of Yd is point identified:

$$F_d(y) \equiv \mathbb{P}[Y_d \leq y] \underbrace{=}_{\text{random assignment}} \mathbb{P}[Y_d \leq y|D=d] = \mathbb{P}[Y \leq y|D=d]$$

- Any parameter that is a function of $\{F_d\}_{d \in \mathcal{D}}$ is also point identified
- Intuitively, conditioning on treatment does not change potential outcomes
  $\rightarrow$ No self-selection, sorting, correlated observables/unobservables, etc.

Random assignment

- Binary treatment is most common: $D \in \{0, 1\}$
- Typical parameters of interest:
  - Average treatment effect (ATE): $\mathbb{E}[Y_1 - Y_0]$
  - Average treatment on the treated (ATT): $\mathbb{E}[Y_1 - Y_0 | D = 1]$
  - Average treatment on the untreated (ATU): $\mathbb{E}[Y_1 - Y_0 | D = 0]$
  - Quantile treatment effect (QTE): $\mathcal{Q}_{Y_1}(t) - \mathcal{Q}_{Y_0}(t)$ for some $t \in (0, 1)$
  - QTE on the treated/untreated (QTT/QTU) defined analogously
- All point identified under random assignment
- Moreover, ATE = ATT = ATU, and QTE = QTT = QTU
- Nothing systematically different about treatment/control groups

The Fundamental Problem of Causal Inference

The problem

- Even with random assignment, joint distributions aren't point id'd
- Sometimes called the **fundamental problem of causal inference**
- Intuition is that we never see both Y_0 and Y_1 for anyone

Implications

- Most features of $Y_1 - Y_0$ are not point identified
- Even with random assignment; so therefore without it as well
- We might care about the proportion of individuals who are hurt:

$$\mathbb{P}[Y_1 - Y_0 \leq 0] \rightarrow \text{ but its not point identified!}$$

- Nor are the quantiles of $Y_1 - Y_0$
- $\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$ is an exception-linearity of expectation

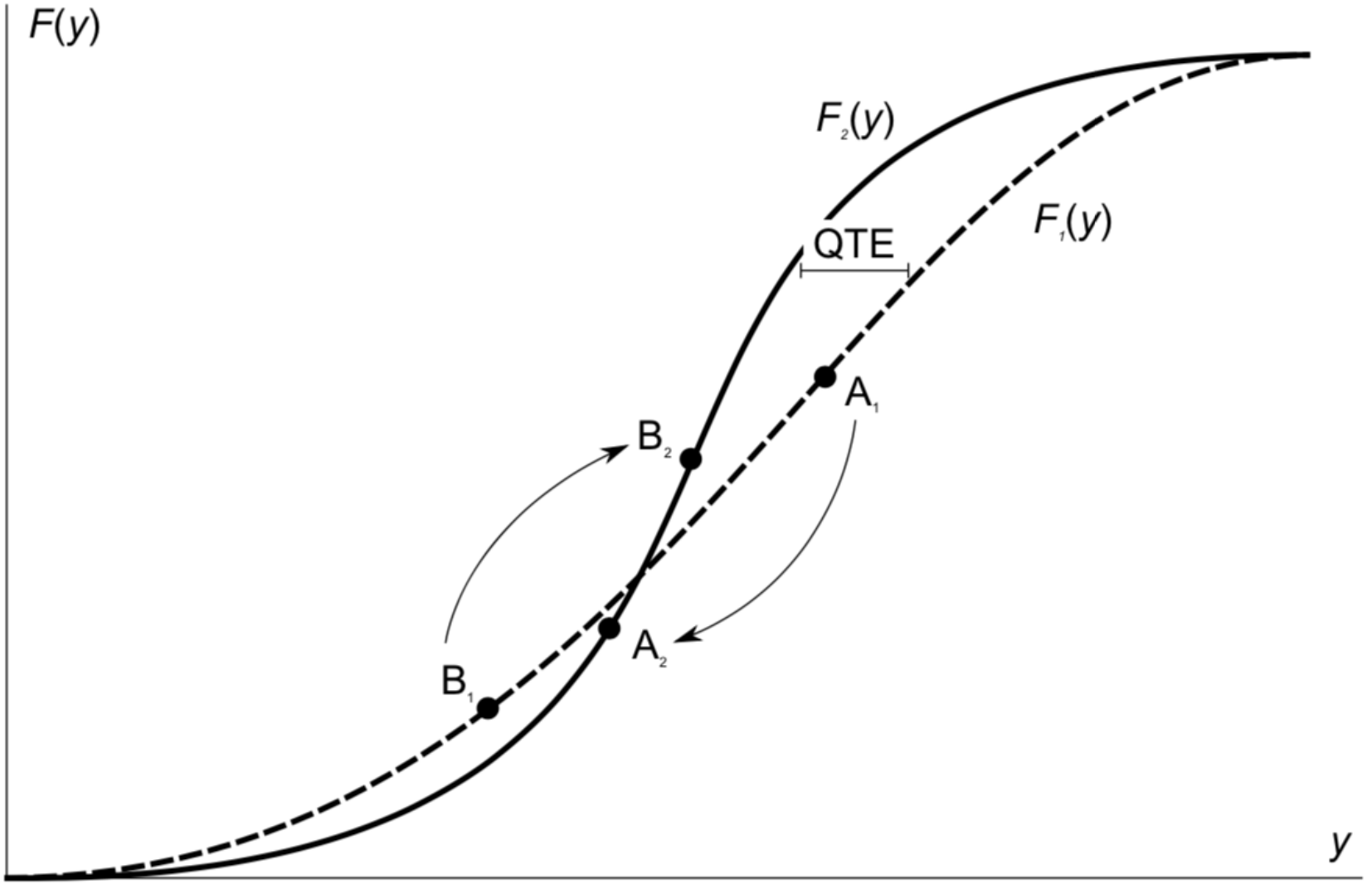
Marginal and joint potential outcome distributions

- The data obtained from an experiment consists of two marginal distributions of outcomes, $F_1(Y_1)$ and $F_0(Y_0)$
- But identification of certain parameters of interest requires knowledge of the joint distribution $F(Y_1 - Y_0)$
- For example, policymakers may care about the effect on the poor, not only the effect on the

Constant effects and rank invariance

- Common assumption in empirical research: $Y_1 - Y_0 = \Delta$ for everyone
- Then experimental data do provide the joint distribution of outcomes in the two states (can you prove this?)
- Alternatively, assumption of rank-invariance allows one to recover quantiles of the treatment effects, not only QTE

Rank invariance



Real life example: What mean impacts miss!

Jobs First program: Seminal RCT study in labor economics

- by Bitler, Gelbach, and Hoynes (2006, American Economic Review)

Random assignment to Jobs First (treatment) or AFDC (control)

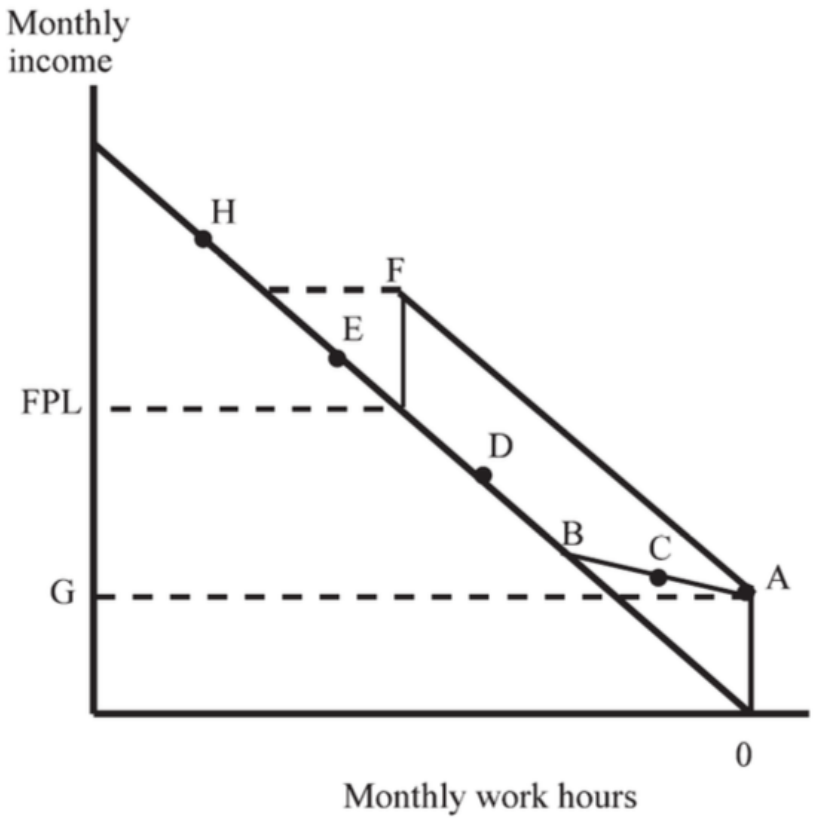
- Two counties in Connecticut: New Haven and Manchester
- Sample of about 4803 welfare recipients

Key features of Job First program:

- Expanded earnings disregard
- Introduced 21 month time limit

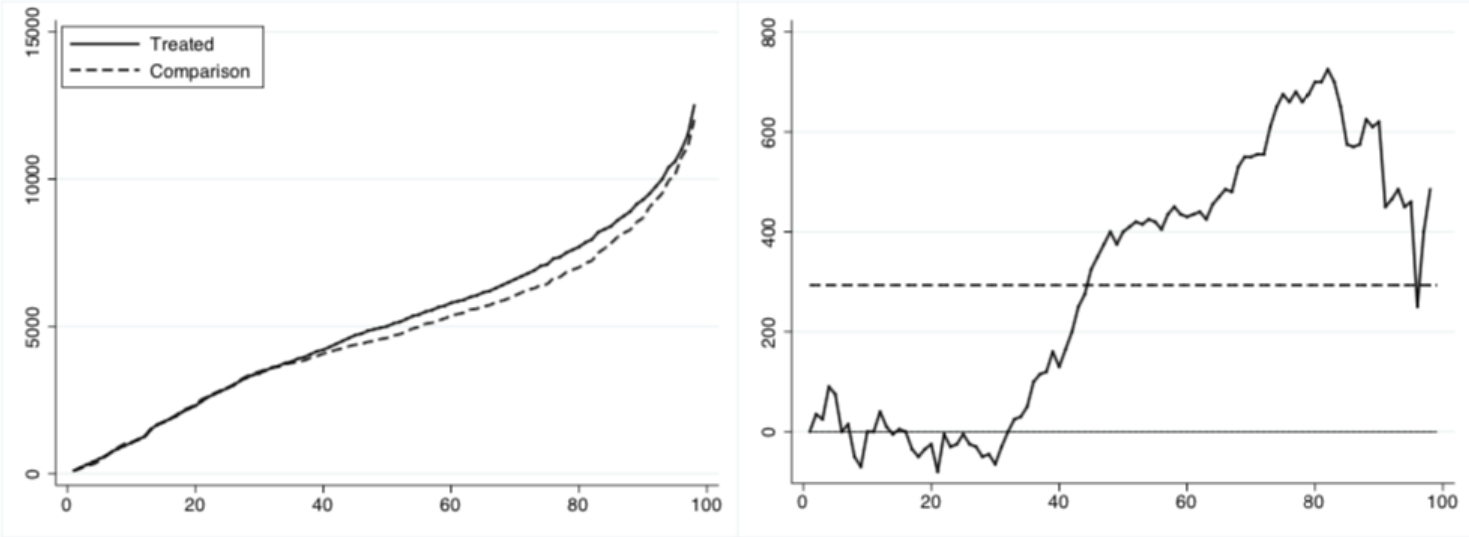
Bitler et al. use the RCT to estimate ATE and QTE

Application: Jobs First Budget constraint



AB = AFDC AF = Jobs First

QTE – average income q1-q16



- Left: The CDFs for the treatment group and the control group
  X: percentile (0-100) Y: Income in USD
- Right: The estimated QTE and ATE.
  X: percentile (0-100) Y: QTE in USD
- QTE might be missing that JF induced some women to earn above the eligibility threshold, while others reduced their earning below the threshold - opt-in behavior (Rank invariance may not hold)

Marginal vs joint: Heckman and Smith (1995)

Figure 1
A Contingency Table

		Untreated: E	Untreated: N	
Treated	E	P_{EE}	P_{EN}	$P_{E\bullet}$
	N	P_{NE}	P_{NN}	$P_{N\bullet}$
		$P_{\bullet E}$	$P_{\bullet N}$	

Constant effects assumption

- From RCT we can estimate row and column totals, giving:
- Switchers from nonemployed to employed due to treatment minus switchers from employed to nonemployed due to treatment
- But can we learn anything about whether program reduced the employment of participants?

Frechet-Hoeffding bounds applied to training programs

Table 1

Employment Percentages and Bounds on the Probabilities P_{EN} and P_{NE}

	Adult Males	*Adult Females*	*Male Youth*	*Female Youth*
% Employed: Treatment	0.72	0.64	0.74	0.57
% Employed: Control	0.71	0.61	0.77	0.58
Bounds on P_{EN}	[.01, .29]	[.03, .39]	[.00, .23]	[.00, .42]
Bounds on P_{NE}	[.00, .28]	[.00, .36]	[.03, .26]	[.01, .43]

Notes: Employment Percentages are based on percentage employed in months 16, 17 and 18 after random assignment. P_{ij} is the probability of having employment status i as a treatment and employment status j as a control, where i and j take on the values of N and E. The Frechet-Hoeffding bounds are then given by

$$P_{ij} \leq \text{FUB}\left(P_{ij}\right) = \min\left\{P_{Nj} + P_{Ej}, P_{iN} + P_{iE}\right\} \text{ and}$$

$$P_{ij} \geq \text{FLB}\left(P_{ij}\right) = \max\left\{\left[P_{Nj} + P_{Ej}\right] + \left[P_{iN} + P_{iE}\right] - 1, 0\right\}.$$

Intuition for bounds:

- Upper bound: $\mathbb{P}$ of the joint event can't exceed $\mathbb{P}$ of the events that compose it
- Lower bound: sum of the 4 individual cells must equal 1

Random Assignment and Covariates

The role of covariates

- Suppose we regress Y on D and predetermined X
- D is randomly assigned, so should be uncorrelated with X
  $\rightarrow$ Common practice to check this as a "balance test"
- Also means variation in coefficient on D will go down
- How much depends on how much X and Y are correlated

Intuitive example

- Y is earnings after the training program
- D is a binary variable for participation in training program
- X is the earnings history, before the experiment
- X probably explains a lot of the variation in Y
- Controlling for X reduces residual variation in Y (but not D)
- This may allow one to estimate the effect of D more precisely

Use and usefulness of random assignment

When is random assignment a good assumption?

- Typically, settings where agents have no control over D
- Less likely: Agents choose D without considering $\{Y_d\}_{d \in \mathcal{D}}$
- Randomized controlled experiments are the leading case
- In economics, common in lab/field experiments (development)
- Random assignment is rarely compelling with observational data
- When agents can control D, we typically expect **selection**

Is RCT the gold standard of evaluation methods?

- No! Different methods generally identify different parameters
- If possible, a RCT can be useful to identify a target parameter
- But many parameters of interest involves choices and self-selection, not simply the effect for a randomly selected person in the population