Econometric Methods for Policy Evaluation

Joan Llull

UAB, MOVE, and Barcelona GSE

Master in Economics and Finance (CEMFI). Winter 2016

joan.llull [at] movebarcelona [dot] eu

Motivation: Structural vs Treatment Effect Approaches

Evaluation of Public Policies

Evaluation of public (and private) policies is very important for efficiency, and ultimately to improve welfare.

Vast **literature** in economics, mostly in public economics, but also in development economics and labor economics, devoted to the evaluation of different programs:

- training programs
- welfare programs
- wage subsidies
- minimum wage laws
- taxation
- Medicaid and other health policies
- school policies, feeding programs
- microcredit and a variety of other forms of development assistance
- ...

These analyses aim at **quantifying** the effects of these policies on different outcomes, and ultimately on welfare.

Structural Approach

Classic approach to quantitative policy eval. **structural approach**:

- specifies a class of theory-based models of individual choice
- chooses the one within the class that best fits the data
- and uses it to evaluate policies through simulation.

Main advantages:

- it allows both *ex-ante* and *ex-post* policy evaluation
- it permits evaluating different variations of a similar policy without need to change the structure of the model or reestimate it (out of sample simulation)

Main critique:

- host of untestable functional form assumptions which have unknown implications for the results (too much discretion)
- too much emphasis on external validity at the expense of a more basic internal validity
- complexity (transparency) & computational cost (difficult to replicate)

Treatment Effects Approach

Last two decades: **treatment effect approach**.

It has **changed** language, priorities, techniques, and practices, and the perception of evidence-based economics among economists, public opinion, and policy makers.

Main goal: **evaluate ex-post** the impact of an existing policy.

Compare distribution of a chosen outcome variable for individuals affected by the policy (the **treatment group**), with the distribution of unaffected individuals (**control group**).

Main challenge: comparison so that the distribution of outcome for the control group serves as a good **counterfactual** for the distribution of the outcome for the treated group in the absence of treatment.

Main focus: understanding of the sources of variation in data with the objective of identifying the policy parameters.

Pros and Cons of Treatment Effects

Main advantage: given its focus on internal validity, the exercise gives transparent and credible identification.

Main disadvantage: estimated parameters are not useful for welfare analysis because they are not deep parameters (they are reduced-forms instead), and as a result, they are not policy-invariant (Lucas, 1976; Heckman and Vytlacil 2005).

In that respect, a treatment effect exercise is less ambitious.

A Link Between the Two

The deep differences between the two approaches has **split the economics profession** into two camps whose research programs have evolved almost independently despite focusing on similar questions.

However, recent developments have changed this trend, as researchers realized about the important **complementarity** between the two (see Chetty, 2009; Todd and Wolpin, 2010).

In **this part** of the course we will review the main designs for policy evaluation under the treatment effect approach.

In the **second part** (with Pedro) you will review structural approaches.

If time permits, I will introduce some **bridges between the two**, which will serve as an introduction to the second part.

Potential Outcomes and Causality: Treatment Effects

Potential Outcomes

Consider the **population** of individuals that are susceptible of a treatment:

- Y_{1i}: outcome for individual i if exposed to the treatment ($D_i = 1$)
- Y_{0i} be the outcome for the same individual if not exposed ($D_i = 0$)
- Treatment indicator: D_i

Note that Y_{1i} and Y_{0i} are **potential outcomes** in the sense that we only observe:

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i).$$

Main challenge of this approach: the treatment effect can not be computed for a given individual.

Our interest is not in treatment effects for **specific individuals** *per se*, but, instead, in some characteristics of their distribution.

Treatment Effects

Most of the time focus on two main parameters of interest:

The first one is the **average treatment effect** (ATE):

$$\alpha_{ATE} \equiv \mathbb{E}[Y_1 - Y_0],$$

The second is **average treatment effect on the treated** (TT):

$$\alpha_{TT} \equiv \mathbb{E}[Y_1 - Y_0 | D = 1].$$

As noted, the main challenge is that we **only observe** Y. The standard measure of association between Y and D (the regression coefficient) is:

$$\begin{aligned}\beta &\equiv \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \\ &= \underbrace{\mathbb{E}[Y_1 - Y_0|D=1]}_{\alpha_{TT}} + (\mathbb{E}[Y_0|D=1] - \mathbb{E}[Y_0|D=0]),\end{aligned}$$

which differs from α_{TT} unless the second term is equal to zero.

The second term indicates the difference in potential outcomes when **untreated for individuals** that are actually treated and individuals that are not.

A nonzero difference may result from a situation in which treatment status is the result of individual decisions where those with low Y_0 choose treatment more frequently than those with high Y_0 (**difference in composition**).

An important assumption of the potential outcome representation is that the effect of the treatment on one individual is **independent of the treatment received by other** individuals. This excludes equilibrium or feedback effects, as well as strategic interactions among agents.

Structural vs Reduced/Form Effects

From a **structural model** of D and Y, one could obtain the implied average treatment effects.

Instead, here, they are defined with respect to the distribution of potential outcomes, so that, relative to the 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 treatment effects provide this **intermediate category** between predictive and structural effects, in the sense that recovered parameters are causal effects, but they are uninterpretable in the same sense as reduced form effects.

Sample Average Treatment Effects

Sample analogs for α_{ATE} and α_{TT} are:

$$\alpha^S_{ATE} \equiv \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}}(Y_{1i} - Y_{0i})$$

$$\alpha^S_{TT} \equiv \frac{1}{\sum_{\mathrm{i=1}}^{\mathrm{N}} D_i}\sum_{\mathrm{i=1}}^{\mathrm{N}} D_i(Y_{1i} - Y_{0i}).$$

If factual and counterfactual potential outcomes were observed, these quantities could be estimated without error. The sample average version of β is given by:

$$\begin{aligned}\beta^S &\equiv \bar{Y}_T - \bar{Y}_C \\ &\equiv \frac{1}{N_1}\sum_{\mathrm{i=1}}^{\mathrm{N}} Y_i D_i - \frac{1}{N_0}\sum_{\mathrm{i=1}}^{\mathrm{N}}(1 - D_i)Y_i,\end{aligned}$$

where $N_1 \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} D_i$ is the number of treated individuals, and $N_0 \equiv N - N_1$ is the number of untreated.

Identification: Independence

Identification of the treatment effects depends on the **assumptions** we make on the relation between potential outcomes and the treatment.

Simplest case is when the distribution of the potential outcomes is **independent** of the treatment (e.g. randomized experiments):

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

When this happens:

$$F(Y_1|D=1) = F(Y_1)$$
$$F(Y_0|D=0) = F(Y_0)$$

which implies that:

$$\mathbb{E}[Y_1] = \mathbb{E}[Y_1|D=1] = \mathbb{E}[Y|D=1]$$
$$\mathbb{E}[Y_0] = \mathbb{E}[Y_0|D=0] = \mathbb{E}[Y|D=0]$$

and, as a result, $\alpha_{ATE} = \alpha_{TT} = \beta \Rightarrow$ **unbiased estimate** of α_{ATE}:

$$\widehat{\alpha}_{ATE} = \bar{Y}_T - \bar{Y}_C = \beta^S.$$

No need to "control" for other **covariates**.

Identification: Conditional Independence

A less restrictive assumption is **conditional independence**:

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

where X is a vector of covariates.

This situation is known as **matching**: for each "type" of individual (i.e. each value of covariates) we match treated and controls.

Conditional independence implies:

$$\mathbb{E}[Y_j|X] = \mathbb{E}[Y_j|D = j, X] = \mathbb{E}[Y|D = j, X] \text{ for } j = 0, 1$$

and, as a result:

$$\alpha_{ATE} = \mathbb{E}[Y_1 - Y_0] = \int (\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X])dF(X),$$

For the treatment effect on the treated:

$$\alpha_{TT} = \int \mathbb{E}[Y_1 - Y_0|D = 1, X]dF(X|D = 1) = \int \mathbb{E}[Y - \mu_0(X)|D = 1, X]dF(X|D = 1)$$

where $\mu_0(X) \equiv \mathbb{E}[Y|D = 0, X]$. The function $\mu_0(X)$ is used as an imputation for Y_0.

Identification: Absence of Independence

Finally, sometimes we cannot assume conditional independence:

$$(Y_1, Y_0) \not\perp D|X.$$

In this case, we will need some variable Z that constitutes an **exogenous** source of variation in D, in the sense that it satisfies the **independence assumption**:

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

and the **relevance condition**:

$$Z \not\perp D|X.$$

As we discuss below, in this context we are only going to be able to identify an average treatment effect for a subgroup of individuals, and we call the resulting parameter a **local average treatment effect**.

Social Experiments

Randomized Experiments

In the treatment effect approach, a **randomized field trial** is regarded as the ideal research design.

In a controlled experiment, **treatment** status is **randomly assigned** by the researcher, which by construction, ensures independence. Hence:

$$\alpha_{ATE} = \alpha_{TT} = \beta$$

Long **history** of randomized field trials in social welfare in the U.S., beginning in the 1960s (see Moffitt (2003) for a review).

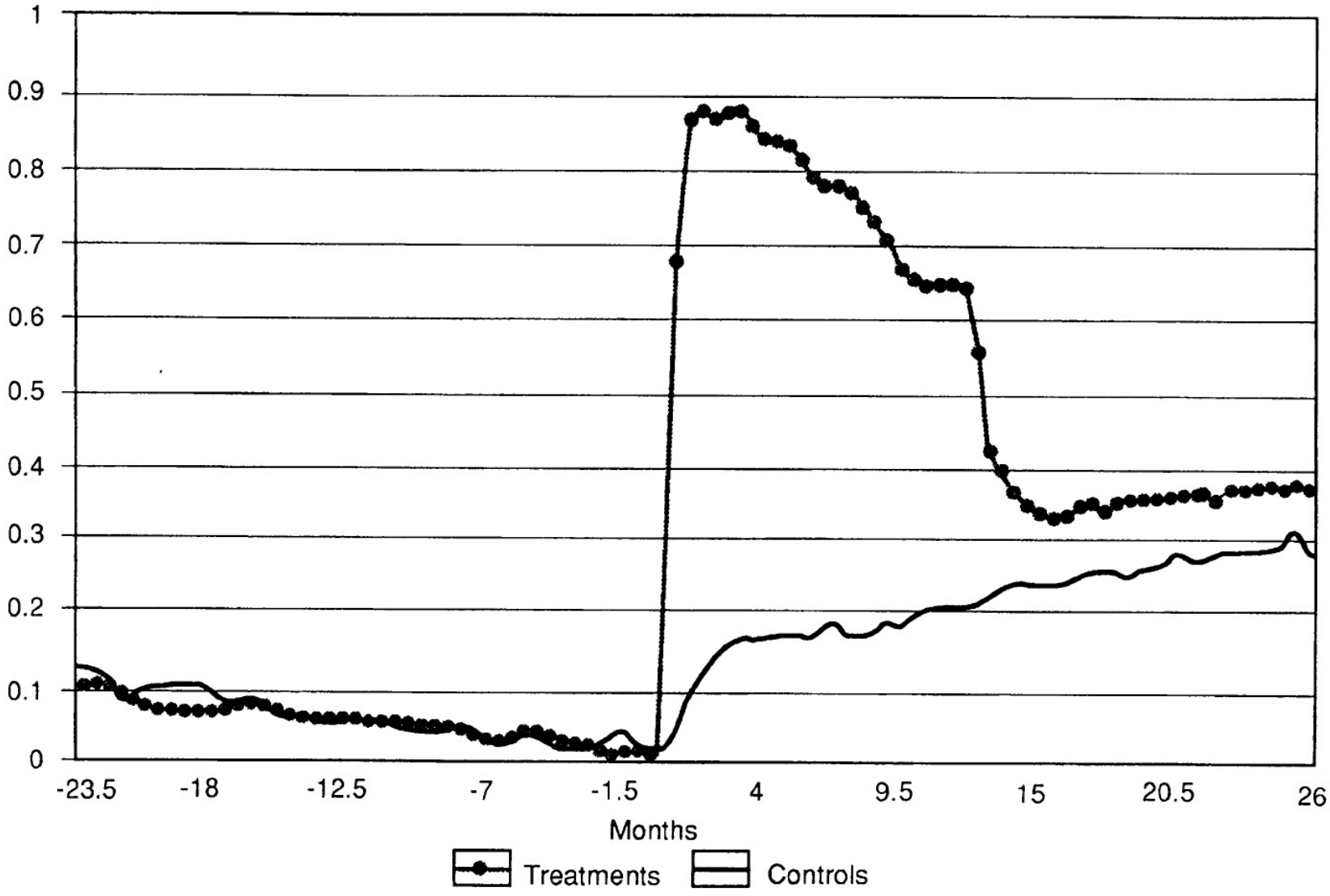
Encouraged by U.S. Federal Government, eventually almost mandatory. Legislation introduced in 1988.

Resistance from many states on ethical grounds (more so in other countries, where treatment groups are often areas for treatment instead of individuals).

Training Program and Unemployment Outcomes

National Supported Work program (NSW):

- designed in the U.S. in the mid 1970s
- training and job opportunities to disadvantaged workers
- NSW guaranteed to treated participants 12 months of subsidized employment (as trainees) in jobs with gradual increase in work standards.
- experimental design on women who volunteered for training
- Requirements: unemployed, a long-term AFDC recipient, and have no preschool children
- Participants were randomly assigned to treatment (275) and control groups (266) in 1976-1977
- Training in 1976, and then followed.
- Ham and LaLonde (1996) analyze the effects of the program.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-23.5
-18
-12.5
-7
-1.5
4
9.5
15
20.5
26
Months
Treatments
Controls

Effects on Unemployment Rates

Thanks to randomization, comparison between employment rates of treatments and controls gives an **unbiased estimate** of the effect of the program on employment at different horizons.

Initially, by construction there is a mechanical effect from the fact that treated women are offered a **subsidized job**.

Compliance with the treatment is decreasing over time, as women can decide to **drop from the subsidized job**.

The **employment growth for controls** is just a reflection of the program's eligibility criteria.

Importantly, after the program ends, a **9 percentage points difference** in employment rates is sustained.

Ham and LaLonde's Additional Point

But Ham and LaLonde (1996) make an important additional point: randomization **does not guarantee independence** for any possible outcomes.

Two examples: wages and unemployment durations (hazards).

Effect of training program on employment rates of the treated $\Rightarrow$ those who are working are a **selected sample**.

Notation: W wages; $Y = 1$ if employed; $\eta = 1$ skilled type.

Suppose:

$$P(Y = 1|D = 1, \eta = j) > P(Y = 1|D = 0, \eta = j), \quad j = 0, 1$$

and:

$$\frac{P(Y = 1|D = 1, \eta = 0)}{P(Y = 1|D = 0, \eta = 0)} > \frac{P(Y = 1|D = 1, \eta = 1)}{P(Y = 1|D = 0, \eta = 1)}.$$

This implies that the **frequency of low skill** will be greater in the group of employed treatments than in the employed controls:

$$P(\eta = 0|Y = 1, D = 1) > P(\eta = 0|Y = 1, D = 0),$$

which is a way to say that η, which is unobserved, is **not independent** of D given $Y = 1$, although, unconditionally, $\eta \perp D$.

Consider the **conditional effects**:

$$\Delta_j \equiv \mathbb{E}[W|Y = 1, D = 1, \eta = j] - \mathbb{E}[W|Y = 1, D = 0, \eta = j], \quad j = 0, 1$$

Our effect of interest is:

$$\Delta_{ATE} = \Delta_0 P(\eta = 0) + \Delta_1 P(\eta = 1),$$

and comparison of average wages between treatments and controls is:

$$\Delta_W = \mathbb{E}[W|Y = 1, D = 1] - \mathbb{E}[W|Y = 1, D = 0] < \Delta_{ATE}.$$

$\Rightarrow$ may not be possible to correctly measure the effect on wages.

Similar problem with exit rates from employment or unemployment $\Rightarrow$ **multi-spell duration** model with **unobserved heterogeneity**.

Matching

Selection Based on Observables and Matching

Experiments are often **too expensive**, **unfeasible**, or **unethical** (e.g. smoking on mortality) $\Rightarrow$ observational data (unlikely to satisfy independence).

Selection into treatment: independence not satisfied.

Selection on observables: independence holds given X (but not unconditionally).

Consider the ATE from above:

$$\alpha_{ATE} = \int (\mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X]) dF(X).$$

Matching: compares individuals with the same characteristics and then integrates over the distribution of characteristics.

Similar arguments follow the expression for α_{TT}.

The common support condition

Essential condition for matching: for each possible value of X, there are individuals in the treatment and control group for which we can average outcomes $\Rightarrow$ **common support condition**:

$$0 < P(D = 1|X) < 1 \quad \text{for all } X \text{ in its support.}$$

Counterexample (with a single covariate):

$$P(D = 1|X) = \begin{cases} 1 & \text{if } X_{min} \leq X < X_0 \\ p \in (0,1) & \text{if } X_0 \leq X \leq X_1 \\ 0 & \text{if } X_1 < X \leq X_{max} \end{cases} .$$

Implication:

- $\mathbb{E}[Y|D = 1, X]$ only identified for values of X in the range (X_{min}, X_1)
- $\mathbb{E}[Y|D = 0, X]$ only identified for values of X in the range (X_0, X_{max})
- $\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X]$ only for values of X in the intersection range $(X_0, X_1) \Rightarrow \alpha_{ATE}$ and α_{TT} not identified

Estimation Methods: Discrete Case

Notation:

- X is discrete and takes on J possible values $\{x_j\}_{j=1}^J$
- N observations $\{X_i\}_{i=1}^N$
- N^j is the number of observations in cell j
- N_ℓ^j be the number of observations in cell j with $D = \ell$
- $\bar{Y}_\ell^j$ be the mean outcome in cell j for $D = \ell$

Note $\bar{Y}_1^j - \bar{Y}_0^j$ is the sample counterpart of $\mathbb{E}[Y|D = 1, X = x_j] - \mathbb{E}[Y|D = 0, X = x_j]$, which can be used to get the following estimates:

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j \right) \frac{N^j}{N}$$

$$\widehat{\alpha}_{TT} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j \right) \frac{N_1^j}{N_1} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - \bar{Y}_0^{j(i)} \right).$$

where $j(i)$ indicates the cell of X_i (note matching interpretation of the second expression for $\widehat{\alpha}_{TT}$).

Estimation Methods: Continuous Case

In the continuous case, a **matching estimator** can be regarded as a way of constructing **imputations** for missing potential outcomes in a similar way, so that gains $Y_{1i} - Y_{0i}$ can be estimated for each unit.

In the **discrete case** we were doing:

$$\widehat{Y}_{0i} = \bar{Y}_0^{j(i)} \equiv \sum_{k \in (D=0)} \frac{\mathbb{1}\{X_k = X_i\} Y_k}{\sum_{\ell \in (D=0)} \mathbb{1}\{X_\ell = X_i\}},$$

and now we can generalize it to:

$$\widehat{Y}_{0i} = \sum_{k \in (D=0)} w(i,k) Y_k,$$

and different estimators will use **different weighting** schemes.

Estimation Methods: Two Classes of Weights

Nearest neighbor matching. Picks the closest observation:

$$w(i,k) = \mathbb{1}\{X_k = \min_i ||X_k - X_i||\},$$

sometimes restricting the sample to cases in which $\min_i ||X_k - X_i|| < \varepsilon$ for some ε.

Typically applied to compute α_{TT}, but also applicable to α_{ATE}.

Kernel matching.

$$w(i,k) = \frac{\kappa\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)}{\sum_{\ell \in (D=0)} \kappa\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)},$$

where $\kappa(.)$ is a kernel that downweights distant observations, and γ_{N_0} us a bandwidth parameter.

Estimation Methods: Propensity Score

Popular method: **propensity score matching**.

Rosenbaum and Rubin (1983) defined the **propensity score** as:

$$\pi(x) \equiv P(D = 1|X),$$

and proved that if $(Y_1, Y_0) \perp D|X$ then:

$$(Y_1, Y_0) \perp D|\pi(X),$$

provided that $0 < \pi(X) < 1$ (do the proof).

Two-step methods: estimate the propensity score, and then create the appropriate weighting.

In the **unconditional independence** case we can write α_{ATE} as:

$$\alpha_{ATE} = \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0] = \frac{\mathbb{E}[DY]}{P(D = 1)} - \frac{\mathbb{E}[(1 - D)Y]}{P(D = 0)}.$$

Thus, under conditional independence we can write:

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|X] &= \mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X] \\ &= \frac{\mathbb{E}[DY|X]}{P(D=1|X)} - \frac{\mathbb{E}[(1-D)Y|X]}{P(D=0|X)} \\ &= \frac{\mathbb{E}[DY|X]}{\pi(X)} - \frac{\mathbb{E}[(1-D)Y|X]}{1-\pi(X)},\end{aligned}$$

which implies:

$$\begin{aligned}\alpha_{ATE} &= \mathbb{E}\left[\mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\middle| X\right]\right] \\ &= \mathbb{E}\left[Y\frac{D-\pi(X)}{\pi(X)[1-\pi(X)]}\right].\end{aligned}$$

Based on the sample analog of the above expression, Hirano et al. (2003) propose the following estimator:

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i\left(\frac{D-\widehat{\pi}(X_i)}{\widehat{\pi}(X_i)[1-\widehat{\pi}(X_i)]}\right).$$

Note that is estimator is of the matching type described above, where the expression in parenthesis is the corresponding weight.

Advantages and Disadvantages of Matching

Regression coefficient including X as controls is a consistent estimate of $\alpha_{ATE} \Rightarrow$ comparison between the two:

Advantages of matching:

- It avoids functional form assumptions
- It emphasizes the common support condition
- Focuses on a single parameter at a time, which is obtained through explicit aggregation

Disadvantages of matching:

- Works under the presumption that for $X = x$ there is random variation in D, so that we can observe both Y_0 and $Y_1 \Rightarrow$ fails if D is a deterministic function of X (i.e. $\pi(X)$ is 0 or 1).
- Good enough X may not have within-cell variation in D, but too much variation in D may be too little X.

Monetary Incentives and Schooling in the UK

Example: Dearden et al. (2009) analyze the effect of a conditional cash transfer on school participation in the UK.

- They participated in the design and did the evaluation
- It was called *Education Maintenance Allowance* (EMA)
- Pilot implementation started in September 2009
- EMA paid youths aged 16-18 that continued in full time education (after 11 compulsory grades) a weekly stipend of £30 to £40, plus bonuses for good results up to £140
- Eligibility (and amounts paid) depends on household characteristics (full payment, <£13,000; >£30,000 not eligible)
- No experimental design; treatment and control areas, including both rural and urban

Question: more education results from this policy? Families fail to decide optimally due to liquidity constraints or missinformation.

Problem: given that individuals in treatment and control areas can differ in characteristics, the unconditional independence may not hold $\Rightarrow$ propensity score matching

Implementation: estimate $\pi(X)$ using a Probit with family, local, and school characteristics. For each treated observation they construct a counterfactual mean using kernel regression and bootstrap standard errors.

Results: EMA increased participation in grade 12 by 5.9% for eligible individuals, and by 3.7% for the whole population. Estimated effects significantly different from zero only for full-payment recipients.

Instrumental Variables (IV)

Identification of Causal Effects in IV Settings

Suppose:

$$(Y_1, Y_0) \not\perp D|X,$$

but there is some variable Z that satisfies **independence condition**:

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

and the **relevance condition**:

$$Z \not\perp D|X.$$

Matching can be regarded as a special case in which $Z = D$, i.e. all the variation in D is exogenous given X.

For simplicity, we do most of the analysis below considering a **single binary instrument** Z, and we abstract from including **covariates**.

Two cases: homogeneous and heterogeneous treatment effects.

Identification: Homogeneous Treatment Eff.

In this case, the causal effect is the **same for every individual**:

$$\widehat{Y}_{1i} - \widehat{Y}_{0i} = \alpha.$$

Availability of an instrumental variable allows us to **identify** α (traditional situation in econometric models — IV regression).

Note that:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = Y_{0i} + \alpha D_i.$$

Taking into account that $Y_{0i} \perp Z_i$:

$$\alpha = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]},$$

Identification **requires** that $\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0] \neq 0$, which is the relevance condition.

All in all, we are getting the effect of D on Y through the effect of Z because Z only affects Y through D (**exclusion restriction**).

Identification: Heterogeneous TE

In the heterogeneous case, the availability of instrumental variables is **not sufficient** to identify a causal effect (e.g. α_{ATE}).

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 the 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, which is known as the **local average treatment effect** (LATE).

Potential Treatment Representation:

Let:

- D_0: D when $Z = 0$
- D_1: D when $Z = 1$

Only observe D_ℓ, for ℓ either equal to one or to zero $\Rightarrow$ **four observable** groups, eight **potential groups**:

Obs. type	Z	D	D_0	D_1	**Latent type**
Type 1	0	0	0	0	Never-taker
				1	Complier
Type 2	0	1	1	0	Defier
				1	Always-taker
Type 3	1	0	0	0	Never-taker
			1		Defier
Type 4	1	1	0	1	Complier
			1		Always-taker

Role of monotonicity:

Now we have:

$$\mathbb{E}[Y|Z=1] = \mathbb{E}[Y_0] + \mathbb{E}[(Y_1 - Y_0)D_1]$$
$$\mathbb{E}[Y|Z=0] = \mathbb{E}[Y_0] + \mathbb{E}[(Y_1 - Y_0)D_0],$$

which implies:

$$\begin{aligned}\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] &= \mathbb{E}[(Y_1 - Y_0)(D_1 - D_0)] \\ &= \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = 1]P(D_1 - D_0 = 1) \\ &\quad - \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = -1]P(D_1 - D_0 = -1).\end{aligned}$$

Thus, $\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$ **could be negative** and yet the causal effect be **positive for everyone**, as long as the probability of *defiers* is sufficiently large.

Example of monotonicity: **eligibility rule**.

Assume:

$$P(D = 1|Z = 0) = 0,$$

(individuals with $Z = 0$ are **denied treatment**).

In this case:

$$\mathbb{E}[Y|Z = 1] = \mathbb{E}[Y_0] + \mathbb{E}[Y_1 - Y_0|D = 1, Z = 1]P(D = 1|Z = 1),$$

and, since $P(D = 1|Z = 0) = 0$:

$$\mathbb{E}[Y|Z = 0] = \mathbb{E}[Y_0].$$

Therefore:

$$\alpha_{TT} = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{P(D = 1|Z = 1)}$$

(where we use $P(Z = 1|D = 1) = 1$.

$\Rightarrow$ if the eligibility condition holds, the **IV coefficient** coincides with the **treatment effect on the treated**.

Local Average Treatment Effects (LATE)

Ruling out defiers (which implies monotonicity):

$$\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]=\mathbb{E}[Y_1-Y_0|D_1-D_0=1]P(D_1-D_0=1),$$

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

Local average treatment effect (LATE):

$$\alpha_{LATE}\equiv\mathbb{E}[Y_1-Y_0|D_1-D_0=1]=\frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[D|Z=1]-\mathbb{E}[D|Z=0]}.$$

Imbens and Angrist (1994) called like this because it is the average treatment effects **on the subsample of compliers**.

$\Rightarrow$ different instrumental variables lead to **different parameters**, even under instrument validity, which is counter to standard GMM thinking.

$\Rightarrow$ need to think of the **group of compliers selected** by the instrument (policy relevant instruments).

This concept changed radically the **way we think** of and understand IV.

Relevance requires presence of compliers.

Conditional Estimation with IV

Assume independence and relevance only cold **conditionally**:

$$(Y_0, Y_1) \perp Z|X \text{ (conditional independence)}$$
$$Z \not\perp D|X \text{ (conditional relevance)} .$$

Example: distance to college, Z, is not randomly assigned but chosen by parents, and this choice may depend on family background, X.

In general, we now have a **conditional LATE** given X:

$$\gamma(X) \equiv \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = 1, X],$$

and a conditional IV estimator:

$$\beta(X) \equiv \frac{\mathbb{E}[Y|Z = 1, X] - \mathbb{E}[Y|Z = 0, X]}{\mathbb{E}[D|Z = 1, X] - \mathbb{E}[D|Z = 0, X]}.$$

Aggregate effect: we proceed differently depending on whether the effects are homogeneous or heterogeneous.

In the **homogeneous** case:

$$Y_1 - Y_0 = \beta(X).$$

In the **heterogeneous** case, it makes sense to consider an average treatment effect for the **overall subpopulation of compliers**:

$$\begin{aligned}\beta_C &\equiv \int \beta(X)\frac{P(compliers|X)}{P(compliers)}dF(X) \\ &= \int \{\mathbb{E}[Y|Z=1,X]-\mathbb{E}[Y|Z=0,X]\}\,\frac{1}{P(compliers)}dF(X),\end{aligned}$$

where:

$$P(compliers) = \int \{\mathbb{E}[D|Z=1,X]-\mathbb{E}[D|Z=0,X]\}\,dF(X).$$

Therefore:

$$\beta_C = \frac{\int \{\mathbb{E}[Y|Z=1,X]-\mathbb{E}[Y|Z=0,X]\}\,dF(X)}{\int \{\mathbb{E}[D|Z=1,X]-\mathbb{E}[D|Z=0,X]\}\,dF(X)},$$

which can be estimated as a **ratio of matching estimators** (Frölich, 2003).

Relating LATE & Parametric Models

Endogenous dummy explanatory variable probit model

The model as usually written in terms of observables is:

$$\begin{array}{l} Y = \mathbb{1}\{\beta_0 + \beta_1 D + U \geq 0\} \\ D = \mathbb{1}\{\pi_0 + \pi_1 Z + V \geq 0\} \end{array} \qquad \begin{pmatrix} U \\ V \end{pmatrix} \Big| Z \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

In this model D is an **endogenous** explanatory variable as long as $\rho \neq 0$. D is exogenous if $\rho = 0$.

In this model, there are only two **potential outcomes**:

$$\begin{aligned} Y_1 &= \mathbb{1}\{\beta_0 + \beta_1 + U \geq 0\} \\ Y_0 &= \mathbb{1}\{\beta_0 + U \geq 0\} \end{aligned}$$

The ATE is given by:

$$\theta = \mathbb{E}[Y_1 - Y_0] = \Phi(\beta_0 + \beta_1) - \Phi(\beta_0).$$

In less parametric specifications, $\mathbb{E}[Y_1 - Y_0]$ may **not be point identified**, but we may still be able to estimate a LATE.

The index model for the treatment equation imposes **monotonicity**.

For example, consider the case in which Z is **binary**, so that there are only **two potential values** of D:

$$D_1 = \mathbb{1}\{\pi_0 + \pi_1 + V \geq 0\}$$
$$D_0 = \mathbb{1}\{\pi_0 + V \geq 0\}.$$

Suppose, without loss of generality, that $\pi_1 \geq 0$. Then:

Group	Conditon			Probability mass
Never-takers	$V < -\pi_0 - \pi_1$	$\Rightarrow$	$D_1 = 0, D_0 = 0$	$1 - \Phi(\pi_0 + \pi_1)$
Compliers	$-\pi_0 - \pi_1 \leq V \leq -\pi_0$	$\Rightarrow$	$D_1 = 1, D_0 = 0$	$\Phi(\pi_0 + \pi_1) - \Phi(\pi_0)$
Always-takers	$V \geq -\pi_0$	$\Rightarrow$	$D_1 = 1, D_0 = 1$	$\Phi(\pi_0)$

We can obtain the average treatment effect for **compliers**:

$$\theta_{LATE} = \mathbb{E}[Y_1 - Y_0 | D_1 - D_0 = 1] = \mathbb{E}[Y_1 - Y_0 | -\pi_0 - \pi_1 \leq V < -\pi_0].$$

We have:

$$\mathbb{E}[Y_1|-\pi_0-\pi_1 \leq V < -\pi_0] = P(\beta_0+\beta_1+U \geq 0|-\pi_0-\pi_1 \leq V < -\pi_0)$$

$$= 1 - \frac{P(U \leq -\beta_0-\beta_1, V \leq -\pi_0) - P(U \leq -\beta_0-\beta_1, V \leq -\pi_0-\pi_1)}{P(V \leq -\pi_0) - P(V \leq -\pi_0-\pi_1)},$$

and similarly:

$$\mathbb{E}[Y_0| - \pi_0 - \pi_1 \leq V < -\pi_0] = P(\beta_0 + U \geq 0| - \pi_0 - \pi_1 \leq V < -\pi_0)$$

$$= 1 - \frac{P(U \leq -\beta_0, V \leq -\pi_0) - P(U \leq -\beta_0, V \leq -\pi_0-\pi_1)}{P(V \leq -\pi_0) - P(V \leq -\pi_0-\pi_1)}.$$

Finally:

$$\theta_{LATE} = \frac{\left\{ \begin{array}{c} [\Phi_2(-\beta_0-\beta_1, -\pi_0-\pi_1; \rho) - \Phi_2(-\beta_0-\beta_1, -\pi_0; \rho)] \\ -[\Phi_2(-\beta_0, -\pi_0-\pi_1; \rho) - \Phi_2(-\beta_0, -\pi_0; \rho)] \end{array} \right\}}{\Phi(-\pi_0) - \Phi(-\pi_0-\pi_1)},$$

where $\Phi_2(r, s; \rho) \equiv P(U \leq r, V \leq s)$ is the standard normal bivariate probability. The nice thing about θ_{LATE} is that it **is identified** also in the **absence of joint normality**. In fact, it does not even require monotonicity in the relationship between Y and D.

Models with additive errors: switching regressions

Consider the **switching regression** model with endogenous switch:

$$Y_i = \beta_0 + \beta_{1i} D_i + U_i$$
$$D_i = \mathbb{1}\{\gamma_0 + \gamma_1 Z_i + \varepsilon_i \geq 0\}.$$

The **potential outcomes** are:

$$Y_{1i} = \beta_0 + \beta_{1i} + U_i \equiv \mu_1 + V_{1i}$$
$$Y_{0i} = \beta_0 + U_i \equiv \mu_0 + V_{0i},$$

so that the treatment effect $\beta_{1i} = Y_{1i} - Y_{0i}$ is heterogeneous.

Traditional models assume β_{1i} is constant or that it varies only with observable characteristics.

Instead, in this model β_{1i} may depend on unobservables and D_i may be correlated with both U_i and β_{1i}.

We assume the **exclusion restriction** holds, in the sense that $(V_{1i}, V_{0i}, \varepsilon_i)$ or $(U_i, \beta_{1i}.\varepsilon_i)$ are independent of Z_i.

In terms of the **alternative notation**:

$$Y_i = \mu_0 + (\mu_1 - \mu_0)D_i + [V_{0i} + (V_{1i} - V_{0i})D_i],$$

and we can write the **ATE** as $\bar{\beta}_1 \equiv \mu_1 - \mu_0$, and $\xi_i \equiv V_{1i} - V_{0i}$, so that $\beta_{1i} = \bar{\beta}_1 + \xi_i$.

Thus:

$$\mathbb{E}[Y_i|Z_i] = \mu_0 + (\mu_1 - \mu_0)\,\mathbb{E}[D_i|Z_i] + \mathbb{E}[V_{1i} - V_{0i}|D_i = 1, Z_i]\,\mathbb{E}[D_i|Z_i].$$

If β_{1i} is **mean independent** of D_i, then $\mathbb{E}[V_{1i} - V_{0i}|D_i = 1, Z_i] = 0$ and:

$$\mathbb{E}[Y_i|Z_i] = \mu_0 + (\mu_1 - \mu_0)\,\mathbb{E}[D_i|Z_i],$$

so that $\bar{\beta}_1 = \mathrm{Cov}(Z, Y)/\,\mathrm{Cov}(Z, D)$, which is the **IV coefficient**.

Otherwise, $\bar{\beta}_1$ **does not coincide** with the IV coefficient.

A special case of mean independence of β_{1i} with respect to D_i occurs when β_{1i} is **constant**.

The failure of IV can be seen as the result of a **missing variable**:

$$Y_i = \beta_0 + \bar{\beta}_1 D_i + \varphi(Z_i) D_i + \zeta_i,$$

where $\varphi(Z_i) \equiv E[V_{1i} - V_{0i} | D_i = 1, Z_i]$, and $\mathbb{E}[\zeta_i | Z_i] = 0$.

When we are doing IV estimation we are **omitting** the variable $\varphi(Z_i) D_i$.

The **average treatment effect on the treated** and the **LATE** are:

$$\alpha_{TT} = \bar{\beta}_1 + \mathbb{E}[V_{1i} - V_{0i} | D_i = 1]$$

$$\alpha_{LATE} = \bar{\beta}_1 + \mathbb{E}[V_{1i} - V_{0i} | -\gamma_0 - \gamma_1 \leq \varepsilon_i \leq -\gamma_0].$$

The model is completed with the **assumption**:

$$\begin{pmatrix} V_{1i} \\ V_{0i} \\ \varepsilon_i \end{pmatrix} \Big| Z_i \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{1\varepsilon} \\ & \sigma_0^2 & \sigma_{0\varepsilon} \\ & & 1 \end{pmatrix} \right].$$

$\Rightarrow$ parametric model, can be estimated by **ML** or **two-step methods**.

Note that:

$$\mathbb{E}[V_{1i} - V_{0i} | D_i = 1, Z_i] = (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\lambda_i,$$

where $\lambda_i \equiv \lambda(\gamma_0 + \gamma_1 Z_i)$ and $\lambda(.)$ is the **inverse Mills ratio**, so that we can do IV estimation in:

$$Y_i = \beta_0 + \bar{\beta}_1 D_i + (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\lambda_i D_i + \zeta_i,$$

or OLS estimation in:

$$Y_i = \beta_0 + \bar{\beta}_1 \Phi_i + (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\phi_i + \zeta_i^*.$$

With a continuous exclusion restriction, $\mathbb{E}[Y_{1i}|X_i]$ and $\mathbb{E}[Y_{0i}|X_i]$ are **identified up to a constant** (X_i denotes controls, omitted so far). However, the constants are important.

Rosen and Willis (1979) provide an illustration for this model.

Continuous Instr: Marginal Treatment Effects

Support of Z not binary: **multiplicity** of causal effects.

Which of these causal effects are **relevant** for evaluating a given policy?

We can define a different LATE parameter for **every pair** (z, z'):

$$\alpha_{LATE}(z, z') \equiv \frac{\mathbb{E}[Y|Z = z] - \mathbb{E}[Y|Z = z']}{\mathbb{E}[D|Z = z] - \mathbb{E}[D|Z = z']}.$$

Multiplicity is even higher when there is **more than one** instrument.

Potential treatment status: as many potential treatment status indicators D_z as possible values z of the instrument.

IV assumptions:

- Relevance: $P(D = 1|Z = z) \equiv P(z)$ is a nontrivial function of z.
- Orthogonality: $(Y_0, Y_1, D_z) \perp Z$.
- Monotonicity: $D_{zi} \geq D_{z'i}$ or $D_{zi} \leq D_{z'i}$ for any (z, z') and for all units in the population.

Alternatively, we can postulate an **index model** for D_z (Heckman and Vytlacil, 2005):

$$D_z = \mathbb{1}\{\mu(z) - U > 0\} \text{ and } U \perp Z,$$

Monotonicity and index model are **equivalent** (Vytlacil, 2002).

Without loss of generality, **we can set** $\mu(z) = P(z)$, and take U as uniformly distributed in the $(0, 1)$ interval. To see this, note that:

$$\mathbb{1}\{\mu(z) > U\} = \mathbb{1}\{F_U(\mu(z)) > F_U(U)\} = \mathbb{1}\{P(z) > \widetilde{U}\},$$

where $\widetilde{U}$ is uniformly distributed.

Analogy when Z is a **binary scalar instrument**: if $P(0) < P(1)$,

- always-takers have $U < P(0)$,
- compliers have U between $P(0)$ and $P(1)$,
- and never-takers have $U > P(1)$.

Similar argument for any pair (z, z') in the case of a **general** Z.

This result connects LATE thinking with **selection models**. Under monotonicity we can always invoke an index equation and assign a value of U to each unit in the population.

Using the **propensity score** $P(Z) \equiv P(D = 1|Z)$ as an instrument:

$$\alpha_{LATE}(P(z), P(z')) = \frac{\mathbb{E}[Y|P(Z) = P(z)] - \mathbb{E}[Y|P(Z) = P(z')]}{P(z) - P(z')}.$$

- **Binary** $Z \Rightarrow$ what we had in the first place
- **Continuous** $Z \Rightarrow$ marginal treatment effect (taking limits as $z \to z'$):

$$\alpha_{MTE}(P(z)) = \frac{\partial \mathbb{E}[Y|P(Z) = P(z)]}{\partial P(z)}.$$

$\alpha_{LATE}(P(z), P(z'))$ gives the ATE for individuals who would change schooling status from changing $P(Z)$ from $P(z)$ to $P(z')$:

$$\alpha_{LATE}(P(z), P(z')) = \mathbb{E}[Y_1 - Y_0|P(z') < U < P(z)].$$

Similarly, $\alpha_{MTE}(P(z))$ gives the ATE for individuals who would change treatment following a marginal change in $P(z)$, i.e. **indifferent between treatment choices** at $P(Z) = P(z)$:

$$\alpha_{MTE}(P(z)) = \mathbb{E}[Y_1 - Y_0|U = P(z)].$$

Integrating $\alpha_{MTE}(U)$ over different ranges of U we can get other ATE measures. For example:

$$\alpha_{LATE}(P(z), P(z')) = \frac{\int_{P(z')}^{P(z)} \alpha_{MTE}(u)du}{P(z) - P(z')}, \text{ and } \alpha_{ATE} = \int_0^1 \alpha_{MTE}(u)du,$$

which makes it clear that to be able to identify α_{ATE} we need identification of $\alpha_{MTE}(u)$ over the **entire** $(0, 1)$ **range**.

Heckman and Vytlacil (2005) suggest to **estimate MTE** by estimating the derivative of the conditional mean $\mathbb{E}[Y|P(Z) = P(z), X = x]$ using kernel-based local linear regression techniques.

Homogeneity **can be tested** checking linearity of the conditional mean outcome on the propensity score:

$$\mathbb{E}[Y|P(Z)] = \mathbb{E}[Y_0|P(Z)] + \mathbb{E}[Y_1 - Y_0|P(Z), D = 1]P(Z).$$

The quantity $\mathbb{E}[Y_1 - Y_0|P(Z), D = 1]$ is constant under homogeneity, so that the conditional mean is linear in $P(Z)$.

Remarks about Unobserved Heterog. in IV

How important is it?

Balance between observed and unobserved heterogeneity depends on how detailed information on agents is available (empirical issue).

The worry is not heterogeneity *per se*, but the fact that heterogeneous gains may affect program participation.

Warnings:

In the absence of an economic model or a clear notional experiment, it is often difficult to interpret what IV estimates estimate.

Knowing that IV estimates can be interpreted as averages of heterogeneous effects is not very useful if understanding the heterogeneity itself is first order.

Heterogeneity of gains vs heterogeneity of treatments:

Heterogeneity of treatments may be more important. For example, the literature has found significant differences in returns to different college majors.

Problem of aggregating educational categ. is that returns are less meaningful.

Sometimes aggregated into just two categories, because some techniques are only well developed for binary explanatory variables.

Example: Non-Compliance in Randomized Trial

In a classic example, Z indicates **assignments to treatment** in an experimental design. Therefore, $(Y_0, Y_1) \perp Z$.

However, the **"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 by construction.

Assignment to treatment is **not a valid** instrument in the presence of **externalities** that benefit members of the treatment group, even if they are not treated themselves. In such case, the exclusion restriction fails to hold.

An example of this situation arises in a study of the effect of deworming on school participation in Kenya using school-level randomization (**Miguel and Kremer, 2004**).

Example: Ethnic Enclaves & Immigr Outcomes

Edin et al. (2003): effect of living in highly concentrated ethnic area (enclave) on labor success.

Sweden: 11% population born abroad, $>$ 40% of those live in enclaves.

Causal effect ambiguous *ex-ante*: segregation lowers local skill acquisition, but offer networks which increase opportunities.

Immigrants in ethnic enclaves have **5% lower earnings**, after controlling for age, education, gender, family background, country of origin, and year of immigration.

Problem: decision to live in an enclave is endogenous $\Rightarrow$ **exogenous source of variation**: in 1985-1991 government assigned initial areas of residence to refugee immigrants (correlated w/ settlements, uncorrelated w/ outcomes).

IV estimates imply a **13% gain for low-skill immigrants** associated with one standard deviation increase in ethnic concentration. For high-skill immigrants, there was no effect.

Example: Vietnam Veterans & Earnings

Angrist (1990): Did military service in Vietnam have a negative effect on earnings?

He uses:

- Draft lottery eligibility as the **instrumental variable**
- Veteran status as **treatment variable**
- Log earnings as the **outcome**

Need for instrumentation: strong selection process in the military during the Vietnam period (some volunteered, others avoided enrollment using student or job deferments).

Administrative records for 11,637 white men born in 1950-1953 linked with March CPS of 1979 and 1981-1985.

Lottery:

- Conducted annually during 1970-1974
- Assigned numbers from 1 to 365 to dates of birth in the cohorts being drafted.
- Men with lowest numbers were called up to a ceiling determined every year by the department of defense.

Regression Discontinuity (RD)

The Fundamental RD Assumption

In regression discontinuity we consider a situation where there is a **continuous** variable Z that is not necessarily a valid instrument (it does not satisfy the exogeneity assumption), but such that **treatment** assignment is a **discontinuous function** of Z:

$$\lim_{z \to z_0^+} P(D = 1|Z = z) \neq \lim_{\to z_0^-} P(D = 1|Z = z)$$

$$\lim_{z \to z_0^+} P(Y_j \leq r|Z = z) = \lim_{z \to z_0^-} P(Y_j \leq r|Z = z) \quad (j = 0, 1)$$

which are **relevance** and **orthogonality** conditions respectively.

Implicit regularity conditions are:

- existence of the limits,
- Z has positive density in a neighborhood of z_0.

For now we abstract from **conditioning covariates** for simplicity.

Sharp and Fuzzy Designs

Early RD literature in Psychology (Cook and Campbell, 1979) distinguishes between:

- **Sharp design**: $D = \mathbb{1}\{Z \geq z_0\}$, with:

$$\lim_{z \to z_0^+} \mathbb{E}[D|Z = z] = 1$$
$$\lim_{z \to z_0^-} \mathbb{E}[D|Z = z] = 0.$$

- **Fuzzy design**: $0 < P(D = 1|Z \geq z_0) < 1$, with:

$$P(D = 1|Z = z_0 - \varepsilon) \neq P(D = 1|Z = z_0 + \varepsilon)$$

Homogeneous Treatment Effects

Suppose that $\alpha = Y_1 - Y_0$ is **constant**, so that $Y_i = \alpha D_i + Y_{0i}$.

Conditional expectations given $Z = z$ and left- and right-side limits:

$$\lim_{z \to z_0^+} \mathbb{E}[Y|Z=z] = \alpha \lim_{z \to z_0^+} \mathbb{E}[D|Z=z] + \lim_{z \to z_0^+} \mathbb{E}[Y_0|Z=z]$$
$$\lim_{z \to z_0^-} \mathbb{E}[Y|Z=z] = \alpha \lim_{z \to z_0^-} \mathbb{E}[D|Z=z] + \lim_{z \to z_0^-} \mathbb{E}[Y_0|Z=z],$$

which leads to the consideration of the following **RD parameter**:

$$\alpha_{RD} = \frac{\lim_{z \to z_0^+} \mathbb{E}[Y|Z=z] - \lim_{z \to z_0^-} \mathbb{E}[Y|Z=z]}{\lim_{z \to z_0^+} \mathbb{E}[D|Z=z] - \lim_{z \to z_0^-} \mathbb{E}[D|Z=z]}.$$

determined by **relevance** and **orthogonality** conditions above.

In the case of a sharp design, the denominator is unity so that:

$$\alpha_{RD} = \lim_{z \to z_0^+} \mathbb{E}[Y|Z=z] - \lim_{z \to z_0^-} \mathbb{E}[Y|Z=z],$$

Sharp corresponds to **matching** and fuzzy corresponds to **IV**.

Intuitively, considering units within a small interval around the cutoff point is similar to a **randomized experiment** at the cutoff point.

Heterogeneous Treatment Effects

Now suppose that: $Y_i = \alpha_i D_i + Y_{0i}$.

In the **sharp** design since $D = \mathbb{1}\{Z \geq z_0\}$ we have:

$$\mathbb{E}[Y|Z = z] = \mathbb{E}[\alpha|Z = z]\,\mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_0|Z = z].$$

$\Rightarrow$ the situation is one of **selection on observables**. That is, letting:

$$k(z) \equiv \mathbb{E}[Y_0|Z = z] + \{\mathbb{E}[\alpha|Z = z] - \mathbb{E}[\alpha|Z = z_0]\}\,\mathbb{1}\{z \geq z_0\}$$

we have:

$$\mathbb{E}[Y|Z = z] = \mathbb{E}[\alpha|Z = z_0]\,\mathbb{1}\{z \geq z_0\} + k(z),$$

where $k(z)$ is continuous at $z = z_0$.

Therefore, the **OLS population coefficient** on D in the equation:

$$Y = \alpha_{RD} D + k(z) + w$$

coincides with α_{RD}, which in turn equals $\mathbb{E}[\alpha|Z = z_0]$.

The **control function** $k(z)$ is **nonparametrically identified**: α_{RD} is identified as above $\Rightarrow$ $k(z)$ is identifiable as the nonparametric regression $\mathbb{E}[Y - \alpha_{RD} D | Z = z]$.

Homogeneous TE: $k(z) = \mathbb{E}[Y_0 | Z = z]$, but not in general.

If $\mu(z) \equiv \mathbb{E}[Y_0 | Z = z]$ was known (e.g. using data from a setting in which no program was present) then we could consider a regression of Y on D and $\mu(z) \Rightarrow \mathbb{E}[\alpha | z \geq z_0]$ (coefficient of D in that regression).

In the **fuzzy design**, D not only depends on $\mathbb{1}\{Z \geq z_0\}$, but also on other unobserved variables. Thus, D is an endogenous variable in the above regression.

We can use $\mathbb{1}\{Z \geq z_0\}$ as an **instrument** for D in such equation to identify α_{RD}, at least in the homogeneous case (connection with IV was first made explicit by van der Klaaw (2002)).

Below we discuss **independence** and **monotonicity** in fuzzy designs.

Conditional independence near z_0:

Weak conditional independence: $D \perp (Y_0, Y_1)|Z = z$ for z near z_0, i.e. for $z = z_0 \pm e$, where e is arbitrarily small positive number, or:

$$P(Y_j \leq r|D = 1, Z = z_0 \pm e) = P(Y_j \leq r|Z = z_0 \pm e) \quad (j = 0, 1).$$

An **implication** is:

$$\mathbb{E}[\alpha D|Z = z_0 \pm e] = \mathbb{E}[\alpha|Z = z_0 \pm e]\,\mathbb{E}[D|Z = z_0 \pm e].$$

Proceeding as before, we have:

$$\lim_{z \to z_0^+} \mathbb{E}[Y|Z = z] = \lim_{z \to z_0^+} \mathbb{E}[\alpha|Z = z]\,\mathbb{E}[D|Z = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_0|Z = z]$$
$$\lim_{z \to z_0^-} \mathbb{E}[Y|Z = z] = \lim_{z \to z_0^-} \mathbb{E}[\alpha|Z = z]\,\mathbb{E}[D|Z = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_0|Z = z].$$

Noting that $\lim_{z \to z_0^+} \mathbb{E}[\alpha|Z = z] = \lim_{z \to z_0^-} \mathbb{E}[\alpha|Z = z] = \mathbb{E}[\alpha|Z = z_0]$:

$$\mathbb{E}[\alpha|Z = z_0] = \mathbb{E}[Y_1 - Y_0|Z = z_0] = \frac{\lim_{z \to z_0^+} \mathbb{E}[Y|Z=z] - \lim_{z \to z_0^-} \mathbb{E}[Y|Z=z]}{\lim_{z \to z_0^+} \mathbb{E}[D|Z=z] - \lim_{z \to z_0^-} \mathbb{E}[D|Z=z]} = \alpha_{RD}.$$

That is, the RD parameter can be interpreted as the **average TE** at z_0.

Monotonicity near z_0:

Alternative assumption: **local monotonicity** (Hahn et al., 2001):

$$D_{z_0+\varepsilon} \geq D_{z_0-\varepsilon} \text{ for all units in the population,}$$

for some $\bar{\varepsilon} > 0$ and any pair $(z_0 - \varepsilon, z_0 + \varepsilon)$ with $0 < \varepsilon < \bar{\varepsilon}$, where D_z is the potential assignment indicator associated with $Z = z$.

In some situations, **conditional independence** can be problematic and **local monotonicity** not.

In such cases, it can be shown that α_{RD} identifies the **local average treatment effect** at $z = z_0$:

$$\alpha_{RD} = \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_1 - Y_0 | D_{z_0+\varepsilon} - D_{z_0-\varepsilon} = 1]$$

that is, the ATE for the units for whom treatment changes discontinuously at z_0.

Estimation Strategies

Hahn et al. (2001): Let $S_i \equiv \mathbb{1}\{z_0 - h < Z_i < z_0 + h\}$ where $h > 0$ denotes the bandwidth, and consider the subsample such that $S_i = 1$, and define $W_i \equiv \mathbb{1}\{z_0 < Z_i < z_0 + h\}$ as an instrument, applied to the subsample with $S_i = 1$:

$$\widehat{\alpha}_{RD} = \frac{\widehat{\mathbb{E}}[Y_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[Y_i|W_i = 0, S_i = 1]}{\widehat{\mathbb{E}}[D_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[D_i|W_i = 0, S_i = 1]}.$$

Alternative by the same authors, **control function**: let:

$$\mathbb{E}[D|Z] = g(Z) + \delta\, \mathbb{1}\{Z \geq z_0\} \text{ and } \mathbb{E}[Y_0|Z] = k(Z).$$

Then consider the following **regression**:

$$Y = \alpha_{RD}\, \mathbb{E}[D|Z] + k(Z) + w$$

Parametric approach: assume functional forms for $g(Z)$ and $k(Z)$.

Semiparametric approach (van der Klaaw, 2002): power series approximation for $k(Z)$.

If $g(Z) = k(Z)$: **2SLS** using $\mathbb{1}\{Z \geq z_0\}$ and $g(Z)$ as IVs, where $g(Z)$ is the **included** instrument and $\mathbb{1}\{Z \geq z_0\}$ is the **excluded** instrument.

These methods of estimation, not local to data points near the threshold, are implicitly predicated on the assumption of **homogeneous TE**.

Conditioning on Covariates

Even if the RD assumption is satisfied unconditionally, conditioning on covariates may **mitigate the heterogeneity in treatment effects**, hence contributing to the relevance of RD estimated parameters.

Covariates may also make the **local conditional exogeneity** assumption more credible.

This would also be true of **within-group estimation** in a panel data context (see application in Hoxby, 2000).

Example: Effect of Class Size on Test Scores

Angrist and Lavy (1999) analyze the effect of class size on test scores using the "Maimonides' rule" in Israel.

The Maimonides rule divides students in classes of less than a given **maximum number of students** (40).

Maimonides' rule allows enrollment cohorts of **1-40** to be grouped in a **single class**, but enrollment groups of **41-80** are split into **two classes** of average size 20.5-40, enrollment groups of **81-120** are split into **three classes** of average size 27-40, etc. (in practice, the rule was not exact: class size predicted by the rule differed from actual size).

Angrist and Lavy (1999) use this **discontinuity** to analyze the effect of class size on school outcomes.

Their **outcome variable** is the average test score at a class i in school s, the **treatment variable** (not binary) is the size of class i, and the **instrument** is the total enrollment at the beginning of an academic year at school s.

Example: Effect of financial aid offers on students' enrollment decisions.

This is the interest of **van der Klaaw (2002)**.

His setting is:

- **Outcome**: the decision of student i to enroll in college a given college (binary).
- **Treatment**: the amount of financial aid offer to student i.
- **Instrument**: the index that aggregates SAT score and high school GPA: applicants for aid were divided into four groups on the basis of the interval the index Z fell into.

Average aid offers as a function of Z contained **jumps at the cutoff points** for the different ranks: those scoring just below a cutoff point received much less on average than those who scored just above the cutoff.

Differences-in-Differences (DID)

Card and Krueger (1994)

March 1992: state of New Jersey increased the **legal minimum wage** by 19%, whereas the bordering state of Pennsylvania kept it constant.

Card and Krueger (1994) evaluate the effect of this change on the **employment of low wage workers**.

Competitive model: increase the minimum wage $\Rightarrow$ **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:

- Large source of employment for low-wage workers;
- Comply with minimum wage regulations (especially franchised restaurants);
- Fairly homogeneous job, so good measures of employment and wages can be obtained;
- Easy to get a sample frame of franchised restaurants (yellow pages) with high response rates (response rates 87% and 73% —less in Pennsylvania, because the interviewer was less persistent).

Differences-in-Differences Coefficient

The **differences-in-differences** coefficient is:

$$\beta = \{\mathbb{E}[Y_2|D=1] - \mathbb{E}[Y_1|D=1]\} - \{\mathbb{E}[Y_2|D=0] - \mathbb{E}[Y_1|D=0]\},$$

where Y_1 and Y_2 denote employment before and after the reform, $D = 1$ denotes a store in New Jersey (**treatment group**) and $D = 0$ denotes one in Pennsylvania (**control group**).

β measures the **difference** between the average employment change in New Jersey and the average employment change in Pennsylvania.

Key assumption for causal interpretation: temporal effect in the two states is the same in the absence of intervention (controls).

Card and Krueger found that rising the minimum wage increased employment in some of their comparisons but in no case caused an employment reduction! (economic and political debate).

DID estimation has become **very popular** (especially US: federal structure provides cross state variation in legislation).

Context of DID comparisons

Before-after instead of DID: it can be contaminated by the effect of events other than the treatment that occurred between the two periods.

To see **identification more formally**, consider the two-period potential outcomes representation with treatment in $t = 2$:

$$Y_1 = Y_0(1)$$
$$Y_2 = (1 - D)Y_0(2) + DY_1(2)$$

Fundamental identifying **assumption**: average changes in the two groups are the same in the absence of treatment:

$$\mathbb{E}[Y_0(2) - Y_0(1)|D = 1] = \mathbb{E}[Y_0(2) - Y_0(1)|D = 0].$$

$Y_0(1)$ is observed but $Y_0(2)$ is **counterfactual** for units with $D = 1$.

Under such identification assumption, DID coefficient coincides with the **average treatment effect for the treated**:

$$\beta = \{\mathbb{E}[Y_1(2)|D=1] - \mathbb{E}[Y_0(1)|D=1]\} - \{\mathbb{E}[Y_0(2)|D=0] - \mathbb{E}[Y_0(1)|D=0]\}.$$

Now, adding and subtracting $\mathbb{E}[Y_0(2)|D=1]$:

$$\begin{aligned}\beta = &\mathbb{E}[Y_1(2) - Y_0(2)|D=1] \\ &+\{\mathbb{E}[Y_0(2) - Y_0(1)|D=1] - \mathbb{E}[Y_0(2) - Y_0(1)|D=0]\},\end{aligned}$$

which as long as the last term vanishes it equals:

$$\beta = \mathbb{E}[Y_1(2) - Y_0(2)|D=1].$$

Comments and Problems

Comments:

- β 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, just cross-sectional data for at least two periods.
- We can estimate β from a regression of outcome changes on the treatment dummy. This is convenient for accounting for dependence between the two periods.

Problems:

- β is obtained from differences in averages in the two periods and two groups. If the composition of the cross-sectional populations change over time, estimates will be biased (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.

Distributional Effects and Quantile Treatment Effects

Distributional Effects under Cond. Indep.

Most of the literature focused on **average effects**, but the results seen in previous sections also hold for **distributional comparisons**.

Under **conditional independence**, the full marginal distributions of Y_1 and Y_0 can be identified.

To see this, first note that we can identify not just α_{ATE} but also $\mathbb{E}[Y_j]$:

$$\mathbb{E}[Y_j] = \int \mathbb{E}[Y_j|X]dF(X) = \int \mathbb{E}[Y|D=j,X]dF(X) \quad \text{for } j=0,1.$$

Next, same applies to expected value of any function of the outcomes $\mathbb{E}[h(Y_j)]$:

$$\mathbb{E}[h(Y_j)] = \int \mathbb{E}[h(Y_j)|X]dF(X) = \int \mathbb{E}[h(Y_j)|D=j,X]dF(X) \quad \text{for } j=0,1.$$

Thus, setting $h(Y_j) = \mathbb{1}\{Y_j \leq r\}$ we get:

$$\mathbb{E}[\mathbb{1}\{Y_j \leq r\}] = P(Y_j \leq r) = \int P(Y \leq r|D=j,X)dF(X), \quad \text{for } j=0,1.$$

$\Rightarrow$ we can identify **quantiles** of Y_1 and Y_0.

Quantile TE are differences in the marginal quantiles of Y_1 and Y_0. **Joint distribution** of (Y_1, Y_0) or the **distribution of gains** Y_1, Y_0 requires stronger assumptions.

Quantile TE Matching Estimator

Firpo (2007): quantile TE estimator under the **matching assumptions**.

Let (Y_1, Y_0) be potential outcomes with marginal cdfs $F_1(r)$ and $F_0(r)$, and quantile functions $Q_{1\tau} = F_1^{-1}(\tau)$ and $Q_{0\tau} = F_0^{-1}(\tau)$. The **QTE** is defined as:

$$\alpha_\tau \equiv Q_{1\tau} - Q_{0\tau}$$

Under **conditional exogeneity** $F_j(r) = P(Y \leq r|D = j, X)dG(X)$ for $j = 0, 1$. Moreover, $Q_{1\tau}$ and $Q_{0\tau}$ satisfy the **moment conditions**:

$$\mathbb{E}\left[\frac{D}{\pi(X)}\,\mathbb{1}\{Y \leq Q_{1\tau}\} - \tau\right] = 0$$
$$\mathbb{E}\left[\frac{1-D}{1-\pi(X)}\,\mathbb{1}\{Y \leq Q_{0\tau}\} - \tau\right] = 0,$$

and

$$Q_{1\tau} = \arg\min_q \mathbb{E}\left[\frac{D}{\pi(X)}\rho_\tau(Y - q)\right]$$
$$Q_{0\tau} = \arg\min_q \mathbb{E}\left[\frac{1-D}{1-\pi(X)}\rho_\tau(Y - q)\right],$$

where $\rho_\tau(u) \equiv [\tau - \mathbb{1}\{u < 0\}]u$ is the **"check" function**.

Firpo's method is a **two-step weighting procedure** in which the propensity score $\pi(X)$ is estimated in a first stage.

Identification in IV Settings

Imbens and Rubin (1997) show that if conditional independence does not hold, but valid instruments that satisfy **monotonicity** are available, not only the **average treatment effect for compliers** is identified but also the entire **marginal distributions** of Y_0 and Y_1 for them.

Interpretation: example college subsidy.

Abadie (2002) gives a **simple proof** that suggests a similar calculation to the one done for ATE. For any function $h(.)$ consider:

$$W \equiv h(Y)D = \begin{cases} W_1 \equiv h(Y_1) & \text{if } D = 1 \\ W_0 \equiv 0 & \text{if } D = 0 \end{cases}$$

$(W_1, W_0, D_1, D_0) \perp Z \Rightarrow$ apply **LATE formula** to W and get:

$$\mathbb{E}[W_1 - W_0 | D_1 - D_0 = 1] = \frac{\mathbb{E}[W|Z=1] - \mathbb{E}[W|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]},$$

or substituting:

$$\mathbb{E}[h(Y_1) | D_1 - D_0 = 1] = \frac{\mathbb{E}[h(Y)D|Z=1] - \mathbb{E}[h(Y)D|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]},$$

If we choose $h(Y) = \mathbb{1}\{Y \leq r\}$, the previous formula gives as an expression for the **cdf of Y_1 for the compliers**.

Similarly, if we consider:

$$V \equiv h(Y)(1-D) = \begin{cases} V_1 \equiv h(Y_0) & \text{if } 1-D=1 \\ V_0 \equiv 0 & \text{if } 1-D=0 \end{cases}$$

then:

$$\mathbb{E}[V_1 - V_0 | D_1 - D_0 = 1] = \frac{\mathbb{E}[V|Z=1] - \mathbb{E}[V|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]},$$

or:

$$\mathbb{E}[h(Y_0)|D_1 - D_0 = 1] = \frac{\mathbb{E}[h(Y)(1-D)|Z=1] - \mathbb{E}[h(Y)(1-D)|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]},$$

from which we get the **cdf of Y_0 for compliers**, for $h(Y) = \mathbb{1}\{Y \leq r\}$.

Intuition: suppose D is exogenous ($Z = D$), then the cdf of $Y|D=0$ coincides with the cdf of Y_0, and the cdf of $Y|D=1$ coincides with the cdf of Y_1. If we regress $h(Y)D$ on D, the **OLS regression coefficient** is:

$$\mathbb{E}[h(Y)D|D=1] - \mathbb{E}[h(Y)D|D=0] = E[h(Y_1)],$$

which for $h(Y) = \mathbb{1}\{Y \leq r\}$ gives us the cdf of Y_1. Similarly, if we regress $h(Y)(1-D)$ on $(1-D)$, the regression coefficient is:

$$\mathbb{E}[h(Y)(1-D)|1-D=1] - \mathbb{E}[h(Y)(1-D)|1-D=0] = E[h(Y_0)].$$

In the **IV case**, we are running similar IV (instead of OLS) regressions using Z as instrument and getting expected $h(Y_1)$ and $h(Y_0)$ for **compliers**.

An IV Quantile TE Estimator

Abadie (2003): weighting that is useful to estimate IV quantile TE.

If Z satisfies the **standard assumptions** given X, for any **measurable function** of (Y, X, D) with finite expectation, $h(Y, X, D)$:

$$\mathbb{E}[h(Y, X, D)|D_1 - D_0 = 1] = \frac{\mathbb{E}[\kappa h(Y, X, D)]}{\mathbb{E}[\kappa]},$$

where:

$$\kappa = 1 - \frac{D(1-Z)}{1 - P(Z = 1|X)} - \frac{(1-D)Z}{P(Z = 1|X)}.$$

The main idea is that the operator κ **"finds compliers"**, given that:

$$\mathbb{E}[\kappa|Y, X, D] = \Pr(D_1 - D_0 = 1|Y, X, D).$$

Intuition: $D(1-Z) = 1$ are classified as *always-takers*; $(1-D)Z = 1$ are classified as *never-takers*; hence, the left-out are the compliers.

Abadie et al (2002): estimator given by the sample analog to:

$$(\alpha_\tau, Q_{0\tau}) = \arg\min_{(a,q)} \mathbb{E}\left[\rho_\tau(Y - aD - q)|D_1 - D_0 = 1\right]$$

$$= \arg\min_{(a,q)} \mathbb{E}\left[\kappa\rho_\tau(Y - aD - q)\right]$$

κ **needs to be estimated** (and standard errors should take this into account, e.g. bootstrapped standard errors).

κ is **negative** when $D \neq Z$ (instead of zero), which makes the regression minimand non-convex. Solution: **law of iterated expectations**:

$$(\alpha_\tau, Q_{0\tau}) = \arg\min_{(a,q)} \mathbb{E}\left[\mathbb{E}[\kappa|Y,X,D]\rho_\tau(Y - aD - q)\right]$$

Note that:

$$\mathbb{E}[\kappa|Y,X,D] = 1 - \frac{D(1-\mathbb{E}[Z|Y,X,D=1])}{1-P(Z=1|X)} - \frac{(1-D)\,\mathbb{E}[Z|Y,X,D=0]}{\Pr(Z=1|X)}.$$

A very simple **two-stage method** consists of the following two steps:

1. **Estimate** $\mathbb{E}[Z_i|Y_i,X_i,D_i]$ with a **Probit** of Z_i on Y_i and X_i separately for $D_i = 0$ and $D_i = 1$ subsamples, and $P(Z_i = 1|X_i)$ with a Probit of Z_i on X_i with the whole sample. Construct $\hat{\mathbb{E}}[\kappa_i|Y_i,X_i,D_i]$ using the fitted values from the previous expressions.
2. Estimate the **quantile regression model** with the standard procedure using these predicted kappas as **weights**.

Regression Discontinuity

For some function $h(.)$, consider the **outcome** W defined as above:

$$W \equiv h(Y)D = \begin{cases} W_1 \equiv h(Y_1) & \text{if } D = 1 \\ W_0 \equiv 0 & \text{if } D = 0 \end{cases}$$

For $h(Y) = \mathbb{1}\{Y \leq r\}$ **RD parameter** for outcome $W(r) = \mathbb{1}\{Y \leq r\}D$:

$$P(Y_1 \leq r | Z = z_0) = \frac{\lim\limits_{z \to z_0^+} \mathbb{E}[W(r)|Z = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[W(r)|Z = z]}{\lim\limits_{z \to z_0^+} \mathbb{E}[D|Z = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[D|Z = z]},$$

under **local conditional independence**. Similarly, for $P(Y_0 \leq r|Z = z_0)$:

$$V \equiv h(Y)(1 - D) = \begin{cases} V_1 \equiv h(Y_0) & \text{if } 1 - D = 1 \\ V_0 \equiv 0 & \text{if } 1 - D = 0, \end{cases}$$

and the **RD parameter** for the outcome $V(r) = \mathbb{1}\{Y \leq r\}(1 - D)$ delivers:

$$P(Y_0 \leq r | Z = z_0) = \frac{\lim\limits_{z \to z_0^+} \mathbb{E}[V(r)|Z = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[V(r)|Z = z]}{\lim\limits_{z \to z_0^+} \mathbb{E}[D|Z = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[D|Z = z]},$$

Comparing the two, we obtain the **distributional treatment effects**.

A Bridge Between Structural and Reduced Form Methods

Ex-Post and Ex-Ante Policy Evaluation

Ex-post policy evaluation:

- Happens after the policy has been implemented.
- Techniques seen so far in the course.
- Makes use of existing policy variation.
- Experimental and non-experimental methods are used.

Ex-ante evaluation:

- Interventions which have not taken place.
- Treatment levels outside of the range of existing programs, other modifications to existing programs, or new programs altogether.
- Requires an extrapolation from (i) existing policy or (ii) policy-relevant variation.

In this section we draw from Todd and Wolpin (2006), Wolpin (2007), and Chetty (2009).

Example: Eval. of School Attendance Subsidy

Consider the *ex-ante* evaluation of a **school attendance subsidy program** in a development country. Consider two possible situations:

1. School **tuition** p **varies exogenously** across countries in range $(\underline{p}, \bar{p})$.
2. **Schools are free**: $p = 0$.

In case 1) it is possible to estimate a relationship between school attendance s and tuition cost p, but in case 2) it is not.

Suppose that s also depends on a set of observed factors X and it is possible to **estimate non-parametrically**:

$$s = f(p, X) + v.$$

$\Rightarrow$ estimate the effect of the subsidy b on s for all households i in which **tuition net of the subsidy** $p_i - b$ is in the support of p.

Some values of net tuition must be **outside of the support** $\Rightarrow$ not possible to estimate **entire response function**, or population estimates of impact of the subsidy without **parametric assumptions**.

In the case 2), we need to look at the **opportunity cost of school**.

Consider a household with one child making a decision about whether to **send the child to school** ($s = 1$) or to work ($s = 0$). Suppose $s = 1$ if:

$$w < w^*$$

where w^* represents the utility gain for the household if the child goes to school. If $w^* \sim \mathcal{N}(\alpha, \sigma^2)$, we get a standard **probit model**:

$$P(s = 1) = 1 - P(w^* < w) = \Phi\left(\frac{\alpha - w}{\sigma}\right)$$

To obtain separate estimates of α and σ we need to observe **child wage offers** (not only the wages of children who work). Under the school subsidy the child goes to school if $w < w^* + b$ so that the probability that a child attends school will increase by:

$$\Phi\left(\frac{b + \alpha - w}{\sigma}\right) - \Phi\left(\frac{\alpha - w}{\sigma}\right).$$

The conclusion is that variation in the opportunity cost of attending school (the child market wage) serves as a **substitute for variation in the tuition** cost of schooling.

Combining Experim. & Structural Estimation

Todd and Wolpin (2006): evaluate the effects of the PROGRESA school subsidy program on schooling of girls in rural Mexico.

Mexican government conducted a **randomized social experiment** between 1997 and 1999, in which 506 rural villages were randomly assigned to either treatment (320) or control groups (186).

Parents of eligible treatment households were offered substantial **payments** contingent on their children's regular **attendance at school**.

The **benefit levels** represented about 1/4 of average family income. The subsidy increased with grade level up to grade 9 (age 15).

Eligibility was determined on the basis of a poverty index.

Experimental treatment effects on school attendance rates one year after the program showed **large gains**, ranging from about 5 to 15 percentage points depending on age and sex.

The experiment alone cannot determine the **size and structure** of the subsidy that achieves the **policy goals at the lowest cost**, or to assess alternative policy tools to achieve the same goals.

Todd and Wolpin use a (dynamic) **structural model** of parental fertility and schooling choices to compare the efficacy of the PROGRESA program with that of alternative policies that were not implemented.

They estimate the model **using control households** only, **exploiting child wage** variation and, in particular, distance to the nearest big city for identification.

They use the **treatment sample for model validation** and presumably also for model selection.

The model is outlined in the notes, and **details of the model** are explained in Pedro's part.

They emphasize that social experiments provide an opportunity for out-of-**sample validation** of models that involve extrapolation outside the range of existing policy variation.

Model Selection and Data Mining

Once the researcher has estimated a model, she can perform diagnostics, like tests of **model fit** and **tests of overidentifying** restrictions.

If the model does not provide a good fit, the researcher will **change the model** in the directions in which the model poorly fits the data.

Formal methods of model selection are no longer applicable because the **model is the result of repeated pretesting**.

Estimating a fixed set of models and employing a model selection criterion (like AIC) is also unlikely to help because **models that result from repeated pretesting** will tend to be very similar in terms of model fit.

Holding Out Data

Policy maker's trade-off: release all data or hold some of it to validate models provided by researchers.

The policy maker **selects several researchers**, each of whose task is to develop a model for ex ante evaluation.

One possibility is to give the researcher **all the data**.

The other possibility is to **hold out the post-program treatment** households, so that the researcher only has access to control households.

Is there any **gain in holding out** the data on the treated households? That is, is there a gain that compensates for the information loss from estimating the model on a smaller sample with less variation?

The problem is that after all the pre-testing associated with model building it is not a **viable strategy** to try to discriminate among models on the basis of within-sample fit because all the models are more or less indistinguishable.

So we need some other **criterion for judging** the relative success of a model. One is assessing a model's predictive accuracy for a hold out sample.

A Bayesian Approach

Alternative: weighting models on the basis of posterior model probabilities in a Bayesian framework $\Rightarrow$ posterior model probabilities carry an automatic penalty for overfitting (Schorfheide and Wolpin, 2012).

The odd **posterior ratio** between two models is given by the odd prior ratio times the likelihood ratio:

$$\frac{P(M_j|y)}{P(M_\ell|y)} = \frac{P(M_j)f(y|M_j)}{P(M_\ell)f(y|M_\ell)},$$

where $f(y|M_j) = \int f(y|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j$.

Schwarz approximation: contains a correction factor for the difference in the number of parameters:

$$\frac{f(y|M_j)}{f(y|M_\ell)} \approx \frac{f(y|\widehat{\theta}_j, M_j)}{f(y|\widehat{\theta}_j, M_j)} \times N^{-[\dim(\theta_j)-\dim(\theta_\ell)]/2}.$$

The **posterior distribution of a treatment effect** or predictor Δ is:

$$P(\Delta|y) = \sum_j P(\Delta|y, M_j)P(M_j|y)$$

where $P(\Delta|y, M_j)$ is the posterior density of Δ calculated under model M_j.

For Bayesian perspective **holdout samples are suboptimal** because posterior probabilities should be based on entire sample y not a subsample.

Problem: the set of models under consideration is incomplete and data dependent (Schorfheide and Wolpin).

Researcher starts with some model, inspects the data, reformulates the model, considers alternative models based on the previous data inspection,...

This is a process of **data mining** (e.g. the Smets and Wouters (2007) DSGE model widely used in macro policy evaluation).

Problem with data mining is prior distribution is shifted towards models that **fit the data well** whereas other models that fit slightly worse are forgotten.

So these data dependent priors produce marginal likelihoods that:

- **overstate** the fit of the reported model
- the posterior distribution **understates** the parameter uncertainty.

There is no **viable commitment** from the modelers not to look at data that are stored on their computers!

A Principal-Agent Model for Holdout

Schorfheide and Wolpin (2012, 2014) develop a **principal-agent framework** to address this trade-off. Data mining generates an **impediment** for the implementation of the ideal Bayesian analysis.

In their analysis there is a **policy maker** (the principal) and **two modelers** (the agents). The modelers can each fit a structural model to whatever data they get from the policy maker and provide predictions of the treatment effect.

The modelers are **rewarded based on the fit** of the model that they are reporting. So they have an **incentive to engage in data mining**.

In the context of a holdout sample, modelers are asked by the policy maker to predict features of the **sample that is held out** for model evaluation.

If the modelers are rewarded such that their payoff is proportional to the log of the reported predictive density for Δ, then they have an **incentive to reveal their subjective beliefs truthfully** (i.e. to report the posterior density of Δ given their model and the data available to them).

They provide a formal **rationale for holding out samples** in situations where the policy maker is unable to implement the full Bayesian analysis.

Sufficient Statistics

Third alternative: **sufficient statistics** (reviewed in Chetty (2009)).

Middle ground between structural and reduced form: sufficient-statistic formulas combining the advantages of reduced-form empirics (**transparent and credible identification**) with an important advantage of structural models (the ability to make **precise statements about welfare**).

Derive formulas for the welfare consequences of policies that are functions of high-level **elasticities rather than deep primitives**.

Even though there are multiple combinations of primitives that are consistent with the inputs to the formulas, all these combinations have the same welfare implications (Chetty, 2009).

School subsidy example: express increase in welfare in terms of the elasticity of school enrollment to changes in tuition (and maybe some other), despite the subsidy can affect later decisions of the individual in terms of future education and employment, as well as parent's fertility decisions.

Provided that the program-evaluation estimates can provide a **value to these elasticities**, this approach allows to give economic meaning to what might otherwise be viewed as atheoretical statistical estimates.

Concluding Remarks

Concluding Remarks

Empirical papers have become more **central** to economics than they used to. This reflects the new possibilities afforded by technical change in research and is a sign of **scientific maturity** of Economics.

In an empirical paper the **econometric strategy** is often paramount, i.e. what aspects of data to look at and how to interpret them.

This typically requires a good understanding of both relevant **theory** and **sources of variation** in data.

Once this is done there is usually a more or less obvious **estimation method** available and ways of **assessing statistical error**.

Statistical issues (quality of large sample approximations or measurement error) may be more or less important, but good empirical papers should **focus on the econometric problems that matter** for the question addressed.

The **quasi-experimental approach** is also having a contribution to reshaping **structural econometric** practice.

It is increasingly becoming standard a reporting style that distinguishes clearly the **roles of theory and data** in getting the results.

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**.

In Applied Micro there has been a **lot of excitement** in recent years in empirically establishing **causal impacts of interventions**. Understandable because causal impacts are more useful for policy than correlations.

However, **increasing awareness** of the limitations due to **heterogeneity** of responses and interactions and **dynamic feedback**. Addressing these matters require more theory.

A good thing of the treatment effect literature is that it has substantially raised the **empirical credibility hurdle**.

Challenge for coming years: more theory-based or structural empirical models that are structural not just because the author has written the model as derived from utility functions but because he/she has been able to establish empirically invariance to a particular class of interventions, which therefore lends credibility to the model for ex ante policy evaluation within this class.