# 1 Empirical Analysis in Economics

Estimate (quantify) causal *effects:* Does $x$ affects $y$?How much?

But we can't run experiments on many important questions.

**Example 1** *Observational health studies (Nurses') vs. randomized trials.*

Need randomized treatment and control groups. In economics, we look for experiment-like events.

**Example 2** *Does Disability Insurance (DI/wage) affect labor force participation? Parsons (1980) vs. Bound (1989). Look at those who never applied for DI or those turned down (healthier than DI recipients).*

## 1.1 Experimental Setup and Solution

Consider two *hypothetical* outcomes *for each person*: $y_{1i}$ earning with training, $y_{0i}$ earnings without training. Observe $y_{1i}$ when $D_i = 1$ (the person applied for and took training), $y_{0i}$ when $D_i = 0$ (ENPs).

Three training effects:

$E[y_{1i} - y_{0i}]$ ATE, $E[y_{1i} - y_{0i}|D_i = 1]$ ATT, $E[y_{1i} - y_{0i}|D_i = 0]$ ATU

Observe only $E[y_{1i}|D_i = 1]$. The rest, $E[y_{0i}|D_i = 1]$ is the counterfactual.

Randomization: take $D = 1$ and randomize into *treatment* $(R = 1)$ and *control* $(R = 0)$. $E[y^*_{1i}|D^*_i = 1, R_i = 1] - E[y^*_{0i}|D^*_i = 1, R_i = 0]$.

Benchmark for non-experimental studies.

## 1.2 Causal or Descriptive Evidence

OLS of $y$ on $x$, controlling for several other variables. Meaning?

Causality or descriptives $E[y|x] = \int_{-\infty}^{\infty} ydF(y|x)$.

$x$ a *determinant* of $y$ when (a) a model says so, and (b) we have exogenous variation in $x$.

In "ignorant" research design, use whatever (sources of) variation in $x$ there is.

**Example 3** *Card (1993): returns to schooling, but ability bias? Use proximity to college as IV.*

**Example 4** *Wage curve studies with cross-sectional or within-region variation.*

## 1.3 Control for $X$

Control for other $Xs$ correlated with your causing variable $x$. (I.e., OLS.) When you fail to find all of $X$, you need an IV.

**Example 5** *Returns to education, ability bias and IQ test scores.*

When is controlling for $X$ enough to identify a causal effect? When is *selection on observables* plausible? When is assignment to treatment as good as random, conditional on $X$?

**Example 6** *Applicants to a college are screened based on* $X$; *,conditional on passing the* $X$ *test, they are accepted based on a first-come/first-serve basis.*

**Example 7** *Applying for a green card; conditional on enough points, cards are assigned in a lottery.*

### 1.3.1 Regression or Matching?

Techniques of controlling for $X$.

Goal: compare outcome for individuals from the treatment and control groups for each value of $X$. Then average the difference in the outcomes using the distribution of $X$ for treatments to obtain the estimate of the overall treatment effect on those who got the training (ATT).

Regression applies different weights and linearity.

A nice way to implement matching is to condition on the unidimensional probability of treatment $P(X)$.

## 1.4 Exogenous Variation (IV)

Run $y = X\beta + \varepsilon$ but $E[\varepsilon|X] \neq 0$. A valid IV $Z$ is correlated with $X$ but not with $\varepsilon$. How do you find an IV? From theory or a "natural" experiment.

**Example 8** *Angrist (1990): Vietnam-era draft lottery.*

**Example 9** *Changes in wage structure"Women, War and Wages" by Acemoglu, Autor and Lyle. Variation in draft causes differences in female labor supply. Get at effect of female labor supply on wage dispersion.*

**Example 10** *Card (1993) returns to schooling and IV with proximity to college. What is the identifying assumption? Test over-identified IV.*

## 1.5 Simultaneous Equations Reminder

IVs can/should come from a model, often in the form of an "exclusion restriction": Consider the *structural* demand and supply system

$$\begin{aligned} q_D &= \alpha_0 + \alpha_1 p + \alpha_2 y + \varepsilon_D \\ q_S &= \beta_0 + \beta_1 p + \qquad + \varepsilon_S \\ q_D &= q_S \end{aligned}$$

Unfortunately $E[\epsilon_D p] \neq 0$. Solve for and estimate the *reduced form*

$$\begin{aligned} p &= \pi_1 y + \upsilon_p \\ q &= \pi_2 y + \upsilon_q \end{aligned}$$

Can't go back from two $\pi$s to 5 $\alpha$s and $\beta$s.

Identify $\beta_1$ by *instrumenting* for $p$ using ?

### 1.5.1 Local Average Treatment Effect interpretation of IV.

What if the effect of $x$ on $y$ differs across groups? IV uses only part of the original variation in $x$ — that predicted by the IV; hence, we are estimating the effect of $x$ on $y$ for the sub-population whose behavior is well explained by the instrument (the compliers).

**Example 11** *Angrist and Krueger (1991) use quarter of birth and compulsory schooling laws to estimate returns to education. Use only a small part of the overall variation in schooling!*

**Example 12** *Angrist (1990) estimates the effect of military service on those drafted. Volunteers?*

This is a general problem.

## 1.6 Group-Level Identification

Variation of interest in $x$ is across groups of individuals. (Need to *correct standard errors!*)

Differences in avg. unobservables across groups? Union/non-union and productivity? Gender segregation and preferences?

With panel data, compare changes instead of levels.

$$y_{it} = UNION_{it}\beta + \alpha_i + \epsilon_{it}$$

$$y_{it} - y_{it-1} = (UNION_{it} - UNION_{it-1})\beta + \triangle\epsilon_{it}$$

Remove time constant unobservables. But are "movers" exogenous?

## 1.7 Difference in Differences

Before/after identification: $y_{it} = \alpha + \beta D_t + \varepsilon_{it}$.

What about underlying trends? => Diff-in-Diffs:

$$\begin{aligned} y_{it}^{j} &= \alpha + \alpha_1 d_t + \alpha^j d^j + \beta d_t^j + \gamma' x_{it}^j + \varepsilon_{it}^j \\ \beta_{DD} &= \overline{y}_1^1 - \overline{y}_0^1 - (\overline{y}_1^0 - \overline{y}_0^0). \end{aligned}$$

The threat is the possibility of an interaction between group and time period.

**Example 13** *Card and Krueger (1994) NJ-PA minimum wage study or Card (1990) Mariel Boatlift study.*

**Example 14** *Topalova (AEJ:AE, 2010) uses the the 1991 Indian trade liberalization to measure the impact of trade liberalization on poverty by exploiting the variation in sectoral composition across districts and liberalization intensity across production sectors in a D-in-Ds approach.*

**Example 15** *DD is fragile. The Mariel Boatlift that wasn't.*

DD best when (a) 0 and 1 similar before treatment; (b) $\widehat{\alpha_1}$ not too large.

DD implemented as fixed effects panel data OLS.

**Example 16** *Union status effect on wages; only movers used.*

**Example 17** *Gould and Paserman (2002) ask if women marry later when male wage inequality increases. U.S. cities with fixed effects.*

**Example 18** *Gonzales and Viitanen (2007): timing of legislation legalizing divorce across Europe to identify the effect of exposure to divorce as a child; there is a significant long run effect.*

**Example 19** *Ashenfelter and Greenstone "Using Mandated Speed Limits to Measure the Value of a Statistical Life"*

$$\ln(\textit{hours of travel})_{srt} = \beta \ln(\textit{miles})_{srt} + \gamma \ln(\textit{fatalities})_{srt} + \alpha_{sr} + \eta_{rt} + \mu_{st} + \nu_{srt}$$

*but there is endogeneity problem in that people adjust travel speed to reduce fatalities when the weather is bad etc. So they use a dummy for having the 65 m.p.h. speed limit as an IV. In the end they get $1.5m per life.*

**Remark 1** *Often use state-time changes as IV, instead of putting the $d_{it}^{j}$ dummies on the RHS.*

**Example 20** *Cutler and Gruber (1995) estimate the crowding out effect of public insurance in a large sample of individuals.*

$$Coverage_i = \beta_1 Elig_i + X_i \beta_2 + \varepsilon_i$$

*To instrument for $Elig_i$ they select a $national$ random sample and assign that sample to each state in each year to impute an average state level eligibility.*