# Empirical Evaluation of Public Policies: The Pursuit of Causality

Manuel Arellano

XII Encuentro de Economía Pública
Palma de Mallorca, 3-4 February 2005

## 1. Structural and treatment effect approaches

- The classic approach to quantitative policy evaluation in economics has been the *structural approach*.
- Its goals are to specify a class of theory-based models of individual choice, choose the one within the class that best fits the data, and use it for ex-post or ex-ante policy simulation.
- During the last 15 years the *treatment effect approach* has established itself as a formidable competitor that has introduced a different language, different priorities, techniques and practices in applied work.
- Not only that, it has also changed the perception of evidence-based economics among economists, public opinion, and policy makers.
- The ambition in a structural exercise is to use data from a particular context to identify, with the help of theory, deep rules of behavior that can be extrapolated to other contexts.

- A treatment effect (TE) exercise is context-specific and addresses less ambitious policy questions.
- The goal is to evaluate the impact of an *existing* policy by comparing the distribution of a chosen outcome variable for individuals affected by the policy (treatment group) with the distribution of unaffected individuals (control group).
- The aim is to choose the control and treatment groups in such a way that membership of one or the other, either results from randomization or can be regarded as if they were the result of randomization.
- In this way one hopes to achieve the standards of empirical credibility on causal evidence that are typical of experimental biomedical studies.

- The TE literature has expressed dissatisfaction with the existing structural approach along several dimensions:
  - (a) Between theory, data, and estimable structural models there is a host of untestable functional form assumptions that undermine the force of structural evidence by:
    - (i) Having unknown implications for results.
    - (ii) Giving researchers too much discretion.
    - (iii) Complexity affects transparency and replicability.
  - (b) By being too ambitious on the policy questions we get very little credible evidence from data. Too much emphasis on "external validity" at the expense of the more basic "internal validity".
- The TE literature sees the role of empirical findings as one of providing bits and pieces of hard evidence that can help the assessment of future policies in an informal way.
- Main gains in empirical research are not expected to come from the use of formal theory or sophisticated econometrics, but from understanding the sources of variation in data with the objective of identifying policy parameters.
- The award of the 2003 Clark Medal to Steven Levitt, for confronting "important empirical questions in the economics of crime and political economy, *by finding new data and devising novel and clever identification schemes*."

- Many policy interventions at the micro level have been evaluated:
  (a) training programs
  (b) welfare programs (e.g. unemployment insurance, worker's sickness compensation)
  (c) wage subsidies and minimum wage laws
  (d) tax-credit programs
  (e) effects of taxes on labor supply and investment
  (f) effects of Medicaid on health
- In this talk I will review the following contexts or research designs of evaluation:
  (a) social experiments
  (b) matching
  (c) instrumental variables
  (d) differences in differences

## 2. Potential outcomes and causality

- Association and causation have always been known to be different, but a mathematical framework for an unambiguous characterization of *statistical causal effects* is surprisingly recent (Rubin, 1974; despite precedents in statistics and economics, Neyman, 1923; Roy, 1951).
- Think of a population of individuals that are susceptible of treatment. Let $Y_1$ be the outcome for an individual if exposed to treatment and let $Y_0$ be the outcome for the same individual if not exposed. The treatment effect for that individual is $Y_1 - Y_0$.
- In general, individuals differ in how much they gain from treatment, so that we can imagine a distribution of gains over the population with mean
$$\alpha_{ATE} = E\left(Y_1 - Y_0\right).$$
- The average treatment effect so defined is a standard measure of the causal effect of treatment $1$ relative to treatment $0$ on the chosen outcome.
- Suppose that treatment has been administered to a fraction of the population, and we observe whether an individual has been treated or not ($D = 1$ or $0$) and the person's outcome $Y$. Thus, we are observing $Y_1$ for the treated and $Y_0$ for the rest:
$$Y = (1 - D)\, Y_0 + DY_1.$$

- Because $Y_1$ and $Y_0$ can never be observed for the same individual, the distribution of gains lacks empirical entity. It is just a conceptual device that can be related to observables.
- This notion of causality is statistical because it is not interested in finding out causal effects for specific individuals. Causality is defined in an average sense.

*Connection with regression*

- A standard measure of association between $Y$ and $D$ is:

$$\beta = E(Y \mid D = 1) - E(Y \mid D = 0)$$
$$= E(Y_1 - Y_0 \mid D = 1) + \{E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)\}$$

- The second expression makes it clear that in general $\beta$ differs from the *average gain for the treated* (another standard measure of causality, that we call $\alpha_{TT}$).
- The reason is that treated and nontreated units may have different average outcomes in the absence of treatment.
- For example, this will be the case if treatment status is the result of individual decisions, and those with low $Y_0$ choose treatment more frequently than those with high $Y_0$.

- From a structural model of $D$ and $Y$ one could obtain the implied average treatment effects, but here $\alpha_{ATE}$ or $\alpha_{TT}$ have been directly defined with respect to the distribution of potential outcomes, so that relative to a structure they are reduced form causal effects.
- Econometrics has conventionally distinguished between reduced form effects (uninterpretable but useful for prediction) and structural effects (associated with rules of behavior).
- The TE literature emphasizes "reduced form causal effects" as an intermediate category between predictive and structural effects.

*Social feedback*

- The potential outcome representation is predicated on the assumption that the effect of treatment is independent of how many individuals receive treatment, so that the possibility of different outcomes depending on the treatment received by other units is ruled out.
- This excludes general equilibrium or feedback effects, as well as strategic interactions among agents.
- So the framework is not well suited to the evaluation of system-wide reforms which are intended to have substantial equilibrium effects.

## 3. Social experiments

- In the TE approach, a randomized field trial is regarded as the ideal research design.
- Observational studies seen as "more speculative" attempts to generate the force of evidence of experiments.
- In a controlled experiment, treatment status is randomly assigned by the researcher, which by construction ensures:

$$(Y_0, Y_1) \perp D$$

  In such a case, $F(Y_1 \mid D = 1) = F(Y_1)$ and $F(Y_0 \mid D = 0) = F(Y_0)$. The implication is $\alpha_{ATE} = \alpha_{TT} = \beta$.
- Analysis of data takes a simple form: An unbiased estimate of $\alpha_{ATE}$ is the difference between the average outcomes for treatments and controls:

$$\widehat{\alpha}_{ATE} = \overline{Y}_T - \overline{Y}_C$$

- If interested in ATE by observed characteristics $X$, randomization ensures that

$$\begin{aligned} \alpha_{ATE}(x) &\equiv E(Y_1 - Y_0 \mid X = x) \\ &= E(Y \mid D = 1, X = x) - E(Y \mid D = 0, X = x), \end{aligned}$$

  so that $\alpha_{ATE}(x)$ can be estimated from cell-mean differences or nonparametric regression.
- In a randomized setting, there is no need to "control" for covariates, rendering multiple regression unnecessary, except if interested in effects for specific groups.

*Experimental testing of welfare programs in the US*

- Long history of randomized field trials in social welfare in the US, beginning in the 1960s.
- Moffitt (2003) provides a lucid assessment.
- Early experiments had many flaws due to lack of experience in designing experiments and in data analysis.
- During the 1980s the US federal government started to encourage states to use experimentation, eventually becoming almost mandatory.
- The analysis of the 1980s experimental data consisted of simple treatment-control differences. The force of the results had a major influence on the 1988 legislation.
- In spite of these developments, randomization encountered resistance from many US states on ethical grounds.
- Even more so in other countries, where treatment groups have often been formed by selecting areas for treatment instead of individuals.
- Randomization is not appropriate for evaluating reforms with major spillovers from which the control group cannot be isolated.
- But it is an effective means of testing incremental reforms and searching for policy designs "that reveal what works and for whom." (Moffitt).

*Example 1: Employment effect of a subsidized job program.*

- The NSW program was designed in the US in the mid 70's to provide training and job opportunities to disadvantaged workers, as part of an experimental demonstration.
- Ham and LaLonde (1996) looked at the effects of the NSW on women that volunteered for training.
- NSW guaranteed to treated participants 12 months of subsidized employment (as trainees) in jobs with gradual increase in work standards.
- Eligibility requirements: To be unemployed, a long-term AFDC recipient, and have no preschool children.
- Participants were randomly assigned to treatment & control groups in 1976-77. Experiment took place in 7 cities.
- Ham–LaLonde data: 275 women in treatment group and 266 controls. All volunteered in 1976. Averages: Age 34, 10 years of schooling, 70% H.S. dropout, 2 children, 65% married, 85% black.
- Thanks to randomization, a simple comparison between the employment rates of treatments and controls gives an unbiased estimate of the effect of the program.
- Figure 1 taken from Ham–LaLonde shows the effects.

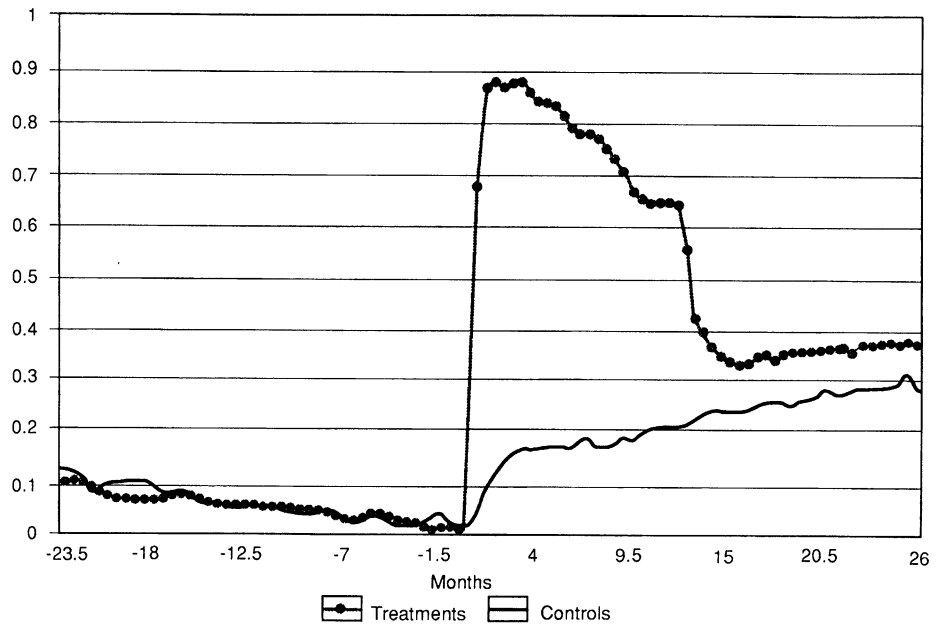


FIGURE 1.—Employment rates of AFDC women in the NSW Demonstration.

experimental evaluation shows that at least in the short run, NSW substantially improved the employment prospects of AFDC participants.

The NSW demonstration achieved these employment gains by helping trainees to hold on to their jobs longer and/or to find jobs faster, thereby increasing the length of their employment spells and/or reducing the length of their unemployment spells. To begin our analysis of these effects of training, we examine the Kaplan-Meier survivor functions for the treatments' and controls' employment and unemployment spells in Table I.[6] The first two columns of the table indicate that 65 percent of the trainees' employment spells lasted six or more months compared with only 57.3 percent of the controls' spells. When we follow standard practice and compare the experience of treatments and controls in fresh unemployment spells in columns three and five of Table I, we see that 73 percent of the treatments are still in an unemployment spell after a duration of 6 months compared to only 61.3 percent of the controls. Thus training appears to be a mixed blessing since it increases the length of both employment and unemployment spells.

Unfortunately, as previously noted, such a simple analysis of the treatments' and controls' employment histories may be misleading First, the possibility that the treatments and controls faced different demand conditions is particularly

[6] In practice many of the employment and unemployment spells are not completed during the sample period (i.e., they are right censored). Therefore, we cannot simply compare their mean durations, especially because the treatments spend on average half the sampling frame in training.

- The growth in the employment rates of the controls is just a reflection of the program's eligibility criteria.
- The conclusion from the experimental evaluation is that, at least in the short run, the NSW substantially improved the employment prospects of participants (a difference of 9 percentage points in employment rates).

*Covariates and job histories*

- At admission time, information collected on age, education, high-school dropout status, children, marital status, race, and labor history for the previous two years.
- Job histories following entry into the program: Treatments and controls were interviewed at 9 month intervals, collecting information on employment status. In this way employment and unemployment spells were constructed for more than two years following the baseline (26 months).

*The Ham–LaLonde critique of experimental data*
*A) Effects on wages*

- A direct comparison of mean wages for treatments and controls gives a biased estimate of the effect of the program on wages. This will happen as long as training has an impact on the employment rates of the treated.
- Let $W$ =wages, let $Y = 1$ if employed and $Y = 0$ if unemployed, $\eta = 1$ if high skill and $\eta = 0$ otherwise.
- Suppose that treatment increases the employment rates of high and low skill workers:
$$\Pr(Y = 1 \mid D = 1, \eta = 0) > \Pr(Y = 1 \mid D = 0, \eta = 0)$$
$$\Pr(Y = 1 \mid D = 1, \eta = 1) > \Pr(Y = 1 \mid D = 0, \eta = 1)$$
- but the effect is of less intensity for the high skill group:
$$\frac{\Pr(Y = 1 \mid D = 1, \eta = 0)}{\Pr(Y = 1 \mid D = 0, \eta = 0)} > \frac{\Pr(Y = 1 \mid D = 1, \eta = 1)}{\Pr(Y = 1 \mid D = 0, \eta = 1)}.$$
- This implies that the frequency of low skill will be greater in the group of employed treatments than in the employed controls:
$$\Pr(\eta = 0 \mid Y = 1, D = 1) > \Pr(\eta = 0 \mid Y = 1, D = 0),$$
i.e. $\eta$ is not independent of $D$ given $Y = 1$, although unconditionally $\eta \perp D$.

- For this reason, a direct comparison of average wages between treatments and controls will tend to underestimate the effect of treatment on wages:

$$\Delta_f = E\left(W \mid Y = 1, D = 1\right) - E\left(W \mid Y = 1, D = 0\right),$$

whereas the effects of interest of $D$ on $W$ are:

* For low skill individuals:

$$\begin{aligned}\Delta_0 \;=\; & E\left(W \mid Y = 1, D = 1, \eta = 0\right) \\ & -E\left(W \mid Y = 1, D = 0, \eta = 0\right),\end{aligned}$$

* for high skill:

$$\begin{aligned}\Delta_1 \;=\; & E\left(W \mid Y = 1, D = 1, \eta = 1\right) \\ & -E\left(W \mid Y = 1, D = 0, \eta = 1\right)\end{aligned}$$

* and the overall effect:

$$\Delta_s = \Delta_0 \Pr\left(\eta = 0\right) + \Delta_1 \Pr\left(\eta = 1\right).$$

- In general, we shall have that $\Delta_f < \Delta_s$.
- It may not be possible to construct an experiment to measure the effect of training the unemployed on subsequent wages. i.e. it does not seem possible to experimentally undo the conditional correlation between D and $\eta$.

*B) Effects on durations*

- Effects on employment duration: similar to wages, the experimental comparison of exit rates from employment may be misleading. Let $T_e$ be the duration of an employment spell. An experimental comparison is

$$\Pr\left(T_e = t \mid T_e \geq t, D = 1\right) - \Pr\left(T_e = t \mid T_e \geq t, D = 0\right)$$

  but we are interested in

$$\Pr\left(T_e = t \mid T_e \geq t, D = 1, \eta\right) - \Pr\left(T_e = t \mid T_e \geq t, D = 0, \eta\right).$$

- $D$ is correlated with $\eta$ given $T_e \geq t$ for various reasons. e.g. If treatment especially helps to find a job those with $\eta = 0$ , the frequency of $\eta = 0$'s in the group $\{T_e \geq t, D = 1\}$ will increase relative to $\{T_e \geq t, D = 0\}$.

- Similar problems arise with unemployment durations. Ham and LaLonde's solution is to use an econometric model of labor histories with unobserved heterogeneity.

- The problem with wages and spells is one of censoring. It could be argued that the causal question is not well posed in these examples.

- Suppose that we wait until every individual completes an employment spell, and we consider the causal effect of treatment on the duration of such spell. This generates the problem that if the spells of controls and treatments tend to occur at different points in time, the economic environment is not held constant by the experimental design.

## 4. Matching

- There are many situations where experiments are too expensive, unfeasible, or unethical. A classical example is the analysis of the effects of smoking on mortality rates.
- Experiments guarantee the independence condition
$$(Y_1, Y_0) \perp D$$
but with observational data it is not very plausible.
- A less demanding condition for nonexperimental data is:
$$(Y_1, Y_0) \perp D \mid X$$
In the TE literature, called selection on observables.
- Conditional independence implies
$$\begin{aligned} E(Y_1 \mid X) &= E(Y_1 \mid D=1, X) = E(Y \mid D=1, X) \\ E(Y_0 \mid X) &= E(Y_0 \mid D=0, X) = E(Y \mid D=0, X). \end{aligned}$$
Therefore, for $\alpha_{ATE}$ we can calculate (and similarly for $\alpha_{TT}$):
$$\begin{aligned} \alpha_{ATE} &= E(Y_1 - Y_0) = \int E(Y_1 - Y_0 \mid X)\, dF(X) \\ &= \int [E(Y \mid D=1, X) - E(Y \mid D=0, X)]\, dF(X). \end{aligned}$$
- Most of the literature focused on average effects, but the matching assumption also works for distributional comparisons.

*Relation with multiple regression*

- If we specify $E(Y \mid D, X)$ as a linear regression on $D$, $X$ and $D \times X$ we have
$$E(Y \mid D, X) = \beta D + \gamma X + \delta DX$$
and
$$E(Y \mid D = 1, X) - E(Y \mid D = 0, X) = \beta + \delta X.$$
$$\begin{aligned} \alpha_{ATE} &= \beta + \delta E(X) \\ \alpha_{TT} &= \beta + \delta E(X \mid D = 1), \end{aligned}$$
which can be easily estimated using linear regression.
- Alternatively, we can treat $E(Y \mid D = 1, X)$ and $E(Y \mid D = 0, X)$ as nonparametric functions of $X$.
- The last approach is closer in spirit to the *matching* literature, which has emphasized direct comparisons, free from functional form assumptions and extrapolation.

*Imputing missing outcomes*

- Suppose that $X$ is discrete and takes on $J$ values $\left\{\xi_j\right\}_{j=1}^{J}$ and we have a sample $\{X_i\}_{i=1}^{N}$. Let
  $N^j$ = number of observations in cell $j$.
  $N_\ell^j$ = number of observations in cell $j$ with $D = \ell$.
  $\overline{Y}_\ell^j$ = mean outcome in cell $j$ for $D = \ell$.

- Thus, $\left(\overline{Y}_1^j - \overline{Y}_0^j\right)$ is the sample counterpart of
$$E\left(Y \mid D = 1, X = \xi_j\right) - E\left(Y \mid D = 0, X = \xi_j\right),$$
which can be used to get the estimates
$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J}\left(\overline{Y}_1^j - \overline{Y}_0^j\right)\frac{N^j}{N}, \quad \widehat{\alpha}_{TT} = \sum_{j=1}^{J}\left(\overline{Y}_1^j - \overline{Y}_0^j\right)\frac{N_1^j}{N_1}$$

- The formula for $\widehat{\alpha}_{TT}$ can also be written in the form
$$\widehat{\alpha}_{TT} = \frac{1}{N_1}\sum_{D_i=1}\left(Y_i - \overline{Y}_0^{j(i)}\right)$$
where $j\,(i)$ is the cell of $X_i$. Thus, $\widehat{\alpha}_{TT}$ matches the outcome of each treated unit with the mean of the nontreated units in the same cell.

- If $X$ is continuous but low dimensional, the idea can be extended by matching observations with similar or discretized values of $X$.

*Methods based on the propensity score*

- Rosenbaum and Rubin called "propensity score" to
$$\pi(X) = \Pr(D = 1 \mid X)$$
and proved that if $(Y_1, Y_0) \perp D \mid X$ then
$$(Y_1, Y_0) \perp D \mid \pi(X)$$
provided $0 < \pi(X) < 1$ for all $X$.
- The result tells us that we can match units with very different values of $X$ as long as they have similar values of $\pi(X)$.
- These results suggest two-step procedures in which we begin by estimating the propensity score.

*The common support condition*

- $X$ can take very different values for treatments & controls.
- Heckman et al. (1997) found that violation of the common support condition for the matching variables ("comparing the incomparable") is an important source of bias.
- Restricting matching to regions of common support $S$, we have:
$$M(S) = \frac{\int_S E(Y_1 - Y_0 \mid X)\, dF(X \mid D = 1)}{\int_S dF(X \mid D = 1)}$$

*Differences between matching and OLS*

- Matching avoids functional form assumptions and emphasizes the common support condition.
- Matching focuses on a single parameter at a time, which is obtained through explicit aggregation.

*The requirement of random variation in outcomes*

- Matching works on the presumption that for $X = x$ there is random variation in $D$, so that we can observe both $Y_1$ and $Y_0$. It fails if $D$ is a deterministic function of $X$.
- There is a tension between the thought that if $X$ is good enough then there may not be within-cell variation in $D$, and the suspicion that seeing enough variation in $D$ for given $X$ is an indication that exogeneity is at fault.

*Example 2: Monetary incentives and schooling in the UK*

- The pilot of the *Education Maintenance Allowance* (EMA) program started in Sept. 1999. EMA paid youths aged 16–18 that continued in full time education (after 11 compulsory grades) a weekly stipend of £ 30 to 40, plus final bonuses for good results up to £140.
- Eligibility (and amounts paid) depends on household characteristics. Eligible for full payments if annual income under £13000. Those above £30000, not eligible.
- Dearden, Emmerson, Frayne & Meghir (2002) participated in the design of the pilot and did the evaluation.
- No experimental design for political reasons, but one defining treatment and control areas, both rural and urban.
- Basic question asked is whether more education results from this policy. The worry is that families fail to decide optimally due to liquidity constraints or misinformation.
- They use propensity scores. Probit estimates of $\pi(X)$ with family, local, and school characteristics. For each treated observation they construct a counterfactual mean using *kernel* regression and *bootstrap* standard errors.
- EMA increased participation in year 12 by 5.9% for eligible individuals, and by 3.7% for the whole population. Only significant results for full-payment recipients.

## 5. Instrumental Variables

- Suppose we have nonexperimental data with covariates, but cannot assume conditional independence as in matching:
$$(Y_1, Y_0) \perp D \mid X.$$
- Suppose, however, that we have a variable $Z$ that is an "exogenous source of variation in $D$" in the sense that it satisfies the *independence assumption*:
$$(Y_1, Y_0) \perp Z \mid X$$
and the *relevance assumption*:
$$Z \text{ dep. } D \mid X.$$
- In a classic example, $Z$ indicates assignment to treatment in an experimental design. Therefore, $(Y_1, Y_0) \perp Z$.
- However, "actual treatment" $D$ differs from $Z$ because some individuals in the treatment group decide not to treat (non-compliers). $Z$ and $D$ will be correlated in general.
- Matching can be regarded as a special case of IV in which $Z = D$, i.e. all variation in $D$ is exogenous given $X$.
- See examples of sources of IVs in Angrist–Krueger *JEP* article Table 1.
- The question is whether this situation identifies causal effects. To answer it, I consider a binary $Z$, and abstract from the fact that the reasoning can be conditional on $X$.

*Table 1*
**Examples of Studies That Use Instrumental Variables to Analyze Data From Natural and Randomized Experiments**

| *Outcome Variable* | *Endogenous Variable* | *Source of Instrumental Variable(s)* | *Reference* |
|---|---|---|---|
| | | *1. Natural Experiments* | |
| Labor supply | Disability insurance replacement rates | Region and time variation in benefit rules | Gruber (2000) |
| Labor supply | Fertility | Sibling-Sex composition | Angrist and Evans (1998) |
| Education, Labor supply | Out-of-wedlock fertility | Occurrence of twin births | Bronars and Grogger (1994) |
| Wages | Unemployment insurance tax rate | State laws | Anderson and Meyer (2000) |
| Earnings | Years of schooling | Region and time variation in school construction | Duflo (2001) |
| Earnings | Years of schooling | Proximity to college | Card (1995) |
| Earnings | Years of schooling | Quarter of birth | Angrist and Krueger (1991) |
| Earnings | Veteran status | Cohort dummies | Imbens and van der Klaauw (1995) |
| Earnings | Veteran status | Draft lottery number | Angrist (1990) |
| Achievement test scores | Class size | Discontinuities in class size due to maximum class-size rule | Angrist and Lavy (1999) |
| College enrollment | Financial aid | Discontinuities in financial aid formula | van der Klaauw (1996) |
| Health | Heart attack surgery | Proximity to cardiac care centers | McClellan, McNeil and Newhouse (1994) |
| Crime | Police | Electoral cycles | Levitt (1997) |
| Employment and Earnings | Length of prison sentence | Randomly assigned federal judges | Kling (1999) |
| Birth weight | Maternal smoking | State cigarette taxes | Evans and Ringel (1999) |
| | | *2. Randomized Experiments* | |
| Earnings | Participation in job training program | Random assignment of admission to training program | Bloom et al. (1997) |
| Earnings | Participation in Job Corps program | Random assignment of admission to training program | Burghardt et al. (2001) |
| Achievement test scores | Enrollment in private school | Randomly selected offer of school voucher | Howell et al. (2000) |
| Achievement test scores | Class size | Random assignment to a small or normal-size class | Krueger (1999) |
| Achievement test scores | Hours of study | Random mailing of test preparation materials | Powers and Swinton (1984) |
| Birth weight | Maternal smoking | Random assignment of free smoker's counseling | Permutt and Hebel (1989) |

*Homogeneous effects*

- If the causal effect is the same for every individual
$$Y_{1i} - Y_{0i} = \alpha$$
the availability of an IV allows us to identify $\alpha$. This is the traditional situation in econometric models with endogenous explanatory variables.

- In general
$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i$$
and in the homogeneous case
$$Y_i = Y_{0i} + \alpha D_i.$$

- Also, taking into account that $Y_{0i} \perp Z_i$
$$\begin{aligned} E(Y_i \mid Z_i = 1) &= E(Y_{0i}) + \alpha E(D_i \mid Z_i = 1) \\ E(Y_i \mid Z_i = 0) &= E(Y_{0i}) + \alpha E(D_i \mid Z_i = 0). \end{aligned}$$

- Subtracting both equations we obtain
$$\alpha = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{E(D_i \mid Z_i = 1) - E(D_i \mid Z_i = 0)}$$
which determines $\alpha$ as long as
$$E(D_i \mid Z_i = 1) \neq E(D_i \mid Z_i = 0).$$

- Intuitively, the effect of $D$ on $Y$ can be measured through the effect of $Z$ because we have assumed that $Z$ only affects $Y$ through $D$.

*Heterogeneous effects*

- In the heterogeneous case the availability of IVs is not sufficient to identify a causal effect.
- An additional assumption that helps to identify $\alpha_{TT}$ is an eligibility rule of the form:
$$\Pr\left(D = 1 \mid Z = 0\right) = 0$$
i.e. individuals with $Z = 0$ are denied treatment.
- An alternative, more general, additional assumption is the following "monotonicity" condition: Any person that was willing to treat if assigned to the control group, would also be prepared to treat if assigned to the treatment group.
- The plausibility of this assumption depends on the context of application.
- Under monotonicity, the IV coefficient coincides with the average treatment effect for those whose value of $D$ would change when changing the value of $Z$ (local average treatment effect or LATE).

*Example 3: Ethnic enclaves and the success of immigrants*

- Interest in the effect of leaving in a highly concentrated ethnic area on labor success. In Sweden 11% of the population was born abroad. Of those, more than 40% live in an ethnic enclave (Edin, Fredriksson and Åslund, 2003).
- The causal effect is ambiguous. Residential segregation lowers the acquisition rate of local skills, preventing access to good jobs. But enclaves act as opportunity-increasing networks by disseminating information to new immigrants.
- Immigrants in ethnic enclaves have 5% lower earnings, after controlling for age, education, gender, family background, country of origin, and year of immigration.
- But this association may not be causal if the decision to live in an enclave depends on expected opportunities.
- Swedish governments of 1985-1991assigned initial areas of residence to refugee immigrants. Motivated by the belief that dispersing immigrants promotes integration.
- Let $Z$ indicate initial assignment (8 years before measuring ethnic enclave indicator $D$). Edin et al. assumed that $Z$ is independent of potential earnings $Y_0$ and $Y_1$.
- IV estimates implied a 13% gain for low-skill immigrants associated with one std. deviation increase in ethnic concentration. For high-skill immigrants there was no effect.

- The instrumental variable method is at the basis of the simultaneous equation theory developed by the econometricians of the 1940s and 50s.
- In the classic simultaneous equation framework, the goal is to determine a structure. In contrast, in the previous example, instrumental variables are used to identify a reduced form causal effect (resulting from the interaction of a variety of underlying effects).

## 6. Differences in differences

*Example 4: minimum wages and employment*

- In March 1992 the state of New Jersey increased the legal minimum wage by 19%, whereas the bordering state of Pennsylvania kept it constant.
- Card and Krueger (1994) evaluated the effect of this change on the employment of low wage workers. In a competitive model the result of increasing the minimum wage is to reduce employment.
- They conducted a survey to some 400 fast food restaurants from the two states just before the NJ reform, and a second survey to the same outlets 7-8 months after.
- Characteristics of fast food restaurants:
  (a) A large source of employment for low-wage workers.
  (b) They comply with minimum wage regulations (especially franchised restaurants).
  (c) Fairly homogeneous job, so good measures of employment and wages can be obtained.
  (d) Easy to get a sample frame of franchised restaurants (yellow pages) with high response rates.
  (e) Response rates 87% and 73% (less in Penn, because the interviewer was less persistent).

- The DID coefficient is

$$\begin{aligned}\beta \;=\; & \left[E\left(Y_2 \mid D=1\right) - E\left(Y_1 \mid D=1\right)\right] \\ & - \left[E\left(Y_2 \mid D=0\right) - E\left(Y_1 \mid D=0\right)\right].\end{aligned}$$

  where $Y_1$ and $Y_2$ denote employment before and after the reform, $D = 1$ denotes a store in NJ (treatment group) and $D = 0$ in Penn (control group).

- $\beta$ measures the difference between the average employment change in NJ and the average employment change in Penn.

- The key assumption in giving a causal interpretation to $\beta$ is that the temporal effect in the two states is the same in the absence of intervention.

- But it is possible to generalize the comparison in several ways, for example controlling for other variables.

- Card and Krueger found that rising the minimum wage increased employment in some of their comparisons but in no case caused an employment reduction.

- This article originated much economic and political debate.

- DID estimation has become a very popular method of obtaining causal effects, especially in the US, where the federal structure provides cross state variation in legislation.

- See Table in Bertrand, Duflo and Mullainathan (2004).

## Table 1: Survey of DD Papers[a]

| | | |
|---|---|---|
| **Number of DD papers** | 92 | |
| Number with more than 2 periods of data | 69 | |
| Number which collapse data into before-after | 4 | |
| **Number with potention serial correlation problem** | 65 | |
| **Number with some serial correlation correction** | 5 | |
| GLS | 4 | |
| Arbitrary variance-covariance matrix | 1 | |
| **Distribution of time-span for papers with more than 2 periods** | **Average** | 16.5 |
| | **Percentile** | **Value** |
| | 1% | 3 |
| | 5% | 3 |
| | 10% | 4 |
| | 25% | 5.75 |
| | 50% | 11 |
| | 75% | 21.5 |
| | 90% | 36 |
| | 95% | 51 |
| | 99% | 83 |
| **Informal manipulations of data** | **Number** | |
| Graph time series of effect | 15 | |
| See if effect persists | 2 | |
| Examine lags of law to see timing of effect | 2 | |
| DDD | 11 | |
| Include trend specific to passing states | 7 | |
| Explicitly include lead to look for effect prior to law | 3 | |
| Include laggged dependent variable | 3 | |
| **Number which have clustering problem** | 80 | |
| Number which deal with it | 36 | |
| **Most commonly used variables** | | |
| Employment | 18 | |
| Wages | 13 | |
| Health/Medical Expenditure | 8 | |
| Unemployment | 6 | |
| Fertility/Teen Motherhood | 4 | |
| Insurance | 4 | |
| Poverty | 3 | |
| Consumption/Savings | 3 | |

[a]Notes: Data comes from a survey of all articles in six journals between 1990 and 2000:*American Economic Review*; *Industrial Labor Relations Review*; *Journal of Labor Economics*; *Journal of Political Economy*; *Journal of Public Economics*; and *Quarterly Journal of Economics*. We define an article as "Difference-in-Difference" if it: (1) examines the effect of a specific interventions and (2) uses units unaffected by the intervention as a control group.

*The context of difference in difference comparisons*

- If we observe outcomes before and after treatment, we could use the treated before treatment as controls for the treated after treatment.
- The problem of this comparison is that it can be contaminated by the effect of events other than the treatment that occurred between the two periods.
- Suppose that *only a fraction* of the population is exposed to treatment. In such a case, we can use the group that never receives treatment to identify the temporal variation in outcomes that is *not due* to exposure to treatment. This is the basic idea of the DID method.
- Two-period potential outcomes with treatment in $t = 2$:

$$\begin{aligned} Y_1 &= Y_0(1) \\ Y_2 &= (1 - D) Y_0(2) + D Y_1(2) \end{aligned}$$

- The *fundamental identifying assumption* is that the average changes in the two groups are the same in the absence of treatment:

$$E(Y_0(2) - Y_0(1) \mid D = 1) = E(Y_0(2) - Y_0(1) \mid D = 0).$$

- $Y_0(1)$ is always observed but $Y_0(2)$ is counterfactual for units with $D = 1$.
- Under such identification assumption, the DID coefficient coincides with the average treatment effect for the treated.

*Comments and problems*

- $\beta$ can be obtained as the coefficient of the interaction term in a regression of outcomes on treatment and time dummies.
- To obtain the DID parameter we do not need panel data (except if e.g. we regard the Card–Krueger data as an aggregate panel with two units and two periods), just cross-sectional data for at least two periods.
- With panel data, we can estimate $\beta$ from a regression of outcome changes on the treatment dummy. This is convenient for accounting for dependence between the two periods.
- Differences in the composition of the cross-sectional populations over time (especially problematic if not using panel data).
- The fundamental assumption might be satisfied conditionally given certain covariates, but identification vanishes if some of them are *unobservable*.

## 7. Concluding remarks

- Empirical work has become more central to economic research in the last decade.
- In labor economics, the fraction of articles with empirical content published in top journals went from 63% in 1985-1987 to 77% in 1995-1997 (Moffitt).
- Experimental and quasi-experimental approaches have an important but limited role to play in policy evaluation.
- There are relevant quantitative policy questions that cannot be answered without the help of economic theory.
- The quasi-experimental approach is also having a contribution to reshaping structural econometric practice.
- It is increasingly becoming standard fare a reporting style that distinguishes clearly the roles of theory and data in getting the results.
- This perspective affects:
  - the choice of estimation methods in structural work (as in Ridder & van den Berg's equilibrium unemployment search models, 2003),
  - the modelling of selection and policy effects (as in Heckman & Vytlacil, 2005),
  - or the econometric theorists' research agendas (as in the recent nonparametric IV literature).