The problem of causality in microeconometrics

Andrea Ichino
European University Institute

This course is an introduction to some conventional and unconventional methods for the identification and estimation of the causal effect of a “treatment” on an “outcome”. The relationship between educational choices and labor market outcomes will offer the main source of examples and applications, but, occasionally, also other fields in economics as well as medical sciences will be considered. The course will also discuss how theoretical reasoning should be combined with these methods to guide the interpretation of estimated effects.

The problem of causality in microeconometrics.

Andrea Ichino
European University Institute

April 15, 2014

Contents

1 The Problem of Causality

- Does fertility depend on the number of storks?
- Does aspirin reduce the risk of heart attacks?
- Does an additional year of schooling increase future earnings?
- Are temporary jobs a stepping stone to permanent employment?
- Does EPL increase unemployment?

The answers to these questions (and to many others which affect our daily life) involve the identification and measurement of causal links: an old problem in philosophy and statistics.

We need a framework to study causality.

1.1 A formal framework to think about causality

We have a population of units; for each unit we observe a variable D and a variable Y.

We observe that D and Y are correlated. Does *correlation* imply *causation*?

In general no, because of:

- confounding factors;
- reverse causality.

We would like to understand in which sense and under which hypotheses one can conclude from the evidence that D *causes* Y.

It is useful to think at this problem using the terminology of experimental analysis.

- i is an index for the units in the population under study.

- D_i is the *treatment* status:

  $D_i = 1$ if unit i has been exposed to treatment;

  $D_i = 0$ if unit i has not been exposed to treatment.

- $Y_i(D_i)$ indicates the potential outcome according to treatment:

  $Y_i(1)$ is the outcome in case of treatment;

  $Y_i(0)$ is the outcome in case of no treatment;

The observed outcome for each unit can be written as:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{1}$$

This approach requires to think in terms of "counterfactuals".

1.2 The fundamental problem of causal inference

Definition 1. Causal effect.
For a unit i, the treatment D_i has a causal effect on the outcome Y_i if the event $D_i = 1$ instead of $D_i = 0$ implies that $Y_i = Y_i(1)$ instead of $Y_i = Y_i(0)$. In this case the causal effect of D_i on Y_i is

$$\Delta_i = Y_i(1) - Y_i(0)$$

The identification and the measurement of this effect is logically impossible.

Proposition 1. The Fundamental Problem of Causal Inference.
It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values $Y_i(1)$ and $Y_i(0)$ and, therefore, it is impossible to observe *the effect of D on Y for unit i (Holland, 1986).*

Another way to express this problem is to say that we cannot infer the effect of a treatment because we do not have the *counterfactual* evidence i.e. what would have happened in the absence of treatment.

1.3 The statistical solution

Statistics approaches the problem by aiming at population parameters like the average causal effect for the entire population or for some interesting sub-groups.

The effect of treatment on a random unit (ATE):

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1) - Y_i(0)\} \\ &= E\{Y_i(1)\} - E\{Y_i(0)\} \end{aligned} \tag{2}$$

The effect of treatment on the treated (ATT):

$$\begin{aligned} E\{\Delta_i \mid D_i = 1\} &= E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \end{aligned} \tag{3}$$

... ATNT, LATE ...

Are these effects interesting from the viewpoint of an economist?

Are these effects identified?

Naive estimator: is the comparison by treatment status informative?

A comparison of output by treatment status gives a biased estimate of the ATT:

$$\begin{aligned} E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} & \qquad (4) \\ = E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} & \\ = E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} & \\ + E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} & \\ = \tau + E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} & \end{aligned}$$

where $\tau = E\{\Delta_i \mid D_i = 1\}$ is the ATT.

The difference between the left hand side (which we can estimate) and τ is the *sample selection bias* equal to the difference between the outcomes of treated and control subjects in the counterfactual situation of no treatment (i.e. at the baseline).

The problem is that the outcome of the treated and the outcome of the control subjects are not identical in the no-treatment situation.

1.4 Randomized experiments

Consider two random samples C and T from the population. Since by construction these samples are statistically identical to the entire population we can write:

$$E\{Y_i(0)|i \in C\} = E\{Y_i(0)|i \in T\} = E\{Y_i(0)\} \tag{5}$$

and

$$E\{Y_i(1)|i \in C\} = E\{Y_i(1)|i \in T\} = E\{Y_i(1)\}. \tag{6}$$

Substituting 5 and 6 in 2 it is immediate to obtain:

$$\begin{aligned} E\{\Delta_i\} &\equiv E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= E\{Y_i(1)|i \in T\} - E\{Y_i(0)|i \in C\}. \end{aligned} \tag{7}$$

Randomization solves the Fundamental Problem of Causal Inference because it allows to use the *control* units C as an image of what would happen to the *treated* units T in the counterfactual situation of no treatment, and vice-versa.

Lalonde (1986) gives a provocative description of the mistakes that a researcher can make using observational data instead of experimental randomized data.

However, randomized experiments are not always a feasible solution for economists because of:

- ethical concerns;
- difficulties of technical implementation;
- external validity and replication (consider instead structural estimation ...).

In these lectures we will study some alternatives to randomized experiments.

Each of these alternatives aims at getting as close as possible to a randomized experiment.

Before doing so we analyse the problem of causality in a more familiar regression framework.

2 Causality in a regression framework

Consider the following specification of outcomes, with or without treatment:

$$\begin{aligned} Y_i(1) &= \mu(1) + U_i(1) \\ Y_i(0) &= \mu(0) + U_i(0) \end{aligned} \tag{8}$$

where $E\{U_i(1)\} = E\{U_i(0)\} = 0$. The causal effect of treatment for an individual is

$$\begin{aligned} \Delta_i &= Y_i(1) - Y_i(0) \\ &= [\mu(1) - \mu(0] + [U_i(1) - U_i(0)] \\ &= E\{\Delta_i\} + [U_i(1) - U_i(0)]. \end{aligned} \tag{9}$$

It is the sum of:

$E\{\Delta_i\} = \mu(1) - \mu(0)$:
the common gain from treatment equal for every individual i;

$[U_i(1) - U_i(0)]$:
the idiosyncratic gain from treatment that differs for each individual i and that may or may not be observed by the individual.

The statistical effects of treatment in this model

i. *The effect of treatment on a random individual (ATE).*

$$\begin{aligned} E\{\Delta_i) &= E\{Y_i(1) - Y_i(0)\} \\ &= E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= \mu(1) - \mu(0) \end{aligned} \tag{10}$$

ii. *The effect of treatment on the treated (ATT)*

$$\begin{aligned} E\{\Delta_i \mid D_i = 1) &= E\{Y_i(1) - Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &= \mu(1) - \mu(0) + E\{U_i(1) - U_i(0) \mid D_i = 1\} \end{aligned} \tag{11}$$

The two effects differ because of the idiosyncratic gain for the treated

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} \tag{12}$$

This is the average gain that those who are treated obtain on top of the average gain for a random person in the population.

A regression with random coefficients

Let D_i indicate treatment: using equation 1 the outcome can be written as:

$$\begin{aligned} Y_i &= \mu(0) + [\mu(1) - \mu(0) + U_i(1) - U_i(0)]D_i + U_i(0) \\ &= \mu(0) + \Delta_i D_i + U_i(0) \end{aligned} \tag{13}$$

where $D_i = 1$ in case of treatment and $D_i = 0$ otherwise.

This is a linear regression with a random coefficient on the RHS variable D_i.

(Figure on board: Differences between treated and control individuals.)

2.1 Specification of the selection into treatment

The model is completed by the specification of the rule that determines the participation of individuals into treatment:

$$D_i^* = \alpha + \beta Z_i + V_i \tag{14}$$

where $E\{V_i\} = 0$ and

$$D_i = \begin{cases} 1 \text{ if } D_i^* \geq 0 \\ 0 \text{ if } D_i^* < 0 \end{cases} \tag{15}$$

D_i^* is the (unobservable) criterion followed by the appropriate decision maker concerning the participation into treatment of individual i. The decision maker could be nature, the researcher or the individual.

Z_i is the set of variables that determine the value of the criterion and therefore the participation status. No randomness of coefficients is assumed here.

Z_i could be a binary variable.

The model in compact form

$$Y_i = \mu(0) + \Delta_i D_i + U_i(0) \tag{16}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{17}$$

$$D_i = \left\{ \begin{array}{l} 1 \text{ if } D_i^* \geq 0 \\ 0 \text{ if } D_i^* < 0 \end{array} \right\} \tag{18}$$

$$\begin{aligned} \Delta_i &= \mu(1) - \mu(0) + U_i(1) - U_i(0) \\ &= E\{\Delta_i\} + U_i(1) - U_i(0) \end{aligned} \tag{19}$$

$$E\{U_i(1)\} = E\{U_i(0)\} = E\{V_i\} = 0 \tag{20}$$

Correlation between U_i and V_i is possible.

2.2 Problems with OLS estimation

2.2.1 Bias for the effect of treatment on a random person

Using 19 we can rewrite equation 16 as:

$$\begin{aligned} Y_i &= \mu(0) + E\{\Delta_i\}D_i + U_i(0) + D_i[U_i(1) - U_i(0)] \\ &= \mu(0) + E\{\Delta_i\}D_i + \epsilon_i \end{aligned} \tag{21}$$

that tells us what we get from the regression of Y_i on D_i.

Problem:

$$E\{\epsilon_i D_i\} = E\{U_i(1) \mid D_i = 1\}Pr\{D_i = 1\} \neq 0 \tag{22}$$

Therefore the estimated coefficient of Y_i on D_i is a biased estimate of $E\{\Delta_i\}$

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i\} + \tag{23}$$

$$E\{U_i(1) - U_i(0) \mid D_i = 1\} + E\{U_i(0) \mid D_i = 1\} - E\{U_i(0) \mid D_i = 0\}$$

The second line is the bias for the ATE

Readjusting the second line of 23, the bias can be written as:

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i\} + \tag{24}$$
$$E\{U_i(1) \mid D_i = 1\} - E\{U_i(0) \mid D_i = 0\}$$

This bias is equal to the difference between two componenents:

- $E\{U_i(1) \mid D_i = 1\}$
  the unobservable outcome of the treated in case of treatment;
- $E\{U_i(0) \mid D_i = 0\}$
  the unobservable outcome of the controls in the case of no treatment.

In general, there is no reason to expect this difference to be equal to zero.

Consider a controlled experiment in which participation into treatment is random because

- assignment to the treatment or control groups is random and
- there is full compliance with the assignment.

Under these assumptions it follows that:

$$\begin{aligned} E\{U_i(1)\} &= E\{U_i(1) \mid D_i = 1\} = 0 \\ E\{U_i(0)\} &= E\{U_i(0) \mid D_i = 0\} = 0 \end{aligned} \tag{25}$$

Hence, under perfect randomization, the treatment and the control groups are statistically identical to the entire population and therefore

$$\begin{aligned} E\{\Delta_i\} &= E\{Y_i(1)\} - E\{Y_i(0)\} \\ &= E\{Y_i(1) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} \\ &= \mu(1) - \mu(0) \end{aligned} \tag{26}$$

But, is the effect of treatment on a random person interesting in economic examples?

2.2.2 Bias for the effect of treatment on a treated person

Adding and subtracting $D_i E\{U_i(1) - U_i(0) \mid D_i = 1\}$ in 21 and remembering from 11 that $E\{\Delta_i \mid D_i = 1\} = E\{\Delta_i\} + E\{U_i(1) - U_i(0) \mid D_i = 1\}$, we can rewrite 21 as:

$$\begin{aligned} Y_i &= \mu(0) + E\{\Delta_i \mid D_i = 1\} D_i + \\ &\quad U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D_i = 1\}] \\ &= \mu(0) + E\{\Delta_i \mid D_i = 1\} D_i + \eta_i \end{aligned} \tag{27}$$

Using 27 we can define the OLS bias in the estimation of $E\{\Delta_i \mid D_i = 1\}$.

$E\{\Delta_i \mid D_i = 1\}$ is the ATT which is equal to the common effect plus *the average idiosyncratic gain*.

The error term is again correlated with the treatment indicator D_i:

$$\begin{aligned} E\{\eta_i D_i\} &= E\{D_i U_i(0) + D_i[U_i(1) - U_i(0) - E\{U_i(1) - U_i(0) \mid D_i = 1\}]\} \\ &= E\{D_i U_i(0)\} \neq 0. \end{aligned} \tag{28}$$

Therefore, the estimated coefficient of Y_i on D_i is biased also with respect to $E\{\Delta_i \mid D_i = 1\}$:

$$E\{Y_i \mid D_i = 1\} - E\{Y_i \mid D_i = 0\} = E\{\Delta_i \mid D_i = 1\} + \tag{29}$$
$$E\{U_i(0) \mid D_i = 1\} - E\{U_i(0) \mid D_i = 0\}$$

The second line in 29 is the bias for the ATT

$$E\{U_i(0) \mid D_i = 1\} - E\{U_i(0) \mid D_i = 0\}$$

is called *mean selection bias* and "tells us how the outcome in the base state differs between program participants and non-participants. Absent any general equilibrium effects of the program on non participants, such differences cannot be attributed to the program." (Heckman, 1997)

This bias is zero only when participants and non-participants are identical in the base state i.e. when $E\{U_i(0)D_i\} = 0$.

Would randomization help in the estimation of the ATT?

3 Heckman's solution for endogenous dummy variable models

Consider the case in which

- $U_i(1) = U_i(0)$: no idiosyncratic gain from treatment)
- $\Delta = \mu(1) - \mu(0)$

and we want to estimate the following model which allows for covariates X_i :

$$\begin{aligned} Y_i &= \mu(0) + \gamma X_i + \Delta D_i + U_i(0) \\ Y_i &= \mu + \gamma X_i + \Delta D_i + U_i \end{aligned} \tag{30}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{31}$$

$$D_i = \left\{ \begin{array}{l} 1 \text{ if } D_i^* \geq 0 \\ 0 \text{ if } D_i^* < 0 \end{array} \right\} \tag{32}$$

where $E\{U_i\} = E\{V_i\} = 0$ but $E\{D_i U_i\} \neq 0$.

This model is commonly called the *endogenous dummy variable* model (see Heckman (1978) and Maddala (1983).

OLS is biased for Δ because

- those who have on average higher unobservable outcomes
- may also be more likely to enter into treatment (or viceversa).

Examples:

- Roy model (Roy, 1951).
- Parental background for returns to schooling (Willis-Rosen, 1979).
- Effects of unions on wages (Robinson, 1989)
- Labor supply of female workers (Heckman, 1978)
- ...

The model rewritten as a switching regression model

We can transform the model in the following way:

$$\text{Regime 1: if } D_i^* \geq 0 \qquad Y_i = \mu + \gamma X_i + \Delta + U_i \tag{33}$$

$$\text{Regime 0: if } D_i^* < 0 \qquad Y_i = \mu + \gamma X_i + U_i \tag{34}$$

or equivalently

$$\text{Regime 1: if } V_i \geq -\alpha - \beta Z_i \qquad Y_i = \mu + \gamma X_i + \Delta + U_i \tag{35}$$

$$\text{Regime 0: if } V_i < -\alpha - \beta Z_i \qquad Y_i = \mu + \gamma X_i + U_i \tag{36}$$

where Regime 1 implies treatment.

This is an endogenous switching regression model in which the intercept differs under the two regimes. More generally we could allow also the coefficient γ to differ in the two regimes.

It would seem feasible to estimate separately the above two equations on the two sub-samples that correspond to each regime and to recover an estimate of $\triangle$ from the difference between the two estimated constant terms. But ...

If $E\{D_i U_i\} \neq 0$, the error terms U_i do not have zero mean within each regime.

$$\text{Regime 1:} \quad E\{U_i \mid V_i \geq -\alpha - \beta Z_i\} \neq E\{U_i\} = 0 \tag{37}$$

$$\text{Regime 0:} \quad E\{U_i \mid V_i < -\alpha - \beta Z_i\} \neq E\{U_i\} = 0 \tag{38}$$

The selection bias takes the form of an omitted variable specification error such that the error term in each regime does not have zero mean.

If we could observe the two espectations in 37 and 38, we could include them in the two regressions and avoid the misspecification.

Heckman's great intution has been to find a way to estimate these expectations

3.1 Some useful results on truncated normal distributions

Assume that U and V are jointly normally distributed with

- zero means,
- standard deviations respectively equal to σ_U and σ_V
- covariance equal to σ_{UV}.

Let $\phi(.)$ be the standard normal density and $\Phi(.)$ its CDF. The:

$$E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} > k_1\right\} = \frac{\phi(k_1)}{1-\Phi(k_1)} \tag{39}$$

$$E\left\{\frac{U}{\sigma_U} \mid \frac{U}{\sigma_U} < k_2\right\} = -\frac{\phi(k_2)}{\Phi(k_2)} \tag{40}$$

$$E\left\{\frac{U}{\sigma_U} \mid k_1 < \frac{U}{\sigma_U} < k_2\right\} = \frac{\phi(k_1)-\phi(k_2)}{\Phi(k_2)-\Phi(k_1)} \tag{41}$$

The ratios on the RHS are the *Inverse Mill's ratios*.

The following results also hold:

$$E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} > k\right\} = \sigma_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > k\right\} \tag{42}$$

$$= \sigma_{UV} \frac{\phi(k)}{1 - \Phi(k)}$$

$$\tag{43}$$

$$E\left\{\frac{U}{\sigma_U} \mid \frac{V}{\sigma_V} < k\right\} = \sigma_{UV} E\left\{\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} < k\right\} \tag{44}$$

$$= -\sigma_{UV} \frac{\phi(k)}{\Phi(k)}$$

These are precisely the results we need to solve the problem.

3.2 The Heckman two-steps procedure

We cannot observe $E\{U_i \mid V_i \geq -\alpha - \beta Z_i\}$ and $E\{U_i \mid V_i < -\alpha - \beta Z_i\}$ but we can estimate them using the participation equation 31.

Without loss of generality we can assume $\sigma_V = 1$ (this parameter is anyway not identified in a probit model). The steps of the procedure are as follows

i. Estimate a probit model for the participation into treatment using 31, and retrieve the estimated absolute values of the *Inverse Mill's Ratios*

$$M_{1i} = \frac{\phi(-\hat{\alpha} - \hat{\beta} Z_i)}{1 - \Phi(-\hat{\alpha} - \hat{\beta} Z_i)} = \frac{\phi(\hat{\alpha} + \hat{\beta} Z_i)}{\Phi(\hat{\alpha} + \hat{\beta} Z_i)} \tag{45}$$

$$M_{0i} = \frac{\phi(-\hat{\alpha} - \hat{\beta} Z_i)}{\Phi(-\hat{\alpha} - \hat{\beta} Z_i)} = \frac{\phi(\hat{\alpha} + \hat{\beta} Z_i)}{1 - \Phi(\hat{\alpha} + \hat{\beta} Z_i)} \tag{46}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated probit coefficients.

ii. Estimate using OLS the equations for the two regimes augmented with the appropriate *Inverse Mill's Ratios* obtained in the first step

$$\text{Regime 1:} \qquad Y_i = \mu + \gamma X_i + \Delta + \lambda_1 M_{1i} + \nu_i \tag{47}$$

$$\text{Regime 0:} \qquad Y_i = \mu + \gamma X_i + \lambda_0 M_{0i} + \nu_i \tag{48}$$

where $\lambda_1 = \sigma_U \sigma_{UV}$, $\lambda_0 = -\sigma_U \sigma_{UV}$ and $E\{\nu_i\} = 0$ since the *Inverse Mill's ratios* have been consistently estimated.

With the above two steps we can get a consistent estimate of the treatment effect Δ

We just need to subract the estimated constant in 48 from the estimated constant in 47.

This two steps procedure generalizes in a full maximum likelihood estimation.

3.3 Comments

- Note that $\hat{\lambda}_1$ is a consistent estimate of $\sigma_U \sigma_{UV}$ while $\hat{\lambda}_0$ is a consistent estimate of $-\sigma_U \sigma_{UV}$. Full maximum likelihood estimation, instead of the two step procedure described above is, possible (and is provided by most of the available software packages).

- Therefore, if the error terms are positively correlated (i.e. those who tend to have higher outcomes are also more likely to participate into treatement) we should expect a positive coefficient on the *Inverse Mill's ratio* in Regime 1 and a negative coefficient in Regime 0.

- If the coefficients on the *Inverse Mill's Ratios* $\hat{\lambda}_1$ and $\hat{\lambda}_0$ are not significantly different form zero, this indicates that there is no endogenous selection in the two regimes. So this procedure provides a test for the existence of endogenous selection.

- Suppose that $Z_i = X_i$, i.e. there is no exogenous variable which determines the selection into treatment and which is excluded from the outcome equation. In this case you could still run the procedure and get estimates of λ_0 and λ_1. But the identification would come only from the distributional assumptions. Only because of the assumptions the *Inverse Mill's ratios* would be a non-linear transformation of the regressors X_i in the outcome equations.

- Therefore this procedure does not avoid the problem of finding a *good instrument*. And if we had one then using IV we could obtain estimates of treatment effects without making unnecessary distributional assumptions.

- Check the performance of this procedure in LaLonde's (1986) results

4 Standard IV and solutions based on control functions

4.1 IV

Let's continue to assume that

- $U_i(1) = U_i(0)$: no idiosyncratic gain from treatment;
- $\Delta = \mu(1) - \mu(0)$

so that the model in compact form is

$$Y_i = \mu(0) + \Delta D_i + U_i \tag{49}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{50}$$

$$D_i = \left\{ \begin{array}{l} 1 \text{ if } D_i^* \geq 0 \\ 0 \text{ if } D_i^* < 0 \end{array} \right\} \tag{51}$$

$$E\{U_i\} = E\{V_i\} = 0 \tag{52}$$

If subjects are not randomly selected into treatment:

$$COV\{U,V\} = E(UV) \neq 0 \tag{53}$$

and OLS gives an inconsistent estimate of Δ.

$$\text{plim}\{\hat{\Delta}_{OLS}\} = \frac{COV\{Y,D\}}{V\{D\}} = \Delta + \frac{COV\{U,D\}}{V\{D\}} \neq \Delta \tag{54}$$

But under the assumptions

$$COV(Z,D) \neq 0 \tag{55}$$

$$COV(U,Z) = 0. \tag{56}$$

satisfied by our compact model, we have that:

$$\frac{COV\{Y,Z\}}{COV\{D,Z\}} = \Delta + \frac{COV\{U,Z\}}{COV\{D,Z\}} = \Delta = \text{plim}\{\hat{\Delta}_{IV}\} \tag{57}$$

Substituting the appropriate sample covariances on the LHS of 57 we get the well known IV estimator $\hat{\Delta}_{IV}$.

We will come back to it later with a different and inspiring perspective

4.2 Control functions

We want to show that under the same identification assumptions one can use a different estimation strategy first described by Heckman and Robb (1985).

This is the "control function" strategy which has advantages and disatvantages with respect to IV depending on the specific model to be estimated.

Consider the reduced form

$$D_i = \pi_0 + \pi_1 Z_i + \epsilon_i \tag{58}$$

where by construction

$$E(Z\epsilon) = 0 \tag{59}$$

Note that the endogeneity of D derives from the fact that

$$E(U\epsilon) \neq 0 \tag{60}$$

because Z is exogenous.

Consider the linear projection of U on ϵ

$$U_i = \rho_0 + \rho_1 \epsilon_i + q_i \tag{61}$$

where by construction and 56

$$\begin{aligned} E(q\epsilon) &= 0 \\ E(qZ) &= 0 \\ E(\epsilon Z) &= 0 \end{aligned} \tag{62}$$

Then, plugging 61 into 49 we obtain:

$$Y_i = \mu(0) + \Delta D_i + \rho_0 + \rho_1 \epsilon_i + q_i \tag{63}$$

where we can consider ϵ_i as an explanatory variable.

But note that now

- given 58, D is a linear function of Z and ϵ;
- given 62, the error term q is uncorrelated with Z, ϵ and thus D;
- the OLS estimator of equation 63 is consistent for Δ.

The control function method

- Estimate the reduced form (first stage):

$$D_i = \pi_0 + \pi_1 Z_i + \epsilon_i \tag{64}$$

  which gives consistent estimate $\hat{\pi}_0$ of π_0 and $\hat{\pi}_1$ of π_1;

- retrieve

$$\hat{\epsilon}_i = D_i - \hat{\pi}_0 - \hat{\pi}_1 Z_i \tag{65}$$

  and note that

$$\epsilon_i = \hat{\epsilon}_i + \text{error}_i \tag{66}$$

  where error_i is sampling error deriving from the estimation of $\hat{\pi}_0$ and $\hat{\pi}_1$;

- plug 66 into 49 and estimate consistently with OLS:

$$Y_i = \mu(0) + \Delta D_i + \rho_0 + \rho_1 \hat{\epsilon}_i + \text{error}_i + q_i \tag{67}$$

The inclusion of the residual $\hat{\epsilon}$ controls for the endogeneity of D in the original equation of interest although it does so with sampling error because $\hat{\pi}_0 \neq \pi_0$ and $\hat{\pi}_1 \neq \pi_1$.

Comments

- See Lecture 6 of the NBER course by Imbens and Wooldridge for more information on the control function method.
- Under the above assumptions, it is possible to test the null that $\rho_1 = 0$;
- In the simple linear case described above, IV and control function estimates are numerically identical.
- If the model is non-linear in D
  - IV and control function estimates no longer coincide numerically;
  - The assumption

    $$E(UZ) = 0$$

    is no longer sufficient (as instead in the IV case) to apply the control function method, and the stronger assumption

    $$E(U|Z) = 0$$

    is needed: U and V must be independent not just uncorrelated.

- The control function approach is likely more efficient but less robust then standard IV. Specifically it might be inconsistent but more precise in cases in which IV is consistent.

- If D is a discrete or count or truncated or censored variable,
  - then appropriate non-linear models are needed to estimate the residuals,
  - these models require additional functional form assumptions that make the control function approach even less robuts but considerably more precise if the assumptions are satisfied;
- In the binary case, one can assume normality and estimate a probit model for the first stage; in this case the control function approach boils down to the Heckman two step procedure that we have seen above and which rests on the normality assumption.
- There are cases however in which the control function approach has no alternatives: for example if the dependent variable is a censored duration and the estimation of non linear hazard models is needed (see example in class).

- As concluded by Imbens and Wooldrigde, except in cases in which
  - D appears linearly in the main equation of interest and
  - in which a linear reduced form can be estimated for it

  the bottom line is that
  - the control function approach imposes extra assumptions not imposed by IV approches
  - but in more complex models it is hard to beat the control function approach.

5 Instrumental Variable interpreted as "quasi-experiments"

Consider the following notation:

- N units denoted by i.
- They are exposed to two possible levels of treatment: $D_i = 0$ and $D_i = 1$.
- Y_i is a measure of the outcome.
- Z_i is a binary indicator that denotes the assignment to treatment; it is crucial to realize that:
  i. assignment to treatment may or may not be random;
  ii. the correspondence between assignment and treatment may be imperfect.

Examples: Willis and Rosen (1979), Angrist (1990), Angrist and Krueger (1991), Card (1995), Ichino and Winter-Ebmer (2004).

Participation into treatment depends on the vector of assignments $\mathbf{Z}$

$$D_i = D_i(\mathbf{Z}) \tag{68}$$

The outcome depends on the vector of assignments $\mathbf{Z}$ and treatments $\mathbf{D}$:

$$Y_i = Y_i(\mathbf{Z}, \mathbf{D}) \tag{69}$$

Note that in this framework we can define three (main) causal effects:

- the effect of assignment Z_i on treatment D_i;
- the effect of assignment Z_i on outcome Y_i;
- the effect of treatment D_i on outcome Y_i.

The first two of these effects are called *intention-to-treat* effects.

The Angrist-Imbens-Rubin Causal model (see Angrist et. al. 1996) defines the minimum set of assumptions that ensures the identification of these effects for a relevant subgroup in the population.

5.1 Assumptions of the Angrist-Imbens-Rubin Causal model

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA).
The potential outcomes and treatments of unit i are independent of the potential assignments, treatments and outcomes of unit $j \neq i$:

i. $D_i(\mathbf{Z}) = D_i(Z_i)$

ii. $Y_i(\mathbf{Z}, \mathbf{D}) = Y_i(Z_i, D_i)$

Given this assumption we can write the *intention-to-treat* effects as:

Definition 2. *The Causal Effect of Z on D for unit i is*

$$D_i(1) - D_i(0)$$

Definition 3. *The Causal Effect of Z on Y for unit i is*

$$Y_i(1, D_i(1)) - Y_i(0, D_i(0))$$

Counterfactual reasoning requires to imagine that for each subject the sets of

- potential outcomes $[Y_i(0,0),\ Y_i(1,0),\ Y_i(0,1),\ Y_i(1,1)]$
- potential treatments $[D_i(0)=0, D_i(0)=1, D_i(1)=0, D_i(1)=1]$
- potential assignments $[Z_i=0, Z_i=1]$

exist, although only one item for each set is actually observed.

Implications of SUTVA for general equilibrium analysis and external validity.

If SUTVA holds, we can classify subjects according to the following useful typology.

Table 1: Classification of units according to assignment and treatment status

		$Z_i = 0$	
		$D_i(0) = 0$	$D_i(0) = 1$
$Z_i = 1$	$D_i(1) = 0$	*Never-taker*	*Defier*
	$D_i(1) = 1$	*Complier*	*Always-taker*

Examples: Willis and Rosen (1979), Angrist (1990), Angrist and Krueger (1991), Card (1995), Ichino and Winter-Ebmer (2004).

Assumption 2. Random Assignment (ignorability).
All units have the same probability of assignment to treatment:

$$Pr\{Z_i = 1\} = Pr\{Z_j = 1\}$$

Given SUTVA and random assignment we can identify and estimate the two *intention to treat* causal effects:

$$E\{D_i \mid Z_i = 1\} - E\{D_i \mid Z_i = 0\} = \frac{COV\{D_i Z_i\}}{VAR\{Z_i\}} \tag{70}$$

$$E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\} = \frac{COV\{Y_i Z_i\}}{VAR\{Z_i\}} \tag{71}$$

Note that the ratio between these effects is the IV estimand

$$\frac{COV\{Y, Z\}}{COV\{D, Z\}} \tag{72}$$

Is this the causal effect of D_i on Y_i?

Assumption 3. Non-zero average causal effect of Z on D.
The probability of treatment must be different in the two assignment groups:

$$Pr\{D_i(1) = 1\} \neq Pr\{D_i(0) = 1\}$$

or equivalently

$$E\{D_i(1) - D_i(0)\} \neq 0$$

This assumption requires that the assignment to treatment is correlated with the treatment indicator.

It is easy to test.

It is the equivalent of the"first stage" in the conventional IV approach.

Assumption 4. Exclusion Restrictions.
The assignment affects the outcome only through the treatment and we can write

$$Y_i(0, D_i) = Y_i(1, D_i) = Y_i(D_i).$$

It cannot be tested because it relates quantities that can never be observed jointly:

$$Y_i(0, D_i) = Y_i(1, D_i)$$

It says that given treatment, assignment does not affect the outcome. So we can define the causal effect of D_i on Y_i with the following simpler notation:

Definition 4. *The Causal Effect of D on Y for unit i is*

$$Y_i(1) - Y_i(0)$$

Are the first four assumptions enough?

We can now establish the relationship *at the unit level* between the *intention to treat* effects of Z on D and Y and the causal effect of D on Y.

$$
\begin{aligned}
Y_i(1, D_i(1)) - Y_i(0, D_i(0)) & \\
&= Y_i(D_i(1)) - Y_i(D_i(0)) \\
&= [Y_i(1)D_i(1) + Y_i(0)(1 - D_i(1))] - \\
&\quad [Y_i(1)D_i(0) + Y_i(0)(1 - D_i(0))] \\
&= (D_i(1) - D_i(0))(Y_i(1) - Y_i(0)) \qquad (73)
\end{aligned}
$$

At the unit level the causal effect of Z on Y is equal to the product of the the causal effect of Z on D times the causal effect of D on Y.

Can we take the expectation of both sides of 73 and identify the average causal effect of D on Y:

$$E(Y_i(1) - Y_i(0))?$$

And the answer is "no", because:

$$
\begin{aligned}
& E\ \{Y_i(1, D_i(1)) - Y_i(0, D_i(0))\} \\
&= E\{(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))\} \\
&= E\{Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1\} Pr\{D_i(1) - D_i(0) = 1\} - \\
&\quad E\{Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = -1\} Pr\{D_i(1) - D_i(0) = -1\}
\end{aligned} \tag{74}
$$

Equation 74 shows that even with the four assumptions that were made so far we still have an identification problem.

What we observe (the left hand side), is equal to the weighted difference between the average effect for *compliers* and the average effect for *defiers*.

To solve this problem we need a further and last assumption.

Table 2: Causal effect of Z on Y according to assignment and treatment status

		$Z_i = 0$	
		$D_i(0) = 0$	$D_i(0) = 1$
$Z_i = 1$	$D_i(1) = 0$	*Never-taker* $Y_i(1,0) - Y_i(0,0) = 0$	*Defier* $Y_i(1,0) - Y_i(0,1) = -(Y_i(1) - Y_i(0))$
	$D_i(1) = 1$	*Complier* $Y_i(1,1) - Y_i(0,0) = Y_i(1) - Y_i(0)$	*Always-taker* $Y_i(1,1) - Y_i(0,1) = 0$

In the previous table:

- Each cell contains the causal effect of Z on Y (the numerator of LATE).
- The SUTVA assumption allows us to write this causal effect for each unit independently of the others.
- The random assignment assumption allows us to identify the causal effect for each group.
- Exclusion restrictions ensure that the causal effect is zero for the *always-* and *never-takers*; it is non-zero only for *compliers* and *defiers* (via D).
- The assumption of strong monotonicity ensures that there are no *defiers* and that *compliers* exist.

All this ensures that the numerator of the LATE estimator is the average effect of Z on Y for the group of *compliers* (absent general equilibrium considerations).

Assumption 5. Monotonicity.
No one does the opposite of his/her assignment, no matter what the assignment is:

$$D_i(1) \geq D_i(0) \quad \forall i \tag{75}$$

This assumption amounts to excluding the possibility of *defiers*.

The combination of Assumptions 3 and 5 is called *Strong Monotonicity*

$$D_i(1) \geq D_i(0) \quad \forall i \text{ with strong inequality for at least some } i \tag{76}$$

and ensures that:

- there is no defier and
- there exists at least one complier.

Since now *defiers* do not exist by assumption, we can use equation 74 to identify the average treatment effect for *compliers*.

5.2 The Local Average Treatment Effect

Equation 74 now is:

$$\begin{aligned} & E\ \{Y_i(1, D_i(1)) - Y_i(0, D_i(0))\} \\ = \ & E\{Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1\} Pr\{D_i(1) - D_i(0) = 1\} \end{aligned} \tag{77}$$

Rearranging this equation, the Local Average Treatment Effect is defined as:

$$E\{Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1\} = \frac{E\{Y_i(1, D_i(1)) - Y_i(0, D_i(0))\}}{Pr\{D_i(1) - D_i(0) = 1\}}$$

Definition 5. LATE.
The Local Average Treatment Effect is the average effect of treatment for those who change treatment status because of a change of the instrument; i.e. the average effect of treatment for compliers.

Equivalent expressions for the LATE estimator:

$$E\{Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0\}$$

$$= \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{Pr\{D_i(1) = 1\} - Pr\{D_i(0) = 1\}} \tag{78}$$

$$= \frac{E\{Y_i \mid Z_i = 1\} - E\{Y_i \mid Z_i = 0\}}{E\{D_i \mid Z_i = 1\} - E\{D_i \mid Z_i = 0\}} \tag{79}$$

$$= \frac{COV\{Y, Z\}}{COV\{D, Z\}} \tag{80}$$

- The IV estimand is the LATE.

- The LATE is the only treatment effect that can be estimated by IV, unless we are willing to make further assumptions.

Table 3: Frequency of each type of unit in the population

		$Z_i = 0$	
		$D_i(0) = 0$	$D_i(0) = 1$
$Z_i = 1$	$D_i(1) = 0$	*Never-taker* $Pr\{D_i(1) = 0, D_i(0) = 0\}$	*Defier* $Pr\{D_i(1) = 0, D_i(0) = 1\}$
	$D_i(1) = 1$	*Complier* $Pr\{D_i(1) = 1, D_i(0) = 0\}$	*Always-taker* $Pr\{D_i(1) = 1, D_i(0) = 1\}$

In the previous table:

- The denominator of the Local Average Treatment Effect is the frequency of *compliers*.

- Note that the frequency of compliers is also the average causal effect of Z on D (see eq 79):

$$E\{D_i \mid Z_i = 1\} - E\{D_i \mid Z_i = 0\} =$$
$$Pr\{D_i = 1 \mid Z_i = 1\} - Pr\{D_i = 1 \mid Z_i = 0\}.$$

- Indeed the LATE-IV estimator is the ratio of the two average *intention-to-treat* effects: the effect of Z on Y divided by the effect of Z on D.

5.3 Comments on the LATE interpretation of IV

i. The AIR approach clarifies the set of assumptions under which the IV estimand is an average causal effect, but shows that this is not the ATT.

ii. To identify the ATT the conventional approach implicitly assumes that the causal effect is the same for all treated independently of assignment.

iii. Translated in the AIR framework this conventional assumption is (see the debate Heckman-AIR in Angrist et al., 1996):

$$E\{Y_i(1)-Y_i(0) \mid Z_i, D_i(Z_i)=1\} = E\{Y_i(1)-Y_i(0) \mid D_i(Z_i)=1\} \quad (81)$$

$$\begin{aligned} E\{Y_i(1)-Y_i(0) \mid D_i(1)=1; D_i(0)=1\} & \quad (82) \\ = E\{Y_i(1)-Y_i(0) \mid D_i(1)=1; D_i(0)=0\} & \end{aligned}$$

i.e., the causal effect of D on Y must be the same for *compliers* and *always-taker*. Typically this assumption cannot be tested and is unlikely to hold in many applications.

iv. The conventional approach hides also the assumption of strong monotonicity.

v. The AIR approach concludes that the only causal effect that IV can identify with a minimum set of assumptions is the causal effect for *compliers*, i.e. the LATE: the effect of treatment for those who change treatment status because of a different assignment.

vi. Intuitively this makes sense because *compliers* are the only group on which the data can be informative :

- *compliers* are the only group with units observed in both treatments (given that *defiers* have been ruled out).
- *always takers* and *never-takers* are observed only in one treatment.
- The LATE is analogous to a regression coefficient estimated in linear models with unit effects using panel data. The data can only be informative about the effect of regressors on units for whom the regressor changes over the period of observation.

vii. The conventional approach to IV, however, argues that the LATE is a controversial parameter because it is defined for an unobservable sub-population and because it is instrument dependent. And therefore it is no longer clear which interesting policy question it can answer.

viii. Furthermore it is difficult to think about the LATE in a general equilibrium context

ix. Hence, the conventional approach concludes that it is preferable to make additional assumptions, in order to answer more interesting and well posed policy questions.

x. Yet there are many relevant positive and normative questions for which the LATE seems to be an interesting parameter in addition to being the only one we can identify without making unlikely assumptions.

6 Regression Discontinuity Designs

In the absence of random assignment, an alternative favorable situation for the identification of treatment effects arises when participation into treatment is determined by a *sharp* Regression Discontinuity Design (RDD)

In this design, assignment to treatment solely depends on whether an observable pre-intervention variables satisfy a set of conditions *known* to the analyst.

For examples, units willing to participate are divided into two groups according to whether or not a pre-intervention measure exceeds a known threshold, but only units scoring above that threshold are assigned to the program.

In a neighborhood of the threshold for selection a $sharp$ RDD presents some features of a pure experiment.

Examples: Angrist and Lavy (1999), Van der Klauuw (2002), Di Nardo and Lee (2004), Lee (2005), Ichino et al. (2013).

The comparison of mean outcomes for participants and non-participants $at\ the\ margin$ allows to control for confounding factors and identifies the mean impact of the intervention *locally* at the threshold for selection.

For identification at the cut-off point to hold it must be the case that any discontinuity in the relationship between the outcome of interest and the variable determining the treatment status is fully attributable to the treatment itself.

The $sharp$ RDD features two main limitations:

- assignment to treatment must depend *only* on observable pre-intervention variables
- identification of the mean treatment effect is possible only at the threshold for selection.

Matters complicate further in the case of a $fuzzy$ RDD, i.e. a situation in which there is imperfect compliance with the assignment rule at the threshold.

6.1 Treatment effects in a RDD

- (Y_1, Y_0) are the two potential outcomes induced, respectively, by participation and non-participation.
- $\beta = Y_1 - Y_0$ is the causal effect of the treatment, which is not observable.
- We consider the general case in which β may vary across units.
- I is the binary variable that denotes treatment status, with $I = 1$ for participants and $I = 0$ for non-participants.
- If the assignment is determined by randomization and subjects comply with the assignment:

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

- Given randomization, we can identify the mean impact

$$E\{\beta\} = E\{Y_1|I = 1\} - E\{Y_0|I = 0\}, \tag{83}$$

Formal characterization of an RDD

Following Battistin and Rettore (2006) and Hahn et al.(2001), a RDD arises when:

- treatment status depends on an *observable* unit characteristic S;
- there exist a *known* point in the support of S where the probability of participation changes discontinuously.

If $\bar{s}$ is the discontinuity point, then a RDD is defined if

$$Pr\{I = 1|\bar{s}^{+}\} \neq Pr\{I = 1|\bar{s}^{-}\}. \tag{84}$$

where $\bar{s}^{+}$ and $\bar{s}^{-}$ refer to units *marginally* above or below $\bar{s}$.

Without loss of generality, we also assume

$$Pr\{I = 1|\bar{s}^{+}\} - Pr\{I = 1|\bar{s}^{-}\} > 0.$$

Sharp and Fuzzy RDD

Following Trochim (1984), the distinction between *sharp* and *fuzzy* RDD depends on the size of the discontinuity in (84).

A $sharp$ design occurs when the probability of participating conditional on S steps from zero to one as S crosses the threshold $\bar{s}$.

In this case, the treatment status depends deterministically on whether units' values of S are above $\bar{s}$

$$I = 1(S \geq \bar{s}). \tag{85}$$

A $fuzzy$ design occurs when the size of the discontinuity at $\bar{s}$ is smaller than one.

In this case the probability of treatment jumps at the threshold, but it may be greater than 0 below the threshold and smaller than 1 above.

6.2 Identification in a sharp RDD

The observed outcome can be written as $Y = Y_0 + I(s)\beta$

The difference of observed mean outcomes marginally above and below $\bar{s}$ is

$$\begin{aligned} E\{Y|\bar{s}^+\} - E\{Y|\bar{s}^-\} & \qquad (86) \\ = E\{Y_0|\bar{s}^+\} - E\{Y_0|\bar{s}^-\} + E\{I(s)\beta|\bar{s}^+\} - E\{I(s)\beta|\bar{s}^-\} \\ = E\{Y_0|\bar{s}^+\} - E\{Y_0|\bar{s}^-\} + E\{\beta|\bar{s}^+\} \end{aligned}$$

where the last equality holds in a sharp design because $I = 1(S \geq \bar{s})$.

It follows that the mean treatment effect at $\bar{s}^+$ is identified if

Condition 1. *The mean value of Y_0 conditional on S is a continuous function of S at $\bar{s}$:*

$$E\{Y_0|\bar{s}^+\} = E\{Y_0|\bar{s}^-\}$$

This condition for identification requires that in the counterfactual world, no discontinuity takes place at the threshold for selection.

Note that condition 1 allows to identify $only$ the average impact for subjects in a $right\text{-}neighborhood$ of $\bar{s}$.

Thus, we obtain a local version of the average treatment effect in (83)

$$E\{\beta|\bar{s}^+\} = E\{Y|\bar{s}^+\} - E\{Y|\bar{s}^-\}.$$

which is the effect of treatment on the treated (ATT) in this context.

The identification of $E\{\beta|\bar{s}^-\}$ (the effect of treatment on the non-treated), requires a similar continuity condition on the conditional mean $E\{Y_1|S\}$.

In practice, it is difficult to think of cases where Condition 1 is satisfied and the same condition does not hold for Y_1.

The sharp RDD represents a special case of selection on observables (on which we will come back in Section 7).

Moreover, assuming that the distribution of (Y_0, Y_1) as a function of S is continuous at the discontinuity point, implies

$$(Y_1, Y_0) \perp I | S = \bar{s}. \tag{87}$$

Because of this property, a sharp RDD is often referred to as a quasi-experimental design (Cook and Campbell, 1979).

If the sample size is large enough, $E\{Y|\bar{s}^+\}$ and $E\{Y|\bar{s}^-\}$ can be estimated using only data for subjects in a neighborhood of the discontinuity point.

If the sample size is not large enough, one can make some parametric assumptions about the regression curve away from $\bar{s}$ and use also data for subjects outside a neighborhood of the discontinuity point.

Typically this involves the parametric estimation of two polynomials of Y as a function of S on the two sides of the discontinuity, measuring how they differ for values of S that approach the discontinuity.

Evidence on the validity of the identification condition

An attractive feature of a RDD is that it allows to test the validity of the identification condition (87).

These tests are based on the idea of comparing units marginally above and below the threshold with respect to variables which:

- *cannot* be affected by the treatment;
- are affected by the same unobservables which are relevant for the outcome.

Finding that the two groups of subjects present systematic differences in the values of these variables would cast serious doubts on the validity of the identification condition (87).

6.3 Identification in a fuzzy RDD

If compliance with the design implied by S and $\bar{s}$ is imperfect, a *fuzzy* RDD arises.

In this case, the continuity of Y_0 and Y_1 at $\bar{s}$ is no longer sufficient to ensure the orthogonality condition in (87).

Now the treatment status depends not only on S but also on unobservables, and the following condition is needed:

Condition 2. *The triple* $(Y_0, Y_1, I(s))$ *is stochastically independent of* S *in a neighborhood of* $\bar{s}$.

The stochastic independence between $I(s)$ and S in a neighborhood of $\bar{s}$ corresponds to *imposing* that assignment at $\bar{s}$ takes place as if it were randomized.

The stochastic independence between (Y_1, Y_0) and S at $\bar{s}$ corresponds to a standard exclusion restriction.

It imposes that in a neighborhood of $\bar{s}$, S affects the outcome only through its effect on the treatment I.

In other words, there is no direct effect of S on the outcome for given treatment status in a neighborhood of the threshold.

If Condition 2 holds we are in the familiar IV framework of Section 5:

- S is the random assignment to treatment and plays the same role of Z.
- I is treatment status and plays the same role of D.
- Y_0, Y_1 are the potential outcomes and Y is the observed outcome.

The categorization of subjects into *always takers*, *never takers*, *compliers* and *defiers* applies as well.

If Condition 2 is satisfied, the outcome comparison of subjects above and below the threshold gives:

$$\begin{aligned} E\{Y|\bar{s}^+\} - E\{Y|\bar{s}^-\} \\ = E\{\beta|I(\bar{s}^+) > I(\bar{s}^-)\}Pr\{I(\bar{s}^+) > I(\bar{s}^-)\} \\ - E\{\beta|I(\bar{s}^+) < I(\bar{s}^-)\}Pr\{I(\bar{s}^+) < I(\bar{s}^-)\}. \end{aligned}$$

The right hand side is the difference between:

- the average effect for $compliers$, times the probability of compliance;
- the average effect for $defiers$, times the probability of defiance.

As in the IV framework:

- $always\ takers$ and $never\ takers$ do not contribute because their potential treatment status does not change on the two sides of the threshold;
- for the identification of a meaningful average effect of treatment an additional assumption of strong monotonicity is needed.

Condition 3. *Participation into the program is monotone around* $\bar{s}$*, that is it is either the case that* $I(\bar{s}^+) \geq I(\bar{s}^-)$ *for all subjects or the case that* $I(\bar{s}^+) \leq I(\bar{s}^-)$ *for all subjects.*

This monotonicity condition excludes the existence of *defiers*, so that the outcome comparison of subjects above and below the threshold gives:

$$E\{\beta | I(\bar{s}^+) \neq I(\bar{s}^-)\} = \frac{E\{Y|\bar{s}^+\} - E\{Y|\bar{s}^-\}}{E\{I|\bar{s}^+\} - E\{I|\bar{s}^-\}}, \tag{88}$$

The right hand side of (88) is the mean impact on those subjects in a neighborhood of $\bar{s}$ who would switch their treatment status if the threshold for participation switched from just above their score to just below it.

It is the analog of the LATE in this context.

The denominator in the right-hand side of (88) identifies the proportion of *compliers* at $\bar{s}$.

6.4 A partially *fuzzy* design

Battistin and Rettore (2001) consider an interesting particular case:

- Subjects with S above a known threshold $\bar{s}$ are $eligible$ to participate in a program but may decide not to participate;
- Unobservables determine participation given eligibility;
- Subjects with S below $\bar{s}$ cannot participate, under any circumstance.

This is a "one-sided" *fuzzy* design, in which the population is divided into three groups of subjects:

- eligible participants;
- eligible non-participants;
- non-eligible.

Despite the *fuzzy* nature of this design, the mean impact for all the treated (ATT) can be identified under Condition 1 only, as if the design were $sharp$.

Condition 1 says that:

$$E\{Y_0|\bar{s}^+\} = E\{Y_0|\bar{s}^-\}. \tag{89}$$

and

$$E\{Y_0|\bar{s}^+\} = E\{Y_0|I=1,\bar{s}^+\}\phi + E\{Y_0|I=0,\bar{s}^+\}(1-\phi),$$

where $\phi = E\{I|\bar{s}^+\}$ is the probability of self-selection into the program conditional on marginal eligibility.

The last expression combined with (89) yields

$$E\{Y_0|I=1,\bar{s}^+\} = \frac{E\{Y_0|\bar{s}^-\}}{\phi} - E\{Y_0|I=0,\bar{s}^+\}\frac{1-\phi}{\phi}. \tag{90}$$

The *counterfactual* mean outcome for marginal participants is a linear combination of *factual* mean outcomes for marginal ineligibles and for marginal eligibles not participants.

The coefficients of this combination add up to one and are a function of ϕ, which is identified from observed data.

Hence, equation (90) implies that the mean impact on participants is identified:

$$E\{\beta|I=1,\bar{s}^{+}\}=E(Y_1|I=1,\bar{s}^{+})-E(Y_0|I=1,\bar{s}^{+}).$$

Note that in this setting, by construction there are no $always\ takers$, although there may be $never\ takers$, who are the eligible non-participants.

All the treated are $compliers$ as in the experimental framework of Bloom (1984).

This result is relevant because such a one-sided $fuzzy$ design is frequently encountered in real application.

Less frequent, however, is the availability of information on eligible non participants, which is necessary for identification.

6.5 A regression framework for a *fuzzy* RDD (Angrist and Lavy, 1999)

Under the assumptions of a *fuzzy* design consider the equation

$$Y = g(S) + \beta T + \epsilon \tag{91}$$

where:

- Y is the observed outcome;
- $g(S)$ is a polinomial in the score S;
- T is a binary indicator that denotes actual exposure to treatment;
- $I = 1(S \geq \bar{s})$ is the side of the threshold on which each subject is located.

The IV-LATE estimate of 91 using I as an instrument is equivalent to the RDD comparison of outcomes for subjects marginally above or below the threshold

Both methods identify the mean impact on those subjects in a neighborhood of $\bar{s}$ who would switch their treatment status if the threshold for participation switched from just above their score to just below it.

6.6 Comments on RDD

- A $sharp$ RDD identifies the mean impact of a treatment for a broader population than the one for which identification is granted by a $fuzzy$ RDD.
- Whether the parameter identified by a $fuzzy$ RDD is policy relevant depends on the specific case.
- A $fuzzy$ RDD requires stronger identification conditions.
- Some of the simplicity of the RDD is lost moving to a $fuzzy$ design.
- Both $sharp$ and $fuzzy$ designs cannot identify the impact for subjects far away from the discontinuity threshold.
- A RDD framework naturally suggests ways to test the validity of the identification assumptions.
- RDDs are promising tools for the identification of causal effects.

7 Selection on observables and matching

Matching methods may offer a way to estimate average treatment effects when:

- controlled randomization is impossible and
- there are no convincing natural experiments providing a substitute to randomization (a RDD, a good instrument ...).

But these methods require the debatable assumption of *selection on observables* (or *unconfoundedness*, or *conditional independence*):

- the selection into treatment is completely determined by variables that can be observed by the researcher;
- "conditioning" on these observable variables, the assignment to treatment is random.

Given this assumption, these methods base the estimation of treatment effects on a "very careful" matching of treated and control subjects.

Apparently it sounds like ... assuming away the problem.

However, matching methods have the following desirable features:

- The observations used to estimate the causal effect are selected *without* reference to the outcome, as in a controlled experiment.

- They dominate other methods based on selection on observables (like OLS), thanks to a more convincing comparison of treated and control units;

- They offer interesting insights for a better understanding of the estimation of causal effects.

- There is some (debated) evidence suggesting that they contribute to reduce the selection bias
  (see Dehejia and Wahba 1999; Dehejia 2005; Smith and Todd 2005a, 2005b).

As a minimum, matching methods provide a convincing way to select the observations on which other estimation methods can be later applied.

7.1 Notation

- i denotes subjects in a population of size N .
- $D_i \in \{0, 1\}$ is the treatment indicator for unit i.
- $Y_i(D_i)$ are the potential outcomes in the two treatment situations.
  - $Y_i(1)$ is the outcome in case of treatment;
  - $Y_i(0)$ is the outcome in case of no treatment.
- the observed outcome for unit i is:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{92}$$

- Δ_i is the causal treatment effect for unit i defined as

$$\Delta_i = Y_i(1) - Y_i(0) \tag{93}$$

  which cannot be computed because only one of the two counterfactual treatment situations is observed.

We want to estimate the average effect of treatment on the treated (ATT):

$$\tau = E\{\Delta_i | D_i = 1\} = E\{Y_i(1) - Y_i(0) | D_i = 1\} \tag{94}$$

The problem is the usual one: for each subject we do not observe the outcome in the counterfactual treatment situation.

Note that this can be viewed as a problem of "missing data".

Matching methods are a way to "impute" missing observations for counterfactual outcomes.

From this viewpoint, their validity stands on the assumption that the counterfactual observations are "missing at random" (Rubin, 1974).

Remember (see Section 1) that in this situation a comparison of output by treatment status gives a biased estimate of the ATT.

The case of random assignment to treatment

If assignment to treatment is random:

$$Y(1), Y(0) \perp D \tag{95}$$

And the missing information does not create problems because:

$$E\{Y_i(0)|D_i = 0\} = E\{Y_i(0)|D_i = 1\} = E\{Y_i(0)\} \tag{96}$$

$$E\{Y_i(1)|D_i = 0\} = E\{Y_i(1)|D_i = 1\} = E\{Y_i(1)\} \tag{97}$$

and substituting 96 and 97 in 94 it is immediate to obtain:

$$\begin{aligned} \tau &\equiv E\{\Delta_i \mid D_i = 1\} \\ &\equiv E\{Y_i(1)|D_i = 1\} - E\{Y_i(0) \mid D_i = 1\} \\ &= E\{Y_i(1)|D_i = 1\} - E\{Y_i(0)|D_i = 0\} \\ &= E\{Y_i|D_i = 1\} - E\{Y_i|D_i = 0\}. \end{aligned} \tag{98}$$

Randomization ensures that the missing information is "missing completely at random" and thus the sample selection bias is zero:

$$E\{Y_i(0) \mid D_i = 1\} - E\{Y_i(0) \mid D_i = 0\} = 0 \tag{99}$$

7.2 Selection on observables

Let X denote a set of pre-treatment characteristics of subjects.

Definition 6. Unconfoundedness
Assignment to treatment is unconfounded given pre-treatment variables if

$$Y(1), Y(0) \perp D \mid X \tag{100}$$

Note that assuming unconfoundedness is equivalent to say that:

- within each cell defined by X treatment is random;
- the selection into treatment depends on observables X up to a random factor.

Pure randomization is a particularly strong version of "unconfoundedness", in which the assignment to treatment is unconfounded independently of pre-treatment variables.

ATT assuming unconfoundedness

$$E\{Y_i(0)|D_i=0,X\} = E\{Y_i(0)|D_i=1,X\} = E\{Y_i(0)|X\} \tag{101}$$

$$E\{Y_i(1)|D_i=0,X\} = E\{Y_i(1)|D_i=1,X\} = E\{Y_i(1)|X\} \tag{102}$$

Using these expressions, we can define for each cell defined by X

$$\begin{aligned} \delta_x &\equiv E\{\Delta_i|X\} \\ &\equiv E\{Y_i(1)|X\} - E\{Y_i(0)|X\} \\ &= E\{Y_i(1)|D_i=1,X\} - E\{Y_i(0)|D_i=0,X\} \\ &= E\{Y_i|D_i=1,X\} - E\{Y_i|D_i=0,X\}. \end{aligned} \tag{103}$$

Using the Law of Iterated expectations, the ATT is given by:

$$\begin{aligned} \tau &\equiv E\{\Delta_i|D_i=1\} \\ &= E\{E\{\Delta_i|D_i=1,X\} \mid D_i=1\} \\ &= E\{\ \ E\{Y_i|D_i=1,X\} - E\{Y_i|D_i=0,X\}\ \ |D_i=1\} \\ &= E\{\delta_x|D_i=1\} \end{aligned} \tag{104}$$

where the outer expectation is over the distribution of $X|D_i=1$.

Matching and regression

Unconfoundedness suggests the following strategy for the estimation of the ATT:

i. stratify the data into cells defined by each particular value of X;

ii. within each cell (i.e. conditioning on X) compute the difference between the average outcomes of the treated and the controls;

iii. average these differences with respect to the distribution of X in the population of treated units.

This strategy (called exact matching) raises the following questions:

- Is this strategy different from the estimation of a linear regression of Y on D controlling non parametrically for the full set of main effects and interactions of the covariates X(i.e. a $fully\ saturated$ regression)?
- Is this strategy feasible?

Note that both matching and regression assume selection on observables.

In which sense matching and regression differ?

The essential difference between

- regression on a fully saturated model and
- exact matching

is the weighting scheme used to take the average of the treatment effects at different values of the covariates.

Regression gives more weights to cells in which the proportion of treated and non-treated is similar.

Matching gives more weights to cells in which the proportion of treated is high.

Angrist (1998) gives an interesting example of the differences between matching and regression.

See Ichino et al. (2014) for an application based on exact matching.

Are matching and regression feasible: the dimensionality problem

Both exact matching and fully saturated regression may not be feasible if the sample is small, the set of covariates is large and many of them are multivalued, or, worse, continue.

- With K binary variables the number of cells is 2^K.
- The number increases further if some variables take more than two values.
- If the number of cells is very large with respect to the size of the sample it is possible that cells contain only treated or only control subjects.

Rosenbaum and Rubin (1983) propose an equivalent and feasible estimation strategy based on the concept of $Propensity\ Score$ and on its properties which allow to reduce the dimensionality problem.

It is important to realize that regression with a non-saturated model is not a solution and may lead to seriously misleading conclusions.

Definition 7. Propensity Score (Rosenbaum and Rubin, 1983)
The propensity score is the conditional probability of receiving the treatment given the pre-treatment variables:

$$p(X) \equiv Pr\{D = 1|X\} = E\{D|X\} \tag{105}$$

Lemma 1. Balancing of pre-treatment variables given the propensity score
If $p(X)$ is the propensity score

$$D \perp X \mid p(X) \tag{106}$$

Lemma 2. Unconfoundedness given the propensity score
Suppose that assignment to treatment is unconfounded, i.e.

$$Y(1), Y(0) \perp D \mid X$$

Then assignment to treatment is unconfounded given the propensity score, i.e

$$Y(1), Y(0) \perp D \mid p(X) \tag{107}$$

Proof of Lemma 1:

First:

$$\begin{aligned} Pr\{D=1|X,p(X)\} &= E\{D|X,p(X)\} \\ &= E\{D|X\} = Pr\{D=1|X\} \\ &= p(X) \end{aligned} \tag{108}$$

Second:

$$\begin{aligned} Pr\{D=1|p(X)\} &= E\{D|p(X)\} \\ &= E\{E\{D|X,p(X)\}|p(X)\} = E\{p(X)|p(X)\} \\ &= p(X) \end{aligned} \tag{109}$$

Hence:

$$Pr\{D=1|X,p(X)\} = Pr\{D=1|p(X)\} \tag{110}$$

which implies that conditionally on $p(X)$ the treatment and the observables are independent. *QED*.

Proof of Lemma 2:

First:

$$\begin{aligned} Pr\{D=1|Y(1),Y(0),p(X)\} &= E\{D|Y(1),Y(0),p(X)\} \\ &= E\{E\{D|X,Y(1),Y(0)\}|Y(1),Y(0),p(X)\} \\ &= E\{E\{D|X\}|Y(1),Y(0),p(X)\} \\ &= E\{p(X)|Y(1),Y(0),p(X)\} \\ &= p(X) \end{aligned} \tag{111}$$

where the step from the second to the third line uses the unconfoundedness assumption. Furthermore, because of Lemma 1

$$Pr\{D=1|p(X)\} = p(X) \tag{112}$$

Hence

$$Pr\{D=1|Y(1),Y(0),p(X)\} = Pr\{D=1|p(X)\} \tag{113}$$

which implies that conditionally on $p(X)$ the treatment and potential outcomes are independent. *QED*.

The propensity score and its properties make it possible to match cases and controls on the basis of a monodimensional variable.

$$E\{Y_i(0)|D_i = 0, p(X_i)\} = E\{Y_i(0)|D_i = 1, p(X_i)\} = E\{Y_i(0)|p(X_i)\}$$
$$E\{Y_i(1)|D_i = 0, p(X_i)\} = E\{Y_i(1)|D_i = 1, p(X_i)\} = E\{Y_i(1)|p(X_i)\}$$

Using these expressions, we can define for each cell defined by $p(X)$

$$\begin{aligned} \delta_{p(x)} &\equiv E\{\Delta_i|p(X_i)\} \\ &\equiv E\{Y_i(1)|p(X_i)\} - E\{Y_i(0)|p(X_i)\} \\ &= E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\} \\ &= E\{Y_i|D_i = 1, p(X_i)\} - E\{Y_i|D_i = 0, p(X_i)\}. \end{aligned} \tag{114}$$

Using the Law of Iterated expectations, the ATT is given by:

$$\begin{aligned} \tau &= E\{\Delta_i|D_i = 1\} \\ &= E\{E\{\Delta_i|D_i = 1, p(X_i)\}|D_i = 1\} \\ &= E\{\ \ E\{Y_i(1)|D_i = 1, p(X_i)\} - E\{Y_i(0)|D_i = 0, p(X_i)\}\ \ |D_i = 1\} \\ &= E\{\delta_{p(x)}|D_i = 1\} \end{aligned} \tag{115}$$

where the outer expectation is over the distribution of $p(X_i)|D_i = 1$.

Implementation of matching based on the pscore

Two sequential steps are needed.

i. *Estimation of the propensity score*
This step is necessary because the "true" propensity score is unknown and therefore the propensity score has to be estimated.

ii. *Estimation of the average effect of treatment given the propensity score*
Ideally in this step, we would like to

- match cases and controls with exactly the same (estimated) propensity score;
- compute the effect of treatment for each value of the (estimated) propensity score (see equation 114).
- obtain the average of these conditional effects as in equation 115.

This is infeasible in practice because it is rare to find two units with exactly the same propensity score.

There are, however, several alternative and feasible procedures to perform this step:

- Stratification on the Score;
- Nearest neighbor matching on the Score;
- Radius matching on the Score;
- Kernel matching on the Score;

7.4 Estimation of the Propensity Score

Apparently, the same dimensionality problem that prevents the estimation of treatment effects should also prevent the estimation of propensity scores.

This is, however, not the case thanks to the *balancing property* of the propensity score (Lemma 1) according to which:

- observations with the same propensity score have the same distribution of observable covariates independently of treatment status;
- for given propensity score assignment to treatment is random and therefore treated and control units are on average observationally identical.

Hence, any standard probability model can be used to estimate the propensity score, e.g. a logit model:

$$Pr\{D_i = 1|X_i\} = \frac{e^{\lambda h(X_i)}}{1 + e^{\lambda h(X_i)}} \tag{116}$$

where $h(X_i)$ is a function of covariates with linear and higher order terms.

The choice of which higher order terms to include is determined solely by the need to obtain an estimate of the propensity score that satisfies the $balancing\ property$.

Inasmuch as the specification of $h(X_i)$ which satisfies the $balancing\ property$ is more parsimonious than the full set of interactions needed to match cases and controls on the basis of observables (as in equations 103 and 104), the propensity score reduces the dimensionality of the estimation problem.

Note that, given this purpose, the estimation of the propensity scores does not need a behavioral interpretation.

An algorithm for the estimation of the propensity score

i. Start with a parsimonious logit or probit function to estimate the score.

ii. Sort the data according to the estimated propensity score (from lowest to highest).

iii. Stratify all observations in blocks such that in each block the estimated propensity scores for the treated and the controls are not statistically different:

   (a) start with five blocks of equal score range $\{0 - 0.2, ..., 0.8 - 1\}$;

   (b) test whether the means of the scores for the treated and the controls are statistically different in each block;

   (c) if yes, increase the number of blocks and test again;

   (d) if no, go to next step.

iv. Test that the *balancing property* holds in all blocks for all covariates:

(a) for each covariate, test whether the means (and possibly higher order moments) for the treated and for the controls are statistically different in all blocks;

(b) if one covariate is not balanced in one block, split the block and test again within each finer block;

(c) if one covariate is not balanced in all blocks, modify the logit estimation of the propensity score adding more interaction and higher order terms and then test again.

Note that in all this procedure the outcome has no role.

See the STATA program pscore.ado downloadable at
http://www.iue.it/Personal/Ichino/Welcome.html

With small variations, this is the algorithm proposed by Dehejia and Wahba 1999.

Some useful diagnostic tools

Propensity score methods are based on the idea that the estimation of treatment effects requires a careful matching of cases and controls.

If cases and controls are very different in terms of observables this matching is not sufficiently close and reliable or it may even be impossible.

The comparison of the estimated propensity scores across treated and controls provides a useful diagnostic tool to evaluate how similar are cases and controls, and therefore how reliable is the estimation strategy.

More precisely, it is advisable to:

- count how many controls have a propensity score lower than the minimum or higher than the maximum of the propensity scores of the treated.
  - Ideally we would like that the range of variation of propensity scores is the same in the two groups.

- generate histograms of the estimated propensity scores for the treated and the controls with bins corresponding to the strata constructed for the estimation of propensity scores.
  - Ideally we would like an equal frequency of treated and control in each bin.

Note that these fundamental diagnostic indicators are not computed in standard regression analysis.

7.5 Estimation of the ATT by Stratification on the Propensity Score

This method is based on the same stratification procedure used for estimating the propensity score. By construction, in each stratum covariates are balanced and the assignment to treatment is random.

Let T be the set of treated units and C the set of control units, and Y_i^T and Y_j^C be the observed outcomes of the treated and control units, respectively.

Letting q index the strata defined over intervals of the propensity score, within each block we can compute

$$\tau_q^S = \frac{\sum_{i \in I(q)} Y_i^T}{N_q^T} - \frac{\sum_{j \in I(q)} Y_j^C}{N_q^C} \tag{117}$$

where $I(q)$ is the set of units in block q while N_q^T and N_q^C are the numbers of treated and control units in block q.

The estimator of the ATT in equation 115 is computed with the following

formula:

$$\tau^S = \sum_{q=1}^{Q} \tau_q^S \frac{\sum_{i \in I(q)} D_i}{\sum_{\forall i} D_i} \tag{118}$$

where the weight for each block is given by the corresponding fraction of treated units and Q is the number of blocks.

Assuming independence of outcomes across units, the variance of τ^S is given by

$$Var(\tau^S) = \frac{1}{N^T} \left[Var(Y_i^T) + \sum_{q=1}^{Q} \frac{N_q^T}{N^T} \frac{N_q^T}{N_q^C} Var(Y_j^C) \right] \tag{119}$$

In the program *atts.ado*, standard errors are obtained analytically using the above formula, or by bootstrapping using the *bootstrap* STATA option. See http://www.iue.it/Personal/Ichino/Welcome.html

7.6 Estimation of the ATT by Nearest Neighbor, Radius and Kernel Matching

Ideally, we would like to match each treated unit with a control unit having exactly the same propensity score and viceversa.

This exact matching is, however, impossible in most applications.

The closest we can get to an exact matching is to match each treated unit with the $nearest$ control in terms of propensity score.

This raises however the issue of what to do with the units for which the nearest match has already been used.

We describe here three methods aimed at solving this problem.

- Nearest neighbor matching with replacement;
- Radius matching with replacement;
- Kernel matching

Nearest and radius matching with replacement for the ATT

The steps for the nearest neighbor matching method are as follows:

- For each treated unit find the nearest control unit.
- If the nearest control unit has already been used, use it again (replacement).
- Drop the unmatched controlled units.
- The algorithm delivers a set of N^T pairs of treated and control units in which control units may appear more than once.

The steps for the radius matching method are as follows:

- For each treated unit find all the control units whose score differs by less than a given tolerance r chosen by the researcher.
- Allow for replacement of control units.
- When a treated unit has no control closer than r take the nearest control.
- The algorithm delivers a set of N^T treated units and N^C control units some of which are used more than once.

Formally, denote by $C(i)$ the set of control units matched to the treated unit i with an estimated value of the propensity score of p_i.

Nearest neighbor matching sets

$$C(i) = \min_j \parallel p_i - p_j \parallel, \tag{120}$$

which is a singleton set unless there are multiple nearest neighbors.

In radius matching,

$$C(i) = \left\{ p_j \mid \parallel p_i - p_j \parallel < r \right\}, \tag{121}$$

i.e. all the control units with estimated propensity scores falling within a radius r from p_i are matched to the treated unit i.

Denote the number of controls matched with observation $i \in T$ by N_i^C and define the weights $w_{ij} = \frac{1}{N_i^C}$ if $j \in C(i)$ and $w_{ij} = 0$ otherwise.

The formula for both types of matching estimators can be written as follows (where M stands for either nearest neighbor matching or radius matching):

$$\tau^M = \frac{1}{N^T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C(i)} w_{ij} Y_j^C \right] \tag{122}$$

$$= \frac{1}{N^T} \left[\sum_{i \in T} Y_i^T - \sum_{i \in T} \sum_{j \in C(i)} w_{ij} Y_j^C \right] \tag{123}$$

$$= \frac{1}{N^T} \sum_{i \in T} Y_i^T - \frac{1}{N^T} \sum_{j \in C} w_j Y_j^C \tag{124}$$

where the weights w_j are defined by $w_j = \Sigma_i w_{ij}$. The number of units in the treated group is denoted by N^T.

To derive the variances of these estimators the weights are assumed to be fixed and the outcomes are assumed to be independent across units.

$$Var(\tau^M) = \frac{1}{(N^T)^2}\left[\sum_{i\in T} Var(Y_i^T) + \sum_{j\in C}(w_j)^2 Var(Y_j^C)\right] \quad (125)$$

$$= \frac{1}{(N^T)^2}\left[N^T Var(Y_i^T) + \sum_{j\in C}(w_j)^2 Var(Y_j^C)\right] \quad (126)$$

$$= \frac{1}{N^T} Var(Y_i^T) + \frac{1}{(N^T)^2}\sum_{j\in C}(w_j)^2 Var(Y_j^C). \quad (127)$$

Note that there is a penalty for over using controls.

In the STATA programs *attnd.ado*, *attnw.ado*, and *attr.ado*, standard errors are obtained analytically using the above formula, or by bootstrapping using the *bootstrap* option. See http://www.iue.it/Personal/Ichino/Welcome.html

Estimation of the treatment effect by Kernel matching

Every treated unit is matched with a weighted average of all control units with weights that are inversely proportional to the distance between the scores.

Formally the kernel matching estimator is given by

$$\tau^K = \frac{1}{N^T} \sum_{i \in T} \left\{ Y_i^T - \frac{\sum_{j \in C} Y_j^C G(\frac{p_j - p_i}{h_n})}{\sum_{k \in C} G(\frac{p_k - p_i}{h_n})} \right\} \tag{128}$$

where $G(\dot{})$ is a kernel function and h_n is a bandwidth parameter. Under standard conditions on the bandwidth and kernel

$$\frac{\sum_{j \in C} Y_j^C G(\frac{p_j - p_i}{h_n})}{\sum_{k \in C} G(\frac{p_k - p_i}{h_n})} \tag{129}$$

is a consistent estimator of the counterfactual outcome Y_{0i}.

In the program *attk.ado*, standard errors are obtained by bootstrapping using the *bootstrap* option. See http://www.iue.it/Personal/Ichino/Welcome.html

7.7 Sensitivity of Matching Estimators to the CIA

Matching estimators crucially rely on the CIA to identify treatment effects.

Suppose that this condition is not satisfied given observables, but would be satisfied if we could observe another variable.

This variable can be simulated in the data and used as an additional matching factor in combination with the preferred matching estimator.

A comparison of the estimates obtained with and without matching on this simulated binary variable tells us to what extent the baseline estimates are robust to this specific source of failure of the CIA.

The simulated values of the binary variable can be constructed to capture different hypotheses on the nature of potential confounding factors.

See Ichino et al (2006b), Nannicini (2006) and http://nuke.tommasonannicini.eu

7.8 Comments on matching methods.

Matching methods should not be applied $just$ because there is no alternative experimental or quasi-experimental solution for the estimation of treatment effects.

They should applied only when the assumption of $selection\ on\ observables$ is plausible.

In any case, their sensitivity to the validity of the CIA should be assessed before drawing conclusions.

One of their most desirable feature is that they force the researcher to design the evaluation framework and check the data before looking at the outcomes.

They dominate other identification strategies that require selection on observables, like OLS, because they involve a more convincing comparison between treated and control subjects.

8 Extended Reference List

Abadie, Alberto, Angrist, Joshua D. and Imbens, Guido (2000), "Instrumental Variables Estimates of the effect of Subsidized Training on the Quantile of Trainee Earnings", *Econometrica*, 70 (1), 91-117.

Angrist, Joshua D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review* 80, 313–336.

Angrist, Joshua D. and Lavy, Victor (1999), "Using Maimonides'rule to estimate the effect of class size on scholastic achievement", *The Quarterly Journal of Economics*, May 1999.

Angrist, Joshua D. (2000), "Estimation of Limited–Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice", *NBER Technical WWorking Paper* 248.

Angrist, Joshua D. and Jinyong Hahn (2004), "When to Control for Covariates? Panel–Asymptotic Results for Estimates of Treatment Effects", *Review of Economics and Statistics*, 86(1), 1-15.

Angrist, Joshua D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica, vol.66, N2-March, 1988* p.249

Angrist, Joshua D., Kathryn Graddy and Guido W. Imbens (2000), "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish", *Review of Economic Studies*, Vol. 67, Issue 3, pp. 499-527

Angrist, Joshua D. and Alan B. Krueger (1991), "Does Compulsory Schooling Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*.

Angrist, Joshua D. and Krueger, Alan B. (1999), "Empirical Strategies in Labor Economics", (Chap.23 in Handbook of Labor, Economics, Vol. 3, Edited by O. Ashenfelter and D. Card, 1999 Elsevier Science B.V.) pp.: 1277-1366.

Angrist, Joshua D. and Guido W. Imbens (1995), "Two–Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity", *Journal of the American Statistical Association* 90, 431–442.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association* 91, 444–472.

Battistin E, and Rettore E. (2006), "Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs", *Journal of Econometrics*, Elsevier, vol. 142(2), pages 715-730, February.

Bloom, H.S. (1984), "Accounting for No-Shows in Experimental Evaluation Designs", *Evaluation Review*, Vol. 8, 225-246

Blundell, R. and Costa Dias, M. (2000), *Evaluation methods for non-experimental data*, Fiscal Studies, Vol. 21, No. 4, 427-468

Bound John, David A. Jaeger and Regina M. Baker (1995), "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak", *Journal of the American Statistical Association* 90, 443–450.

Card, David (1995), "Using Geographic Variation in College Proximity to Estimate the Returns to Schooling", in: L.N. Christofides, E.K. Grant, and R. Swidinsky (eds.), *Aspects of Labout Market Behaviour: Essays in Honour of John Vanderkamp*. Toronto: University of Toronto Press, 201–222,.

Card, David (March 1998), "The Causal Effect of Education on Earnings", in: Orley Ashenfelter and David Card (eds.). in: *Handbook of Labor Economics* Vol. 3, North–Holland, Amsterdam.

Cook, T.D. and Campbell, D.T. (1979), *Quasi-Experimentation. Design and Analisys Issues for Field Settigs*, Boston: Houghton Mifflin Company

Dehejia, R.H. and S. Wahba (2002), "Propensity Score Matching Methods for Non-Experimental Causal Studies", *The Review of Economics and Statistics*, February 2002, 84(1): 151161.

Dehejia, R.H., (2005) "Practical propensity score matching: a reply to Smith and Todd", *Journal of Econometrics*, 125, 355364

Dehejia, R.H. and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94, 448, 1053-1062.

Di Nardo, J. and D. Lee (2004), "Economic impacts of new unionization on private sector employers: 1984-2001", *Quarterly Journal of Economics*, November, 1383-1441.

Hahn, Jinyong (1998), " ON the role of the propensity score in efficient semiparamentric estimation of average treatment effects ", *Econometrica*, 66,2,315-331.

Hahn, J. Todd,P. and Van der Klaauw,W.(2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design", *Econometrica* ol.69, N1 (January, 2001), 201-209.

Heckman, James J. (1978), "Dummy Endogenous Variable in a Simultaneous Equations System", *Econometrica* 46, 4, 931–60.

Heckman, James J. (1979), "Sample Specification Bias as a Specification Error", *Econometrica*.

Heckman, James J. (1990), "Varieties of Selection Bias", *AEA Papers and Proceedings* 80(2), 313–318.

Heckman, James J. (1996), "Randomization as an Instrumental Variable", *Review of Economics and Statistics*, Notes, 336–341.

Heckman, James J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations", *Journal of Human Resources* XXXII, 441–462.

Heckman, James J. (2001), "Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programs", *The Economic Journal*, Volume 111, Issue 475, pages 654699, November 2001.

Heckman, James J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *The Quarterly Journal of Economics*, Vol. 115, No. 1 (Feb., 2000), pp. 45-97

Heckman, James J. H. Ichimura, and P. Todd (1997), " Matching as an econometric evaluation estimator: Evidence from Evaluating a Job Training program ", *Review of Economic Studies*, 65, 261-294.

Heckman, James J. H. Ichimura, and P. Todd (1998), " Matching as an econometric evaluation estimator ", *Review of Economic Studies*, 65, 261-294.

*Heckman, James J. and H. Ichimura, H. Smith and P. Todd (1999?), " Sources of selection bias in evaluating programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method", *Econometrica*

Heckman, James J. and H. Ichimura, H. Smith and P. Todd (1996), "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method,*Proceeding of the National Academy of Science*, Vol. 93, pp. 1341613420, November 1996.

Heckman, James J. J. Hotz, (1989), " Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training ", *Journal of the American Statistical Association*, 84,408, 862-880 (including the comments by P. Holland and R. Moffit).

Heckman, James J. and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions", in: J. Heckman and B. Singer, *Longitudinal Analysis of Labor Market Data*. New York: Wiley, 1985, 156–245.

Heckman, James J. and Guilherme Sedlacek (1985), "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self–Selection in the Labor Market", *Journal of Political Economy* 93, 1077–1125.

Heckman, James J. and Jeffrey A. Smith (1999), "The Pre–Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies", *The Economic Journal*, 109 (July), 313-348, 1999.

Heckman, James J., Justin L. Tobias and Edward J. Vytlacil (2000), "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Returns to Schooling", *NBER Working Paper* 7950.

*Heckman, James J., Justin L. Tobias and Edward J. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent Variable Framework ", *The Review of Economics and Statistics*, August 2003, 85(3): 748755.

Heckman, James J. and Edward J. Vytlacil (March 1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling when the Return is Correlated with Schooling", mimeo (University of Chicago).

Heckman, James J. and Edward J. Vytlacil (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects", *Proceedings of the National Academy of Sciences, USA*, 96, 4730-4734.

Heckman, James J. and Edward J. Vytlacil (2000), "Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluations of Social Programs", mimeo, University of Chicago.

Heckman, James J.(2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture", in: Journal of Political Economy, N4, Vol.109, August 2001, p. 673-748.

Heckman, James J., Robert LaLonde and Jeffrey A. Smith (September 1999), "The Economics and Econometrics of Active Labor Market Programs", in O. Ashenfelter and D. Card (ed.), 1999. "Handbook of Labor Economics," Handbook of Labor Economics, Elsevier, edition 1, volume 3, number 3.

Heckman, James J., Lance Lochner and Christopher Taber, "General–Equilibrium Cost–Benefit Analysis of Education and Tax Policies", in: G. Ranis and L.K. Raut (eds.), *Trade, Growth, and Development*, ch. 14. Elsevier Science B.V., 1999.

Hirano, K., G.W. Imbens and G. Ridder (2000), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score", *Econometrica*, Vol. 71, No. 4 (July, 2003), 11611189

Hirano, K., G.W. Imbens, D.B. Rubin and X.-Hua Zhou (2000), "Assessing the Effect of an Influenza Vaccine in an Encouragement Design", *Biostatistics*, 1, 1, 69–88.

Holland, Paul W. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association* 81, 945–970.

mimeo.

Ichino, Andrea and Rudolf Winter–Ebmer (2004), "The Long Run Educational Cost of World War II: An Example of Local Average Treatment Effect Estimation", *Journal of Labour Economics*

Ichino, Andrea and Rudolf Winter–Ebmer (1999), "Lower and Upper Bounds of Returns to Schooling: An Exercise in IV Estimation with Different Instruments", *European Economic Review* 43, 889–901.

Ichino A., F. Mealli and T. Nannicini (2008), " From Temporary Help Jobs to Permanent Employment: What Can we Learn from Matching Estimators and their Sensitivity?, *Journal of Applied Econometrics* 23: 305327 (2008).

Ichino A., G. Schwerdt, R. Winter–Ebmer and J. Zweimüller, (2013) "Too Old to Work, Too Young to Retire?", Department of Economics, Johannes Kepler University of Linz, Working Paper 1313/2013.

Ichino A., P. Garibaldi, F. Giavazzi and E. Rettore, (2006c) "College cost and time to obtain a degree: Evidence from tuition discontinuities." *The Review of Economics and Statistics*, 94.3 (2012): 699711

Imbens, G.W. (2004), *Semiparametric Estimation of Average Treatment Effects under Exogeneity: a Review*, The Review of Economics and Statistics

Imbens, Guido W. and Joshua D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica* 62, 467–475.

Imbens, Guido W. and Donald B. Rubin (1997b), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models", *Review of Economic Studies* 64, 555–574.

Lalonde, Robert (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review* 76,4, 604–620.

Lee, David S. (2008), "Randomized Experiments from Non-random Selection in U.S. House Elections", $Journal\ of\ Econometrics$ vol. 142, issue 2, pages 675-697

Nannicini, T. (2006) "College cost and time to obtain a degree: Evidence from tuition discontinuities.", mimeo.

Pearl, Judea (2000), "Causality. Models, Reasoning and Inference." Cambridge University Press.

Rosenbaum, P.R. and D.B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70, 1, 41–55.

Rosenbaum, P.R. and D.B. Rubin (1984), "Reducing Bias in Observational Studies using Subclassification on the Propensity Score", *Journal of the American Statistical Association* 79, 387, 147–156.

Roy, Andrew D. (1951), "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*.

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66, 5, 688–701.

Smith, J and Todd p. (2005a), " Does matching overcome LaLondes critique of nonexperimental estimators?", *Journal of Econometrics*, 125, 305353.

Smith, J and Todd p. (2005b), " Rejoinder", *Journal of Econometrics*, 125, 365375.

Staiger, Douglas and J. Stock (1997), "Instrumental Variable Regressions with Weak Instruments", *Econometrica*.

Thistlethwaite, D.L. and Campbell, D.T (1960), "Regression discontinuity analysis: an alternative to the ex post facto experiment", $Journal\ of\ Educational\ Psycology$, Vol. 51, No. 6, 309-317

Trochim, W. (1984), *Research Design for Program Evaluation: the Regression-Discontinuity Approach*, Beverly Hills: Sage Publications

Van der Klauuw, (2002), "Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach", *International Economic Review*, 43, 4 November 2002.

Willis, R. and S. Rosen (1979), "Education and Self Selection", *Journal of Political Economy* (Supplement).