Empirical Analysis III
Spring 2019

(Magne Mogstad & James J. Heckman)

Simon Sangmin Oh

University of Chicago

Note:

These lecture notes are based on Professor Mogstad and Heckman's lectures in Empirical Analysis III, Spring Quarter.

Contents

Part I
Workhorse Models of Empirical Analysis

1 Summary

RDD vs. DiD They have a similar flavor in that you can illustrate identification with a figure, but it has some key differences. In particular, assignment to treatment is NOT random (since assignment is based on value of a running variable).

2 Randomized Control Trials and Their Limitations

2.1 Three Steps of Good Empirical Work

It is useful to separate identification from statistical inference:

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

Identification is logically the first thing to consider: if you can't recover a parameter from the population, you can't recover it from the sample.

2.1.1 Step 1: Define the target parameters

This amounts to specifying a counterfactual question. Thought experiments define the parameter of interest.

▷ For example, the following are valid target parameters:

$$\begin{aligned} \mathbb{E}\left[Y_1\right] &: \text{ average earnings if everyone were trained} \\ \mathbb{E}\left[Y_1 - Y_0\right] &: \text{ average effect of the program} \end{aligned}$$

▷ They can also be framed as questions:

* What would happen to employment if government increased minimum wages?
* What would happen to prices if two firms merged?

2.1.2 Step 2: Learn the target parameters from data (= Identification)

The target parameter is a function of the unobservables, so it naturally begs the question of what we can learn about this entity from observed data.

▷ Identification is a binary property: it is either identified or unidentified.

▷ A parameter is said to be **identified** if under the stated assumptions, alternative values of the parameter implies different distributions of observed data.

2.1.3 Step 3: Statistical Inference

In practice, we only see a finite sample of the observables. Thus, we use statistical inference – use samples to learn about the population.

2.2 Step 1: Defining Target Parameters

Formal models are useful to be precise about target parameters, identification and inference. We will use this framework to formally define the target parameters of interest.

2.2.1 Models and Notation

Different researchers use different notation and models.

1. Potential outcome model: Neyman-Fisher-Quandt-Rubin model
2. Economic choice model: Roy model and generally latent variables model

Explicit choice models can be useful to define counterfactuals and economically interpret assumptions and results.

2.2.2 Potential Outcome Notation

Setup

- ▷ $\mathcal{D}$: mutually exclusive and exhaustive set of states ("treatments")
- ▷ For each $d \in \mathcal{D}$, there is a potential outcome Y_d which is a random variable.
- ▷ We observe (1) actual state, a random variable $D \in \mathcal{D}$ and (2) outcome Y that is related to potential outcomes as:

$$Y = \sum_{d \in \mathcal{D}} Y_d 1\left\{D = d\right\} = Y_d$$

The key point is that $Y = Y_D$ is observed, but Y_d for $d \neq D$ are unobserved.

Example 2.1. (Binary Treatment) Consider the "switching regression" of the form $Y = DY_1 + (1 - D)\, Y_0$.

- ▷ Treatment D may be dependent Y_0 or Y_1.
  - * Selection Bias: D is correlated with with Y_0
  - * Selection on the gains: D is correlated with $Y_1 - Y_0$
- ▷ Model does not preclude possibility that individuals choose treatment with knowledge of (Y_0, Y_1)
- ▷ Choice equation: $D = ZD_1 + (1 - Z)\, D_0$ where Z is a binary instrument.

Target Parameters Possible target parameters include:

- ▷ What would average earnings be if everyone were trained? – $\mathbb{E}\left[Y_1\right]$
- ▷ What is the average effect of the program? – $\mathbb{E}\left[Y_1 - Y_0\right]$
- ▷ What is the average effect of the program for those who are trained? – $\mathbb{E}\left[Y_1 - Y_0 | D = 1\right]$

2.2.3 Latent Variable Notation

We can replace potential outcome equation with the latent variable notation.

- ▷ Empirical models in economics often take the form $Y = g\left(D, V\right)$ where g is a function and V are unobserved characteristics
  - * For example, earnings is a function of the program (D) and other characteristics (V)
- ▷ A **causal interpretation** of this model is implicitly $Y_d = g\left(d, V\right), \forall d \in \mathcal{D}$

  - * This is a thought experiment – for a fixed value of d, what do we get out of the model?

We can use latent variable notation to describe choices

$$D = D_1 Z + (1-Z) D_0 \quad \Leftrightarrow \quad D = 1 [U \leq \nu (W)]$$

- ▷ U is the latent variable and $\nu(W)$ is an unknown function of $W \equiv (Z, X)$ is observable with Z being the instrument and X being the covariate.
- ▷ Intuitively, people with low U are more likely to be treated. Using this intuition and the assumption that U is uniformally distributed in $[0,1]$, we can map U back to potential choices:
  - * $u < \underline{U} \Leftrightarrow D_0 = D_1 = 1$
  - * $\underline{U} \leq u \leq \bar{U}$ is equivalent to $D_1 > D_0$
  - * $u > \bar{U}$ is equivalent to $D_0 = D_1 = 0$
- ▷ Using the assumption on U, we have
$$\nu(W) = p(D|W) \equiv p(W)$$
which is referred to as the **propensity score**.

Combined with $Y = Y_1 D + Y_0 (1-D)$, we call this the **Generalized Roy Model**.

2.2.4 Roy Model

Setup A common version of the Roy model is the following:

- ▷ Outcome equations: $Y = DY_1 + (1-D) Y_0$ where $Y_0 = X'\beta_0 + V_0, \quad Y_1 = X'\beta_1 + V_1$
- ▷ Selection equation: $D = 1[U \leq W'\gamma]$

where (V_0, V_1, U) are unobservable and $W \equiv (X, Z)$ are observable.

- ▷ This framework allows for both observed and unobserved heterogeneity:
$$Y_1 - Y_0 = \underbrace{X'(\beta_! - \beta_0)}_{\text{observed}} + \underbrace{V_1 - V_0}_{\text{unobserved}}$$
- ▷ This also implies a random coefficient specification for the observed outcome:
$$\begin{aligned} Y &= DY_1 + (1-D) Y_0 \\ &= \underbrace{(V_1 - V_0)}_{\text{random coefficient}} D + X'\beta_0 + DX'(\beta_1 - \beta_0) + V_0 \end{aligned}$$

2.2.5 Marginal Treatment Effect (MTE)

We define the **marginal treatment effect (MTE)** as

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

which is the average across all individuals with $U = u$.

- ▷ $MTE(u)$ is the ATE for those agents with first stage unobservable u .

  - * Recall that u takes value between 0 and 1. Those with small u (close to 0) often choose $D = 1$, and those with large u (close to 1) rarely choose $D = 1$.

In other words, u provides a single dimension on which we can organize heterogeneity. Furthermore, using the MTE, we can back out the ATE, which is the unweighted average of the MTEs:

$$ATE = \mathbb{E}\left[\mathbb{E}\left[Y_1 - Y_0 | U\right]\right] = \int_0^1 MTE\left(u\right) du$$

since u is uniformly distributed. However, ATE is not that informative, so we are interested in ATT and the ATU:

- ▷ ATT can be written as: "What happes if I remove the program?"

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

  - * Those with low values of u – those more likely to take treatment – are more highly weighted.

- ▷ ATU can be written as: "What happens if I make it mandatory for those who did not go?"

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

  - * Those with high values of u – those less likely to take treatment – are more highly weighted.

The punchline is that different estimates have differnet implications.

2.2.6 Policy Relevant Treatment Effect

Previously, we discussed the MTE framework which partitions all agents in a clear way. But ATE or ATT may not be the target parameter of interest.

Example 2.2. Let $D \in \{0, 1\}$ be an indicator for attending a four-year college. In this case, the ATE (average effect of forcing college) is not interesting. ATT (effet on college-goers of shutting down college) may also not be interesting. We're interested in the effects via D of adjusting tuition Z.

Heckman and Vytlacil (2011) formalize this ideas a policy-relevant treatment effects (PRTE), the aggregate effect on Y of a change in the propensity score / instrument. This corresponds to a policy that affects treatment choice. More formally:

- ▷ Let $p^*\left(Z^*\right)$ be the propensity score / instrument under a new policy and D^* the treatment choice under the new policy:

$$D^* = 1\left[U \leq p^*\left(Z^*\right)\right]$$

  and the new outcome under the new policy:

$$Y^* = D^* Y_1 + \left(1 - D^*\right) Y_0$$

- ▷ Then HV define the PRTE as

$$\beta_{PRTE} \equiv \frac{\mathbb{E}\left[Y^*\right] - \mathbb{E}\left[Y\right]}{\mathbb{E}\left[D^*\right] - \mathbb{E}\left[D\right]}$$

  which is the mean effect (per net person) of the policy change.

▷ One can show (via problem set) that

$$\beta_{PRTE} = \int_0^1 MTE\left(u\right)\omega_{PRTE}\left(u\right)du$$

where

$$\omega_{PRTE} = \frac{F_P^-\left(u\right) - F_{P^*}^-\left(u\right)}{\mathbb{E}\left[P^*\right] - \mathbb{E}\left[P\right]}$$

This setup is useful because instead of contrasting with the status quo, you can compare between two policies by defining D^a, Y^a, D^b, Y^b and considering

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

Remark 2.1. (ATE, ATT, ATU, and PRTE) Essentially, the four quantities contrast in the following manner. ATE is the mean of the MTEs. ATT gives more weight to the lower values of u but still uses the full sample. ATU gives more weight to the higher values of u but still uses the full sample. PRTE uses the Z to select a subsample and computes the mean (in a ATE-like fashion) in that subsample

Example 2.3. We illustrate an example through the model of college education. Suppose D is college education, and for each individual you observe realized wage

$$Y = DY_1 + \left(1 - D\right)Y_0$$

where

$$\begin{aligned} Y_1 &= X\beta_1 + U_1 \\ Y_0 &= X\beta_0 + U_0 \\ D &= 1\left[Y_1 > Y_0\right] \end{aligned}$$

and

$$\left[\begin{array}{c} U_1 \\ U_0 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{array}\right]\right)$$

▷ The error difference is

$$U_1 - U_0 \sim \mathcal{N}\left(0, \sigma^2 + 1 - 2\rho\sigma\right)$$

with $Cov(U_1, U_0 - U_1) = \rho\sigma - \sigma^2$ and $Cov\left(U_0, U_0 - U_1\right) = 1 - 2\rho\sigma$

▷ The decision rule can be rewritten as

$$\begin{aligned} D &= 1\left[Y_1 > Y_0\right] \\ &= 1\left[X\left(\beta_1 - \beta_0\right) > U_0 - U_1\right] \end{aligned}$$

which implies

$$P\left(D = 1|X\right) = \Phi\left(\frac{X\left(\beta_1 - \beta_0\right)}{\sqrt{\sigma^2 + 1 - 2\rho\sigma}}\right)$$

Given this, we can now compute expression for the ATE and the ATT.

▷ ATE can be written as

$$\mathbb{E}\left[Y_1 - Y_0|X\right] = X\left(\beta_1 - \beta_0\right)$$

▷ ATT can be written as

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

which implies those who select into college benefit from it

▷ ATU can be written as

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

which implies those who do not select into college do so because they do not benefit from it.

2.3 Step 2: Identification

Having defined target parameters, the next step is to think of identification. To fix ideas, we will begin with randomized controlled trials which are considered the "gold standard" by many academics including Banerjee, Imbens, and Duflo.

2.3.1 Motivation

The parameter of interest is a function of the unobservables $\{Y_d\}$ and we are interested in seeing what we can learn about this function from the observables (Y, D).

Suppose we care about the average effect of the program on participants:

$$ATT = \mathbb{E}\left[Y_1 - Y_0 | D = 1\right] = \mathbb{E}\left[Y_1 | D = 1\right] - \mathbb{E}\left[Y_0 | D = 1\right]$$

▷ $\mathbb{E}\left[Y_1 | D = 1\right]$ is a function of the population distribution, which we must use the sample to understand this from data. Note that

$$\mathbb{E}\left[Y_1 | D = 1\right] = \mathbb{E}\left[Y | D = 1\right]$$

since

$$Y = \sum_{d \in \mathcal{D}} 1\left\{D = d\right\} Y_d$$

▷ $\mathbb{E}\left[Y_0 | D = 1\right]$ is a function of unobservables, which depends on our assumptions.

2.3.2 Random Assignment

One way to understand $\mathbb{E}\left[Y_0|D=1\right]$ is to perform a RCT. A random assignment is the assumption that $\{Y_d\} \perp D$ i.e. the treatment state D is independent of potential outcomes.

- ▷ In this case, the distribution of Y_d is point-identified since

$$F_d\left(y\right) \equiv P\left(Y_d \leq y\right) = P\left(Y_d \leq y|D=d\right)$$

- ▷ Generally, any parameter that is a function of $\{F_d\}$ is also point identified.
- ▷ Under random assignment, we have

$$ATE = ATT = ATU, \quad QTE = QTT = QTU$$

When is random assignment a good assumption?

- ▷ Typically settings where agents have no control over D. When agents can control D, we typically expect selection.

2.3.3 Limits of Random Assignment

Even with random assignment, joint distributions are not point identified, since we never see both Y_1 and Y_0 for anyone. For example, we might care about the proportion of individuals who are hurt:

$$P\left(Y_1 - Y_0 \leq 0\right)$$

but this is not point identified. $\mathbb{E}\left[Y_1 - Y_0\right]$ is an exception thanks to the linearity of expectation.

- ▷ More formally, the data obtained from an experiment consists of two marginal distribution of outcomes, $F_1\left(Y_1\right)$ and $F_0\left(Y_0\right)$ but the identification of certain parameters of interest requires knowledge of the joint distribution $F\left(Y_1 - Y_0\right)$

If we assume $Y_1 - Y_0 = \Delta$ for everyone, then experimental data provides the joint distribution of outcomes in the two states. Alternately, one could assume rank-invariance to recover quantiles of the treatment effects.

Example 2.4. (Heckman and Smith, 1995) From the RCT, we can estimate the row and column totals of the contingency tables, but to learn anything about whether program reduced the employment of participants, it requires additional assumption on one of the cells.

Another way to get better inference is to use natural upper and lower bounds on probabilities. For example, $\mathbb{P}$ of the joint event cannot exceed $\mathbb{P}$ of the events that compose it, and sum of the four individual cells must equal 1.

2.4 Guest Lecture: Lessons for Designing Mechanism Experiments

Leonardo Bursztyn is a rising star in experiments. He discusses the experiment looking at peer effects

2.4.1 Context

The goal of the experiment was to understand social influence. In particular, the setting was investment decisions to understand mechanisms: social learning vs. keeping up with the Jonesese. The research question is:

> Do peers' decisions have a causal impact on investment choices? If so, through which channel?

The channels differ in the following respect:

- ▷ Social learning: change in beliefs that results from observing another's revealed preference
- ▷ Social utility: direct effect on one's utility caused by another person's possession of an asset

There are numerous examples in economics – this experiment specifically looks at a financial application.

2.4.2 Identification Challenges

Bottomline: identifying peer effects is notoriously difficult.

- ▷ One needs an exogenous variation in a peer's purchase of an asset to identify the peer effect.

Identifying the *channel* through which peer effects work is even more difficult for two reasons:

1. Revealed preference decision of the peer to purchase the asset
2. Peer's possession of the asset

Observational data is ill-suited for this purpose since revealed preference and possession are typically observed together. In other words, we want to separate the willingness to pay and the actual purchase of the asset.

2.4.3 Experiment

Field experiment that identifies peer effects and disentangles the channels. The key ingredient is the lottery that determines whether individuals who choose to purchase an asset are actually allowed to make the investment.

- ▷ If lottery determines that Magne cannot buy the product, this would be the social learning effect.
- ▷ If lottery determines that Magne can buy the product, this would be social learning + social utility effect.

What this *cannot* do is just separating the social utility effect. One way to do this is to randomly give out assets to people.

Requirements for the New Asset There are a bunch of conceptual requirements that this asset needs to fulfill. Examples include:

1. There needs to be a possibility of learning.
2. Sufficient demand for the asset
3. Limited entry to the fund to justify a lottery
4. No other opportunity to purchase the asset
5. No secondary market
6. No joint learning

Permutation Tests When you don't have much sample size, many people ask for different ways of dealing with inference. Permutation tests are becoming more common. The idea is to randomly assign the treatment and computing the treatment effects to get a distribution of treatment effects. In other words, instead of testing:

$$\mathbb{E}\left[Y_1 - Y_0\right] \quad vs. \quad 0$$

we are now testing

$$Y_1 - Y_0 \quad vs. \quad 0$$

and this is increasingly becoming more popular. The price you pay is that the null hypothesis is stronger.

Evaluating Alternative Hypotheses Potential alternative hypotheses include: possibility of side payments, desire to match peer's inferred portfolio, effets of the lottery to authorize investments, changes in supply-side behavior. This becomes a central part of the publication process.

2.4.4 Takeaways

The focus here was to use field experiments to "qualitatively" test theory. Note that field experiments are very distinct from:

- lab experiments: issues of external validity, sample representativeness, stakes, awareness of intervention, and distorted magnitude of mechanisms
- structural estimation + experiments: interesting and popular now, but issues of scale are important: how much faith can we have in a specific number coming from a specific setting with small sample?

3 Selection on Observables

3.1 Conditional Independence & Controls

We focus on observational data – a single cross-section – and think about different assumptions that guide our inference.

3.1.1 Selection in Treatment

There is **selection into treatment state** D if $Y_d|D=d$ is distributed differently from $Y_D|D=d'$ for $d' \neq d$.

- ▷ This is not the case under the random assignment assumption, and it is expected to occur if agents choose D with knowledge of $\{Y_d\}_{d\in\mathcal{D}}$

Naturally, we can define selection bias as:

$$\begin{aligned}\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0] &= \underbrace{(\mathbb{E}[Y_1|D=1]-\mathbb{E}[Y_0|D=1])}_{ATT}+\underbrace{(\mathbb{E}[Y_0|D=1]-\mathbb{E}[Y_0|D=0])}_{\text{Selection Bias}} \\ &= \underbrace{(\mathbb{E}[Y_1|D=0]-\mathbb{E}[Y_0|D=0])}_{ATU}+\underbrace{(\mathbb{E}[Y_1|D=1]-\mathbb{E}[Y_1|D=0])}_{\text{Selection Bias}}\end{aligned}$$

3.1.2 Selection on Observables

A simple relaxation of random assignment is **selection on observables**.

- ▷ Suppose we observe (Y, D, X) where X are covariates.
- ▷ The seletion on observables assumption is that

$$\{Y_d\}_{d\in\mathcal{D}} \perp D|X$$

  i.e. conditional on X, the treatment is as-good-as randomly assigned.

3.1.3 Conditional Independence

When might conditional independence hold?

- ▷ You have detailed information about the assignment mechanism
- ▷ You have very rich data: personal traits, histories, etc

For selection on observables to be plausible, X should be predetermined.

- ▷ If we accidentally include Y as part of X, then clearly we will not have $(Y_1, Y_0) \perp D|X$.
- ▷ This is why you shouldn't include earnings after the program in X.

Problem 3.1. What happens if your controls are affected by treatment? (Neale and Johnson, JPE 1996)?

Solution. In their paper, they examine the role of premarket facts in Black-White wage gap by regressing adult earnings (Y) on race $(D = 1$ if white) and some covariates. A key finding is that

$$\mathbb{E}\left[Y|D=1\right]-\mathbb{E}\left[Y|D=0\right]\gg\mathbb{E}\left[Y|D=1,T=1\right]-\mathbb{E}\left[Y|D=0,T=1\right]$$

which authors interpret to mean that conditioned on pre-market human capital, the racial wage gap is small. However, this can happen due to selection bias since the RHS can re-expressed as:

$$\begin{aligned}
&\mathbb{E}\left[Y|D=1,T=1\right]-\mathbb{E}\left[Y|D=0,T=1\right]\\
=&\mathbb{E}\left[Y_1|D=1,T=1\right]-\mathbb{E}\left[Y_0|D=0,T=1\right]\\
=&\mathbb{E}\left[Y_1|T_1=1\right]-\mathbb{E}\left[Y_0|T_0=1\right]\qquad(\because)\,D\perp(Y_1,Y_{0,}T_1,T_0)\\
=&\quad\underbrace{\mathbb{E}\left[Y_1-Y_0|T_1=1\right]}_{\text{causal effect conditional on test score}}+\underbrace{\left(\mathbb{E}\left[Y_0|T_1=1\right]-\mathbb{E}\left[Y_0|T_0=1\right]\right)}_{\text{selection bias}}
\end{aligned}$$

so if selection bias is sufficiently negative, we can have their result. It would be negative if blacks need to be "smarter" to achieve the same test score. To be fair, the authors do have a section explaining that test score is unaffected by race.

Problem 3.2. Does bias go down if you control for more? (Heckman and Navarro-Lozano, 2004)).

Solution. Adding information need not reduce bias.

3.2 Propensity Score Matching

Using a propensity score is a very common practice. It can help us reduce dimensionality.

3.2.1 Propensity Score

Consider the binary treatment case $D \in \{0,1\}$. The **propensity score** is defined as

$$p(x)\equiv P\left[D=1|X=x\right]$$

and further define P be the random variable $P\left[D=1|X\right]$.

▷ Rosenbaum and Rubin (1983) argue that selection on observables implies $(Y_0,Y_1)\perp D|p(X)$ which implies that we can condition on $p(X)$ instead of X.

This allows us to rewrite the traditional target parameters as a weighted average of Y:

$$ATE(x)=\mathbb{E}\left[\frac{Y(D-p(x))}{p(x)(1-p(x))}|X=x\right]$$

and averaging over x to get:

$$ATE=\mathbb{E}\left[\frac{Y(D-p(X))}{p(X)(1-p(X))}\right]We$$

We can derive analogous expression for ATT and ATU.

Problem 3.3. Why do we have different identification arguments?

Solution. So far, we have seen three identification results for ATE: match on X, match on P, weight using p, each of which shows that ATE is point identified and derived under the same assumptions. The three different estimators may have different properties: efficiency, rates of convergence, finite sample performance, etc.

3.2.2 Curse of Dimensionality

Matching requires estimation of $\mathbb{E}\left[Y|D=0,X\right]$, which is not feasible if we want to do this non-parameterically. To overcome this, people have suggested regression analysis, matching on propensity score, propensity score weighting, and inexact matching.

3.2.3 Matching vs. Linear Regression

Consider the following saturated-in-X_i regression model:

$$\begin{aligned} Y_i &= \sum_{k=1}^{K} d_{ik}\alpha_k + \beta D_i + U_i \\ d_{ik} &= 1\left[X_i = x_k\right] \end{aligned}$$

Then it can be shown that

$$\begin{aligned} \beta &= \sum_{k=1}^{K} \left\{\mathbb{E}\left[Y_i|D_i=1,X_i=x_k\right] - \mathbb{E}\left[Y_i|D_i=0,X_i=x_k\right]\right\} w_k \\ w_k &= \frac{P\left[D_i=1|X_i=x_k\right]\left(1-P\left[D_i=1|X_i=x_k\right]\right)P\left(X_i=x_k\right)}{\sum_{k=1}^{K}\left[D_i=1|X_i=x_k\right]\left(1-P\left[D_i=1|X_i=x_k\right]\right)P\left(X_i=x_k\right)} \end{aligned}$$

i.e. the weights depend on the conditional variance of treatment status.

Matching and ATT Suppose X takes on values $x_1, ..., x_K$. Then the matching estimator $\hat{\beta}^{ATT}$ can be written as

$$\sum_{k=1}^{K} \left\{\mathbb{E}\left[Y_i|D_i=1,X_i=x_k\right] - \mathbb{E}\left[Y_i|D_i=0,X_i=x_k\right]\right\} P\left[X_i=x_k|D_i=1\right]$$

Essentially we are partitioning the treated and control sample in K cells by X, calculating the mean outcome difference in each cell, take a weighted average of the mean differences where the weight is the fraction of treated observations in each cell as the weights.

Matching and ATE Similarly, we have :

$$\hat{\beta}^{ATE} = \sum_{k=1}^{K} \left\{\mathbb{E}\left[Y_i|D_i=1,X_i=x_k\right] - \mathbb{E}\left[Y_i|D_i=0,X_i=x_k\right]\right\} P\left[X_i=x_k\right]$$

Essentially we are partitioning the treated and control sample in K cells by X, calculating the mean outcome difference in each cell, take a weighted average of the mean differences where the weight is the fraction of total sample in each cell as the weights.

3.2.4 Propensity Score Matching

We are now interested in estimating $\mathbb{E}\left[Y|D=0,p\left(X\right)\right]$.

1. Estimate $p\left(X\right)$.
   - Choice of X: guided by economic theory, a priori considerations, institutional set-up
   - Estimation: probit or logit

2. Match treated to controls based on some metric

3. Check common support

4. Check balancing of X's on common support

    ▷ If it holds, then compute the treatment effect
    ▷ If not, reiterate until the balancing condition holds.

Matching There is an entire class of matching estimators:

▷ Traditional matching estimators: one-to-one (nearest neighbor) matching

  * To each treated unit, match only one non-treated unit
  * You could do it with or without replacement

▷ Simply smoothed matching estimators: K-nearest neighbor

▷ Weighted smoothed matching estimators

  * Kernel matching: the matched outcome for treated i is a kernel-weighted average of the non-treated outcomes
  * Local linear regression matching: For each treated i, estimate $\hat{Y}_0 \equiv \mathbb{E}\left[Y|D=0, p\left(X\right)=p\left(X_i\right)\right]$ non-parametrically by

    - Fit a line estimated on a local neighborhood of $p\left(X_i\right)$
    - Apply a weighting scheme with weights:

$$\min_{\theta_0,\theta_1} \sum_{j\in C^0} \left(Y_j-\theta_0, \theta_1\left(p\left(X_i\right)-p\left(X_j\right)\right)\right)^2 K\left(\frac{p\left(X_i\right)-p\left(X_j\right)}{h}\right)$$

3.2.5 Mahalanobis-metric Matching

This is an alternative to Mahalanobis-metric matching.

▷ Combine the Xs into a distance measure and then match on the resulting scalar:

$$d\left(i,j\right)=\left(X_i-X_j\right)^T V^{-1}\left(X_i-X_j\right)$$

with V pooled within-sample covariance matrix of X

3.3 Many Ways to Skinning the Statistical Cat

There are many ways to implement selection on observables.

3.3.1 Discrete Non-Parametric Estimation: Binning Estimator

Suppose X is discrete. Then a non-parametric binning estimator is very natural:

$$\hat{\mu}_d\left(x\right)=\frac{1}{N_{d,x}}\sum_{i:D_i=d,X_i=x}^{N} Y_i, \qquad N_{d,x}=\sum_{i=1}^{N} 1\left\{D_i=d|X_i=x\right\}$$

Limitations include:

▷ It only works if X is discrete

▷ Poor finite sample performance if small bins

3.3.2 Non-parametric Smoothing Regression

Two main approaches in smoothing include:

1. Kernel regression: Take a sample mean of Y over $D_i = d, X_i \in [x-h, x+h]$ where $h > 0$ is a bandwidth parameter to be chosen
   - ▷ Bias goes down as $h \to 0$ and variance goes up
2. Series/sieve approximation: The idea is to write $\mu_d(x) = \sum_{k=1}^{K} \theta_k b_k(x)$ for some basis function b_k
   - ▷ For example, regress Y on $1, X, X^2, \cdots$
   - ▷ Bias goes down as we include more terms, and variance goes up

In these approaches, curse of dimensionality kicks in again.

- ▷ Estimator quality rapidly deteriorates with the dimension of X
- ▷ Formally, the rate of convergence of the estimators goes down
- ▷ Each $\hat{\mu}_d$ will have a slow rate of convergence, but $A\hat{T}E$ can still be $\sqrt{N}$ (intuition is that by averaging, we are again using all N observations)

Smoothing with the propensity score:

- ▷ Idea: Instead of estimating $\mu_d(x)$, we could estimate

$$\nu_x(p) \equiv \mathbb{E}[Y|D=d, P=p]$$

  but now we have to estimate p non-parametrically, which is the same issue

- ▷ Parametric p is arguably better than parametric μ_d: p is just a matter of fit, whereas μ_d is a counterfactual object

3.3.3 Blocking

Blocking (subclassification) is a type of non-parametric smoothing regression.

1. Divide $[0,1]$ into $\{b_0, b_1, ..., b_J\}$ and define $B_j = 1$ if $p(X) \in (b_{j-1}, b_j)$
2. Estimate ATE_j per block and average ATE_j by block size to get ATE

The key question is how to construct the blocks. Imbens (2015) suggests combining blocking with linear regression: construct the blocks, then run a linear regression separately in each block, and then average up

- ▷ This could potentially reduce both bias and variance. Variance is obvious (since accounting for X reduces variance) but bias less so.

3.4 Relevant Papers

3.4.1 LaLonde (1986)

This paper sparked a big debate whether matching works or not. They wanted to know the impact of being offered training $(D=1)$ on applicants. In the NSW data, applying/being offered job training is randomly assigned. Suppose, however, that we did not have this, in which case we need to infer $\mathbb{E}[Y_0|D=1]$. The author tries this with PSID/CPS using a variety of different methods, and they find that they cannot get close to the experimental benchmark.

3.4.2 Dehejia and Wahba (1999)

The authors repeat the exercise using a huge array of fancy selection on observable methods that came later, and they find that selection on observables seem to "work well" as in matching the experimental benchmark.

3.4.3 Heckman et al. & Other Studies

How much should we trust DW's findings?

- ▷ Heckman et al. (1997a) and Heckman et al. (1996, 1998a) conclude that for matching estimators to have low bias, it is important that the data include:
    - * variables affecting program part. and labor market outcomes
    - * non-experimental comparison group be drawn from the same local labor markets as the participants
    - * Comparable data for participants and non-participants
- ▷ All these conditions fail to hold in the NSW Data

3.4.4 Smith and Todd (2005)

The authors argue that DW results are not robust.

- ▷ Changing sample from LaLonde (1986) is important
- ▷ Income is measured differently in NSW/CPS

3.4.5 Fagerang et al. (2019)

They are interested in why wealthy parents have wealthy children.

- ▷ There are multiple possible causal channels.
- ▷ Authors investigate the role of family background in determining children's wealth accumulation and investor behavior by linking Korean-born children who were adopted at infancy by Norwegian parents.

They argue selection on observables by pointing out that conditional on time of adoption, the assignment is as good as random.

- ▷ As part of this exercise, they regress gender & age at adoption on a bunch of regressors.

3.4.6 Angrist. (1988)

This is a bad use of selection on observables.

- ▷ Setting: Y is a labor market outcome (employment / earnings) and D is a veteran status. X are socioeconomic variables.
- ▷ Selection on Observables: Given X, military participation is as-if randomly assigned.
- ▷ Why is this problematic? It's unlikely that observationally similar people randomly join the military. It ignores first-order issues such as outside employment options which are unobserved.

3.5 Concluding Comments

1. Arguing for selection on observables is difficult.

    ▷ Many of these are unobservables: preferences, private information, expectations...

    ▷ A large X does not make selection on observables more likely. Even if it were, it raises an uncomfortable friction with overlap, since if we can perfectly explain D with X then $P\left[D=1|X\right]=0$ or 1.

2. To argue for selection on observables, you need to say:

    ▷ Obervationally identical people behave differently due to A, and A is like a coin flip because of B.

3. Alternatively, we can find an instrument or make a functional form / parameteric assumptions about the relationship between the observed data and the unobservables.

    ▷ Heckman (2008) shows that under a specific parameteric assumption, it's not necessary to observe a variable in order to compute its conditional expectation with respect to another variable.

4 Instrumental Variables

4.1 LATE and Instruments

We will discuss how instruments are useful in identiying LATE.

4.1.1 Defining Instruments

Consider the heterogeneous potential outcome setup, but this time we have a Z which affects the variable of interest D which in turn affects Y. We call Z the instrument variable if the following four assumptions hold:

1. Random assignment: $Y_{d,z}, D_z \perp Z$ for all d, z

    ▷ Note that this is sufficient to identify average causal effect of Z on Y (and of Z on D).

2. Exclusion restriction: $Y_{d,1} = Y_{d,0} = Y_d$ i.e. any effect of Z on Y must be via an effect of Z on D.

    ▷ And this tells us that the causal effect of Z on Y is only due to the effect of Z on D.

3. Monotonicity: $D_1 \geq D_0$ or vice versa i.e. all those affected by the instrument are affected in the same direction.

    ▷ This is needed to avoid offsetting effects.

4. First-stage: $\mathbb{E}\left[D_1 - D_0\right] = 0$ i.e. the instrument Z needs to have some effect on the average probability of treatment.

    ▷ This is needed to have treatment variation in the sample.

Then the Wald estimand gives the Local Average Treatment Effect

$$\beta_{IV} = \mathbb{E}\left[\beta | D_1 = 1, D_0 = 0\right]$$

which is the average treatment effect for those affected by the instrument.

4.1.2 Interpreting LATE

We can divide the population ito four groups:

1. Compliers: $D_1 = 1$ and $D_0 = 0$
2. Always-takers: $D_1 = 1$ and $D_0 = 1$
3. Never-takers: $D_1 = 0$ and $D_0 = 0$
4. Defiers: $D_1 = 0$ and $D_0 = 1$

This helps us refine our understanding of LATE.

▷ We can show that the Wald estimand is equal to

$$\frac{\mathbb{E}\left[Y | Z = 1\right] - \mathbb{E}\left[Y | Z = 0\right]}{\mathbb{E}\left[D | Z = 1\right] - \mathbb{E}\left[D | Z = 0\right]}$$

* The average causal effect of Z on Y can be written as weighted average of the causal effects of the four subpopulations:

$$\mathbb{E}\left[Y_{D_1} - Y_{D_0}\right] = \sum_{\text{types}} \mathbb{E}\left[Y_{D_1} - Y_{D_0} | \text{ type }\right] P\left(\text{type}\right)$$

  * Only the compliers remain since:
    - Defiers are ruled out by monotonicity.
    - Always-takers and Never-takers have $Y_{D_1} - Y_{D_0} = 0$
  * From independence, we also have $\mathbb{E}\left[Y_{D_1} - Y_{D_0}\right] = \mathbb{E}\left[Y|Z=1\right] - \mathbb{E}\left[Y|Z=0\right]$. Thus it follows that

$$\mathbb{E}\left[Y|Z=1\right] - \mathbb{E}\left[Y|Z=0\right] = \mathbb{E}\left[Y_{D_1} - Y_{D_0}|\text{ compliers}\right]$$

  * Furthermore, by independence and monotonicity, we can show

$$\mathbb{E}\left[D|Z=1\right] - \mathbb{E}\left[D|Z=0\right] = P\left(D_1 = 1, D_0 = 0\right)$$

▷ Thus with heterogeneous effects, IV estimates the average causal effect for compliers.

We can't identify the compliers because we never observe D_0 and D_1, but we can describe them:

$$\begin{aligned}\frac{P\left(X=x|D_1 > D_0\right)}{P\left(D=1\right)} &= \frac{P\left(D_1 > D_0|X=x\right)}{P\left(D_1 > D_0\right)} \\ &= \frac{\mathbb{E}\left[D|Z=1, X=x\right] - \mathbb{E}\left[D|Z=0, X=x\right]}{\mathbb{E}\left[D|Z=1\right] - \mathbb{E}\left[D|Z=0\right]}\end{aligned}$$

4.2 Extensions

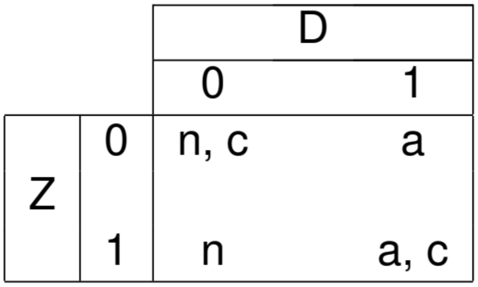
Previously, we saw that we can estimate the average causal effect for compliers. But can we do more?

4.2.1 Counterfactual Distributions

Imbens & Rubin (1997) show that we can estimate more than average causal effects for compliers. In fact, they show how to recover the complete marginal distributions of the outcome under different treatment cases. Specifically, we know that

$$\begin{aligned}P\left(D=1|Z=0\right) &= p_a \\ P\left(D=0|Z=1\right) &= p_n\end{aligned}$$

And since there are no defiers, we know $p_c = 1 - p_a - p_n$. Graphically:



		D	
		0	1
Z	0	n, c	a
	1	n	a, c

Now denote $f_{zd}\left(y\right)$ as the observed marginal distribution of Y conditional on D and Z:

$$f_{zd}\left(y\right) \equiv f\left(y|Z=z, D=d\right)$$

Then we can map $f_{zd}(y)$ to the marginal distributions of each type:

$$\begin{aligned}
f_{10}(y) &= g_n(y) \\
f_{01}(y) &= g_a(y) \\
f_{00}(y) &= g_{c0}(y)\frac{p_c}{p_c+p_n} + g_n(y)\frac{p_n}{p_c+p_n} \\
f_{11}(y) &= g_{c1}(y)\frac{p_c}{p_c+p_a} + g_a(y)\frac{p_a}{p_c+p_a}
\end{aligned}$$

where $g_{c0}(y)$ and $g_{c1}(y)$ are the counterfactual distributions for the compliers. Note that we have $g_{n1} = g_{n0}$ and $g_{a0} = g_{a1}$ since they are not affected by the instrument.

▷ We can rearrange the above equations to back out the counterfactual distributions for the compliers:

$$\begin{aligned}
g_{c0}(y) &= f_{00}(y)\frac{p_c+p_n}{p_c} - f_{10}(y)\frac{p_n}{p_c} \\
g_{c1}(y) &= f_{11}(y)\frac{p_c+p_a}{p_c} - f_{01}(y)\frac{p_a}{p_c}
\end{aligned}$$

We can use the above discussion as a test for instrument validity: under the IV assumptions, the complier distribution should actually be a distribution.

▷ Kitagawa (2015) develops a formal statistical test based on these implications.

4.2.2 Multiple Instruments

Suppose you have two mutually exclusive instruments and run 2SLS to obtain

$$\beta_{2SLS} = \frac{Cov\left(Y, \hat{D}\right)}{Cov\left(D, \hat{D}\right)}, \quad \hat{D} = \pi_1 Z_1 + \pi_2 Z_2$$

Expanding this gives

$$\begin{aligned}
\beta_{2SLS} &= \pi_1 \frac{Cov(Y, Z_1)}{Cov\left(D, \hat{D}\right)} + \pi_2 \frac{Cov(Y, Z_2)}{Cov\left(D, \hat{D}\right)} \\
&= \psi \beta_{Z_1} + (1-\psi)\beta_{Z_2}
\end{aligned}$$

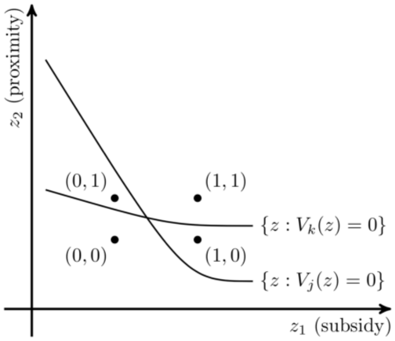
where

$$\psi \equiv \frac{\pi_1 Cov(D, Z_1)}{\pi_1 Cov(D, Z_1) + \pi_2 Cov(D, Z_2)}$$

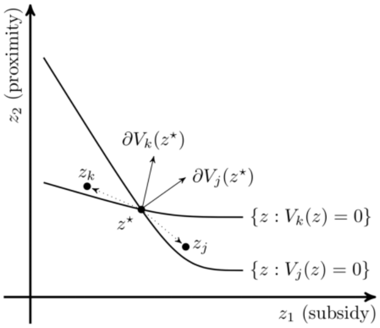
i.e. the relative strength of Z_1 in the first stage. Thus, the 2SLS estimate is an instrument-strength weighted average of the instrument specific LATEs.

▷ How should we interpret monotonicity with multiple instruments?

- * Specifically, we require for all $z, z' \in Z$ that either $D_z \geq D_{z'}$ or $D_z \leq D_{z'}$.
- * For example, IA monotonicity does not permit individuals to differ in responses. Thus we cannot get a graph that looks like this for binary instruments:



* For continuous instruments, we require that the marginal rates of substitution to be the same:



4.2.3 Variable Treatment Intensity

Assume treatment is no longer binary but varies in its level.

▷ We need to define potential outcomes indexed by the level of treatment (Y_S, S_Z).

▷ This allows us to write the obseved outcome as the following:

$$Y = \sum_{s=0}^{J} Y_s \mathbb{I}\{S = s\} = Y_0 + \sum_{s=1}^{J} (Y_s - Y_{s-1}) \mathbb{I}\{S \geq s\}$$

which implies that the average effect of the sth year of schooling is then $\mathbb{E}[Y_s - Y_{s-1}]$ and we have J treatment effects.

We can proceed similarly as before.

▷ For three treatment intensities, the Wald estimate turns out to be

$$\frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[S|Z=1] - \mathbb{E}[S|Z=0]} = \sum_{s=1}^{J} \omega_s \mathbb{E}[Y_s - Y_{s-1} | S_1 \geq s > S_0]$$

where

$$\omega_s = \frac{Pr(S_1 \geq s > S_0)}{\sum_{j=1}^{J} Pr(S_1 \geq j > S_0)}$$

Angrist and Imbens call this the average causal response (ACR).

- ▷ We cannot estimate $\mathbb{E}\left[Y_s - Y_{s-1}|S_1 \geq s > S_0\right]$ for the different local complier groups, but we can estimate their weights in the ACR (the numerator) since

$$\begin{aligned} P\left(S_1 \geq s > S_0\right) &= P\left(S_1 \geq s\right) - P\left(S_0 \geq s\right) \\ &= P\left(S_0 < s\right) - P\left(S_1 < s\right) \\ &= P\left(S < s|Z=0\right) - P\left(S < s|Z=1\right) \end{aligned}$$

- ▷ Important note here is that the averaging takes place across potentially overlapping segments. We cannot express this as a positive weighted average of causal effects across mutually exclusive groups.

Example: Angrist & Krueger (1991) Angrist & Krueger (1991) uses quarter of birth as an instrument for schooling.

- ▷ $D=1$ of education is at least high school, and $Z=1$ if born in the 4th quarter, $Z=0$ if born in the 1st quarter.

4.3 Covariates

Often one wants covariates X to help justify the exogeneity of Z, reduce residual noise in , and/or look at observed heterogeneity in treatment effects.

4.3.1 Instrument Conditions

We can correspondingly adjust the assumptions to be conditional on X:

- ▷ Exogeneity: $\left(Y_0, Y_1, D_0, D_1\right) \perp Z|X$
- ▷ Relevance: $P\left[D=1|X, Z=1\right] \neq P\left[D=1|X, Z=0\right]$ a.s.
- ▷ Monotonicity: $P\left[D_1 \geq D_0|X\right] = 1$ a.s.
- ▷ Overlap: $P\left[Z=1|X\right] \in (0,1)$ a.s.

There are multiple ways to do this.

1. Non-parametric IV with Covariates.

    - ▷ Suppose we can estimate stratified LATEs:

    $$\beta\left(x\right) = \cdots = \mathbb{E}\left[Y_1 - Y_0|D_1 - D_0 = 1, X = x\right]$$

    We have to go from here to some population averaged LATE using some weights.

2. 2SLS Regression with Covariates

    - ▷ Consider the following staurated 2SLS estimation:

    $$\begin{aligned} Y &= \beta D + \alpha_X + \epsilon \\ D &= \pi_X Z + \gamma_X + u \end{aligned}$$

    i.e. x-dummies in both stages and π_x is the size of the first stage for given value of x.

- ▷ Angrist & Imbens (1985) show that $\beta = \mathbb{E}\left[\beta(x)\,\omega(x)\right]$ where $\beta(x)$ is the x-specific LATE and

$$\omega(x) = \frac{\sigma^2_{\hat{D}}(x)}{\mathbb{E}\left[\sigma^2_{\hat{D}}(x)\right]} = \frac{\pi_x^2 \sigma_Z^2(x)}{\mathbb{E}\left[\pi_x^2 \sigma_Z^2(x)\right]}$$

  i.e. the weighting depends on the square of the local complier share and instrument variance.

3. Abadie's (2003) κ

- ▷ This is a more elegant approach. The idea is to run regressions only on the compliers.
- ▷ Abadie shows that for any function $G = g(Y, X, D)$, we have

$$\mathbb{E}\left[G|T=c\right] = \frac{1}{P\left[T=c\right]}\mathbb{E}\left[\kappa G\right], \qquad \kappa = 1 - \frac{D(1-Z)}{P\left[Z=0|X\right]} - \frac{Z(1-D)}{P\left[Z=1|X\right]}$$

- ▷ Intuition: On average, κ only applies positive weights to compliers, so on average κG is only positive for compliers.
- ▷ To implement this, one must estimate κ i.e. $P(Z=1|X)$.
  - * If $P(Z=1|X)$ is linear, then the following κ-weighted regression

$$\min_{\alpha,\beta} \mathbb{E}\left[\kappa\left(Y - \alpha D - X'\beta\right)^2\right]$$

    equals TSLS.

Example: Angrist and Evans (1998) They are interested in the relationship between fertility decisions and female labor supply.

- ▷ Setup: Y is a labor market outcome for the woman and D is an indicator for having more than 2 children (as opposed to just two, since they restrict the sample to only women). $Z = 1$ if the first two children had the same sex.
- ▷ Assumptions:
  - * Exogeneity: This requires the assumption that sex at birth is randomly assigned. Authors also conduct balance tests to support this.
  - * Monotonicity: This restricts preference heterogeneity in unattractive ways (some families may want two boys or girls then stop)
- ▷ They find that OLS is quite different from IV (consistent with endogeneity).

Interestingly, Abadie's κ method results in estimates similar to TSLS. But this is not always true, so it's weird that they use this to promote TSLS.

4.4 Multiple Unordered Treatments

In many cases, individuals are often choosing between multiple unordered treatments: education types, occupations, locations etc. What does 2SLS identify in this case?

4.4.1 Example with 3 Field Choice

Suppose students are choosing between three fields $D \in \{0,1,2\}$ and we are interested in interpreting the IV estimates of

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \epsilon$$

where Y is observed earnings. Furthermore, individuals are assigned to one of three groups $Z = \{0,1,2\}$ which allows us to write:

$$\begin{aligned} Y &= Y^0 + \left(Y^1 - Y^0\right) D_1 + \left(Y^2 - Y^0\right) D_2 \\ D_1 &= D_1^0 + \left(D_1^1 - D_1^0\right) Z_1 + \left(D_1^2 - D_1^0\right) Z_2 \\ D_2 &= D_2^0 + \left(D_2^1 - D_2^0\right) Z_1 + \left(D_2^2 - D_2^0\right) Z_2 \end{aligned}$$

where $D_j^z \equiv 1$ if individual chooses field j for a given value of Z .

▷ The standard IV assumptions are:

  * $Y^{d,z} = Y^d$ for all d, z
  * $Y^0, Y^1, Y^2, D^0, D^1, D^2 \perp Z$
  * Rank $\mathbb{E}\left[Z'D\right] = 3$
  * $D_1^1 \geq D_1^0$ and $D_2^2 \geq D_2^0$

▷ IV uses the following moment conditions:

$$\mathbb{E}\left[\epsilon Z_1\right] = \mathbb{E}\left[\epsilon Z_2\right] = \mathbb{E}\left[\epsilon\right] = 0$$

Expressing these in potential earnings and choices yields:

$$\begin{aligned} \mathbb{E}\left[\left(\Delta^1 - \beta_1\right)\left(D_1^1 - D_1^0\right) + \left(\Delta^2 - \beta_2\right)\left(D_2^1 - D_2^0\right)\right] = 0 \\ \mathbb{E}\left[\left(\Delta^1 - \beta_1\right)\left(D_1^2 - D_1^0\right) + \left(\Delta^2 - \beta_2\right)\left(D_2^2 - D_2^0\right)\right] = 0 \end{aligned}$$

where $\Delta^j = Y^j - Y^0$.

What can IV identify?

▷ If you solve for β_1 and β_2, it turns out that β_j is a linear combination of Δ^1 (payoff of field 1 respect to field 0), Δ^2 (payoff of field 2 compared to field 0), and $\Delta^2 - \Delta^1$ (payoff of field 2 compared to 1). But it's weird that β_1 includes $\Delta^2 - \Delta^1$.

▷ If you assume constant effects and solve for β_1 and β_2, we have $\beta_1 = \Delta^1$ and $\beta_2 = \Delta^2$.

4.5 Weak Instruments and Many Instruments

Both weak instruments and many instruments are an issue.

4.5.1 Weak Instruments

An instrument variable is weak if its correlation with the included endogenous regressor is small.

- ▷ Since IV = reduced form / first stage, it is essentially a divided by zero problem.

The standard practice is to report the F-stat for instruments and proceed if it exceeds some arbitrary number. Magne's recommendation is that we:

1. Report and interpret reduced form
2. Think hard about why instrument could be weak
3. Report weak instrument robust confidence sets

4.5.2 Many Instruments

Many instruments are also a problem – if you fit first-stage perfectly, then you will get back OLS estimates.

- ▷ In Angrist and Krueger (1991), they worry it is a weak instrument so they interact the instrument with many control variables and find the coefficient to be similar to that from the OLS.
- ▷ Essentially, S and $\hat{S}$ are essentially the "same" and since the true S is endogenous, this means that $\hat{S}$ is also endogenous.

5 Regression Discontinuity

RDD is another way to identify causal effect of some treatment on outcome Y. It makes use of treatment assignment that **isn't random**, but where process follows some known and arbitrary cutoff rule.

There are two types of RDD: sharp and fuzzy. Sharp RDD is when treatment is deterministic and only depends on the running variable R; fuzzy RDD is when treatment is stochastic and probability of treatment has a discontinuity at $R = c$. Formal estimators are similar but different; fuzzy RDD is really just an IV.

5.1 Introduction

The basic idea is the following: observations are "treated" based on known cutoff, which creates a discontinuity. Researcher is interested in how this treatment affects outcome variable of interest y.

5.1.1 Notation

Outcome (Y) and binary treatment $(D \in \{0, 1\})$. There is also a variable R that has a discontinuous relationship with D at $R = c$:

$$P\left[D = 1 | R = r\right] \text{ has a discontinuity at } r = c$$

R is called the running, forcing, or assignment variable.

5.1.2 Intuition

The idea is to compare individuals on different sides of the cutoff point.

- ▷ We are assuming that Y_d (potential outcome) varies continuously with R at $R = c$, whereas D varies discontinuously at $R = c$.
- ▷ This implies that changes in Y at $R = c$ should be causally due to D.
- ▷ We do not believe that either D or R is exogenous!

Through the RDD, we are identifying effects for compliers with $R = c$.

5.1.3 Randomization Assumption

Assignment to treatment and control isn't random, but whether individual observation is treated is assumed to be normal.

- ▷ In other words, researcher assumes that observations (e.g. firms, person) cannot perfectly manipulate the value of the running variable.
- ▷ Therefore, whether an observation's R falls immediately above or below key cutoff is random.

5.1.4 Sharp RDD vs. Fuzzy RDD

How are they different?

- ▷ Assignment of variables
  - * Sharp RDD: Assignment to treatment *only* depends on R i.e. if $R \geq c$, you are treated with probability 1.
  - * Fuzzy RDD: Having $R \geq c$ only increases the probability of treatment; other factors (besides R) will influence whether you are actually treated or not.

▷ Estimating the Causal effect

- Sharp RDD: Compare average Y immediately above and below $R = c$
- Fuzzy RDD: Computing the average change in Y around the threshold **underestimates** the causal effect.
  - The comparison assumes all observations were treated, but this isn't true. If all observations had been treated, the observed change in Y would be even larger.

5.2 Sharp RD

Sharp RDD refers to the setting in which D changes deterministically from 0 to 1 when R crosses c.

5.2.1 Intuition

Sharp design implies selection on observables holds for R, since D is deterministic given R. But overlap may not necessarily hold since:

$$P\left[D = 1 | R = r\right] = \begin{cases} 1 & \text{if } r \geq c \\ 0 & \text{if } r < c \end{cases}$$

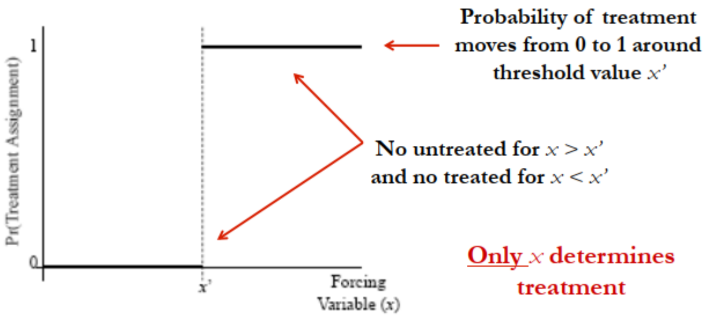
The only place overlap almost holds is exactly at $R = c$.

5.2.2 Assumptions

1. Assignment to treatment occurs through known and **deterministic** decision rule:

$$d = d\left(R\right) = \begin{cases} 1 & \text{if } R \geq c \\ 0 & \text{otherwise} \end{cases}$$

It is important that there exist Rs around the threshold value.



Probability of treatment moves from 0 to 1 around threshold value x'

No untreated for $x > x'$ and no treated for $x < x'$

Only x determines treatment

▷ For example, PSAT score above a certain value means that the student received national merit scholarship. Thistlewaithe and Campbell (1960) used this to study effect of scholarship on career plans.

2. Local Continuity: Potential outcomes Y_0 and Y_1, conditional on forcing variable R, are continuous at the threshold $R = c$.

▷ In words: Y would be a smooth function around threshold absent treatment.

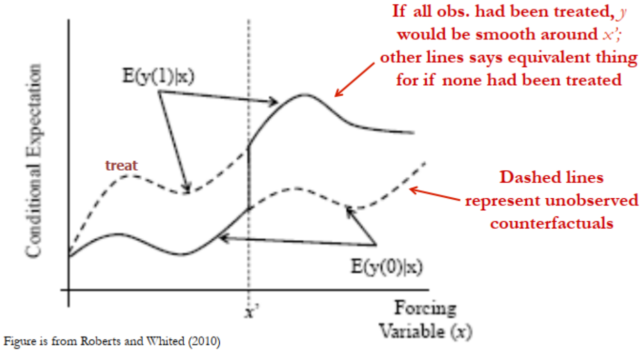


Figure is from Roberts and Whited (2010)

- ▷ Mathematically, $\mathbb{E}\left[Y_d|R=r\right]$ is continuous at $r=c$ for $d=0,1$.

5.2.3 Why does OLS Not Work?

Suppose you estimate the causal effect through

$$Y=\beta_0+\beta_1 D+U$$

then we will not get the causal effect. Why?

- ▷ D is correlated with R, and if R affects Y, we have an omitted variable.

Then can we fix it by adding R to the regression:

$$Y=\beta_0+\beta_1 D+\beta_2 R+U$$

Why may this be problematic?

- ▷ This assumes that the effect of R is linear.
- ▷ It does not really make use of random assignment, which is really occurring near the threshold.

5.2.4 Identification

Argument:

- ▷ Taking the limit of $\mathbb{E}\left[Y|R=r\right]$ as $R\downarrow c$ and $R\uparrow c$:

$$\lim_{r\downarrow c}\mathbb{E}\left[Y|R=r\right]\underbrace{=}_{\text{sharp design}}\lim_{r\uparrow c}\mathbb{E}\left[Y_1|R=r\right]\underbrace{=}_{\text{continuity}}\mathbb{E}\left[Y_1|R=c\right]$$

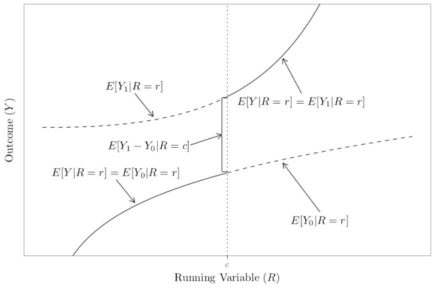
$$\lim_{r\uparrow c}\mathbb{E}\left[Y|R=r\right]\underbrace{=}_{\text{sharp design}}\lim_{r\uparrow c}\mathbb{E}\left[Y_0|R=r\right]\underbrace{=}_{\text{continuity}}\mathbb{E}\left[Y_0|R=c\right]$$

- ▷ Subtracting the second term from the first term:

$$\mathbb{E}\left[Y_1-Y_0|R=c\right]=\lim_{r\downarrow c}\mathbb{E}\left[Y|R=r\right]-\lim_{r\uparrow c}\mathbb{E}\left[Y|R=r\right]$$

  - * LHS is the expression for ATE at the cutoff
  - * RHS can be estimated from the data

Graphically:



- ▷ $\mathbb{E}\left[Y_d|R=r\right]$ is non-parametrically identified only where the lines are solid.
- ▷ With continuity, we are able to point identify $\mathbb{E}\left[Y_1-Y_0|R=c\right]$.

5.3 Fuzzy RD

In a fuzzy RDD, $P\left[D=1|R=r\right]$ is discontinuous at c. This is more general than a sharp design. It turns out that fuzzy designs are IV designs

5.3.1 Intuition

Define

$$Z\equiv\mathbb{I}\left[R\geq c\right]=P\left[D=1|R=r\right]$$

where $\mathbb{I}\left[R\geq c\right]$ is a strong (but not exact) predictor of D .

- ▷ Z is exogenous: Fuzzy RDDs are IV designs: $(Y_0,Y_1)\perp Z|R$ since Z is a function of R.
- ▷ Z is relevant, local to $R=c$, since:

$$\lim_{r\downarrow c}P\left[D=1|Z=1,R=r\right]\neq\lim_{r\uparrow c}P\left[D=1|Z=0,R=r\right]$$

In constrast, Sharp RDD was like having a perfect control variable.

Example 5.1. Angrist & Lavy (QJE, 1999)

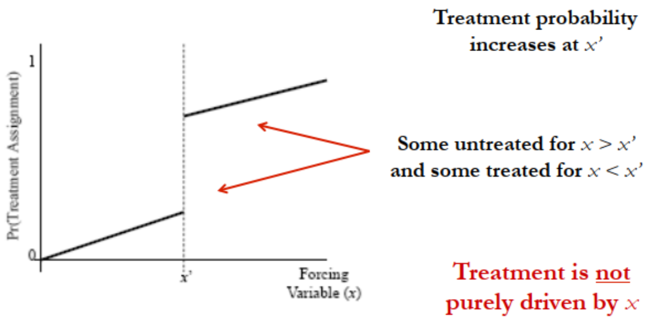
- ▷ They seek to estimate the effects of class size (D) on student achievement (Y) in Israel. They have a class size rule that produces a systematic but discontinuous relationship between the (number of students in a school) and the (average number of students in a class).
- ▷ They compute class size predicted by the school size and use it as an instrument for observed class size.

5.3.2 Assumptions

1. Assignment to treatment is stochastic in that only the probability of treatment has known discontinuity at $R = c$:

$$0 < \lim_{R \downarrow c} P\left(D = 1|R\right) - \lim_{R \uparrow c} P\left(D = 1|X\right) < 1$$

In the fuzzy RDD case, treatment is not purely driven by R.



Treatment probability increases at x'

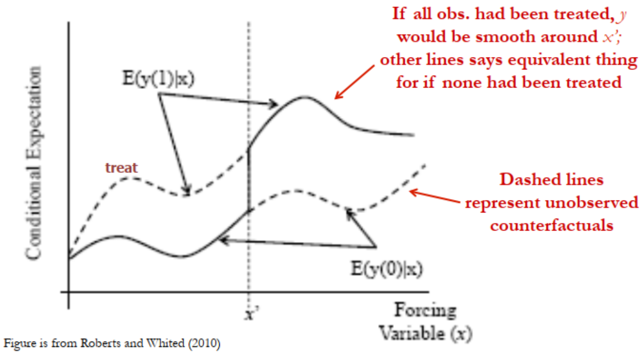
Some untreated for $x > x'$ and some treated for $x < x'$

Treatment is not purely driven by x

▷ FICO score > 620 increases the likelihood of loan being securitized. But the extent of loan documentation, lender, etc will matter as well.

2. Local Continuity: Potential outcomes Y_0 and Y_1, conditional on forcing variable R, are continuous at the threshold $R = c$.

▷ In words: Y would be a smooth function around threshold absent treatment.



If all obs. had been treated, y would be smooth around x'; other lines says equivalent thing for if none had been treated

Dashed lines represent unobserved counterfactuals

Figure is from Roberts and Whited (2010)

▷ Mathematically, we need $\mathbb{E}\left[Y_d|R = r, T = t\right]$ and $P\left[T = t|R = r\right]$ to be continuous at $r = c$.

5.3.3 Identification

Assumptions: $Z \equiv \mathbb{I}\left[R \leq c\right]$

▷ Monotonicity

▷ $\mathbb{E}\left[Y_d|R = r, T = t\right]$ and $P\left[T = t|R = r\right]$ are continuous at $r = c$

Argument:

- ▷ The limiting Wald estimand as $R \to c$ is a LATE:

$$\frac{\lim_{r\downarrow c}\mathbb{E}\left[Y|R=r\right]-\lim_{r\uparrow c}\mathbb{E}\left[Y|R=r\right]}{\lim_{r\downarrow c}\mathbb{E}\left[D|R=r\right]-\lim_{r\uparrow c}\mathbb{E}\left[D|R=r\right]}=\mathbb{E}\left[Y_1-Y_0|R=c,T=cp\right]$$

- ▷ This is a very "local" parameter – not only $R=c$ but also $T=cp$.

Interpretation:

- ▷ Average causal effect for individuals whose treatment status is shifted if we marginally change the cutoff

5.4 Implementation Details

We cannot use data only at the cut-off:

- ▷ We need data away from the cut-off to produce estimates of $\mathbb{E}\left[Y_1|R=c\right]$ and $\mathbb{E}\left[Y_0|R=c\right]$
- ▷ We use parametric and/or non-parametric regressions to flexibly estimate $\mathbb{E}\left[Y|R\right]$ separately for $R\geq c$ and $R<c$.

5.4.1 Checking RD Assumptions

If variation in treatment near the cut-off is as good as random:

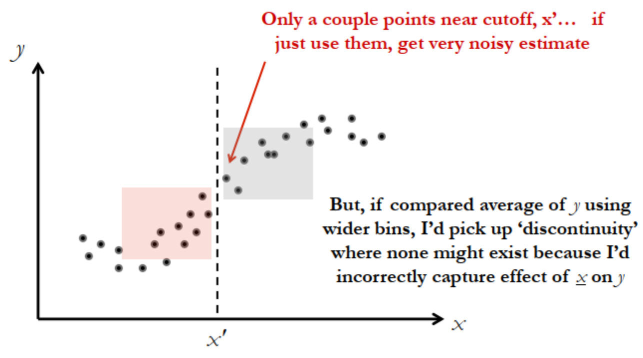
1. Pre-determined characteristics (X) should have the same distribution (just above and just below the cut-off)
2. Density of the assignment variable should not change discontinuously around the cut-off.

Thus distribution of X and density of R are used to informally examine the validity of the RD design. So check for:

- ▷ Discontinuities in covariates
- ▷ Discontinuity in the distribution of the running variable
- ▷ Discontinuities in outcomes at values of R other than c
- ▷ Specification checks and sensitivity to bandwidth choice

5.4.2 Implementation of Sharp RDD

There is a general trade-off between bias and noise, which yields two general approaches to implementing RDD:



Only a couple points near cutoff, x'… if just use them, get very noisy estimate

But, if compared average of y using wider bins, I'd pick up 'discontinuity' where none might exist because I'd incorrectly capture effect of x on y

1. Use all data but control for effect of R on Y in a very general and rigorous way
2. Use less rigorous controls for effect of R but only use data in a small window around the threshold.

Using All Data This approach uses all the data available

1. Estimate two separate regressions:

   $$\begin{aligned}[1] : Y &= \beta^b + f(R-c) + U^b \\ [2] : Y &= \beta^a + g(R-c) + U^a\end{aligned}$$

   where [1] is estimated using only data below $R = c$ and [2] is estimated using only data above $R = c$.

2. Compute the treatment effect: $\beta^a - \beta^b$

The functions $f(\cdot)$ and $g(\cdot)$ are included to control for the underlying effect of R on Y. You can do all this in one step:

$$Y = \alpha + \beta D + f(R-c) + D \times g(R-c) + U$$

and β will equal $\beta^a - \beta^b$.

- ▷ If you drop $D \times g(R-c)$, then you are assuming that the same function form between Y and R is same above and below R.
- ▷ High-order polynomials are used for $f(\cdot)$ and $g(\cdot)$. Since finding the correct order is difficult, it's best to show robustness by illustrating that findings are robust to different polynomial orders.

Using Window We do the same RDD estimate as before, but we restrict analysis to smaller windows around $R = c$ and use lower polynomial order controls.

- ▷ This is less subject to the risk of bias because correctly controlling for relationship between R and Y is less important in the smaller window.

5.4.3 Implementation of Fuzzy RDD

We estimate the model of the following form:

$$Y = \alpha + \beta D + f(R-c) + U$$

where we use $\mathbb{I}[R \geq c]$ as an instrument for D using 2SLS regression.

- ▷ Once again, $f(\cdot)$ is a polynomial function, but unlike sharp RDD it is not easy to allow the functional form to vary above & below. .
- ▷ Can be estimated on a local neighborhood of c
- ▷ Other covariates can be added in straightforward manner and give standard errors.

Sometimes RD involves K discontinuities, which gives K differnt effects. Sometimes, we can get a pooled estimate (analogous to using many instruments)

5.4.4 Bandwidth Choice

While RD is identified locally, estimation uses data away from the discontinuity, so the inferences are sensitive to bandwith choice. There are several options:

- ▷ Cross validation (Ludwig & Miller, 2007)
- ▷ Direct plug-in rules
  - * Imbens & Kalyanaraman (2012): Find the bandwith that minimizes a first-order approximation of the MSE of the estimated treatment effect
  - * Calonico, Cattaneo, and Titiunik (2014): Add bias correction and derive a new MSE optimal bandwith.

5.4.5 Non-parametric Regression

We have to estimate $\lim_{r\downarrow c}\mathbb{E}\left[Y|R=r\right]$ and $\lim_{r\uparrow c}\mathbb{E}\left[Y|R=r\right]$.

- ▷ Kernel regression = local constant estimators
- ▷ Local linear regression (Has low boundary bias)

For example, do local linear regression for both the outcome and the treatment indicator:

$$\begin{aligned}\left(\hat{\alpha}_{yl},\hat{\beta}_{yl}\right) &= \arg\min_{\alpha_{yl},\beta_{yl}}\sum_{i:c-h\leq R_i<c}\left(Y_i-\alpha_{yl}-\beta_{yl}\left(R_i-c\right)\right)^2\\ \left(\hat{\alpha}_{tl},\hat{\beta}_{tl}\right) &= \arg\min_{\alpha_{yl},\beta_{yl}}\sum_{i:c-h\leq R_i<c}\left(Y_i-\alpha_{yl}-\beta_{yl}\left(R_i-c\right)\right)^2\end{aligned}$$

and similarly on the right side of the discontinuity. Then the FRD estimator is

$$\hat{\tau}=\frac{\hat{\tau}_y}{\hat{\tau}_t}=\frac{\hat{\alpha}_{yr}-\hat{\alpha}_{yl}}{\hat{\alpha}_{tr}-\hat{\alpha}_{tl}}$$

5.4.6 Robustness Tests for Internal Validity

We already discussed the following:

- ▷ Showing graphical analysis
- ▷ Making sure findings are robust to chosen polynomial
- ▷ Making sure findings are robust to chosen bandwidth

Here are additional ones:

- ▷ Is there any reason to believe that the threshold $R=c$ was chosen because of some pre-existing discontinuity in Y or lack of comparability above and below $R=c$?
  - * If so, this is a violation of the local continuity assumption.
- ▷ Is there any way or reason why subjects might manipulate R around threshold?
  - * If so, this is a violation of the local continuity assumption since Y might exhibit jump around $R=c$ absent treatment due to manipulation

Example 5.2. Example: In Keys et al. (QJE 2010), default rate of loans at FICO = 620 might jump regardless if weak borrowers manipulate their FICO to get the low interest rates that one gets immediately with FICO above 620.

- * If they cannot perfectly manipulate it, then there will still be randomness in treatment. In small enough bandwidth around $R = c$, there will still be randomness because idiosyncratic shocks will push some abovve and some below threshold even if they try to manipulate!
- * To test this, look for bunching of observations immediately above or below threshold. This, however, may not be a perfect test since it assumes manipulation is monotonic i.e. all subjects try to get above or below $R = c$ which may not be true in all scenarios.
- * McCrary (2008) proposed a test of manipulation. The idea is that manipulation results in an unusually large density on one side of $R = c$.

- ▷ Balance Tests: RDD assumes observations near but on opposite sides of cutoffs are comparable, so we need to check this.
  - * Make sure that other observable factors that might affect Y do not exhibit jumps at the threshold $R = c$.
  - * You could also add other covariates as controls. If RDD is internally valid, these additional controls will only affect the precision of estimate.
- ▷ Falsification Tests: Make sure there's no effect in years where there was no discontinuity or firms where there isn't supposed to be an effect

5.4.7 External Validity

1. Identification relies on observations close to the cutoff threshold.
   - ▷ Effect of treatment might be different for observations further away from this threshold.
2. In fuzzy RDD, treatment is estimated using only compliers
   - ▷ In other words, we will only pick up the effect of those where discontinuity is what pushes them into the treatment.

   Example 5.3. Suppose you study the effect of PhD on wages using GRE score > c with a fuzzy RDD. If the discontinuity matters only for students with mediocre GPA, then you only estimate the effect of PhD for those students.

5.4.8 Practical Advice

1. Motivate validity of design: Why can individuals not manipulate assignment variables?
2. Test validity of design, bot graphically and formally
3. Show robustness of RD estimates with respect to specification of f_r and f_l as well as the choice of bandwidth.
4. If fuzzy RD, think of it as IV.

5.5 Labor Examples

Example 5.4. Køstol and Mogstad (2014) look at return-to-work reform for disability insurance (DI) in Norway.

Example 5.5. Kirkeboen et al. (2016) examines the payoff to different types of post-secondary education, including field and institution of study.

6 Difference-in-Difference

6.1 Introduction

Omitted variables pose a substantial hurdle in our ability to make causal inferences. What's worse is that many of them are inherently unobservable to researchers. Panel data can help us with a particular type of unobserved variable – It helps us with any unobserved variable that doesn't vary within groups of observations.

6.1.1 Setup

There are two periods: before the intervention $(t = 1)$ and after the intervention $(t = 2)$. There are also two group: treatment and control. We use D_{it} to denote whether the person i is treated in period t.

- Before the intervention $(t = 1)$, neither group is treated by the intervention:

$$D_{i1} = 0, \forall i$$

- After the intervention $(t = 2)$, the treatment group is treated but the control group is unaffected:

$$\begin{aligned} D_{i2} &= 0, \forall i \text{ is in treated group} \\ D_{i2} &= 0, \forall i \text{ is in control group} \end{aligned}$$

Associated with this setup is the potential outcomes for each person i:

- Y_{i1}^0: potential outcome in period 1 (superscript is 0 since neither group is treated)
- Y_{i2}^1: potential outcome in period 2 if in treatment group
- Y_{i2}^0: potential outcome in period 2 if in control group
- Therefore, the observed outcome in period t is given as

$$Y_{it} = D_{it} Y_{it}^1 + \left(1 - D_{it}\right) Y_{it}^0 = D_{it}\left(Y_{it}^1 - Y_{it}^0\right) + Y_{it}^0$$

6.1.2 DiD Estimand

Using the previous expression, write the change in individual i's observed outcome from period 1 to period 2:

$$Y_{i2} - Y_{i1} = D_{i2}\left(Y_{i2}^1 - Y_{i2}^0\right) + Y_{i2}^0 - Y_{i1}^0$$

Differencing for the treatment group yields:

$$\begin{aligned} \mathbb{E}\left[Y_{i2} - Y_{i1} | D_{i2} = 1\right] &= \mathbb{E}\left[D_{i2}\left(Y_{i2}^1 - Y_{i2}^0\right) + Y_{i2}^0 - Y_{i1}^0 | D_{i2} = 1\right] \\ &= \mathbb{E}\left[D_{i2}\left(Y_{i2}^1 - Y_{i2}^0\right) | D_{i2} = 1\right] + \mathbb{E}\left[Y_{i2}^0 - Y_{i1}^0 | D_{i2} = 1\right] \\ &= \mathbb{E}\left[Y_{i2}^1 - Y_{i2}^0 | D_{i2} = 1\right] + \mathbb{E}\left[Y_{i2}^0 - Y_{i1}^0 | D_{i2} = 1\right] \end{aligned}$$

Differencing for the control group yields:

$$\mathbb{E}\left[Y_{i2} - Y_{i1} | D_{i2} = 0\right] = \mathbb{E}\left[Y_{i2}^0 - Y_{i1}^0 | D_{i2} = 0\right]$$

The DiD estimand is the difference of these differences:

$$\mathbb{E}\left[Y_{i2} - Y_{i1} | D_{i2} = 1\right] - \mathbb{E}\left[Y_{i2} - Y_{i1} | D_{i2} = 0\right]$$

- ▷ Under the assumption of the commond trend in the absence of intervention:

$$\mathbb{E}\left[Y_{i2}^0 - Y_{i1}^0 | \underline{D_{i2} = 1}\right] = \mathbb{E}\left[Y_{i2}^0 - Y_{i1}^0 | \underline{D_{i2} = 0}\right]$$

  Then DiD estimand identifies the ATT:

$$\mathbb{E}\left[Y_{i2}^1 - Y_{i2}^0 | D_{i2} = 1\right]$$

Note that the common trend assumption allows for:

- ▷ Selection on non-treatment levels:

$$\mathbb{E}\left[Y_{it}^0 | D_{i2} = 1\right] \neq \mathbb{E}\left[Y_{it}^0 | D_{i2} = 0\right], \forall t = 1, 2$$

- ▷ Selection on gains:

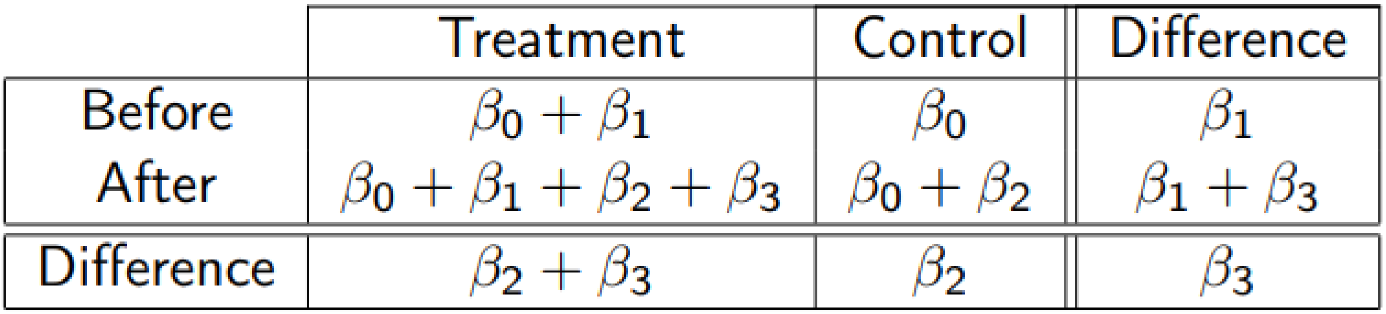
$$\mathbb{E}\left[Y_{i2}^1 - Y_{i2}^0 | D_{i2} = 1\right] \neq \mathbb{E}\left[Y_{i2}^1 - Y_{i2}^0 | D_{i2} = 0\right]$$

6.1.3 DiD as Regression

The following regression delivers our DiD estimate:

$$Y_{it} = \beta_0 + \beta_1 \text{treat}_i + \beta_2 \text{after}_t + \beta_3 \text{treat}_i \times \text{after}_t + \epsilon_{it}$$

The β_3 is our diff-in-diff estimate since:



	Treatment	Control	Difference
Before	$\beta_0 + \beta_1$	β_0	β_1
After	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$	$\beta_1 + \beta_3$
Difference	$\beta_2 + \beta_3$	β_2	β_3

Advantages

- ▷ Convenient way to obtain standard errors
- ▷ Easy to add covariates to control for confounding trends and reduce residual variance
- ▷ Possible to extend to multiple groups & multiple time periods
- ▷ Treatment can be continuous and implemented at different groups at different times (at least under constant effects)

Example 6.1. Card & Krueger (AER 1994): They use diff-in-diff to estimate the effect of increase in minimum wage on employment.

- ▷ Outcome (Y_{igt}): employment in restaurant i in state g at time t
- ▷ Common trend assumption: In absence of intervention employment in NJ would have had same downward trend as PA
- ▷ They find that employment increased as consequence of increase in minimum wage (significant at 5% level)

6.2 Challenges in Implementations

The art of DID lies in the choice of an appropriate control group.

6.2.1 Compositional Changes

The treatment under consideration may affect the composition of the treatment and the control groups.

- ▷ For example, suppose a state lowers welfare benefits. This may induce poor families to move to other states.
- ▷ To resolve this issue, one may (re)define group such that it is not affected by treatment. But then we are no longer estimating ATT since some families move away from the treatment.

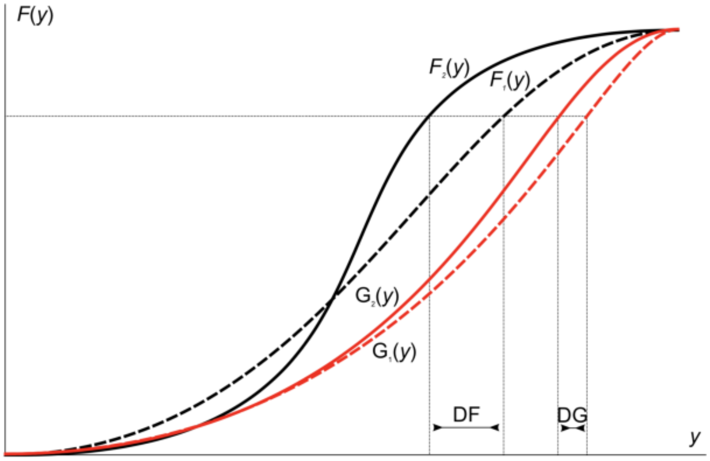
6.2.2 Non-linearity

DiD suffers from non-invariance to transformations of the outcome variable. In other words, identification depends on the transformation used.

- ▷ This is because common trend implies that growth in outcome of a group does not depend on its level.
- ▷ For example, the same DID strategy cannot apply to both wages and log wages.

As a solution, Athey and Imbens (2006) develop a model that is immune to this critique. They call this “change-in-change” (CIC), which is invariant with respect to the monotonic transformations of outcome.

Quantile DiD vs. CiC We first describe how quantile DiD works:

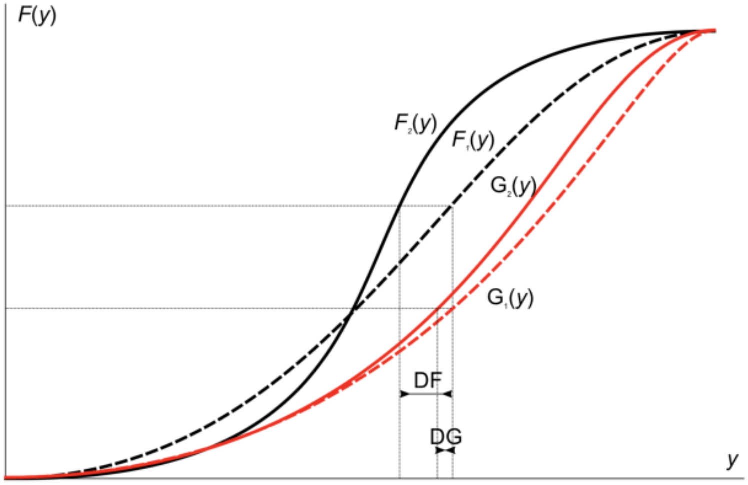


1. Fix the quantile of y in the pre-reform outcome distribution of the treatment group, $F_1(Y) = \tau$ where τ is the desired quantile.
2. Compute the counterfactual post-reform outcome at quantile τ in the treatment group:

$$\begin{aligned} k^{QDID}(\tau) &= F_1^{-1}(\tau) + \Delta^{QDID} \\ &= F_1^{-1}(F_1(y)) + \left[G_2^{-1}(F_1(y)) - G_1^{-1}(F_1(y))\right] \end{aligned}$$

3. The QTE estimate at quantile τ is then:
$$F_2^{-1}(\tau) - k^{QDID}(\tau)$$

Constrast this with the CiC:



1. Fix the outcome level y giving the quantiles in the two groups pre-reform: $F_1(y)$ and $G_1(y)$
2. Compute the counterfactual post-reform outcome at y in the treatment group:
$$\begin{aligned} k^{CIC}(y) &= F_1^{-1}(F_1(y)) + \Delta^{CIC} \\ &= y + \left(G_2^{-1}(G_1(y)) - G_1^{-1}(G_1(y))\right) \\ &= G_2^{-1}(G_1(y)) \end{aligned}$$
3. The CIC estimate at level y is then
$$F_2^{-1}(F_1(y)) - k^{CIC}(y)$$

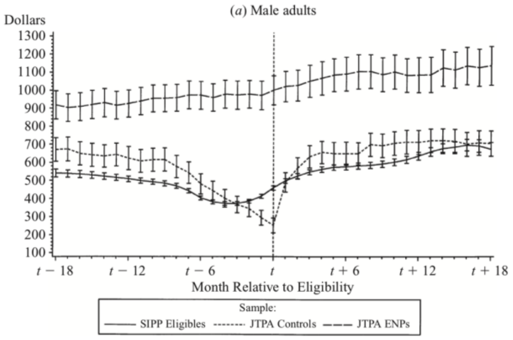
6.2.3 Differential Underlying Trends

Recall that the key identifying assumption was that of common trend: in absence of the intervention, the treatment and control group should have common trends in the outcome variable. Note that the assumption is formally untestable.

Ashenfelter's Dip This is a well-known reason for the violation of common trend assumption.

▷ Ashenfelter (1978) noted that enrollment in a training program is more likely if temporary dip in earnings occurs just before start of the program. As a consequence, earnings growth after enrollment likely different for participants even without treatment. The violation of common trend assumption implies that DID estimator overestimates the effect of treatment.

▷ Heckman & Smith (1999) investigate earnings growth for randomized-out participants of Job Training Partnership Act program and show that randomized-out participants show larger earnings growth than non-participants.



Dealing with Violation of Common Trend When people think common trend assumption is unlikely to hold, there are a few options:

1. Include time-varying covariates and/or group-specific time trends
2. DiDiD
3. DiD + IV
4. Semi-parametric DiD (Abadie 2005) – This is essentially commond trend conditional on covariates
5. Matching on pre-trends (Blundell et al., 2001) – This is DiD for controls with similar pre-trends / levels
6. Synthetic control group approach

6.3 Inference

While obtaining the point estimate is devilishly simple, getting the correct standard errors is non-trivial.

6.3.1 DID Standard Errors

Within Group-Time Correlation Consider the standard DID regression:

$$Y_{igt} = \beta D_{gt} + \alpha_g + \lambda_t + \epsilon_{igt}$$

Observations in group g at time t are unlikely to be independent, i.e.

$$\mathbb{E}\left[\epsilon_{igt}\epsilon_{jgt}\right] \neq 0$$

▷ Group-time period random shocks can cause correlation between residuals within a group-time period.

To deal with this issue, we can do one of the following:

1. Assume $\epsilon_{itg} = v_{gt} + e_{igt}$ and use WLS
2. Use cluster-robust standard errors (This is robust to any correlation structure and heteroskedasticity)
3. Block bootstrap – You resample group-time periods intead of individual observations
4. Use group-time period averages:

$$\bar{Y}_{gt} = \beta \cdot D_{gt} + \alpha_g + \lambda_t + \bar{\epsilon}_{gt}$$

Serial Correlation For Did with many groups and more than 2 time periods, serial correlation is likely an issue. Bertrand, Duflo, and Mullainathan (QJE 2004) investigate the consequences of ignoring serial correlation in DiD and show that one can find a significant effect in 45% of the placebo interventions.

- ▷ Serial correlation is particularly relevant for DiD since Y_{igt} is typically highly positively serially correlated.

Clustered Standard Errors When clustering standard errors, we are using the following estimate of the covariance matrix:

$$\hat{W} = \left(V'V\right)^{-1}\left(\sum_{g=1}^{G} \hat{u}_g \hat{u}_g'\right)\left(V'V\right)^{-1}$$

where V is the matrix with year dummies, state dummies, treatment dummay, and $\hat{u}_g$ is the vector of pooled OLS residuals multiplied by the vector of independent variables for group g.

- ▷ With less than (about) 50 groups, the standard errors are incorrect.
- ▷ This is because clustering is an asymptotic argument, so something needs to go to infinity.

Possible solutions to the few clusters problem include (also see Cameron & Miller 2015):

- ▷ Wild bootstrap: This is widely used based on simulation evidence, but it requires strong assumptions to work (See Canay, Santos, and Shaikh 2018)
- ▷ Permutation inference (See Canay et al. 2017)
- ▷ Small sample approximations derived under parametric assumptions

6.4 Extensions

A few extensions of DiD are used when common trend assumption is unlikley to hold:

1. Time-varying covariates and/or group-specific time trends
2. Difference-in-Difference-in-Difference
3. Difference-in-Difference + Instrumental Variables

6.4.1 Time-varying Covariates and/or group-specific time trends

Common trend assumptions can be relaxed by including time-varying covariates (X_{gt}) and group-specific time trends $(u_g \times t)$:

$$Y_{igt} = \beta \cdot D_{gt} + X_{gt}'\pi + u_g \cdot t + \alpha_g + \lambda_t + \epsilon_{igt}$$

Note that we need at least three time periods to estimate model with group-specific time trends.

Example 6.2. Besley & Burgess (2004) study effect of labor market regulation on manufacturing performance in Indian states. Their outcome variable (Y_{gt}) is log manufacturing output per capita in state g in year t, and treatment (D_{gt}) is the labor regulation that takes on values in $\{-1, 0, 1\}$ depending on whether the amendments to the Industrial Disputes Act were pro-worker, neutral, or pro-employer.

- ▷ Defining D_{gt} to take on three values is problematic because it's assuming that the effect of going from $D_{gt} = 1 \to 0$ is the same in magnitude as the effect of going from $D_{gt} = 0 \to -1$.
- ▷ They include covariates such as state population and state-specific trends.

6.4.2 Difference-in-Difference-in-Difference

Sometimes, a third difference might work when you suspect violation of common trend. For example, consider a case where a state implements change in health care policy for people 65 and older.

- ▷ Concerns about violation of common trend
  - * Approach #1: Outcome is the data on health, and people ≥ 65 are the treatment group, whereas people ≤ 65 are the control group. Then one may suspect different trends between old & young people.
  - * Approach #2: Outcome is the data on health, and people ≥ 65 are the treatment group, whereas people ≥ 65 in other states are control group. Then one may suspect that two states might have different trends.
- ▷ In this setting, one can
  - * Implement DiD using Approach #1 (only in the treated state!)
  - * Implement DiD using Approach #1 (only in the control states)
  - * Take the difference of the estimates

In the above example, identification comes from the differences in the differential trends between the age groups in the treated versus control states. In the regression setting, the treatment effect is coefficient on triple interaction with levels and double interactions as controls.

6.4.3 Difference-in-Difference + IV

The concern is that the treatment variable is potentially related to group-specific trends. Thus, we need to find a variable that affects treatment but is unrelated to group-specific trends. For example, consider a case where you're interested in the effect of supply of schools on pupils test scores. (Haan et al. 2018)

- ▷ Concerns:
  - * Municipalities with many schools are different from municipalities with fewer schools.
- ▷ To mitigate this concenr, one can investigate changes in number of schools within municipalities over time (using municipality fixed effects). However, change in number of schools can be related to change in (unobserved) municipality characteristics (violation of common trend)
- ▷ Thus, the DiD + IV approach here involves including municipality and year fixed effects (DiD) AND isolating change in number of schools which is due to a reform (IV).

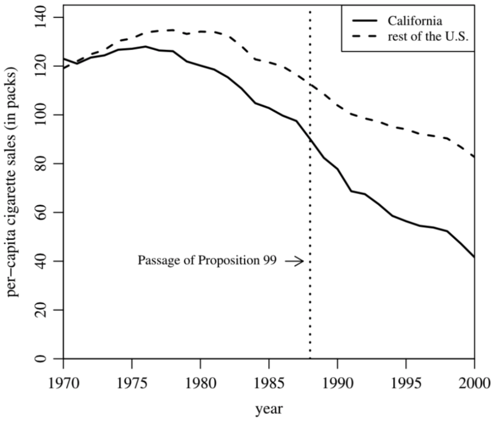
6.5 Synthetic Controls

This is an idea thanks to Abadie and Gardeazabal (2003), Abadie et al. (2010, 2015). It is increasingly used but concerns about identification and inference still remain.

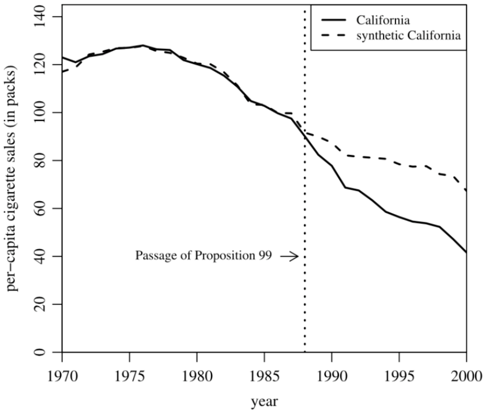
6.5.1 Overview

Suppose you have a panel data with one treated and many untreated units with many time periods. Also, common trends is unappealing for all untreated units. The idea is to construct a weighted average of untreated units to use for counterfactuals.

Example 6.3. Abadie, Diamond, and Hainmueller (2010) – ADH (2010) apply synthetic control approach to estimate the effect of a large scale tobacco control program implemented in California in 1988. They had the following pre-trends and saw that common trend assumption was likely to not hold.



- ▷ Synthetic control is essentially a weighted average of available control groups, where the weights are chosen to minimize the difference between treatment and control group prior to the intervention. It also requires a weight matrix V (diagonal matrix) to put more weight on variables with large predictive power for the outcome of interest.
- ▷ Using the synthetic controls, we obtain the following graph:



What are the assumptions behind synthetic controls?

- ▷ The reason we use synthetic controls is weakening common trends – otherwise, we will just use standard DID method.
- ▷ Since they use a factor model to justify the weighting procedure, does this place a further restriction on potential outcomes?

One major criticism is that there are conceptual difficulties with the statistical inference.

- ▷ Abadie et al. (2010) only argue for (approximate) unbiasedness.
- ▷ Eliminating variance to achieve consistency depends on sampling assumptions. Abadie et al. (2010) use a permutation test, but it's unclear whether this is justified.

6.6 Event Study

In the standard DID, the treated units all receive treatment at the same time, whereas here units receive treatment at different times, where being treated is still an absorbing state.

6.6.1 Setup

Time runs from $t = 0, ..., T$ so the treatment is $D_{it} \in \{0, 1\}$. Denote E_i as i's first treatment date, and since treatment is absorbing, $D_{it} = 1$ if and only if $t \geq E_1$. $Y_{it}(e)$ is the potential outcome where

$$e \in \{0, 1, ..., T, \infty\}$$

so $Y_{it}(E_i)$ is what we actually observe at each t. Note that D_i is now a vector:

$$\begin{aligned} E_i = e &\Leftrightarrow D_i = (0, ..., 0, 1, ..., 1) \\ E_i = \infty &\Leftrightarrow D_i = (0, ..., 0) \end{aligned}$$

Assumptions

1. Common Trends: For any $t \geq e > s$, we need:

$$\mathbb{E}\left[Y_{it}(\infty) | E_i = e\right] = \mathbb{E}\left[Y_{is}(\infty) | E_i = e\right] + \mathbb{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | E_i = \infty\right]$$

   ▷ LHS is the desired counterfactual.

   ▷ First term in RHS is not in the data, and second term in RHS is the observed change in the data.

2. No Anticipation: $Y_{is}(e) = Y_{is}(\infty)$ for all untreated periods $(s < e)$.

   ▷ This is a strong statement about the behavior.

6.6.2 Identification

Using the common trends and no anticipation assumptions, we have

$$\begin{aligned} ATE_t(e) &\equiv \mathbb{E}\left[Y_{it}(e) - Y_{it}(\infty) | E_i = e\right] \\ &= \mathbb{E}\left[Y_{it} | E_i = e\right] - \mathbb{E}\left[Y_{is} | E_i = e\right] \\ &\quad + \mathbb{E}\left[Y_{it} | E_i = e'\right] - \mathbb{E}\left[Y_{is} | E_i = e'\right] \end{aligned}$$

where $e' > t \geq e > s$.

▷ $ATE_t(e)$: average effect on Y_{it} of being treated at time e (vs. never)

6.6.3 Implementation via Regression

In a typical regression framework, we regress Y_{it} on unit dummites and time dummies, amd D_{it} where

$$D_{it} \equiv \mathbb{I}\left[t \geq E_i\right]$$

This is the same specification as in DiD with multiple groups/times. The coefficient on D_{it} is the object of interest, which we denote as β_{fe}.

- ▷ Abraham and Sun (2018) show that

$$\beta_{fe} = \sum_{e=0}^{T} \sum_{t=e}^{T} w_t(e)\, ATE_t(e)$$

where the weights are identified, sum to one, but can be negative if causal effects are heterogenous. This is because the regression specification makes contrasts we wouldn't non-parametrically.

An alternate implementation is to regress Y_{it} on unit dummites, time dummies, and $\left\{R_{it}^{l}\right\}$ where

$$R_{it}^{l} \equiv \mathbb{I}\left[l = t - E_i\right]$$

where l denotes the leads and lags. $l > 0$ captures dynamic treatment effects and $l < 0$ checks on the validity of the design. Coefficients on $\left\{R_{it}^{l}\right\}$s are treated as the objects of interest.

- ▷ Abraham and Sun (2018) show a similar weighting results. This time, lead and lag weights are different.
- ▷ The issue with this approach, however, is that pre-trends may not be detectable with lead estimates. Alternatively, zero pre-trends might be estimated spuriously.

The problems discussed stem from the regression specification, so why not estimate them directly and then construct whatever weighted average you care about?

6.6.4 Example: Employment and Lottery Decisions

Mogstad shows an example with employment and earnings for 1999 to 2015 and lottery participation and winning. See slides for more details.

7 Panel Data

7.1 Motivation

Omitted variables pose a substantial hurdle in our ability to make causal inferences. What's worse is that many of them are inherently unobservable to researchers. Panel data can help us with a particular type of unobserved variable – It helps us with any unobserved variable that doesn't vary within groups of observations.

7.1.1 Panel Data

Panel data is when you have multiple observations per unit of observation. For example, you observe each firm over multiple years. Consider the model

$$y_{i,t} = \alpha + \beta x_{i,t} + \delta f_i + u_{i,t}$$

where f is the unobserved, time-invariant variable f and we assume

$$\begin{aligned} \mathbb{E}\left[u_{i,t}\right] &= 0 \\ \rho\left(x_{i,t}, f_i\right) &\neq 0 \text{ (which gets us omitted variable bias)} \\ \rho\left(x_{i,t}, u_{i,s}\right) &= 0, \forall s, t \end{aligned}$$

7.1.2 Fixed Effects

Notice that if we take the population mean of the dependent variable for each unit of observation i, we have

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \delta f_i + \bar{u}_i$$

So substracting $\bar{y}_i$ from $y_{i,t}$, we have:

$$y_{i,t} - \bar{y}_i = \beta\left(x_{i,t} - \bar{x}_i\right) + \left(u_{i,t} - \bar{u}_i\right)$$

f_i has disappeared since it is time-invariant, and thanks to the assumption of strict exogeneity, $\left(x_{i,t} - \bar{x}_i\right)$ is uncorrelated with the new disturbance. This will yield a consistent estimate of β.

Remark 7.1. When we use the FE estimator in programs like Stata, it does the within transformation for you. You should not do it on your own because the degrees of freedom sometimes needs to be adjusted down by the number of Panels.

Another way to do the FE estimation is by adding indicator variables. We can create a dummy variable for each group i and add it to the regression. Note that using dummies and fixed effects are identical.

▷ This is because the demeaned variables are the residuals from a regression of them onto the group dummies, so using the partial regression results we can establish the equivalence.

The reported R^2 will be larger when using dummies.

▷ This is because R^2 for the FE estimator only reports what proportion of the within variation in y is explained by the within variation in x.

Benefits of the FE Estimator

1. It allows for arbitrary correlation between each fixed effect f_i and each x within group i.

   In general, it is very general and does not impose much structure on what the underlying data may look like. It also offers a very intuitive interpretation, and coefficient is identified using only changes within cross-section.

2. It is very flexible and can help us control for many types of unobserved heterogeneities.

   - We can add year-FE if worried about unobserved heterogeneity across time (e.g. macroeconomic shocks)
   - We can add CEO FE if worried about unobserved heterogeneity across CEOs. (e.g. talent, risk aversion)
   - We can add industry-year-FE if worried about unobserved heterogeneity across industries over time (e.g. investment opportunities, demand shock

3. It is very general and can apply to any scenario where observations can be grouped together. And once you are able to construct groups, you can remove any unobserved group-level heterogeneity by adding group FE.

Costs of the FE Estimator

1. It cannot identify variables that do not vary within group.

   To see this, consider the following CEO-level estimation:

$$\begin{aligned}\ln\left(\text{total pay}\right)_{ijt} &= \alpha + \beta_1 \ln\left(\text{firm size}\right) + \beta_2 \left(\text{volatility}\right)_{ijt} \\ &\quad + \beta_3 \left(\text{female}\right)_i + \delta_t + f_i + \lambda_j + u_{ijt}\end{aligned}$$

   where $\ln\left(\text{total pay}\right)$ is for CEO i, firm j, and year t and the estimation includes year, CEO, and firm FE.

   - In this regression, β_3 cannot be estimated. This is because being female does not vary within the group of each CEO's observations.

2. Measurement error of independent variable (and resulting biases) can be amplified.

   We can think of there being two types of variation: good (meaningful) variation and noise variation. Adding FE can sweep out a lot of the good variation, and thus the fraction of remaining variation coming from the noise will go up.

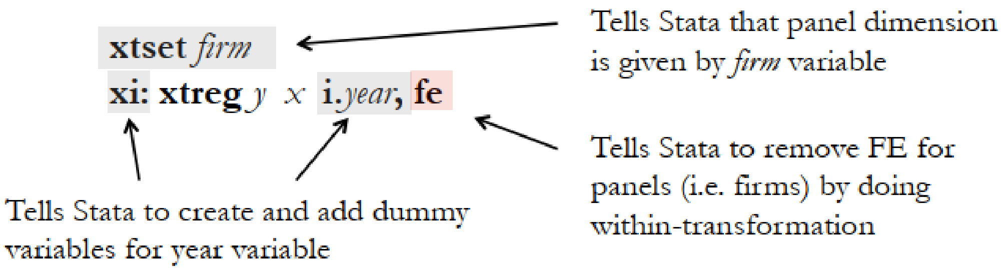
   - In this case, attenuation bias on mismeasured variable will go up.

3. Estimating a model with multiple types of FE can be computationally difficult.

   Consider the model:

$$y_{i,t} = \alpha + \beta x_{i,t} + \delta_t + f_i + u_{i,t}$$

   To estimate this in Stata, we will use a command something like the following:



```
xtset firm
xi: xtreg y x i.year, fe
```

Tells Stata that panel dimension is given by *firm* variable

Tells Stata to create and add dummy variables for year variable

Tells Stata to remove FE for panels (i.e. firms) by doing within-transformation

   Here the dummies not swept away in within-transformation are actually estimated. If we had to estimate 1000s of firm FE, this may be a problem! (This is in fact why we sweep away the firm FE rather than the year FE.)

Issues

1. Sometimes the predicted value of unobserved FE is of interest.

   This can be obtained by using $\hat{f}_i = \bar{y}_i - \hat{\beta}\bar{x}_i$.

   ▷ Bertrand and Schoar (QJE 2003) did this to back out CEO fixed effects, where they show that the CEO FE are jointly statistically significant from zero, suggesting that CEOs have "styles" that affect their firms.

   ▷ While these values are unbiased, they are inconsistent! This is called the Incidental Parameters Problem.

   ▷ Furthermore, doing an F-test to show they are statistically different from zero is only valid under rather strong assumptions, which requires that us are distributed normally, homoskedastic, and serially uncorrelated.

2. In practice, Logit, Tobit, and Probit should not be estimated with many fixed effects.

   This is because they only give consistent estimates under rather strong and unrealistic assumptions.

   ▷ Probit FE requires that unobserved f_is are distributed normally and f_i and $x_{i,t}$ to be independent (which is surely not true in CF).

   ▷ Logit FE requires no serial correlation of y after conditoning on the observable x and unobserved f (which is probably unlikely in many CF settings).

7.1.3 First Differences

First differencing is another way to remove unobserved heterogeneities. Rather than subtracting off the group mean of the variable from each variable, you instead subtract the lagged observation.

▷ This can be also done even when observations within groups aren't ordered by time. Just order the data within groups in whatever way you want, and take "differences."

Fixed Effects vs. First Differences

1. They are identical if there are just two observations per group. In other cases, the difference is generally about efficiency.

   ▷ FE is more efficient if the disturbances $(u_{i,t})$ are serially correlated.

   ▷ FD is more efficient if the disturbances $(u_{i,t})$ follow a random walk.

2. If strict exogeneity is violated – i.e. $x_{i,t}$ is correlated with $u_{i,s}$ for some $s \neq t$ – then FE might be better.

   This is because as long as we believe $x_{i,t}$ and $u_{i,t}$ are uncorrelated, the FE's inconsistency shrinks to 0 at rate $1/T$. But if y and x are spuriously correlated and N is small with T large, then FE can be quite bad.

3. Therefore, it is not a bad idea to try both. If they are very different, you should try to understand why. Usually, with an omitted variable or measurement error, you will get different answers with FD and FE.

   ▷ Griliches and Hausman (1986) shows that because measurement error causes predictably different biases in FD and FE, you can use the biased estimates to back out the true parameter.

7.1.4 Random Effects

RE is very similar to FE with one big difference: it assumes that the unobserved heterogeneity f_i and the observed x's are uncorrelated.

- ▷ This is not a realistic assumption for corporate finance. In fact, the violation of this assumption is the whole motivation behind why we do FE estimation.
- ▷ This assumption means that OLS will give you a consistent estimate of β. So why bother? Because of a potential efficiency gain relative to FE.

In practice, RE model is not very useful. As Angrist-Pischke write, relative to FE estimation, RE requires stronger assumptions to hold. Even if right, the asymptotic gain is likely modest, and the finite sample properties can be worse.[1]

[1] Gormley: "Bottom line, don't bother with it."

Part II
Alternative Approaches to Empirical Analysis

8 Multiple Goals and Approaches to Empirical Analysis

"I'll leave Eyo to discuss this tomorrow... he's a nice depository." – James Heckman

8.1 Styles of Empirical Research

We consider a categorization of studies and their features.

8.1.1 Comparison of Studies

We contrast four main ways to conducting studies:

1. Descriptive Studies: they describe "just the facts" and link to precisely formulated models is very vague. Use of primary data is central, and computationally the studies are simple.
   - ▷ There are lots of testing and robustness checks that involve use of multiple datasets.
   - ▷ Obviously, sources of identification is not a concern.
2. Causal Analysis: people try to estimate "the effect" but question being addressed is often not formulated within a precise economic model. Linear models are heavily favored with IV central.
   - ▷ Use of multiple dataset is encouraged, but styles vary widely.
   - ▷ See Fogle's biography of Kuznetz for a description of how Kuznetz did his empirical work
   - ▷ Source of identification: IV intuition (search for instrument or exclusion; randomization is an instrument)
3. Structural Analysis: there is a tight link to models, and the computational complexity depends on the problem, ranging from simple and modern game theory to dynamic contract theory
   - ▷ Replication is difficult given computational costs.
   - ▷ Source of identification is unclear – like all approaches, it requires external variation and cross-equation restrictions
4. Calibration: Used in macroeconomics and the use of primary data is more casual, since people pick a few "relevant" momenets for the data. Economic models are nonetheless complex.

We will explore methods somewhere between (2) and (3).

8.2 Abducting Economics: How to Learn from Surprises

It turns out that there is a fifth way of doing economics, which addresses a dimension that aforementioned four approaches do not address.

8.2.1 Motivation

One thing to realize is that data *never* speaks for themselves. This is an old fallacy.

- ▷ All analysts approach data with preconceptions, and sometimes preconceptions are encoded in precise models.
- ▷ Sometimes, they are just intuitions that analysts seek to confirm and solidify.
- ▷ Therefore, a central question is how to revise these preconceptions in the light of new evidence.

Noting this limitation, we can recognize that the aforementioned approaches all lack a formal guideline for taking the next step and learning from surprising findings. In other words, there is no established practice for dealing with surprise, even though surprise is an everyday occurrence. This begs the question – what is the best way to respond to data surprises?

8.2.2 Notion of Abduction

Abduction is the process of generating revising models, hypotheses, and data analyzed in response to surprising findings. Peirce[2] (1934) would describe this process of learning as the following: "The surprising fact C is observed. But if A were true, C would be a matter of course. Hence, there is reason to suspect that A is true." Note that this doesn't *prove* that A is true, but we *suspect*.

Example 8.1. (Deduction vs. Induction vs. Abduction) We contrast the following approaches:

- ▷ Deduction
  - * Rule: All the beans from a bag are white.
  - * Case: These beans are from that bag.
  - * Therefore, Result: These beans are white.
- ▷ Induction: Look at a bunch of repeated events
  - * Case: These beans are from this big.
  - * Result: These beans are white.
  - * Therefore, Rule: All the beans from this bag are white.
- ▷ Abduction: We don't know where these beans came from, but we know that a certain bag contains only white beans.
  - * Rule: All the beans from a bag are white.
  - * Result: These beans are white.
  - * Therefore, Case: These beans are from that bag.

We will argue that this sense of doubt lingering in abduction is what we do everyday.

Abduction is different from falsification or corroboration (Popper, 1959). It moves descriptions of the world forward rather than just confirming or falsifying hypotheses. The Popperian notion, on the other hand, is that you take hypotheses and test it out in the real world.

[2] Charles Sanders Peirce was an American philosopher, logician, mathematician, and scientist who is sometimes known as "the father of pragmatism". He was educated as a chemist and employed as a scientist for thirty years. Today he is appreciated largely for his contributions to logic, mathematics, philosophy, scientific methodology, semiotics, and for his founding of pragmatism.

- ▷ What if your matrix of elasticities is not semi-definite? What are you going to do? A Popperian view cannot answer these question.

It is part of a process of discovery where model reformulation, revision of hypotheses, and addition of new information are part of the process. This whole notion is based on "suspect." Suspicion opens the door to creativity.

The abductive model for learning from data follows more closely the methods of Sherlock Holmes than those of textbook econometrics. This approach uses many different kinds of clues of varying trustworthiness, weights them, puts them together, and tells a plausible story of the ensemble.

8.2.3 Abductive Approach vs. Standard Approach

The rich literature on adjustment of test statistics for multiple hypothesis testing does not account for the addition of hypothesis to the mix through the process of abduction. When we go back and forth with data – learning from it, revising hypotheses in light of it, augmenting it with fresh data and fresh theoretical insights – there is significant benefit.

Surprisingly, one of the worst examples of frequentist dogma in action is the common practice in government-sponsored research requiring that investigators specify all of their models in advance of looking at the data. But this is not a crime! It should be encouraged to explore data and amend hypotheses, as long as we report the standard errors to include these trial and errors.

Bayesians can deal with this. They want you to look at the data and for posteriors. The people in-between – the likelihood people – would also be fine with this approach.

Example 8.2. Suppose you got the data and sent it to New York as opposed to California. Should we include this fact in the test statistics? Intuitively no, but we do a version of these in everyday life.

8.2.4 Identification Problem

When we take a peak at the data, construct a model based on this quick observation and test it against the data, isn't this circular? A natural extension is to test other possible models that can arise from the peak and argue that our model is indeed correct. However, the dimension of all possible models is too big. Where do we draw the line? And more generally, how should we learn new things?

8.3 Example: Causality in the Time of Cholera

John Snow wanted to convincingly provide a causal mechanism, but even with ovewhelming evidence and strong analysis, Snow failed to convince the medical establishment, the public, or the authorities. Guess lecture by Thomas S. Coleman.

In 1849, he developed a theory of infection & transmission based on medical knowledge and study of single events. Basically, he had an abductive approach.

8.4 Example: Self-Employment over the Life Cycle

What are the factors that contribute to self-employment? This guy has a huge dataset from Sweden that has information on people's history. This is work by John Eric Humphries.

8.4.1 Overview

He did a machine learning exercise where he identified seven patterns characteristic of self-employed individuals. This is very inductive, puerely a data description procedure. He uses the notion of getting the minimal set of paths to describe this data with five states.

8.4.2 Nota Bene

He proceeded in three steps that reflects an abductive appraoch to empirical economics:

1. He looked at the data and generated some patterns.

   Using machine learning methods to summarize the patterns of self-employment behavior observed in the data, he finds that careers involving self-employment fit into a small number of economically distinct groups.

2. He constructed a very simple, price-theoretic two-period model (Roy model).

   Guided by the descriptive results, he builds an intuitive two-period model of self-employment decisions to interpret and rationalize the patterns reported in the previous section.

3. He built a full-blown structural model and use it to evaluate policies designed to promote self-employment.

   He extends the simple model in which self-employment decisions depend on factors such as cognitive and non-cognitive skills, prior work experience, the cost of capital, and other labor market opportunities.

Heckman notes that step 1 is often not reported in the paper, but Humphries chose to include this in the paper as well.

8.5 Chicago Approach to Empirical Economics

Friedman created a Chicago tradition and a Chicago approach to empirical work. He reacted strongly and negatively to Cowles econometrics, which is also known as “structural econometrics.” His objection was not the same as objections from treatment effect economists.

8.5.1 Cowles Commission

The Cowles Commission’s approach to the identification problem:

1. Define models (a priori)

2. Identify models in principle

3. Isolate which (if any) are cosnistent with the data (estimation and inference)

For Friedman, this was an artificial process, and he rejected the rigid separation of model formulation and model testing. It lacked a fourth stage: what happens when there’s a failure?

8.5.2 Friedman on Abduction

When commenting on Schultz’s *The Theory and Measurement of Demand*, Friedman excluded it from a list of scientific studies in economics because he “always tried to wrench the data into a pre-existing theoretical scheme, no matter how much of a wrench was required.

8.5.3 Measurement Without Theory Debate

Burns and Mitchell were collecting data to measure the business cycle. Koopmans criticized them for measuring without theory. Friedman and Koopmans never debated in person, but it soon launched into an *implicit* debate through a series of papers.

8.6 Samples

We discuss examples of non-random sampling:

- ▷ Truncated samples
- ▷ Censored samples
- ▷ Length-biased samples
- ▷ Choice-based samples
- ▷ Size-biased samples

8.6.1 Sampling rule

A sampling rule can be interpreted as producing a non-negative weighting function of $\omega\left(y,x\right)$ that alters the population density. The density of the sampled data is written as

$$g\left(y^{*},x^{*}\right)=\frac{\omega\left(y^{*},x^{*}\right)f\left(y^{*},x^{*}\right)}{\int\omega\left(y^{*},x^{*}\right)f\left(y^{*},x^{*}\right)dydx}$$

Using this framework, we can describe the following types of samples. Alternatively, define

$$\omega^{*}\left(y^{*},x^{*}\right)=\frac{\omega\left(y^{*},x^{*}\right)}{\int\omega\left(y^{*},x^{*}\right)f\left(y^{*},x^{*}\right)dydx}$$

to write:

$$g\left(y^{*},x^{*}\right)=\omega^{*}\left(y^{*},x^{*}\right)f\left(y^{*},x^{*}\right)$$

Furthermore, define $\Delta=1$ the occurrence of the event “a potential observation is sampled i.e. the values of y and x are observed, and $\Delta=0$ if not. Then a *truncated sample* is one for which $P\left(\Delta=1\right)$ is not know and cannot be identified. A censored sample is one for which $P\left(\Delta=1\right)$ is known or can be identified.

8.6.2 Truncated Samples and Random Variables

Let $f\left(X\right)$ be the density of a random variable. We observe X if $X<R$ (right truncation) or $X>L$ (left truncation). In a similar vein, we can define a truncated random variable. Suppose the weighting rule $\omega\left(y^{*},x^{*}\right)$ is known. If $\omega\left(y,x\right)$ is non-zero, $f\left(x,y\right)$ can be recovered:

$$\frac{g\left(y^{*},x^{*}\right)}{\omega\left(y^{*},x^{*}\right)}=\frac{f\left(y^{*},x^{*}\right)}{\int\omega\left(y^{*},x^{*}\right)f\left(y^{*},x^{*}\right)dy^{*}dx^{*}}$$

The requirement is that (1) the support of $\left(y,x\right)$ is known, and (2) $\omega\left(y,x\right)$ is non-zero. In many important problems, (2) is not often satisfied.

Example 8.3. (Truncated sample) Data are collected on incomes of individuals whose income Y exceeds a certain value c. So we observe Y if $Y > c$ and thus

$$\omega(y) = \begin{cases} 1 & \text{if } y > c \\ 0 & \text{if } y \leq c \end{cases}$$

Knowledge of the sampling rule does ont suffice to recover the population distribution. We do not observe values of Y below c.

Example 8.4. (Linear Regression) Let $Y = X\beta + U$ where the data are collected on incomes of persons for whom Y exceeds c. What can we identify?

▷ As before, let $Y^* = Y$ if $Y > c$ and define $\Delta = 1$ to be the event in which $Y > c$. Then:

$$P(\Delta = 1 | X = x)$$

To get the untruncated distribution, then you may try to parametrize the functional form. If it's normal, you can do it; if it's Pareto, you can't do it. The point is that it depends on some strict assumptions.

8.6.3 Censored Samples and Random Variables

We observe X^* as before, but we know the number of observations outside the interval. Usually, we encounter two types of censoring:

1. Type I censoring

   We only observe a variable if it lies in a range, and the number of values of Y outside the range is known.

2. Type II censoring

   In this case, a fixed proportion of the sample is censored in advance. For example, we stop observing light bulb burnout when we have a proportion.

The joint density of y^*, x^*, Δ for the case of a censored sample is obtained as:

$$\begin{aligned} g(y^*, x^*, \delta) = & \left[\frac{\omega(y^*, x^*) f(y^*, x^*)}{\int \omega(y^*, x^*) f(y^*, x^*) dy^* dx^*}\right]^\delta \\ & \times \left[\int i(y, x) f(y, x) dydx\right]^\delta \qquad \text{(probability of observing)} \\ & \times [1]^{1-\delta} \left[\int (1 - i(y, x)) f(y, x) dydx\right] \end{aligned}$$

Note that $i(y, x)$ and $\omega(y, x)$ cancel each other.

Example 8.5. (Linear Regression) Let Y_1 be the wage of a woman, and Y_2 be the index of a women's propensity to work. Wages of women are observed only if women work.

Example 8.6. (Linear Regression) Let Y_1 be the wage of a woman, and Y_2 be the index of a women's propensity to work. Wages of women are observed only if women work.

8.6.4 General Stratified Sampling

If the weighting rule $\omega\left(y^*, x^*\right)$ is non-zero and known, then we can recover $f\left(y^*, x^*\right)$ since:

$$\frac{g\left(y^*, x^*\right)}{\omega\left(y^*, x^*\right)} = \frac{f\left(y^*, x^*\right)}{\int \omega\left(y^*, x^*\right) f\left(y^*, x^*\right) dy^* dx^*}$$

8.6.5 Length-biased Sample

We illustrate using an example.

- ▷ Let T be the duration of a certain event (completed unemployment spell or a completed duration of a job with an employer) with distribution $F\left(t\right)$ and $f\left(t\right)$. The sampling rule here is that individuals are sampled at random. Importantly, data are recorded on a completed spell *provided that* at the time of interview, the individual is experiencing the event.

- ▷ Mathematically, we can decompose any spell T into a component that happens before the survey $\left(T_b\right)$ and a component that appears after the survey $\left(T_a\right)$, which yields $T = T_a + T_b$. For a person to be sampled, we also have the requirement that $T_b > 0$. Therefore, the density of T given $T_b = t_b$ is

$$f\left(t|t_b\right) = \frac{f\left(t\right)}{1 - F\left(t_b\right)}, \quad t \geq t_b$$

  i.e. the hazard rate.

With this formulation, we can characterize the unemployment rate.

- ▷ Suppose the environment is stationary, and the population entry rate into the state at each instant of time is k. From each vintage of entrants into the state distinguished by their distance from the survey date t_b, only $1 - F\left(t_b\right)$ survive.

- ▷ Therefore, aggregating across all cohorts yields:

$$P = \int_0^\infty k\left(1 - F\left(t_b\right)\right) dt_b$$

- ▷ is the fraction of people who are unemployed.

Now, we can show that longer spells are oversampled when the requirement is imposed that a spell be in progress at the time the survey is conducted $\left(T_b > 0\right)$.

- ▷ The density of T_b^* – sampled presurvey duration – is

$$g\left(t_b^*|t_b^* > 0\right) = \frac{k\left(1 - F\left(t_b^*\right)\right)}{P}$$

- ▷ The density of sampled completed durations is thus:

$$g\left(t^*\right) = \int_0^{t^*} f\left(t^*|t_b^*\right) g\left(t_b^*|t_b^* > 0\right) dt_b^* = \cdots = k\frac{t^* f\left(t^*\right)}{P}$$

Example 8.7. (Waiting Time Paradox) When waiting for a bus that comes on average every 10 minutes, your average waiting time will be 10 minutes. Naïvely, you might expect that if buses are coming every 10 minutes and you arrive at a random time, your average wait would be something like 5 minutes. In reality, though, buses do not arrive exactly on schedule, and so you might wait longer.

8.6.6 Choice-based Sample

Let D be a discrete valued random variable which assumes a finite number of values I . States are mutually exclusive. The interest is in estimating a population choice model:

$$P\left(D=i|X=x\right), \quad i=1,...,I$$

and the population density of $\left(D,X\right)$ is

$$f\left(d,x\right)=P\left(D=d|X=x\right)h\left(x\right)$$

where $h\left(x\right)$ is the density of the data.

- ▷ For example, interviews about transportation preferences conducted at train stations tend to over-sample train riders and under-sample bus riders.
- ▷ Interviews about occupational choice preferences conducted at leading universities over-sample those who select professional occupations.

Why does non-random sampling occur here?

- ▷ In choice-based sampling, selection occurs solely on the D coordinate of $\left(D,X\right)$. Then it can be shown that the sampled $\left(D^{*},X^{*}\right)$ satisfies:

$$g\left(d^{*}|x^{*}\right)=f\left(d^{*}|x^{*}\right)\times\underbrace{\left\{\left[\frac{\omega\left(d^{*}\right)}{\sum_{i=1}^{I}\omega\left(i\right)f\left(i\right)}\right]\left[\frac{1}{\sum_{i=1}^{I}f\left(i|x^{*}\right)\frac{g(i)}{f(i)}}\right]\right\}}_{[1]}$$

  i.e. the bias that results from using choice-based samples to make inferences about $f\left(d^{*}|x^{*}\right)$ is a consequence of neglecting [1].

  - f is the population density of choice; g is sample density of choice.
  - Note that if the data are generated by a random sampling rule (i.e. $w\left(d^{*}\right)=1$) then the term inside the braces is equal to 1 and we have $f=g$.

8.6.7 Size-biased Sample

Size-bias implies that we are more likely to over-sample large families.

8.7 Hypothesis Testing

Meaningful hypothesis testing requires that "significance levels" decrease with sample size.

8.7.1 Power and Sample Size

Is what sense does increasing sample size always lead to rejection of an hypothesis? We motivate this point with an example.

Example 8.8. Consider a one-tail normal test about a mean:

$$H_0 : \bar{X} \sim N\left(\mu_0, \sigma^2/T\right), \quad H_a : \bar{X} \sim N\left(\mu_A, \sigma^2/T\right)$$

where we assume σ^2 is known. Then for any c we get

$$P\left(\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/T}} > \frac{c-\mu_0}{\sqrt{\sigma^2/T}}\right) = \alpha\left(c\right)$$

and for a fixed α, we can solve for $c\left(\alpha\right)$:

$$c\left(\alpha\right) = \mu_0 - \frac{\sigma}{\sqrt{T}}\Phi^{-1}\left(\alpha\right)$$

Now consider the power of the test, i.e. the probability of rejecting μ_0 when μ_A is true. Using the same value of c as before:

$$\begin{aligned} P\left(\frac{\bar{X}-\mu_A}{\sqrt{\sigma^2/T}} > \frac{c-\mu_A}{\sqrt{\sigma^2/T}}\right) &= P\left(\frac{\bar{X}-\mu_A}{\left(\sigma/\sqrt{T}\right)} > \frac{\mu_0-\mu_A-\frac{\sigma}{\sqrt{T}}\Phi^{-1}\left(\alpha\right)}{\left(\sigma/\sqrt{T}\right)}\right) \\ &= P\left(\frac{\bar{X}-\mu_A}{\left(\sigma/\sqrt{T}\right)} > \frac{\mu_0-\mu_A}{\left(\sigma/\sqrt{T}\right)} - \Phi^{-1}\left(\alpha\right)\right) \\ &= \alpha \text{ when } \mu_0 = \mu_A \end{aligned}$$

If $\mu_A > \mu_0$, this probability goes to one.

Typically, a test statistic is some sort of average whose long term behaviour is governed by the strong and/or weak law of large numbers. As the sample size gets large, the distribution of the test statistic approaches that of a point mass --- under either the null or the alternative hypotheses.

Thus, as n gets large, the acceptance region gets smaller and closer to the value of the null. Intuitively, probable outcomes under the null and probable outcomes under the alternative no longer overlap - meaning that the rejection probability approaches 1 (under H_A) and 0 under H_0. Intuitively, increasing sample size is like increasing the magnification of a telescope. From a distance, two dots might seem indistinguishably close: with the telescope, you realize there is space between them. Sample size puts "probability space" between the null and alternative.

I can imagine situations where things don't work: if the number of nuisance parameters increases with sample size, things can fail to converge. In time series estimation, if the series is "insufficiently random" and the influence of the past fails to diminish at a reasonable rate, problems can arise as well.

The **punchline** is that if we measure X with the slightest error and the errors do not have mean zero, then we always reject H_0 for T big enough. If you want to test $\mu \in$ some set that contains zero, then the frequentist test of $H_0 : \mu = 0$ is bad test since for a sufficiently large sample size, you'll always reject.

Design of Sample Size Suppose that we fix the power $=\ \beta$ and pick $c\left(\alpha\right)$. Which sample size produces the desired power? We can find this by solving for:

$$P\left(\frac{\bar{X}-\mu_A}{\left(\sigma/\sqrt{T}\right)} > \frac{\mu_0-\mu_A}{\left(\sigma/\sqrt{T}\right)} - \Phi^{-1}\left(\alpha\right)\right) = \beta$$

which yields

$$\sqrt{T} = \frac{\Phi^{-1}\left(\beta\right) - \Phi^{-1}\left(\alpha\right)}{\Delta/\sigma}$$

8.7.2 Alternative Approaches

Classical inference is ex-ante; if designs an ex-ante rule that on average works well.

- For example, 5% of the time in repeated trials, we make an error of rejecting the null for a 5% significance level. This entails a hypothetical set of trials and is based on a long-run justification.
- Consistency of an estimator is an example of this mindset. But consistency shouldn't be a goal of in itself – if we use OLS for first 10^{100} observations and then use IV, it will likely have poor small sample properties, but on a long run frequency justification, it's just fine.

There are many examples in which people get very unhappy about classical testing procedures.

1. Consider (X_1, X_2) with $X_1 \perp X_2$ and we have

$$P_{\theta_0}\left(X_i = \theta_0 - 1\right) = P_{\theta_0}\left(X_i = \theta_0 + 1\right) = \frac{1}{2}, \forall i = 1, 2$$

One possible confidence set for θ_0 is

$$C\left(X_1, X_2\right) = \begin{cases} \frac{1}{2}\left(X_1 + X_2\right) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2 \end{cases}$$

in which $C\left(X_1, X_2\right)$ contains θ_0 75% of the time. Yet if $X_1 \neq X_2$, we are certain that the confidence interval exactly covers the true value 100% of the time. So after seeing the dat, we can get the exact value.

The **likelihood approach**, on the other hand, is based on the likelihood principle – all of the information is in the sample, and look at the likelihood as best summary of the sample. The Bayesian also looks at the likelihood, but it also uses the prior.

The **punchline** here is that Bayesians use the sample size to adjust "critical region" or the "rejection region." In the classical case, on the other hand, we had that with α fixed, the power of the test goes to 1.

9 Causal Analysis and Structural Analysis

9.1 Econometric Causality

An econometric approach to causality develops explicit models of outcomes where the causes of effects are investigated. We focus on why interventions work, which facilitates the design of estimators to solve selection and evaluation problems. The treatment effect model focuses on effects of causes,whereas an economic model examines the causes of the effects.

9.1.1 Structural Model

Consider

$$[1] : Y = X_b\beta_b + X_p\beta_p + U$$

where X_b is a set of background variables and X_p is a set of policy variables. In this case, if (β_b, β_p) are invariant to shifts in (X_b, X_p) and variables that cause such shifts, then we say $[1]$ is **structural**.

More generally speaking, consider

$$[2] : Y = G\left(X, \theta, U\right)$$

We say $[2]$ is structural if G is invariant to shifts in X.

9.1.2 Policy Evaluation Problems and Criteria of Interest

Consider the following three questions:

1. Evaluating the impact of implemented interventions on outcomes

   This focuses on impacts on a particular population – internal validity.

2. Forecasting the impact of interventions implemented in one environment to other environments

   This takes a treatment parameter or a set of parameters identified in one environment to another environment – external validity.

3. Forecasting the impact of interventions never historically experienced

   This entails structural models with new ingredients, and this is a problem that policy analysts solve daily.

9.1.3 Roy Model (1951)

Roy Model is a prototypical economic model for causal analysis and policy evaluation. Agents face two potential outcomes (Y_0, Y_1) characterized by distribution $F_{Y_0,Y_1}\left(y_0, y_1\right)$ where (y_0,y_1) are particular values of random variables (Y_0, Y_1). There are some immediate challenges in this framework:

- ▷ Evaluation problem: Analysts observe either Y_0 or Y_1, but not both for any given person.
- ▷ Selection problem: Values of Y_0 or Y_1 that are observed are not necessarily a random sample of the potential Y_0 or Y_1 distributions.
  - * For example, an agent may select into sector 1 if $Y_1 > Y_0$, which implies $D = \mathbf{1}\left(Y_1 > Y_0\right)$.

Example To see an example, let C be the cost of going from "0" to "1":

$$D = \mathbf{1}\left(Y_1 - Y_0 > C\right)$$

where C can depend on cost shifters such as Z:

$$\mathbb{E}\left[C|Z\right] = \mu_C\left(Z\right)$$

Z plays the role of instruments if $Z \perp (Y_0, Y_1)$. The observed outcome Y is $Y = DY_1 + (1-D)\,Y_0$. In advance of participation, the agent may be uncertain about all components of (Y_1, Y_0, C) so we subjectively evaluates the expected benefit:

$$I_D = \mathbb{E}\left[Y_1 - Y_0 - C|\mathcal{I}\right]$$

Implications Economic policies can operate through changing (Y_0, Y_1) or C, which can be brought about by changing both the X and Z. The structural approach considers policies affecting both returns and costs.

9.1.4 Treatment Effects vs. Policy Effects

The traditional parameters of interest include:

- ▷ Average Treatment Effect (ATE): $\mathbb{E}\left[Y_1 - Y_0\right]$
- ▷ Average Treatment on the Treated (ATT): $\mathbb{E}\left[Y_1 - Y_0|D = 1\right]$
- ▷ Average Treatment on the Untreated (ATUT): $\mathbb{E}\left[Y_1 - Y_0|D = 0\right]$

In policy, we want to determine the marginal returns to a policy, which would be $\mathbb{E}\left[Y_1 - Y_0|I_D = 0\right]$.

Policy Relevant Treatment Effects (PRTE) PRTE extends the ATE by accounting for voluntary participation in programs. Consider two policy regimes – baseline "b" and a policy "a" and the associated outcomes as

$$\left(Y_0^a, Y_1^a, C^a\right), \left(Y_0^b, Y_1^b, C^b\right)$$

Consider a form of policy invariance that keeps the potential outcomes unchanged for each person:

$$Y_0^a = Y_0^b, Y_1^a = Y_1^b, \qquad C^a \neq C^b$$

This invariance rules out social effects including peer effects and general equilibrium effects affecting possible outcomes. Furthermore, let D^a and D^b be the choices taken under each policy regime. Invoking the invariance of potential outcomes, the observed outcomes under each policy regime are:

$$\begin{aligned} Y^a &= Y_0 D^a + Y_1\left(1 - D^a\right) \\ Y^b &= Y_0 D^b + Y_1\left(1 - D^b\right) \end{aligned}$$

Then PRTE is defined to be

$$PRTE \equiv \mathbb{E}\left[Y^a - Y^b\right]$$

which is essentially a Benthamite comparison of aggregate outcomes under policies "a" and "b."

9.1.5 Econometric Approach vs. Treatment Effect Approach

The econometric approach examines the causes of effects – how Y_1 and Y_0 vary as X varies. The treatment effect approach looks at the effects of causes and does not examine choice mechanisms.

9.2 Causal Frameworks

We compare and contrast different causal frameworks.

1. Causal model based on potential outcomes – "Rubin-Holland" causal model.

   This is not really a causal model, so we will use this as

2. Causal model based on autonomous (≈ structural) equations) inspired by Haavelmo (1944)

3. Other causal frameworks based on Local Markov Conditions (LMC)

   This includes Pearl's Do-calculus (based on framework of structural equations with weird calculus) and the hypothetical framework of Heckman and Pinto.

9.2.1 Rubin-Holland Causal Framework

The potential outcome Y of agent ω fixed $T = t$ is $Y_\omega(t)$, and the causal effects of t' versus t is $Y_\omega(t) - Y_\omega(t')$. The observed outcome is then given by the switching regression framework, which says that we only say one realization of possible opportunities.

Example #1: Randomized Control Trials Identification requires the assumption that

$$Y(t) \perp T | X$$

in which case the average treatment effect (ATE) is

$$\mathbb{E}\left[Y(t_1) - Y(t_0)\right] = \int \left[Y_\omega(t_1) - Y_\omega(t_0)\right] dF(\omega)$$

Example #2: Matching The previous assumption we made:

$$Y(t) \perp T | X$$

also motivates matching, since agents ω are comparable when conditioned on observed values X. In RCT, we obtain exogenous variation of T under X by *design*; in matching, we obtain exogenous variation of T under X by *assumption*.

Example #3: Mediation Model Suppose we have an outcome (Y) such as earnings and schooling (T) as our treatment variable. We also have family background (M), which motivates:

$$Y = \alpha_0 + \alpha_2 T + \alpha_3 M + U$$

T can affect both Y directly and also indirectly through M, in which case M is a mediator for F. In this setup, the average direct effect is:

$$ADE(t) = \mathbb{E}\left[Y(t_1, M(t)) - Y(t_0, M(t))\right]$$

We can also decompose the total effect into direct effects and indirect effects:

$$\begin{aligned} TE &\equiv \mathbb{E}\left[Y(t_1, M(t_1)) - Y(t_0, M(t_0))\right] \\ &= \underbrace{\mathbb{E}\left[Y(t_1, M(t_1)) - Y(t_0, M(t_1))\right]}_{DE(t_1)} + \underbrace{\mathbb{E}\left[Y(t_0, M(t_1)) - Y(t_0, M(t_0))\right]}_{IE(t_0)} \end{aligned}$$

Essentially, we have the following relationship

$$T \to M \to Y$$

where we make the following statistical assumption known as **sequential ignorability**:

$$\begin{aligned}
[1] &: \left(Y\left(t', m\right), M\left(t\right)\right) \perp T | X \\
[2] &: Y\left(t', m\right) \perp M\left(t\right) | \left(T, X\right)
\end{aligned}$$

▷ [1] says that T is exogenous conditional on X.

▷ [2] says that M is exogenous conditioned on X and T, which is stronger than randomization.

These are very strong assumptions that are not testable.

Example #4: Instrumental Variables Model The statistical assumptions are:

$$\begin{aligned}
[1] &: Y\left(t\right) \perp Z \\
[2] &: Z \not\perp T
\end{aligned}$$

The Imbens and Angrist (1994) Monotonicity condition is that

$$T_{\omega}\left(z_0\right) \leq T_{\omega}\left(z_1\right), \forall \omega$$

Note that the exclusion restrictions are necessary but not sufficient to identify causal effect.

Criticisms

1. Not a proper causal framework, since it does not access causal relationships.

   Instead, it postulates conditional independence relationships. The causal relationships are implied, but never formally articulated.

2. This does not allow for unobserved variables.

   The method is defined only on the basis of observed variables.

3. Sequential ignorability does ont hold under the presence of either unobserved confounders or unobserved mediators. (See Heckman and Pinto, 2015a).

Structural equations can address this problems.

9.2.2 Causal (Structural) Model

The causal model is defined by four components:

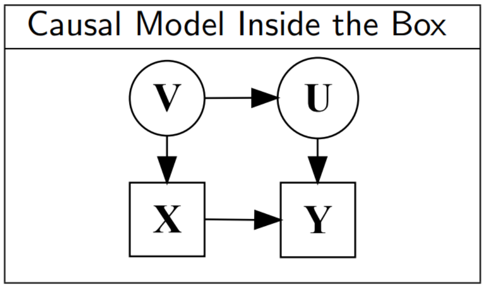
1. Random variables that are observed and/or unobserved by the analysts: $T = \{Y, U, X, V\}$
2. Error terms that are mutually independent $(\epsilon_Y, \epsilon_U, \epsilon_X, \epsilon_V)$
3. Structural equations that are autonomous i.e. deterministic and invariant to changes in the arguments
4. Causal relationships that maps the inputs causing each variable

We illustrate this using a few examples.

Example 9.1. Consider the model

$$
\begin{aligned}
Y &= f_Y\left(X, U, \epsilon_Y\right) \\
X &= f_X\left(V, \epsilon_X\right) \\
U &= f_U\left(V, \epsilon_U\right) \\
V &= f_V\left(\epsilon_V\right)
\end{aligned}
$$

Using a DAG, we can represent this as the following, ignoring the error terms:



- Children: variables that are caused by other variables. For example, $Ch\left(V\right) = \{U, X\}$.
- Descendants: variables that directly or indirectly cause other variables. For example, $De\left(V\right) = \{U, X, Y\}$.
- Parents: variables that directly cause other variables. For example, $Pa\left(Y\right) = \{X, U\}$.
- Recursive property: No variable is descendant of itself.

This representation is useful because we can apply **Bayesian Network tools**, which translates causal links into dependence relationships using graphical tools. It rests on two assumptions:

1. **Local Markov Condition (LMC)**

   This says that a variable is independent of its non-descendants conditioned on its parents:

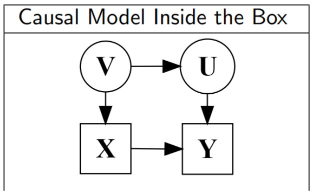
   $$Y \perp V \backslash \left(D\left(Y\right) \cup Y\right) | Pa\left(Y\right), \forall Y \in V$$

   i.e. you only have to worry about the parents.

2. **Graphoid Axioms (GA)**

   These rules define the whole set of operations defined for these parents and children.

Using this toolkit, we can take the example above and obtain the following representation:


Causal Model Inside the Box

Causal Model	LMC Relationships
$V = f_V(\epsilon_V)$	$V \perp\!\!\!\perp \emptyset \vert \emptyset$
$U = f_U(V, \epsilon_U)$	$U \perp\!\!\!\perp X \vert V$
$X = f_X(V, \epsilon_X)$	$X \perp\!\!\!\perp U \vert V$
$Y = f_Y(X, U, \epsilon_Y)$	$Y \perp\!\!\!\perp V \vert (U, X)$

- ▷ In other words, causal model ⇔ set of LMCs (one for each variable).

Fixing Operator Take f_Y to be autonomous, and consider a thought experiment in which fix $X = x$. This is different from conditioning X on x.

- ▷ If we condition on x, then we have

$$\begin{aligned} P\left(Y, V, U | X = x\right) &= P\left(Y | U, V, X = x\right) P\left(U | V, X = x\right) P\left(V | X = x\right) \\ &= P\left(Y | U, V, X = x\right) P\left(U | V\right) \underline{P\left(V | X = x\right)} \end{aligned}$$

where the last step follows from $U \perp X | V$ by LMC.

- ▷ If we fix X at x, then we have

$$P\left(Y, V, U | X \text{ fixed at } x\right) = P\left(Y | U, V, X = x\right) P\left(U | V\right) \underline{P\left(V\right)}$$

In data, V and X are dependent, but if I fix $X = x$, then any value of V should be possible.

Why does this distinction matter? When we condition on $X = x$, we are affecting other dependencies. But when we fix $X = x$, we are not! We can also represent the model through their error terms:

Standard Model	Model under Fixing
$V = f_V(\epsilon_V)$	$V = f_V(\epsilon_V)$
$U = f_U(f_V(\epsilon_V), \epsilon_U)$	$U = f_U(f_V(\epsilon_V), \epsilon_U)$
$X = f_X(f_V(\epsilon_V), \epsilon_X)$	$X = \mathbf{x}$

- ▷ Conditioning imposes term restriction on values error terms.
- ▷ Fixing imposes no restriction on values assumed by the error terms.

Fixing is a causal operator, not a statistical operator. Fixing does not affect the distribution of its ancestors, whereas conditioning affects the distribution of all variables. Y when conditioned on X is $Y | X = x$, whereas Y when X is fixed as x is $Y\left(x\right) = f_Y\left(x, U, \epsilon_Y\right)$.

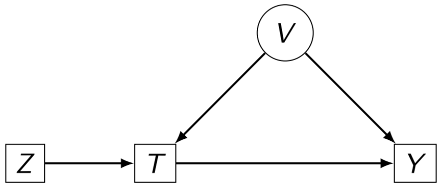
- ▷ Conditioning is a statistical exercise that considers the dependence structure of the data generating process, which yields:
$$\mathbb{E}\left[Y|X=x\right]=x\beta+\mathbb{E}\left[U|X=x\right];\quad \mathbb{E}\left[\epsilon_Y|X=x\right]=0$$
- ▷ Fixing is a causal exercise that hypothetically assigns values to inputs of the autonomous equation we analyze:
$$\mathbb{E}\left[Y\left(x\right)\right]=x\beta+\mathbb{E}\left[U\right];\quad \mathbb{E}\left[\epsilon_Y\right]=0$$
in which case the average treatment effect is
$$ATE=\mathbb{E}\left[Y\left(x\right)\right]-\mathbb{E}\left[Y\left(x'\right)\right]$$

Bayesian Networks Bayesian Networks represents a causal model as DAGs. Causal links are directed arrows; observed variables are displayed as squares and unobserved variables by circles. For example, consider the IV model:

$$[1]:Y\left(t\right)\perp Z$$
$$[2]:Z\not\perp T$$

LMC implies that $Y\perp Z|V,T$ and under fixing $T=t$, $Y\left(t\right)\perp T|V$ since when we fix $T=t$, it will no longer be dependent on V. Thus, V is a matching variable which generates a matching conditional independence relation. This is also called the "common cause" model. Note that $Y\left(t\right)$ can be random variable because of ϵ_Y which is independent of T and V .

Figure 1: DAG for the IV Model

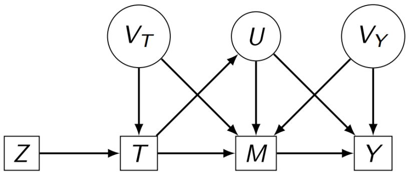


Remark 9.1. (Counterfactual conditions for LATE) When we are doing instrumental variables, we have T, Z, and Y where we are interested in the effect of T on Y in the presence of V that causes both T and Y. For individual ω , we have

$$Y_\omega=Y_\omega\left(1\right)T\left(1\right)+Y_\omega\left(2\right)T\left(2\right)$$

Here is another example. Consider the following model:

Figure 3: DAG for the Mediation Model with IV and Confounding Variables



▷ Let Z be the instrument. We can consider the following three IV regressions:

- * We are interested in the effect of T on Y where Z is an instrument for T. Note that $Z \not\perp T$.
  - Then, from LMC and the axioms, we have the exclusion restriction that $Z \perp Y(t)$.
- * We are interested in the effect of T on M where Z is an instrument for T. Note that $Z \not\perp T$.
  - Then, from LMC and the axioms, we have the exclusion restriction that $Z \perp M(t)$.
- * We are interested in the effect of M on Y where Z is an instrument for M. Note that $Z \not\perp M|T$.
  - Then, from LMC and the axioms, we have the exclusion restriction that $Z \perp Y(m)|T$. x

10 Some General Principles of Estimators

We will discuss some of the estimators we saw earlier and discuss the general principles that will help unify the literature. In a lot of the literature, it gives you a list of things, but students and analysts using this do not understand what are the general principles that underlie these ideas. We will try to unify the cross-sectional estimators, as well as the time-series estimators as well.

10.1 Assumptions in Estimators

▷ The intention-to-treat (ITT) effect, albeit often reported in journals, is not the most ideal.

▷ Least-squares and matching are less conservative than control functions or replacement functions approach since they rule out a major identification problem that stems from the possible asymmetry in information between the agents making participation decisions and the observing economist.

10.1.1 Randomization

We write randomization as $(Y_0, Y_1) \perp D$. When will this be satisfied?

1. Agents (decision makers whose choices are being analyzed) pick outcomes that are random with respect to (Y_0, Y_1) .

   Agents may not know (Y_0, Y_1) at the time of their choice, so we have

   $$P(D = 1|X, Y_0, Y_1) = P(D = 1|X)$$

   This can be illustrated using the Roy framework.

2. Individuals are randomly assigned to treatment status even if they would choose to self select into no-treatment status, and they comply with the randomization protocols.

   This is the more standard case in randomization. Let ζ be randomized assignment status, which satisfies $(Y_0, Y_1) \perp \xi$ under random assignment. Furthermore, let A denote the actual treatment status. If the randomization has full compliance among participants, we have $\zeta \Rightarrow A$.

   ▷ In this case, we have the well-known result that

   $$TT = TUT = ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

   under either (1) constant effects, or (2) the case in which

   $$Y_1 - Y_0 = M_1(x) - M_0(x) + (U_1 - U_0)$$

It is critical to observe that even with random assignment of treatment and full compliance, one *cannot* in general identify the distribution of the treatment effects. One can identify the joint distributions if we assume $Y_1 - Y_0 = \Delta(x)$, a constant given $X = x$. In general, however, the joint distribution is not identified unless the analyst can pin down the dependence across $(Y_{0,}Y_1)$. For example, the Athey-Imbens method for quantile effects is resting on strong assumptions.

10.1.2 Illustration via Structural Model

Now consider the following model for (Y_0, Y_1) :

$$Y_1 = \mu_1(X) + U_1$$
$$Y_0 = \mu_0(X) + U_0$$

where (μ_1, μ_0) are structural. What does randomization identify in this case?

▷ We get:

$$\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X] = \overbrace{\mu_1(X) - \mu_0(X) + \underbrace{\mathbb{E}[U_1|X] - \mathbb{E}[U_0|X]}_{\text{constructed over population}}}^{\text{ATE}}$$

i.e. the ATE consists of the structural parameters and the secondary term.

Now suppose a *common coefficient model* where $U = U_0 = U_1$ and $Y = D(1-D)Y_0 = Y_0 + D(Y_1 - Y_0)$. Plugging in:

$$Y = \mu_0(X) + D(\mu_1(X) - \mu_0(X)) + U$$

Assuming D is randomized with perfect compliance, we have

$$\mathbb{E}[Y|D, X] = \mu_0(X) + D(\mu_1(X) - \mu_0(X)) + \underbrace{\mathbb{E}[U|D, X]}_{=\mathbb{E}[U|X]}$$

in which case $D = \mu_1(X) - \mu_0(X)$ is identified. Randomization is thus balancing the bias.

10.1.3 Imperfect Compliance

If the treatment is chosen by self-selection $(D \Rightarrow A)$ and there is imperfect compliance $(\xi = 1 \not\Rightarrow A = 1)$, then we write $A = \xi D$. In this case, we can identify the intention to treat (ITT):

$$ITT = \mathbb{E}[Y|R = 1, D = 1] - \mathbb{E}[Y|R = 0, D = 0]$$

where D is the choice in the absence of randomization, and R is the result of the randomization. This is not the best approach.

▷ Note that with perfect compliance, we would have

$$P(D = 1|R = 1) = 1, \quad P(D = 0|R = 1) = 0$$
$$P(D = 1|R = 0) = 1, \quad P(D = 0|R = 0) = 1$$

In short, ITT mixes choices (preferences – subjective evaluation) with objective outcome. Although ITT is reported in many journal articles (especially QJE), this is not an ideal metric.

10.1.4 Intention to Treat (ITT)

An intention-to-treat (ITT) analysis of the results of an experiment is based on the initial treatment assignment and not on the treatment eventually received. Intention to treat analyses are done to avoid the effects of crossover and dropout, which may break the random assignment to the treatment groups in a study.

▷ Mathematically, the unconditional ITT is

$$\mathbb{E}[Y|R = 1] - \mathbb{E}[Y|R = 0]$$

i.e. the average effect of assigning treatment to an individual.

▷ In general, the unconditional ITT represents the average effect of assigning treatment to an individual.

10.1.5 Matching

Matching is hailed as a solution to achieve randomization when the analyst has access to X that effectively produces a randomization of D with respect to (Y_0, Y_1) given X. We usually invoke the following two conditions that justify matching:

$$\begin{aligned}
[M1] &: (Y_0, Y_1) \perp D|X \\
[M2] &: 0 < P(D = 1|X = x) < 1
\end{aligned}$$

where $[M2]$ is required for *any* evaluation estimator that compares treated and untreated persons. Assumption $[M1]$ is strong, and many economists do not have enough faith in their data to invoke it. Assumption $[M2]$ is testable and requires no act of faith.

- ▷ From $[M1]$ and $[M2]$, it is possible to identify $F_1(Y_1|x = x)$ from the observed data since

$$\begin{aligned}
F_1(Y_1|D = 1, X = x) &= F_1(Y_1|X = x) \quad (\text{from } M1) \\
&= F_1(Y_1|D = 0, X = x)
\end{aligned}$$

  and similarly:

$$F_0(Y_0|D = 0, X = x) = F_0(Y_0|D = 1, X = x)$$

- ▷ What's the issue here? Since the pair of outcomes (Y_0, Y_1) is not identified for anyone, the joint distributions of (Y_0, Y_1) given X or of $Y_1 - Y_0$ given X are not identified without further information.
- ▷ At values of X that fail to satisfy $[M2]$, there is no variation in D given X.

In short, analysts using matching make strong information assumptions in terms of the data available to them.

10.1.6 Information Asymmetry

To analyze the informational assumptions invoked in matching and other econometric evaluation strategies, it is helpful to introduce five distinct information sets and establish some relationships among them. Recall:

$$\begin{aligned}
[M1] &: (Y_0, Y_1) \perp D|X \\
[M2] &: 0 < P(D = 1|X = x) < 1
\end{aligned}$$

1. Relevant Information Set: $\sigma(I_{R^*})$ is an information set associated with random variable that satisfies conditional independence $(M1)$
2. Minimal Information Set: $\sigma(I_R)$ is the minimal information set that satisfies $(M1)$.
3. $\sigma(I_A)$ denotes the information set available to the agent at the time of participation.
4. $\sigma(I_{E^*})$ denotes the information set available to the economist.
5. $\sigma(I_E)$ denotes the information used by the economist in conducting an empirical analysis.

Why is this notion important? The methods of control functions, replacement functions, proxy variables, and instrumental variables all recognize the possibility of asymmetry in information between the agent being studied and the econometrician. In matching, however, assumption $[A1]$ rules out a major identification problem that stems from the possible asymmetry in information between the agents making participation decisions and the observing economist. This motivates the following:

$$\begin{aligned}
[U1] &: (Y_0,Y_1) \perp D|X, Z, \theta \\
[U2] &: (Y_0, Y_1) \not\perp D|X, Z
\end{aligned}$$

where Z is different from X and akin to an instrumental variable.

- ▷ Under $[U2]$, these approaches model the relationships of the θ to Y_1, Y_0, D in various ways.
- ▷ For example, the control function principle specifies the exact nature of the dependence of the relationship between observables and unobservables in a non-trivial fashion that is consistent with economic theory.

10.2 Alternative Estimators

In this sectino, we discuss the estimators that make weaker assumptions than least-squares and matching.

10.2.1 Instrumental Variables (IV)

Recall the assumptions for IV:

$$\begin{aligned}[IV1] &: (Y_0, Y_{1,} D(z)) \perp Z|X \\ [IV2] &: \mathbb{E}[D|X,Z] = P(X,Z) \text{ is a non-degenerate function of } Z \text{ given } X\end{aligned}$$

where $[IV1]$ is not testable and $[IV2]$ is empirically testable. Also recall the assumptions for matching:

$$\begin{aligned}[M1] &: (Y_0, Y_1) \perp D|X \\ [M2] &: 0 < P(D=1|X=x) < 1\end{aligned}$$

- ▷ Comparing $[IV1]$ and $[M1]$:

  $$\begin{aligned}[IV] &: (Y_0, Y_1) \perp Z|X \\ \text{[Matching]} &: (Y_{0,} Y_1) \perp D|X\end{aligned}$$

  we see that Z plays the role of D in the matching condition.
- ▷ Comparing $[IV2]$ and $[M2]$:
  - * In matching, D varies conditional on X and this is the source of identifying information.
  - * In IV, the choice probability varies with Z conditional on X.
- ▷ Opposite roles for D
  - * In matching, the variation in D that arises after conditioning on X provides the source of randomness that switches people across treatment status.
  - * In IV, the variation in $P(X,Z)$ produces variations in D that switches treatment status. The components of variation in D not predictable by (X,Z) is a problem for IV, while for matching it is the source of identification.

10.2.2 Replacement Functions

Recall the assumptions:

$$\begin{aligned}[U1] &: (Y_{0,} Y_1) \perp D|X, Z, \theta \\ [U2] &: (Y_0, Y_1) \not\perp D|X, Z\end{aligned}$$

Replacement functions (Heckman and Robb, 1985) proxy θ and they substitute out for θ using observables.

- ▷ Basically, if θ is ability and τ is a test score, and if we think that

  $$\tau = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$$

  then we can control for τ, X, Q, Z to control for θ. Notice that one does not need to know the coefficients to implement this method. It only requires conditioning on these variables, which then gives us

  $$(Y_0, Y_1) \perp D | X, Z, Q$$

  which is a version of matching.

This method has been used in the economics of education for decades as we saw in the example above.

10.2.3 Factor Models

In many applications, θ is measured with error. This motivates the factor model that is represented as a system of equations:

$$\begin{aligned} Y_1 &= g_1 \left(X, Z, Q, \theta, \epsilon_1\right) \\ Y_0 &= g_0 \left(X, Z, Q, \theta, \epsilon_0\right) \end{aligned}$$

Alternatively, a linear factor model separable in the unobservables are:

$$\begin{aligned} Y_1 &= g_1 \left(X, Z, Q\right) + \alpha_1 \theta + \epsilon_1 \\ Y_0 &= g_0 \left(X, Z, Q\right) + \alpha_0 \theta + \epsilon_0 \end{aligned}$$

where $(X, Z) \perp (\theta, \epsilon_j), \epsilon_j \perp \theta, \quad j = 0, 1$ and the ϵ_js are mutually independent.

- ▷ In the above equations, Y_j controlling for X and mZ are only imperfect proxies of θ since ϵ_js are present.

10.2.4 Control Function

Most models that are linear in parameters are estimated using standard IV methods – either 2SLS method or GMM. An alternative is the control function (CF) approach, which relies on the same kinds of identification conditions.

- ▷ In the standard case where endogenous explanatory variables appear linearly, CF approach collapses to the usual 2SLS estimator. For models non-linear in parameters, the CF approach offers some distinct advantages.

The basic idea is that when the control function is inserted into the outcome equation (implicitly subtracted from U_i) the purged disturbance $(U_i - \kappa_i (X_i, Z_i))$ is orthogonal of the RHS variables in the new outcome euqation.

- ▷ For example, define $\mathbb{E}[U_i | X_i, d_i = 1] = \kappa(X_i, Z_i, \lambda)$ and write:

  $$Y_i = X_i \beta + \kappa (X_i, Z_i, \lambda) + (U_i - \kappa_i (X_i, Z_i))$$

Example 10.1. (IV vs. Control Function) [3]Consider a training program application in which Y_{it} represents earnings; X_{it} is characteristics; and d_i is dummy if person i received training. This allows us to write:

$$\begin{aligned} Y_{it} &= X_{it} \beta + d_i \alpha + U_{it}, \quad t > k \\ Y_{it} &= X_{it} \beta + U_{it}, \quad t \leq k \end{aligned}$$

We need to model who receives training:

$$Y_i^* = Z_i \gamma + V_i$$

where Y_i^* is a latent varaible representing payoffs to training person i, and $d_i = 1 \Leftrightarrow Y_i^* > 0$.

[3] This example is taken from https://www.ssc.wisc.edu/~walker/wp/wp-content/uploads/2013/09/E718Lec7slides.pdf.

- ▷ In this setup, sample selection occurs if $\mathbb{E}\left[U_{it}d_i\right] \neq 0$. There are two possible sources: (1) dependence between U_{it} and V_i or (2) dependence between U_{it} and Z_i.
- ▷ IV estimator is consistent for α under either source of sample selection.
- ▷ In the CF approach, K_{it} is a control function for the earnings equation if we have

$$\mathbb{E}\left[(U_{it} - K_{it}) \cdot d_i\right] = 0$$
$$\mathbb{E}\left[(U_{it} - K_{it}) \cdot X_i\right] = 0$$
$$\mathbb{E}\left[(U_{it} - K_{it}) \cdot K_i\right] = 0$$

where we can substitute into the earnings equation to obtain:

$$Y_{it} = X_{it} + d_j\alpha + K_{it} + (U_{it} - K_{it})$$

Note that the CF approach requires that K is a known function. Since CF assumptions are stronger than IV, it is typically less robust than IV. And most importantly, CF extends easily to non-linear models.

10.2.5 Summary

The following slides summarize the comparisons:

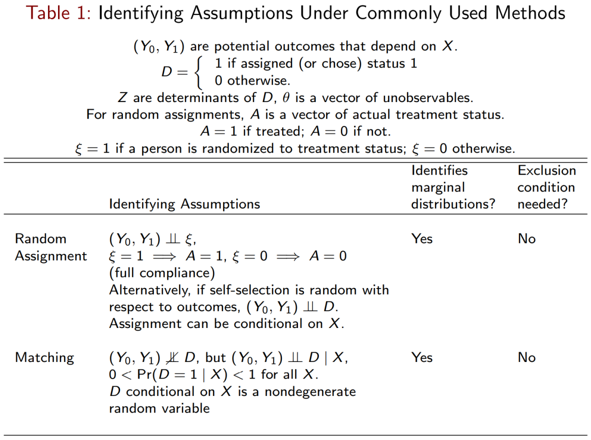


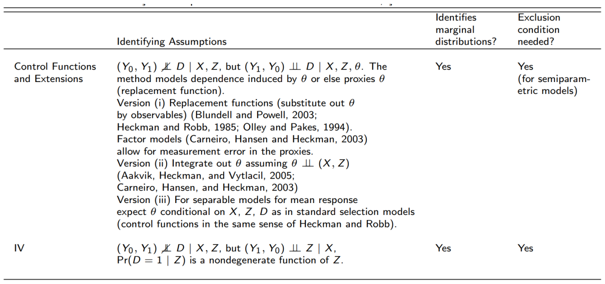
Table 1: Identifying Assumptions Under Commonly Used Methods

(Y_0, Y_1) are potential outcomes that depend on X.

$$D = \begin{cases} 1 \text{ if assigned (or chose) status 1} \\ 0 \text{ otherwise.} \end{cases}$$

Z are determinants of D, θ is a vector of unobservables.
For random assignments, A is a vector of actual treatment status.
$A = 1$ if treated; $A = 0$ if not.
$\xi = 1$ if a person is randomized to treatment status; $\xi = 0$ otherwise.

	Identifying Assumptions	Identifies marginal distributions?	Exclusion condition needed?
Random Assignment	$(Y_0, Y_1) \perp\!\!\!\perp \xi$, $\xi = 1 \implies A = 1$, $\xi = 0 \implies A = 0$ (full compliance) Alternatively, if self-selection is random with respect to outcomes, $(Y_0, Y_1) \perp\!\!\!\perp D$. Assignment can be conditional on X.	Yes	No
Matching	$(Y_0, Y_1) \not\perp\!\!\!\perp D$, but $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$, $0 < \Pr(D = 1 \mid X) < 1$ for all X. D conditional on X is a nondegenerate random variable	Yes	No



	Identifying Assumptions	Identifies marginal distributions?	Exclusion condition needed?
Control Functions and Extensions	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp D \mid X, Z, \theta$. The method models dependence induced by θ or else proxies θ (replacement function). Version (i) Replacement functions (substitute out θ by observables) (Blundell and Powell, 2003; Heckman and Robb, 1985; Olley and Pakes, 1994). Factor models (Carneiro, Hansen and Heckman, 2003) allow for measurement error in the proxies. Version (ii) Integrate out θ assuming $\theta \perp\!\!\!\perp (X, Z)$ (Aakvik, Heckman, and Vytlacil, 2005; Carneiro, Hansen, and Heckman, 2003) Version (iii) For separable models for mean response expect θ conditional on X, Z, D as in standard selection models (control functions in the same sense of Heckman and Robb).	Yes	Yes (for semiparam-etric models)
IV	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp Z \mid X$, $\Pr(D = 1 \mid Z)$ is a nondegenerate function of Z.	Yes	Yes

11 Generalized Roy Model as a Fundamental Economic Model

We will go through this in more detail and emphasize certain aspects.

11.1 Basic Selection Model

In a truncated model, we only have observations of a sample selected by the dependent variable. Therefore, OLS is inconsistent, and it is imperative that we take this into account.

11.1.1 Some Examples

http://www.eco.uc3m.es/~ricmora/mADE/materials/Heckman_MADE_handout.pdf

11.2 Generalized Roy Model

This was developed to deal with selection. An example is when we want to look at wages when not everyone works.

11.2.1 Two Potential Outcome Model

This is a model of comparative advantage with compositional effects and selection models. This was originally in a wage setting, since people respond to skill prices in different sectors by moving to the sector that gives him the highest return.

11.3 Connecting LATE and the Roy Model

In this section, we synthesize the concepts from Magne's class with the framework developed in Heckman's class.

11.3.1 Recap of LATE

Three assumptions define LATE:

1. $(Y_0, Y_{1,} \{D(z)\}) \perp Z|X$
2. $P(D=1|Z=z)$ is non-trivial function of z conditional on X .
3. $D\left(z^1\right) \geq D\left(z^2\right)$ for all persons OR $D\left(z^1\right) \leq D\left(z^2\right)$ for all persons.

   Note that this is a statement across people, so for any person, $D(z)$ need not be monotonic in z.

LATE does not identify which people are induced to change their treatment status by the change in te instrument. Therefore, it leaves unanswered many policy questions.

11.3.2 Identifying Policy Parameters

Consider the following setup:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0 \\ C &= \mu_C(Z) + U_C \end{aligned}$$

where (X, Z) are observed by the analyst where Z includes all of X and U_0, U_1, U_C are unobserved. Further define:

$$\begin{aligned} I_D &= \mathbb{E}\left[Y_1 - Y_0 - C|\mathcal{I}\right] = \mu_D\left(Z\right) - V \\ \mu_D\left(Z\right) &= \mathbb{E}\left[\mu_1\left(X\right) - \mu_0\left(X\right) - \mu_C\left(Z\right)|\mathcal{I}\right] \\ V &= -\mathbb{E}\left[U_1 - U_0 - U_C|\mathcal{I}\right] \end{aligned}$$

in which the choice equation becomes:

$$D = 1\left\{\mu_D\left(Z\right) > V\right\}$$

▷ We can manipulate the choice equation under the assumption that $Z \perp V$:

$$\begin{aligned} P\left(z\right) &= P\left(D = 1|Z = z\right) \\ &= P\left(\mu_D\left(z\right) \geq V\right) \\ &= P\left(\frac{\mu_D\left(z\right)}{\sigma_V} \geq \frac{V}{\sigma_V}\right) \\ &= P\left(F_{\frac{V}{\sigma_V}}\left(\frac{\mu_D\left(z\right)}{\sigma_V}\right) \geq \underbrace{F_{\frac{V}{\sigma_V}}\left(\frac{V}{\sigma_V}\right)}_{\equiv U_D}\right) \\ &= P\left(P\left(z\right) \geq U_D\right) \end{aligned}$$

i.e. $P\left(z\right)$ is the $P\left(z\right)$th quantile of U_D where U_D is Uniform$(0, 1)$.

▷ This allows us to write:

$$\begin{aligned} \mathbb{E}\left[Y|Z = z\right] &= \mathbb{E}\left[Y_0 + D\left(Y_1 - Y_0\right)|Z = z\right] \\ &= \mathbb{E}\left[Y_0\right] + \mathbb{E}\left[Y_1 - Y_0|D = 1, Z = z\right]P\left(D = 1|Z = z\right) \\ &= \mathbb{E}\left[Y_0\right] + \mathbb{E}\left[Y_1 - Y_0|P\left(z\right) \geq U_D\right]P\left(z\right) \end{aligned}$$

11.4 Separating Heterogeneity from Uncertainty

This section pertains to the discussion about the information set on which agents act on i.e. make choices based on. The discussion starts from Carneiro et al. (2003) who separate earnings heterogeneity (information about future earnings known to agents and acted on in their choices) vs. forecastable uncertainty.

11.4.1 Intuition

The method uses choice information to extract ex-ante or forecast components of earnings and distinguishes them from realized earnings. The difference between forecast and realized earnings allows us to identify the distributions of the components of uncertainty facing agents at the time they make their schooling decisions.

11.4.2 Model

They model the following components:

1. Earnings Equations: They use the Roy model (1951).
2. Choice Equations: The cost is $C = Z\gamma + U_C$.

3. Cognitive Ability: M_k (agent's score on the kth test) can be expressed in terms of the conditinoing variables X^M.
4. Heterogeneity and Uncertainty: $Y_{s,t} = \mathbb{E}\left[Y_{s,t}|\mathcal{I}_t\right] + V_{s,.t}$
5. Factor models: Model the unobservables as $U_{s,t} = \theta\alpha_{s,t} + \epsilon_{s,t}$
6. Test Score Equations: $M_k = X^M\beta_k^M + \theta_1\alpha_k^M + \epsilon_k^M, \quad k = 1, ..., K$

12 Simultaneous Equations and Causality

12.1 Simultaneous Causality

Conditioning – i.e. least squares – can break down in much more general cases.

12.1.1 Setup

Consider a linear model in terms of the parameters (Γ, B), observables (Y, X) and unobservables (U):

$$\Gamma Y + BX = U, \quad \mathbb{E}[U] = 0$$

where Y is a vector internal and interdependent variables; X is external and exogenous; and Γ is a full rank matrix. To simplify the setup, consider a two-agent model of social interactions. This is the **structural form**

$$\begin{aligned} Y_1 &= \alpha_1 + \gamma_{12} Y_2 + \beta_{11} X_1 + \beta_{12} X_2 + U_1 \\ Y_2 &= \alpha_2 + \gamma_{21} Y_1 + \beta_{21} X_1 + \beta_{22} X_2 + U_2 \end{aligned}$$

whereas the **reduced form** is:

$$\begin{aligned} Y_1 &= \pi_{10} + \pi_{11} X_1 + \pi_{12} X_2 + \mathcal{E}_1 \\ Y_2 &= \pi_{20} + \pi_{21} X_1 + \pi_{22} X_2 + \mathcal{E}_2 \end{aligned}$$

We allow U_1 and U_2 to be freely correlated. We further assume that

$$\mathbb{E}[U_1 | X_1, X_2] = 0, \quad \mathbb{E}[U_2 | X_{1,} X_2] = 0$$

12.1.2 Interpretation

In the setup above, the causal effect of Y_2 on Y_1 is γ_{12} and of Y_1 on Y_2 is γ_{21}. Can we identify them?

- ▷ Traditional Argument: Conditioning (i.e. using least squares) in general fails to identify these causal effects because U_1 and U_2 are correlated with Y_1 and Y_2.
- ▷ Our Argument: Even if $U_1 = 0, U_2 = 0$, least squares breaks down because Y_2 is perfectly predictable by X_1 and X_2. This implies that we **cannot** simultaneously vary $Y_2, X_{1,}$ and X_2.

12.1.3 Analysis through Reduced-Form

Given the **structural form:**

$$\begin{aligned} Y_1 &= \alpha_1 + \gamma_{12} Y_2 + \beta_{11} X_1 + \beta_{12} X_2 + U_1 \\ Y_2 &= \alpha_2 + \gamma_{21} Y_1 + \beta_{21} X_1 + \beta_{22} X_2 + U_2 \end{aligned}$$

we can express the **reduced form** as:

$$\begin{aligned} Y_1 &= \pi_{10} + \pi_{11} X_1 + \pi_{12} X_2 + \mathcal{E}_1 \\ Y_2 &= \pi_{20} + \pi_{21} X_1 + \pi_{22} X_2 + \mathcal{E}_2 \end{aligned}$$

- ▷ Simple algebra yields:

$$\pi_{11} = \frac{\beta_{11} + \gamma_{12}\beta_{21}}{1 - \gamma_{12}\gamma_{21}}, \quad \pi_{12} = \frac{\beta_{12} + \gamma_{12}\beta_{22}}{1 - \gamma_{12}\gamma_{21}}$$

Without further information on the variances of (U_1, U_2) and their relationship to the causal parameters, we cannot identify the causla effects γ_{12} and γ_{21} from the reduced-form regression coefficients.

▷ This is because in this setup:

$$Y_1 = \alpha_1 + \gamma_{12}Y_2 + \beta_{11}X_1 + \beta_{12}X_2 + U_1$$
$$Y_2 = \alpha_2 + \gamma_{21}Y_1 + \beta_{21}X_1 + \beta_{22}X_2 + U_2$$

Holding X_1, X_2, U_1 and U_2 fixed in the first equation, we cannot vary Y_2 because it is an exact function of $X_{1,}X_{2,}U_{1,}U_2$. This depends even if $U_1 = 0$ and $U_2 = 0$ so that there are no unobservables.

If we assume $(\beta_{12} = 0)$ or $(\beta_{21} = 0)$ or both, we can then identify the ceteris paribus causal effects of Y_2 on Y_1 and of Y_1 on Y_2.

12.2 Dummy Endogenous Variables

This model relies critically on the notion that discrete endogenous variables are generated by continuous latent variables crossing thresholds. We will see that this class of statistical models provides a natural framework for generating simultaneous equatino models with both discrete and continuous random variables.

12.2.1 Setup

Consider a pair of simultaneous equations for continuous latent random variables y_{1i}^* and y_{2i}^*:

$$[1a] : y_{1i}^* = X_{1i}\alpha_1 + d_i\beta_1 + y_{2i}^*\lambda_1 + U_{1i}$$
$$[1b] : y_{2i}^* = X_{2i}\alpha_2 + d_i\beta_2 + y_{1i}^*\lambda_2 + U_{2i}$$

where dummy variable d_i is endogenously defined by

$$[1c] : d_i = \begin{cases} 1 & \text{iff } y_{2i}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

In the special case where $\beta_1 = \beta_2 = 0$ and both y_{1i}^* and y_{2i}^* are observed, this conforms to the classical simultaneous equation model where standard methods are available to estimate all of the parameters. However in this case, the model is cast in terms of latent variables y_{1i}^* and y_{2i}^*, which may or may not be directly observed. Even if y_{2i}^* is never observed, the event $y_{2i}^* > 0$ is observed and its occurrence is recorded by setting a dummy variable d_i equal to one.

▷ Interpretation #1: $[1a]$ is the demand curve and $[1b]$ is the supply curve, where y_{1i}^* is quantity and y_{2i}^* is price observed at some market. $[1c]$ says that if the price exceeds some threshold, the government takes some actions that shift both the supply curve (by β_2) and the demand curve (by β_1) – subsidy to consumers and a per unit subsidy to producers.

▷ Interpretation #2: y_{1i}^* is the measured income of blacks in state i while y_{2i}^* is an unmeasured variable that reflects the population's sentiment towards blacks. If y_{2i}^* is high enough, the state may enact antidiscrimination legislation, where the presence of such legislation can be measured by a dummy variable $d_i = 1$.

12.2.2 Applications

The above setup includes a wide variety of interesting econometric models.

1. Classical SEM

   This model arises when (y_{1i}^*, y_{2i}^*) are observed, and there is no structural shift in the equations $(\beta_1 = \beta_2 = 0)$.

2. Classical SEM with Structural Shift

3. Multivariate Probit Model

   This model arises when (y_{1i}^*, y_{2i}^*) are not observed, but the events $y_{1i}^* > 0, y_{2i}^* > 0$ are observed. This requires adding two dummy variables. No structural shift is permitted, i.e. $\beta_1 = 0$ and $\beta_2 = 0$.

4. Multivariate Probit Model with Structural Shift

5. Hybrid Model

   This model arises when (y_{1i}^*) is observed but (y_{2i}^*) is not, while the event $y_{2i}^* > 0$ is observed. No structural shift permitted.

6. Hybrid Model with Structural Shift

13 Discrete Choice

We will come up with a particular example of discrete choice and discuss the underlying ideas. We will talk about how they came about and the problems they try to solve.

13.1 Basic Discrete choice

To model discrete choices, we need to think of the ingredients that give rise to choices. There are two dominant modeling approaches:

1. Luce Model (1953) ⇔ McFadden Conditional Logit Model
2. Thurstone-Quandt Model (1929, 1930s) i.e. Multivariate probit/normal model

In general, our goal is to find a probabilistic choice model that (1) has a flexible functional form, (2) is computationally practical; (3) allows for flexibility in representing substitution patterns among choices; and (4) is consistent with a random utility model, thereby exhibiting a structural interpretation.

13.1.1 Luce Model / McFadden Conditional Logit Model

Denote X as the universe of objects of choice and S as the universe of attributes of persons. B is the feasible choice set: $x \in B \subseteq X$. A behaviora rule maps attributes into choices:

$$h\left(B, S\right) = x$$

Note that h may be random since:

- ▷ In observations, we may lose some informatino governing choices (unobserved characteristics);
- ▷ There can be random variations in choices due to unmeasured psychological factors

Furthermore, define $P\left(x|S, B\right)$ as the probability that an individual drawn randomly from the population with attributes S and an alternative set B chooses x. We will impose some restrictions on $P\left(x|S, B\right)$ and derive implications for the functional form of P.

Axioms We introduce the following two axioms:

1. Independence of Irrelevant Alternatives:
$$\frac{P\left(x|s, \{x, y\}\right)}{P\left(y|s, \{x, y\}\right)} = \frac{p\left(x|s, B\right)}{P\left(y|s, B\right)}$$
2. $P\left(y|s, B\right) > 0, \forall y \in B$ i.e. eliminates zero-probability choices.
3. Separability:
$$\tilde{v}\left(s, x, z\right) = v\left(s, x\right) - v\left(s, z\right)$$
where
$$\tilde{v}\left(s, x, z\right) = \ln\left(\frac{P_{xz}}{P_{zx}}\right), \quad P_{xz} = P\left(x|s, \{xz\}\right)$$

These axioms together yields:

$$P\left(x|s, B\right) = \frac{e^{v(s,x)}}{\sum_{y \in B} e^{v(s,y)}}$$

i.e. we get the logistic model from the Luce Axioms. Thus, we are able to link the model to familiar models in economics.

Debreu (1960) Criticism of the Model Suppose the $N+1$th alternative is identical to the first. Then the introduction of an identical good changes the probability of riding a bus. This is **not** an attractive result which stems from the need to make an i.i.d. assumption on the new alternative.

13.1.2 Thurstone's Random Utility Models

Assume the utility from choosing alternative j is

$$u_j = v(s, x_j) + \epsilon(s, x_j)$$

Then the probability j is in the set B is:

$$\begin{aligned} & P(u(s, x_j) \geq u(s, x_l)) \\ =& P(v(s, x_j) - v(s, x_l) \geq \epsilon(s, x_l) - \epsilon(s, x_j)) \\ =& \int_{-\infty}^{\infty} F_j(v_j - v_1 + \epsilon_j, ..., v_j - v_{j-1} + \epsilon_j, \epsilon_j, ..., v_j - v_J + \epsilon_j) \, d\epsilon_j \end{aligned}$$

where we specified the cdf $F(\epsilon_1, ..., \epsilon_J)$. Assuming i.i.d., the probability simplifies to:

$$\int_{-\infty}^{\infty} \left[\prod_{\substack{i=1 \\ i \neq j}}^{n} F_i(v_j - v_i + \epsilon_j) \right] f_j(\epsilon_j) \, d\epsilon_j$$

Binary Example $(N = 2)$ Then we have

$$P(1|s, B) = \int_{-\infty}^{\infty} \int_{-\infty}^{v_1 - v_2 + \epsilon_1} f_1(\epsilon_1, \epsilon_2) \, d\epsilon_1 d\epsilon_2$$

If ϵ_1, ϵ_2 are normal, then $\epsilon_1 - \epsilon_2$ is also normal so $P(v_1 - v_2 \geq \epsilon_1 - \epsilon_2)$ is normal. If ϵ_1, ϵ_2 are Weibull, then $\epsilon_1 - \epsilon_2$ is logistic. Recall that $\epsilon \sim$Weibull iff

$$P(\epsilon < c) = e^{-e^{-c+\alpha}}$$

i.e. a "double exponential." In this case, we have

$$P(v_1 + \epsilon_1 > v_2 + \epsilon_2) = \frac{e^{v_1 - v_2}}{1 + e^{v_1 - v_2}} = \frac{e^{v_1}}{e^{v_1} + e^{v_2}}$$

In other words, assuming that the errors follow a Weibull distribution yields the same logic model derived from the Luce Axioms. This link was established by Marshak (1959).

13.2 Expanding Logit

We can expand logit to accomodate multiple nesting levels.

13.2.1 Multinomial Probit Models (MPM) vs. Multinomial Logit (MLM)

Punchline: Relaxing the IIA condition is one of the main reasons why alternative-specific multinomial probit model is preferred over the multinomial logit model.